RESEARCH CENTER
Rennes - Bretagne-Atlantique

FIELD

Activity Report 2016

Section New Results

Edition: 2017-08-25

ANJA Team

6. New Results

6.1. Legal aspects of systems designed to judicial risk quantification

Participants : Jérôme Dupré

Within the ANJA team systems designed to calculate judicial risk using machine learning technology (AI) have been developed. A former French magistrate is one of the team member who has participated to these researches. In the meantime, he endeavored to contribute to design a legal framework applicable to this activity.

Artificial intelligence (AI), particularly when applied to justice, is liable to encounter rules of law, which are applicable even in the absence of a specific law. As with any new field of activity (eg the Internet), the notion of “legal vacuum” must not be confused with that of “legislative void”. It is therefore necessary to identify how to protect these technologies, what is the responsibilities of each, whether designer or / and user, that could already be applicable.

The two main concerns are relating to property and liability.

1.Regarding property, one can observe that predictive/quantitative solutions based on artificial intelligence result from a combination of technical criterion, databases, algorithms and software, each subject to specific legal protections. These elements may hence be protected by the copyright (for technical criterion); the database law, copyright and unfair competition (for databases); the trade secret (for algorithms), the copyright (for software). It may be questioned whether it would be irrelevant to create, ultimately, a unified legal status specific to this complex reality. But it is probably too early to legislate.

At the heart of the solution is the algorithm, an immaterial element which, in France, is the least well protected by law (it belongs to the domain of “ideas”), justifying its secret nature.

2. Considering liability aspects, we observe that this “black box” - the secret being a consequence of complexity and investments made - may be at the origin of a prejudice, either because of a bad use, or because it is not correctly designed.

The French law offers a range of solutions to the victim, depending on the origin of the damage (see 6.2).

Trust in results is also a factor to be considered. Thus, in the absence of a technical problem peculiar to the solution, misuse by the legal professional providing legal advice may justify, for example, his/her contractual liability. From this standpoint, the reliance granted to the technology and the way it is presented are essential. It justifies a specific attention to the way contracts relating to these services are drafted.

The designer of a defective solution may be required to guarantee against the hidden defects of Article 1641 of the French Civil Code. (When there is no contract, one can also be liable on the grounds of Article 1242 paragraph 1 of the same Code).

Standardization of algorithms, which could be tested by an independent body and subject to secrecy, is also an option, but presents a risk of possible paralysis of a promising market in the field of mathematics.

More generally, it seems necessary to comply with the CNIL (the French Data Protection Authority) provisions relating to personal data (Law No 78-17 of 6 January 1978, spec. article 10, and soon Regulation EU 2016/679 of the European Parliament and of the Council of 27 April 2016 applicable in 2018) and with Privacy Law... A huge amount of data is indeed likely to reveal information not resulting from each data taken separately. But this risk is probably more present in the field of big data than in algorithms, the data used for learning being "dissolved" in the formula.

6.2. Liability and ethics

Participants : Jacques Lévy Véhel, Jérôme Dupré

Some legal issues are specifically related to the calculation error. The system user may not have entered the data correctly and then will be responsible for the displayed result (but it is possible to display the entered data with the results to limit this risk).

The user may also have entered the data correctly and: -he/she is aware of the error of the system, in which case the user regains a share of responsibility, -he/she is not aware of the error of the system and the error is not easily detectable, then the responsibility should move towards the authors of the system conception (with its different steps that result from each other: definition of the search criteria, creation of database and creation of the algorithm, development and integration of the software, etc.). -he/she is not aware of the error of the system and the error is easily detectable, in which case the user should regain at least some responsibility.

The use of systems designed to judicial risk quantification entails paradigm shift The probabilistic approach, where all things are equal, seems very far from the hierarchy of legal norms and legal causality. In the absence of any upheaval in the law, we can at least expect an upheaval in legal practice.

Ultimately, the use of predictive tools, which favors the discovery of correlations, may lead to less attention to the causes of events, hence the need to maintain a vigilance on this aspect. The human who "delegates everything" to the machine should not avoid responsibility.

From an ethical perspective, is it acceptable to calculate the sentence which is likely to be pronounced in penal law, e.g. for a crime? Should we accept to profile judges using all their past rulings?

More generally, a reflection as well as a study on the place left to the human at all the stages of the process of elaboration and use (the decision-making) of the predictive tools seems necessary.

6.3. Statistical inference methods for panel of random-coefficient AR(1) data

Participants: A. Philippe and D. Surgailis R. Leipus and V. Pilipauskaité (Vilnius university)

We study the statistical inference methods for panel of random-coefficient AR(1) data [17] . We propose a nonparametric estimation of the distribution function of random coefficients by the empirical distribution of lag 1 sample correlations of individual AR(1) processes. Consistency and asymptotic normality of the empirical distribution function and a class of kernel density estimators is established under some regularity conditions on G(x) as N and n increase to infinity. An extension of this work consists in testing the presence of long memory. The procedure is based on the tail index of G and the théorie of extreme values. In the same direction, a new frequency-domain test statistic is introduced to test for short memory versus long memory, see [23]

6.4. New Bayesian approach for chronological modeling

Participants: Anne. Philippe and Marie-Anne Vibet in collaboration with IRAMAT - Université Bordeaux Montaigne

We have been working on the construction of new Bayesian approach for chronological modeling: this is an important issue in archaeology and paleo-environmental sciences. The proposed solution is based on the "event model'. We define the Event as the date of an archeological context determined from a collection of contemporaneous artifacts. We obtain a robust approach with respect to outliers due to the sampling in the field or the measurement process in the laboratory.

In [25] We propose new tools to analyse the chronologies especially regarding phases. They are implemented in R package 'RChronoModel'.

In [22], [18], [5], we propose bayesian models for optically stimulated luminescence dating. It consists in estimating a central equivalent dose from a set of luminescence measurements. Then a calibration step is required to convert equivalent dose into calendar date.

6.5. Self-regulated processes

Participants : Jacques Lévy Véhel, Anne Philippe, Caroline Robet

We wish to construct various instances of processes Z such that, at each point t, almost surely, the pointwise Hölder exponent of Z at t, denoted $\alpha_Z(t)$, verifies

$$\alpha_Z(t) = g(Z(t))$$

where $g \in \mathcal{C}^1(\mathbb{R}, [a, b])$ is a deterministic function. Then, we would estimate the function g which control the regularity.

The pointwise Hölder exponent at t of a function or a process $f : \mathbb{R} \to \mathbb{R}$, which is $\mathcal{C}^1$ nowhere, is the real $\alpha_f(t)$ such that :

$$\alpha_f(t) = \sup\{\beta, \limsup_{h\to 0} \frac{\mid f(t+h) - f(t) \mid}{\mid h \mid^\beta} = 0\}$$

We worked first on pathwise integrals :

Theorem 1 Let $g \in \mathcal{C}^1(\mathbb{R}, [a, b])$, $0 < a < b < 1$. Provided $\|g'\|_\infty$ is small enough, there exists a unique continuous process Z verifying almost surely on $[0, T]$

$$Z_t = \int_0^t (t-u)^{g(Z_u)-1} W_u \, du$$

where W is an almost surely continuous process.

A random condition ($\|g'\|_\infty \|W(\omega)\|_\infty C(a, T) < 1$) appears in the application of Banach fixed point theorem (in $(\mathcal{C}^0([0, T]; \mathbb{R}), \|.\|_\infty)$). It implies that it is possible to have existence et uniqueness only on $[0, t']$, $t' < T$. We simulated pathwise integrals and showed some cases without uniqueness. We studied some easier processes in order to find the regularity of Z.

Theorem 2 Let $h \in]0, 1[$ and U defined on $[0, T]$ by

$$U_t = \int_0^t (t-u)^{h-1} W_u \, du$$

Then $\forall t \in [0, T]$, $\alpha_U(t) \geq h$.

Theorem 3 Let $g \in \mathcal{C}^1(\mathbb{R}, [a, b])$, $0 < a < b < 1$. Provided $\|g'\|_\infty$ is small enough, there exists a unique continuous process Y verifying almost surely on $[0, T]$

$$Y_t = \int_0^t (t-u)^{g(Y_t)-1} W_u \, du$$

where W is an almost surely continuous process. Furthermore, $\forall t \in [0, T]$, $\alpha_Y(t) \geq g(Y_t)$

Then, we adapted the multifractional Brownian Motion [50], [31] (which a representation is $B_t = \int_0^t K_{H(t)}(t, u)W(du)$, W Brownian Motion et H$\in \mathcal{C}^1$) to construct the modified multifractional Brownian Motion : $Z_t = \int_0^t K_{H(u)}(t, u)W(du)$. We expect obtain a self-regulated process $Y_t = \int_0^t K_{g(Y_u)}(t, u)dW(u)$.

Theorem 4 Let $g \in \mathcal{C}^1(\mathbb{R}, [a, b])$, $0 < a < b < 1$. Provided $\|g'\|_\infty$ is small enough, there exists a unique continuous adapted process Y include in $\mathcal{C}^0([0, T]; L^2(\Omega))$ verifying almost surely on $[0, T]$

$$Y_t = \int_0^t K_{g(Y_u)}(t, u)dW(u)$$

where W is the Brownian motion.

6.6. Causal inference by independent component analysis with Application of to American macro-economic data

Participants : Jacques Lévy-Vehel, Anne Philippe, Marie-Anne Vibet

The aim of this work is to study the causal relationships existing among macro-economic variables under investigation, and trace out how economically interpreted random shocks affect the system. Structural vector of autoregressive models (SVAR) are usually applied in this kind of study and the causal structure is driven by the data. In this work, independent component analysis (ICA) is implemented in order to guaranty the identifiability of the causal structure. However, the use of ICA can only be done under the hypothesis that the residuals are non-Gaussian, an hypothesis easily verified with economic data.

The vector of autoregressive (VAR) model has the following reduced representation :

$$Y_t = A_1 Y_{t-1} + ... + A_p Y_{t-p} + u_t \text{ , for } t = 1, ..., T$$

where, Y_t, is the vector of contemporaneous variables of dimension $K \times 1$, p is the number of autoregressive variables, A_j, for $j = 1, ..., p$, are matrices of dimension $K \times K$ estimated by the model and u_t is the vector of random disturbances of dimension $K \times 1$ and assumed to be a zero-mean white noise process, $u_t \sim N(0_K, \Sigma_u)$. Given enough data, both Σ_u and all matrices A_j can be corretly estimated by the VAR model.

However, the VAR model is not sufficient for policy analysis. Indeed, using the Moving Average representation of a stable VAR :

$$Y_t = \sum_{j=0}^{\infty} \Phi_j u_{t-j} \tag{1}$$

where $\Phi_0 = I_K$ and $\Phi_j, j \geq 1$, are the coefficients matrices representing the impulse responses of the elements of Y_t to the disturbances u_{t-j}. This representation is not unique.

The structiral VAR (SVAR) is essentially a VAR equipped with a particular choice of a matrix P so that $Y_t = \sum_{j=0}^{\infty} \Phi_j P P^{-1} u_{t-j} = \sum_{j=0}^{\infty} \Psi_j \epsilon_{t-j}$

where ϵ_{t-j} are independent random shocks economically interpreted. To this aim, the ICA procedure is then used to find the proper matrix P using the hypothesis that the residuals, ϵ_{t-j}, are non-Gaussian.

We used the VAR-LINGAM procedure developped by Moneta *et al* [28] and their package written for R software. We started by testing this procedure with a series of simulations study. We tackled the following questions : Are the coefficients of the matrices B and A well estimated by the VARLINGAM procedure ? Is the bootstrap function appropriate, and in particular, does it estimate properly the standard error of the coefficients of matrices A and B ? And how long should the economic data be in order to estimate correctly the coefficients of the matrices B ?

As the conclusion to all these studies were correct enough, we went on analysing our real data that consists of 6 weekly time series US macro-economic data, reported from the first week of January 1996 to April 2016 : The BofA Merrill Lynch US Broad Market Index, The Bofa Merril Lynch US Corporate Index, Equity Indices S&P, 500, Federal Funds Rates,Treasury Bills, Other Factors Draining Reserve Balances.

The conclusions of this work is in discussion with economists and a paper will soon be written.

6.7. SAR image denoising using an irregularity-preserved denoising technique based on the global Höllder exponent

Participants : Jacques Lévy-Vehel, Yue Huang

This work addresses the speckle noise reduction for SAR images by using the irregularity-preserved denoising technique proposed in [34]. This irregularity preserving denoising scheme in [34] may be summarized as a three-step process in the following:

1. Apply a Discrete Wavelet Transform (DWT) on the noisy signal and represent the resulting coefficients distribution over scales. Estimate the cut-off scale and the global Hölder exponent α_f using linear regression of $\max_k \left(\log_2 | \langle f, \psi_{j,k} \rangle | \right)$ at larger scales.
2. Extrapolate the larger scale regression line to smaller scales and limit coefficient at smaller scales ($j \geq j_{\text{cut-off}}$) to the boundary value obtained from the linear regression
3. Reconstruct the filtered signal from the set of modified coefficients

where f is the signal under analysis, $\psi_{j,k}$ is the wavelet basis, and $\langle f, \psi_{j,k} \rangle$ is the wavelet coefficient of f at scale j and location k. As it has been shown by simulations in [34], to retrieve irregular signals affected by additive noise, this technique outperforms conventional denoising techniques that apply hard or soft thresholding to the wavelet coefficients.

Considering a speckle-affected SAR image, a complex SAR signal may be represented by:

$$y(l) = s(l)u(l)$$

where l represents one of L realizations, and the noise term $u(l)$ follows a complex circular centered Gaussian white distribution with unit variance, i.e. $u \sim \mathcal{N}_C(0,1)$, $E(u(i)u^*(j)) = \delta_{(i-j)}$. The texture of SAR image significantly depends on the backscattering power $\sigma(l) = |y(l)|^2$.

We aim to use the irregularity-preserved denoising technique to denoise SAR image and enhance its texture. We tested firstly on the simulated signals affected by multiplicative noise and then on real SAR images. This denoising scheme showed potential to reduce the speckle noise, preserve the irregularity of image texture and enhance target signature.

Although the results have been compared with other SAR speckle filtering techniques, we still need more efforts for validation. As long as the results are validated, the work will be written in a paper.

6.8. Underfoliage object imaging using SAR tomography and wavelet-based sparse estimation methods

Participants : Yue Huang, Jacques Lévy-Vehel

Hybrid environments refer to a scenario of deterministic objects embedded in a host natural random environment and their scattering patterns consist of a complex mixture of diverse mechanisms, like, in the case of this study, volume scattering from the canopy, double bounce reflection between the ground and under-foliage objects as well as between objects and trunks, surface scattering from the underlying ground, etc. The resulting SAR information is characterized by a strong complexity, and its analysis using 2-D images or even data acquired in InSAR configuration remains difficult. Using Multi-baseline(MB) InSAR data, SAR tomography can be applied to reconstruct in 3-D the measured scattering responses and polarimetric patterns. Natural volumes, such as forest canopies, being composed of a large number of scatterers whose responses cannot be discriminated at the resolution of analysis, their scattering patterns are generally considered as a vertical density of random or speckle-affected reflectivity. On the other hand, localized objects, such as artificial targets on the ground are associated to point-like contributions, that may be separable in the vertical direction. The global response of under-foliage objects with a deterministic scattering response embedded in surrounding distributed environments, can be described by a mixed spectrum. Conventional tomographic techniques like the Capon and Beamforming methods, estimate continuous Power Spectral Density (PSD) and hence are well adapted to the characterization of continuous volumetric media, but cannot discriminate closely-spaced scatterers, e.g. scattering responses from trucks, due to limited spatial resolution. Conventional high-resolution methods like MUSIC and subspace fitting estimators as well as sparse estimation techniques such as LASSO

[52] and FOCUSS [40], are well adapted to the characterization of discrete scatterers like truck top, truck-ground interaction and calibrators over bare soils, or buildings over urban areas [53], but cannot properly handle the high dimensionality of the scattering responses of natural volumes. Usual tomographic techniques cannot simultaneously cope with both types of spectrum, and not able to deal with mixed spectral estimation problems, characteristic of underfoliage object imaging scenario.

Wavelet-based techniques present a high potential for such applications, as they permit to parameterize in a sparse way continuous functions, i.e. canopy PSDs in the present case. Wavelet-based tomographic techniques have been used for tomographic imaging of forested areas [27], and for such regular signals, large wavelet coefficients being often concentrated in the approximation space, scale thresholding may be implemented to extract the most significant wavelet coefficients for an accurate volume signal recovery [27]. In the underfoliage object scenario, discrete scatterers embedded in a continuous medium, result in a mixed vertical PSD that may be associated to an irregular signal with wavelet coefficients distributed both in the approximation and detail spaces, and a simple scale cut-off is hence not adapted to separate the wavelet coefficients of discrete scatterers from those of continuous media. Therefore, we propose a new wavelet-based method to extract underfoliage objects from their speckle-affected distributed environment and characterize them with a high resolution.

For an MB-InSAR configuration with M acquisition positions, considering an azimuth-range resolution cell containing a mixture of backscattering contributions from object (o) and volume (v) scatterers located at different heights z, the observed data vector at lth realization can be represented by:

$$\mathbf{y}(l) = \mathbf{A}_o(\mathbf{z}_o)\mathbf{s}_o(l) + \mathbf{A}_v(\mathbf{z}_v)\mathbf{s}_v(l) + \mathbf{n}(l) \tag{2}$$

where the steering matrix, $\mathbf{A}_x(\mathbf{z}_x)$, contains the interferometric phase information associated to the InSAR responses of the scatterers located at the unknown elevation positions $\mathbf{z}_x = [z_{x_1}, \cdots, z_{x_{N_x}}]$ above the reference focusing plane, and the source signal vector, $\mathbf{s}_x = [s_{x_1} \cdots s_{x_{N_x}}]^T \in \mathbb{C}^{N_x \times 1}$, contains the unknown complex backscattering coefficients of the N_x source scatterers. The vertical reflectivity function can be represented as $\mathbf{p}_x = E(|\mathbf{s}_x|^2)$ $(x = o, v)$.

Over speckle-affected environments, unknown reflectivity and elevation parameters are generally estimated from second-order statistics, i.e. from the covariance matrix $\widehat{\mathbf{R}} \in \mathbb{C}^{M \times M}$ of the observed MB-InSAR data $\mathbf{y} \in \mathbb{C}^{M \times 1}$. The proposed tomographic processing technique is based on the minimization of the Least-Square (LS) fitting between the observed and modeled data covariance $||\mathbf{R} - \widehat{\mathbf{R}}||_F$. The modeled covariance matrix is composed by the covariances of object and volume contributions $\mathbf{R} = \mathbf{R}_o + \mathbf{R}_v$, each of them being simply related to its discretized vertical density of reflectivity $\mathbf{p}_x$ through $\mathbf{R}_x = \mathbf{A}(\mathbf{z}_x)\,\mathrm{diag}\,(\mathbf{p}_x)\mathbf{A}^H(\mathbf{z}_x) \in \mathbb{C}^{M \times M}$. The proposed method can be represented by a l_1 norm minimization in a transformed space subject to quadratic constraints between the observed and modeled data covariance:

$$\min_{\mathbf{p}} ||\mathbf{B}\mathbf{p}||_1 \text{ subject to } \quad ||\mathbf{R} - \widehat{\mathbf{R}}||_F \leq \epsilon \tag{3}$$

where

- $\mathbf{p} = [\mathbf{p}_o^T \quad \mathbf{p}_v^T]^T \in \mathbb{R}^{+N_s \times 1}$ stands for vertical backscattering power distribution for the resolution cell under analysis,
- $\mathbf{B} = \begin{bmatrix} \mathbf{I}_{(N_o \times N_o)} & \mathbf{0} \\ \mathbf{0} & \Psi_{(N_v \times N_v)} \end{bmatrix} \in \mathbb{R}^{(N_s \times N_s)}$ represents the hybrid sparsifying basis with the wavelet basis Ψ

This tomographic technique is suitable for the mixed-spectrum estimation problem, because it maintains the spectral continuity for the backscattering power of forest canopies and the high-resolution for the vertical reflectivity of objects. The effectiveness of this new approach is demonstrated using L-band airborne tomographic SAR data aquired by the DLR over Dornstetten, Germany. The undeniable performance can be shown by the results in [21] and [20].

This work has been presented in European SAR conference 2016 . Some refined results have been presented in IGARSS conference 2016 as an invited talk. By extending this work in details, a journal paper [24] has been submitted to IEEE Geoscience and Remote Sensing Letters (GRSL) and is currently under reviewing.

6.9. Detection of objects concealed beneath forest canopies using Time-Frequency techniques

Participants : Yue Huang, Jacques Lévy-Vehel

In the scenario of hybrid environments where objects with a deterministic response are embedded in a speckle affected environment, the parameter estimation for this type of scatterers becomes a problem of mixed-spectrum estimation. To isolate and characterize these different scattering contributions, a novel method proposed by Huang et al. was used to extract isolated scatterers (IS) from their surrounding distributed environments, named IS extraction in [42]. Incorporating the Weighted Subspace Fitting (WSF) estimator, this method estimated scattering responses within one resolution cell and then distinguishes isolated scatterers from distributed ones by calculating the cross-correlation between the measured data and the estimated scattering responses. Moreover, to compare the detection performance for coherent scatterers, two statistical methods have been applied to analyse hybrid environments in [43]: GLRT (generalized likelihood ratio test)-based and SSF (weighted Signal Subspace Fitting)-based detection procedures. However, the above mentioned methods based on discrete high-resolution tomographic estimation, require to preselect the number of scattering contributions, which may induce reliability issues due to model order selection.

This paper proposes a new tomographic estimator based on Time-Frequency (TF) techniques using Multi-baseline Polarimetric and Interferometric SAR data. The coherent TF analysis of polarimetric SAR has been introduced in [38], [39] for the study of anisotropic scattering behaviors and then applied in [37], [36] for dense urban environment characterization. Time-frequency techniques can represent spectral properties around specific spatial locations or spatial features at specific spectral positions, leading to describe local variations of spectral or spatial features. Considering SLC SAR images, the spectral locations can be linked to azimuth looking angle and illumination frequency in such a way:

$$w_{az} = \frac{4\pi}{c} f_c v_{SAR} \sin\phi, \quad w_{rg} = \frac{4\pi}{c}(f - f_c)$$

with f_c central frequency and ϕ azimuth looking angle. The TF technique can be used to analyze scattering behaviors at different illuminated positions and frequency components during SAR integration. Based on the correlation between different spectral positions, the TF indicator proposed in [37] can extract coherent components in complex random SAR responses. Polarimetric TF indicator has been developed in [41] for ship discrimination. In this paper, the new tomographic estimator extends 2-D TF analysis to 3-D, which provides an efficient cancellation for clutters from speckle-affected random scattering environments, and discriminates the deterministic responses from coherent scatterers in 3-D space. The effectiveness of this new tomographic approach is demonstrated by using L-band MB-PolInSAR data set acquired over the test site of Dornstetten where the underfoliage objects are set up. The fully polarimetric version of this TF tomographic estimator is also developed to improve the detection efficiency. This work has been accepted for oral presentation at the Polinsar 2017 Workshop and the final paper will be written by the end of Workshop.

ASAP Project-Team

6. New Results

6.1. Theory of Distributed Systems

6.1.1. t-Resilient Immediate Snapshot is Impossible

Participant: Michel Raynal.

Immediate snapshot is the basic communication object on which relies the read/write distributed computing model made up of n crash-prone asynchronous processes, called iterated distributed model. Each iteration step (usually called a round) uses a new immediate snapshot object, which allows the processes to communicate and cooperate. More precisely, the x-th immediate snapshot object can be used by a process only when it executes the x-th round. An immediate snapshot object can be implemented by an $(n-1)$-resilient algorithm, i.e. an algorithm that tolerates up to $(n-1)$ process crashes (also called wait-free algorithm). Considering a t-crash system model (i.e. a model in which up to t processes are allowed to crash), this work [46] is on the construction of an extension of immediate snapshot objects to t-resiliency. In the t-crash system model, at each round each process may be ensured to get values from at least $n-t$ processes, and t-immediate snapshot has the properties of classical immediate snapshot (1-immediate snapshot) but ensures that each process will get values form at least $n-t$ processes. Its main result is the following. While there is a (deterministic) t-resilient read/write-based algorithm implementing t-immediate snapshot in a t-crash system when $t = n-1$, there is no t-resilient algorithm in a t-crash model when $t \in [1..(n-2)]$. This means that the notion of t-resilience is inoperative when one has to implement immediate snapshot for these values of t: the model assumption "at most $t < n-1$ processes may crash" does not provide us with additional computational power allowing for the design of genuine t-resilient algorithms (genuine meaning that such a t-resilient algorithm would work in the t-crash model, but not in the $(t+1)$-crash model). To show these results, the paper relies on well-known distributed computing agreement problems such as consensus and k-set agreement.

This work was done in collaboration with Carole Delporte, Hugues Fauconnier, and Sergio Rajsbaum, and appeared at SIROCCO 2016.

6.1.2. Two-Bit Messages are Sufficient to Implement Atomic Read/Write Registers in Crash-Prone Systems

Participant: Michel Raynal.

Atomic registers are certainly the most basic objects of computing science. Their implementation on top of an n-process asynchronous message-passing system has received a lot of attention. It has been shown that $t < n/2$ (where t is the maximal number of processes that may crash) is a necessary and sufficient requirement to build an atomic register on top of a crash-prone asynchronous message-passing system. Considering such a context, this work [49] presents an algorithm which implements a single-writer multi-reader atomic register with four message types only, and where no message needs to carry control information in addition to its type. Hence, two bits are sufficient to capture all the control information carried by all the implementation messages. Moreover, the messages of two types need to carry a data value while the messages of the two other types carry no value at all. As far as we know, this algorithm is the first with such an optimality property on the size of control information carried by messages. It is also particularly efficient from a time complexity point of view.

This work was done in collaboration with Achour Mostefaoui, and appeared at PODC 2016.

6.2. Network and Graph Algorithms

6.2.1. Vertex Coloring with Communication and Local Memory Constraints in Synchronous Broadcast Networks

Participants: Hicham Lakhlef, Michel Raynal, Francois Taiani.

This work [41] considers the broadcast/receive communication model in which message collisions and message conflicts can occur because processes share frequency bands. (A collision occurs when, during the same round, messages are sent to the same process by too many neighbors. A conflict occurs when a process and one of its neighbors broadcast during the same round.) More precisely, this work considers the case where, during a round, a process may either broadcast a message to its neighbors or receive a message from at most m of them. This captures communication-related constraints or a local memory constraint stating that, whatever the number of neighbors of a process, its local memory allows it to receive and store at most m messages during each round. This work defines first the corresponding generic vertex multi-coloring problem (a vertex can have several colors). It focuses then on tree networks, for which it presents a lower bound on the number of colors K that are necessary (namely, $K = \lceil \frac{\Delta}{m} \rceil + 1$, where Δ is the maximal degree of the communication graph), and an associated coloring algorithm, which is optimal with respect to K.

6.2.2. Optimal Collision/Conflict-Free Distance-2 Coloring in Wireless Synchronous Broadcast/Receive Tree Networks

Participants: Davide Frey, Hicham Lakhlef, Michel Raynal.

We studied the problem of decentralized distance-2 coloring in message-passing systems where communication is (a) synchronous and (b) based on the "broadcast/receive" pair of communication operations. "Synchronous" means that time is discrete and appears as a sequence of time slots (or rounds) such that each message is received in the very same round in which it is sent. "Broadcast/receive" means that during a round a process can either broadcast a message to its neighbors or receive a message from one of them. In such a communication model, no two neighbors of the same process, nor a process and any of its neighbors, must be allowed to broadcast during the same time slot (thereby preventing message collisions in the first case, and message conflicts in the second case). From a graph theory point of view, the allocation of slots to processes is know as the distance-2 coloring problem: a color must be associated with each process (defining the time slots in which it will be allowed to broadcast) in such a way that any two processes at distance at most 2 obtain different colors, while the total number of colors is "as small as possible". In this context, we proposed a parallel message-passing distance-2 coloring algorithm suited to trees, whose roots are dynamically defined. This algorithm, which is itself collision-free and conflict-free, uses $\Delta + 1$ colors where Δ is the maximal degree of the graph (hence the algorithm is color-optimal). It does not require all processes to have different initial identities, and its time complexity is $O(d\Delta)$, where d is the depth of the tree. As far as we know, this is the first distributed distance-2 coloring algorithm designed for the broadcast/receive round-based communication model, which owns all the previous properties. We published these results in [39].

6.2.3. Efficient Plurality Consensus, or: The Benefits of Cleaning Up from Time to Time

Participant: George Giakkoupis.

Plurality consensus considers a network of n nodes, each having one of k opinions. Nodes execute a (randomized) distributed protocol with the goal that all nodes adopt the *plurality* (the opinion initially supported by the most nodes). Communication is realized via the random phone call model. A major open question has been whether there is a protocol for the complete graph that converges (w.h.p.) in polylogarithmic time and uses only polylogarithmic memory per node (local memory). We answered this question affirmatively.

In [22], we propose two protocols that need only mild assumptions on the bias in favor of the plurality. As an example of our results, consider the complete graph and an arbitrarily small constant multiplicative bias in favor of the plurality. Our first protocol achieves plurality consensus in $O(\log k \cdot \log\log n)$ rounds using $\log k + O(\log\log k)$ bits of local memory. Our second protocol achieves plurality consensus in $O(\log n \cdot \log\log n)$ rounds using only $\log k + 4$ bits of local memory. This disproves a conjecture by Becchetti et al. (SODA'15) implying that any protocol with local memory $\log k + O(1)$ has worst-case runtime $\Omega(k)$. We provide similar bounds for much weaker bias assumptions. At the heart of our protocols lies an *undecided state*, an idea introduced by Angluin et al. (Distributed Computing'08).

This work was done in collaboration with Petra Berenbrink (SFU), Tom Friedetzky (Durham University), and Peter Kling (SFU).

6.2.4. Bounds on the Voter Model in Dynamic Networks

Participants: George Giakkoupis, Anne-Marie Kermarrec.

In the *voter model*, each node of a graph has an opinion, and in every round each node chooses independently a random neighbour and adopts its opinion. We are interested in the *consensus time*, which is the first point in time where all nodes have the same opinion. In [23], we consider dynamic graphs in which the edges are rewired in every round (by an adversary) giving rise to the graph sequence $G_1, G_2, \cdots$, where we assume that G_i has conductance at least ϕ_i. We assume that the degrees of nodes don't change over time as one can show that the consensus time can become super-exponential otherwise. In the case of a sequence of d-regular graphs, we obtain asymptotically tight results. Even for some static graphs, such as the cycle, our results improve the state of the art. Here we show that the expected number of rounds until all nodes have the same opinion is bounded by $O(m/(\delta \cdot \phi))$, for any graph with m edges, conductance ϕ, and degrees at least δ. In addition, we consider a *biased* dynamic voter model, where each opinion i is associated with a probability P_i, and when a node chooses a neighbour with that opinion, it adopts opinion i with probability P_i (otherwise the node keeps its current opinion). We show for any regular dynamic graph, that if there is an $\epsilon > 0$ difference between the highest and second highest opinion probabilities, and at least $\Omega(\log n)$ nodes have initially the opinion with the highest probability, then all nodes adopt w.h.p. that opinion. We obtain a bound on the convergence time, which becomes $O(\log n/\phi)$ for static graphs.

This work was done in collaboration with Petra Berenbrink (SFU), and Frederik Mallmann-Trenn (SFU).

6.2.5. How Asynchrony Affects Rumor Spreading Time

Participant: George Giakkoupis.

In standard randomized (push-pull) rumor spreading, nodes communicate in synchronized rounds. In each round every node contacts a random neighbor in order to exchange the rumor (i.e., either push the rumor to its neighbor or pull it from the neighbor). A natural asynchronous variant of this algorithm is one where each node has an independent Poisson clock with rate 1, and every node contacts a random neighbor whenever its clock ticks. This asynchronous variant is arguably a more realistic model in various settings, including message broadcasting in communication networks, and information dissemination in social networks.

In [35] we study how asynchrony affects the rumor spreading time, that is, the time before a rumor originated at a single node spreads to all nodes in the graph. Our first result states that the asynchronous push-pull rumor spreading time is asymptotically bounded by the standard synchronous time. Precisely, we show that for any graph G on n-nodes, where the synchronous push-pull protocol informs all nodes within $T(G)$ rounds with high probability, the asynchronous protocol needs at most time $O(T(G) + \log n)$ to inform all nodes with high probability. On the other hand, we show that the expected synchronous push-pull rumor spreading time is bounded by $O(\sqrt{n})$ times the expected asynchronous time.

These results improve upon the bounds for both directions shown recently by Acan et al. (PODC 2015). An interesting implication of our first result is that in regular graphs, the weaker push-only variant of synchronous rumor spreading has the same asymptotic performance as the synchronous push-pull algorithm.

This work was done in collaboration with Yasamin Nazari and Philipp Woelfel from the University of Calgary.

6.2.6. Amplifiers and Suppressors of Selection for the Moran Process on Undirected Graphs

Participant: George Giakkoupis.

In [47] we consider the classic Moran process modeling the spread of genetic mutations, as extended to structured populations by Lieberman et al. (Nature, 2005). In this process, individuals are the vertices of a connected graph G. Initially, there is a single mutant vertex, chosen uniformly at random. In each step, a random vertex is selected for reproduction with a probability proportional to its fitness: mutants have fitness $r > 1$, while non-mutants have fitness 1. The vertex chosen to reproduce places a copy of itself to a uniformly random neighbor in G, replacing the individual that was there. The process ends when the mutation either reaches fixation (i.e., all vertices are mutants), or gets extinct. The principal quantity of interest is the probability with which each of the two outcomes occurs.

A problem that has received significant attention recently concerns the existence of families of graphs, called strong amplifiers of selection, for which the fixation probability tends to 1 as the order n of the graph increases, and the existence of strong suppressors of selection, for which this probability tends to 0. For the case of directed graphs, it is known that both strong amplifiers and suppressors exist. For the case of undirected graphs, however, the problem has remained open, and the general belief has been that neither strong amplifiers nor suppressors exist. In this work we disprove this belief, by providing the first examples of such graphs. The strong amplifier we present has fixation probability $1 - \widetilde{O}(n^{-1/3})$, and the strong suppressor has fixation probability $\widetilde{O}(n^{-1/4})$. Both graph constructions are surprisingly simple. We also prove a general upper bound of $1 - \widetilde{\Omega}(n^{-1/3})$ on the fixation probability of any undirected graph. Hence, our strong amplifier is existentially optimal.

6.3. Scalable Systems

6.3.1. Cache locality is not enough: High-Performance Nearest Neighbor Search with Product Quantization Fast Scan

Participants: Fabien Andre, Anne-Marie Kermarrec.

Nearest Neighbor (NN) search in high dimension is an important feature in many applications (e.g., image retrieval, multimedia databases). Product Quantization (PQ) is a widely used solution which offers high performance, i.e., low response time while preserving a high accuracy. PQ represents high-dimensional vectors (e.g., image descriptors) by compact codes. Hence, very large databases can be stored in memory, allowing NN queries without resorting to slow I/O operations. PQ computes distances to neighbors using cache-resident lookup tables, thus its performance remains limited by (i) the many cache accesses that the algorithm requires, and (ii) its inability to leverage SIMD instructions available on modern CPUs. In this paper, we advocate that cache locality is not sufficient for efficiency. To address these limitations, in [19] we design a novel algorithm, PQ Fast Scan, that transforms the cache-resident lookup tables into small tables, sized to fit SIMD registers. This transformation allows (i) in-register lookups in place of cache accesses and (ii) an efficient SIMD implementation. PQ Fast Scan has the exact same accuracy as PQ, while having 4 to 6 times lower response time (e.g., for 25 million vectors, scan time is reduced from 74ms to 13ms).

6.3.2. Toward an Holistic Approach of Systems-of-Systems

Participants: Simon Bouget, David Bromberg, Francois Taiani.

Large scale distributed systems have become ubiquitous, from on-line social networks to the Internet-of-Things. To meet rising expectations (scalability, robustness, flexibility,...) these systems increasingly espouse complex distributed architectures, that are hard to design, deploy and maintain. To grasp this complexity, developers should be allowed to assemble large distributed systems from smaller parts using a seamless, high-level programming paradigm. We present in [24] such an assembly-based programming framework, enabling developers to easily define and realize complex distributed topologies as a construction of simpler blocks (e.g. rings, grids). It does so by harnessing the power of self-organizing overlays, that is made accessible to developers through a high-level Domain Specific Language and self-stabilizing run-time. Our evaluation further shows that our approach is generic, expressive, low-overhead and robust.

6.3.3. Speed for the Elite, Consistency for the Masses: Differentiating Eventual Consistency in Large-Scale Distributed Systems

Participants: Davide Frey, Pierre-Louis Roman, Francois Taiani.

Eventual consistency is a consistency model that emphasizes liveness over safety; it is often used for its ability to scale as distributed systems grow larger. Eventual consistency tends to be uniformly applied to an entire system, but we argue that there is a growing demand for differentiated eventual consistency requirements.

We address this demand with UPS [34], a novel consistency mechanism that offers differentiated eventual consistency and delivery speed by working in pair with a two-phase epidemic broadcast protocol. We propose a closed-form analysis of our approach's delivery speed, and we evaluate our complete mechanism experimentally on a simulated network of one million nodes. To measure the consistency trade-off, we formally define a novel and scalable consistency metric that operates at runtime. In our simulations, UPS divides by more than 4 the inconsistencies experienced by a majority of the nodes, while reducing the average latency incurred by a small fraction of the nodes from 6 rounds down to 3 rounds.

This work was done in collaboration with Achour Mostefaoui and Matthieu Perrin from the LINA laboratory in Nantes.

6.3.4. Bringing Secure Bitcoin Transactions to your Smartphone

Participants: Davide Frey, Pierre-Louis Roman, Francois Taiani.

To preserve the Bitcoin ledger's integrity, a node that joins the system must download a full copy of the entire Bitcoin blockchain if it wants to verify newly created blocks. At the time of writing, the blockchain weights 79 GiB and takes hours of processing on high-end machines. Owners of low-resource devices (known as thin nodes), such as smartphones, avoid that cost by either opting for minimum verification or by depending on full nodes, which weakens their security model.

In this work [33], we propose to harden the security model of thin nodes by enabling them to verify blocks in an adaptive manner, with regards to the level of targeted confidence, with low storage requirements and a short bootstrap time. Our approach exploits sharing within a distributed hash table (DHT) to distribute the storage load, and a few additional hashes to prevent attacks on this new system.

This work was done in collaboration with Marc X. Makkes and Spyros Voulgaris from Vrije Universiteit Amsterdam (The Netherlands).

6.3.5. Multithreading Approach to Process Real-Time Updates in KNN Algorithms

Participants: Anne-Marie Kermarrec, Nupur Mittal, Javier Olivares.

K-Nearest Neighbors algorithm is the core of a considerable amount of online services and applications, like recommendation engines, content-classifiers, information retrieval systems, etc. The users of these services change their preferences and evolve with time, aggravating the computational challenges of KNN more with the ever evolving data to process. In this work [48], we present *UpKNN*: an efficient thread-based approach to take the updates of users preferences into account while it computes the KNN efficiently, keeping a check on the wall-time.

By using an efficient thread-based approach, *UpKNN* processes millions of updates online, on a single commodity PC. Our experiments confirm the scalability of *UpKNN*, both in terms of the number of updates processed and the threads used. *UpKNN* achieves speedups ranging from 13.64X to 49.5X in the processing of millions of updates, with respect to the performance of a non-partitioned baseline. These results are a direct consequence of reducing the number of disk operations, roughly speaking, only 1% disk operations are performed as compared to the baseline.

6.3.6. The Out-of-Core KNN Awakens: The Light Side of Computation Force on Large Datasets

Participants: Anne-Marie Kermarrec, Javier Olivares.

K-Nearest Neighbors (KNN) is a crucial tool for many applications, e.g. recommender systems, image classification and web-related applications. However, KNN is a resource greedy operation particularly for large datasets. We focus on the challenge of KNN computation over large datasets on a single commodity PC with limited memory. We propose a novel approach [27] to compute KNN on large datasets by leveraging both disk and main memory efficiently. The main rationale of our approach is to minimize random accesses to disk, maximize sequential accesses to data and efficient usage of only the available memory.

We evaluate our approach on large datasets, in terms of performance and memory consumption. The evaluation shows that our approach requires only 7% of the time needed by an in-memory baseline to compute a KNN graph.

6.3.7. Partial Replication Policies for Dynamic Distributed Transactional Memory in Edge Clouds

Participant: Francois Taiani.

Distributed Transactional Memory (DTM) can play a fundamental role in the coordination of participants in edge clouds as a support for mobile distributed applications. DTM emerges as a concurrency mechanism aimed at simplifying distributed programming by allowing groups of operations to execute atomically, mirroring the well-known transaction model of relational databases. In spite of recent studies showing that partial replication approaches can present gains in the scalability of DTMs by reducing the amount of data stored at each node, most DTM solutions follow a full replication scheme. The few partial replicated DTM frameworks either follow a random or round-robin algorithm for distributing data onto partial replication groups. In order to overcome the poor performance of these schemes, this work [36] investigates policies to extend the DTM to efficiently and dynamically map resources on partial replication groups. The goal is to understand if a dynamic service that constantly evaluates the data mapped into partial replicated groups can contribute to improve DTM based systems performance.

This work was performed in collaboration with Diogo Lima and Hugo Miranda from the University of Lisbon (Portugal).

6.3.8. Being Prepared in a Sparse World: The Case of KNN Graph Construction

Participants: Anne-Marie Kermarrec, Nupur Mittal, Francois Taiani.

Work [25] presents KIFF, a generic, fast and scalable KNN graph construction algorithm. KIFF directly exploits the bipartite nature of most datasets to which KNN algorithms are applied. This simple but powerful strategy drastically limits the computational cost required to rapidly converge to an accurate KNN solution, especially for sparse datasets. Our evaluation on a representative range of datasets show that KIFF provides, on average, a speed-up factor of 14 against recent state-of-the art solutions while improving the quality of the KNN approximation by 18

This work was done in collaboration with Antoine Boutet from CNRS, Laboratoire Hubert Curien, Saint-Etienne, France.

6.3.9. Exploring the Use of Tags for Georeplicated Content Placement

Participants: Stephane Delbruel, Davide Frey, Francois Taiani.

A large portion of today's Internet traffic originates from streaming and video services. Such services rely on a combination of distributed datacenters, powerful content delivery networks (CDN), and multi-level caching . In spite of this infrastructure, storing, indexing, and serving these videos remains a daily engineering challenge that requires increasing efforts on the part of providers and ISPs. In this work [30], we explore how the tags attached to videos by users could help improve this infrastructure, and lead to better performance on a global scale. Our analysis shows that tags can be interpreted as markers of a video's geographic diffusion, with some tags strongly linked to well identified geographic areas. Based on our findings, we demonstrate the potential of tags to help predict distribution of a video's views, and present results suggesting that tags can help place videos in globally distributed datacenters. We show in particular that even a simplistic approach based on tags can help predict a minimum of 65.9% of a video's views for a majority of videos, and that a simple tag-based placement strategy is able to improve the hit rate of a distributed on-line video service by up to 6.8% globally over a naive random allocation.

6.3.10. Mignon: A Fast Decentralized Content Consumption Estimation in Large-Scale Distributed Systems

Participants: Stephane Delbruel, Davide Frey, Francois Taiani.

Although many fully decentralized content distribution systems have been proposed, they often lack key capabilities that make them difficult to deploy and use in practice. In this work [31], we look at the particular problem of content consumption prediction, a crucial mechanism in many such systems. We propose a novel, fully decentralized protocol that uses the tags attached by users to on-line content, and exploits the properties of self-organizing kNN overlays to rapidly estimate the potential of a particular content without explicit aggregation.

6.4. Privacy in User Centric Applications

6.4.1. Hybrid Recommendations with Dynamic Similarity Measure

Participants: Anne-Marie Kermarrec, Nupur Mittal.

This project aims to combine the classical methods of content based and collaborative filtering recommendations, in addition to dynamic similarity computations. The objective is to exploit the varied item-data available from the world wide web, to overcome trivial problems like that of cold-start. In this work, we have designed a new similarity metric inspired from the existing DICE similarity that takes into account changing item/user behavior to compute updated similarity values for the purpose of recommendations. The work leverages the idea of content based recommendations as a first step to create vivid user and item profiles that are iteratively updated.

This work was done in collaboration with Rachid Guerraoui (EPFL, Switzerland), Rhicheek Patra (EPFL, Switzerland).

6.4.2. Lightweight Privacy-Preserving Averaging for the Internet of Things

Participants: Davide Frey, George Giakkoupis, Julien Lepiller.

The number of connected devices is growing continuously, and so is their presence into our everyday lives. From GPS-enabled fitness trackers, to smart fridges that tell us what we need to buy at the grocery store, connected devices—things—have the potential to collect and make available significant amounts of information. On the one hand, this information may provide useful services to users, and constitute a statistical gold mine. On the other, its availability poses serious privacy threats for users. In this work, we designed two new protocols that make it possible to aggregate personal information collected by smart devices in the form of an average, while preventing attackers from learning the details of the non-aggregated data. The first protocol exploits randomness and decomposition into shares as techniques to obfuscate the value associated with each node and lightweight encryption techniques to withstand eavesdropping attacks. The second exploits only randomness and encryption. We carried out a preliminary evaluation and published the results related to the first protocol in [18].

This work was done in collaboration with Tristan Allard from the DRUID Team at IRISA, Rennes.

6.4.3. Collaborative Filtering Under a Sybil Attack: Similarity Metrics do Matter!

Participants: Davide Frey, Anne-Marie Kermarrec, Antoine Rault, Florestan de Moor.

Whether we are shopping for an interesting book or selecting a movie to watch, the chances are that a recommendation system will help us decide what we want. Recommendation systems collect information about our own preferences, compare them to those of other users, and provide us with suggestions on a variety of topics. But is the information gathered by a recommendation system safe from potential attackers, be them other users, or companies that access the recommendation system? And above all, can service providers protect this information while still providing effective recommendations? In this work, we analyze the effect of Sybil attacks on collaborative-filtering recommendation systems, and discuss the impact of different similarity metrics in the trade-off between recommendation quality and privacy. Our results, on a state-of-the-art recommendation framework and on real datasets show that existing similarity metrics exhibit a wide range of behaviors in the presence of Sybil attacks. Yet, they are all subject to the same trade off: Sybil resilience for recommendation quality. We therefore propose and evaluate a novel similarity metric that combines the best of both worlds: a low RMSE score with a prediction accuracy for Sybil users of only a few

percent. A preliminary version of this work was published at EuroSec 2015 [57]. This year, we significantly extended the work during the summer internship of Florestan De Moor. Specifically, we considered new attacks that specifically target our novel similarity metric and showed that regardless of the attack configuration, our metric can preserve the privacy of users without hampering recommendation quality. A new paper with these new results was submitted to PETS 2017.

6.4.4. Privacy-Preserving Distributed Collaborative Filtering

Participants: Davide Frey, Anne-Marie Kermarrec.

In this work, we propose a new mechanism to preserve privacy while leveraging user profiles in distributed recommender systems. Our mechanism relies on (*i*) an original obfuscation scheme to hide the exact profiles of users without significantly decreasing their utility, as well as on (*ii*) a randomized dissemination protocol ensuring differential privacy during the dissemination process.

We compare our mechanism with a non-private as well as with a fully private alternative. We consider a real dataset from a user survey and report on simulations as well as planetlab experiments. We dissect our results in terms of accuracy and privacy trade-offs, bandwith consumption, as well as resilience to a censorship attack. In short, our extensive evaluation shows that our twofold mechanism provides a good trade-off between privacy and accuracy, with little overhead and high resilience.

This work was done with Antoine Boutet and Arnaud Jegou when they were part of the team, and in collaboration with Rachid Guerraoui from EPFL. But the complete results were published this year in [15].

ASCOLA Project-Team

7. New Results

7.1. Software composition and programming languages

Participants: Walid Benghrabit, Ronan-Alexandre Cherrueau, Rémi Douence, Hervé Grall, Florent Marchand de Kerchove de Denterghem, Jacques Noyé, Jean-Claude Royer, Mario Südholt.

This year we have published a number of new results in the domains of software composition and programming languages that range from pragmatic ones like modularity issues to formal studies in the domain of dependent type theory via static analysis and formal verification.

7.1.1. Formal Methods, logics and type theory

Concerning verification and formal semantics, we have defined the semantics of our dependent interoperability framework and we propose the notion the partial type equivalences as a key feature. We have also studied proofs in dependent type theory and synthesized call-by-value and call-by-name translations.

7.1.1.1. Verified Dependent Interoperability.

Full-spectrum dependent types promise to enable the development of correct-by-construction software. However, even certified software needs to interact with simply-typed or untyped programs, be it to perform system calls, or to use legacy libraries. Trading static guarantees for runtime checks, the dependent interoperability framework provides a mechanism by which simply-typed values can safely be coerced to dependent types and, conversely, dependently-typed programs can defensively be exported to a simply-typed application. In [22], we give a semantic account of dependent interoperability. Our presentation relies on and is guided by a pervading notion of type equivalence, whose importance has been emphasized in recent works on homotopy type theory. Specifically, we develop the notion of partial type equivalences as a key foundation for dependent interoperability. Our framework is developed in Coq; it is thus constructive and verified in the strictest sense of the terms. Using our library, users can specify domain-specific partial equivalences between data structures. Our library then takes care of the (sometimes, heavy) lifting that leads to interoperable programs. It thus becomes possible, as we shall illustrate, to internalize and hand-tune the extraction of dependently-typed programs to interoperable OCaml programs within Coq itself.

7.1.1.2. Forcing in Type Theory.

In [26], we study forcing translations of proofs in dependent type theory, through the Curry-Howard correspondence. Based on a call-by-push-value decomposition, we synthesize two simply-typed translations: i) one call-by-value, corresponding to the translation derived from the presheaf construction as studied in a previous paper; ii) one call-by-name, whose intuitions already appear in Krivine and Miquel's work. Focusing on the call-by-name translation, we adapt it to the dependent case and prove that it is compatible with the definitional equality of our system, thus avoiding coherence problems. This allows us to use any category as forcing conditions, which is out of reach with the call-by-value translation. Our construction also exploits the notion of storage operators in order to interpret dependent elimination for inductive types. This is a novel example of a dependent theory with side-effects, clarifying how dependent elimination for inductive types must be restricted in a non-pure setting. Being implemented as a Coq plugin, this work gives the possibility to formalize easily consistency results, for instance the consistency of the negation of Voevodsky's univalence axiom.

7.1.2. Programming languages

In the domain of programming languages we have presented new results on constraint programming, development of correct programs by construction and better controls for computational effects and modularity for JavaScript.

7.1.2.1. Constraint programming

Constraint programming (CP) relies on filtering algorithms in order to deal with combinatorial problems. Global constraints offer efficient algorithms for complex constraints. In particular a large family of global constraints can be expressed as constraints of finite state automata with counters. We have generalized these automata constraints in order to compose them as transducers [16]. We have also extended these results with different techniques [20]. First, we have improved the automaton synthesis to generate automata with fewer accumulators. Second, we have shown how to decompose a constraint specified by an automaton with accumulators into a conjunction of linear inequalities, for use by a MIP (Mixed-Integer Programming) solver. Third, we have generalized the implied constraint generation to cover the entire family of time-series constraints. The newly synthesized automata for time-series constraints outperform the old ones, for both the CP and MIP decompositions, and the generated implied constraints boost the inference, again for both the CP and MIP decompositions.

7.1.2.2. Program correctness

Most IDEs provide refactoring tools to assist programmers when they modify the structure of their software. However the refactoring facilities of many popular tools (Eclipse, Visual Studio, IntelliJ, etc.) are currently not reliable : they occasionally change the program semantics in unexpected ways, and, as a result, the programmers systematically have to re-test the resulting code. We have build a refactoring tool for C programs which core operation is proved correct by construction [21]. To do that, we build an AST transformation with Coq (based on the CompCert C implementation) and we prove that this transformation preserves the external behavior of programs. The code of the transformation is then extracted to OCaml and is then embedded in a traditional parse/transform/pretty-print setting to provide a working prototype.

7.1.2.3. Effect Capabilities

Computational effects complicate the tasks of reasoning about and maintaining software, due to the many kinds of interferences that can occur. While different proposals have been formulated to alleviate the fragility and burden of dealing with specific effects, such as state or exceptions, there is no prevalent robust mechanism that addresses the general interference issue. Building upon the idea of capability-based security, we propose in [18] effect capabilities as an effective and flexible manner to control monadic effects and their interferences. Capabilities can be selectively shared between modules to establish secure effect-centric coordination. We further refine capabilities with type-based permission lattices to allow fine-grained decomposition of authority. We provide an implementation of effect capabilities in Haskell, using type classes to establish a way to statically share capabilities between modules, as well as to check proper access permissions to effects at compile time. We first exemplify how to tame effect interferences using effect capabilities by treating state and exceptions. Then we focus on taming I/O by proposing a fine-grained lattice of I/O permissions based on the current classification of its operations. Finally, we show that integrating effect capabilities with modern tag-based monadic mechanisms provides a practical, modular and safe mechanism for monadic programming in Haskell.

7.1.2.4. Extensible JavaScript Modules

As part of the SecCloud project, we have studied how to modularly extend JavaScript interpreters with dynamic security analyses in particular information flow analyses. This has led us to study ways to improve on the standard JavaScript module pattern. This pattern is commonly used to encapsulate definitions by using closures. However, closures prevent module definitions from being extended at runtime. We have proposed a simple pattern that not only opens the module, but allows one to extend the module definitions in layers [39]. The pattern leverages the with construct and the prototype delegation mechanism of JavaScript to mimic a form of dynamic binding, while minimizing the changes made to the module code.

Florent Marchand's PhD thesis [13] details the proposal further and shows its application to the modular extension of Narcissus, a full-blown JavaScript interpreter, with several dynamic analyses, including the information flow of Austin and Flanagan based on multiple facets. A comparison with a previous ad hoc implementation of the analysis illustrates the benefits of the proposal.

7.1.3. Software Security and Privacy

In the area of security we have focused on expressing advanced security concerns with abstract and formal languages and the study of policy monitoring and the detection of conflicts.

7.1.3.1. Runtime verification of advanced logical security properties.

Monitoring or runtime verification means to observe the system execution and to check if it deviates or not from a predefined contract. Our contract is a formula written in AAL (Abstract Accountability Language) expressing the expected behavior of a system, the audit steps as well as punishment and compensation. We choose to use the rewriting approach with the three valued logic as many other existing approaches. The monitoring problem raised a validity question, if we start with a formula neither true nor false are we sure to conclude? The response is no and this is a completeness problem and all published solutions are incomplete. For LTL, mixing the standard semantics, the rewriting principle and coinduction we are able to define a complete monitoring mechanism. A first implementation has been done into our AccLab tool support and sketched in [38]. We are investigating the extension of our LTL rewriting mechanism to cope with the first-order case.

7.1.3.2. Specification of advanced security and privacy properties.

Security and privacy requirements in ubiquitous systems need a sophisticated policy language with features to express access restrictions and obligations. Ubiquitous systems involve multiple actors owning sensitive data concerning aspects such as location, discrete and continuous time, multiple roles that can be shared among actors or evolve over time. Conflict management is an important problem in security policy frameworks. In [31] we present an abstract language (AAL) dedicated to accountability. We show how to specify most of these security and privacy features and compare it with the XACML approach. We also classified the existing conflict detection for XACML like approaches in dynamic, testing, or static detection. A thorough analysis of these mechanisms reveals that they have several weaknesses and they are not applicable in our context. We advocate for a classic approach using the notion of logical consistency to detect conflicts in AAL.

7.1.3.3. Composition of privacy-enhancing and security mechanisms.

As part of his PhD thesis [11], Ronan Cherrueau's has defined a language for the composition of three privacy-enhancing and security mechanisms: symmetric key encryption, database fragmentation and on-client computations. The language allows the expression of distributed programs that protect data by applying compositions of the three mechanisms to them. The language ensures basic privacy and security properties by a type system based on dependent types. This type system ensures, for example, that data that has been encrypted and stored in a database fragment cannot be accessed in plain form and from another location than that fragment. Furthermore, the language comes equiped with four major additional results. First, a calculus that allows for the semi-automatic derivation of distributed privacy-preserving and secure programs from an original non-distributed one. Second, a transformation from the language to the π-calculus. Third, a transformation into an input specification to the Proverif model checker for security properties. Fourth, two implementations on the basis of, respectively, the Scala and Idris languages that harness their corresponding dependent type systems.

7.2. Distributed programming and the Cloud

Participants: Frederico Alvares, Bastien Confais, Simon Dupont, Md Sabbir Hasan, Adrien Lebre, Thomas Ledoux, Guillaume Le Louët, Jean-Marc Menaud, Jonathan Pastor, Rémy Pottier, Anthony Simonet, Mario Südholt.

7.2.1. Cloud applications and infrastructures

Complex event processing. We presented this year the evolution of SensorScript towards a language for complex event processing dedicated to sensor networks. While the model mainly relies on previous works, we highlighted how the new language builds on the multitree in order to provide complex event processing mechanisms. We are able to balance the syntactic concision of the language with a real-time complex event processor for sensor networks. By providing flexible selections over the nodes, with the possibility to filter

them on complex conditions, possibly over a time window, we offer a strong alternative to traditional SQL used in the literature. Moreover, SensorScript does not focus only on data access. In fact it provides the possibility to widen the scope of the methods accessible on nodes to other features than sensors monitoring, including but not limited to addressing actuators functions. Finally we showed that SensorScript is able to address examples proposed in the literature, with simpler results than SQL, while highlighting its limitations, especially on history management. [24]

Secure cloud storage. The increasing number of cloud storage services like Dropbox or Google Drive allows users to store more and more data on the Internet. However, these services do not give users enough guarantees in protecting the privacy of their data. In order to limit the risk that the storage service scans user documents for commercial purposes, we propose a storage service that stores data on several cloud providers while preventing these providers to read user documents. TrustyDrive is a cloud storage service that protects the privacy of users by breaking user documents into blocks in order to spread them on several cloud providers. As cloud providers only own a part of the blocks and they do not know the block organization, they can not read user documents. Moreover, the storage service connects directly users and cloud providers without using a third-party as is generally the practice in cloud storage services. Consequently, users do not give critical information (security keys, passwords, etc.) to a third-party. [30]

7.2.1.1. Service-level agreement for the Cloud.

Quality-of-service and SLA guarantees are among the major challenges of cloud-based services. In [19], we first present a new cloud model called SLAaaS — SLA aware Service. SLAaaS considers QoS levels and SLA as first class citizens of cloud-based services. This model is orthogonal to other SaaS, PaaS, and IaaS cloud models, and may apply to any of them. More specifically, we make three contributions: (i) we provide a domain-specific language that allows to define SLA constraints in cloud services; (ii) we present a general control-theoretic approach for managing cloud service SLA; (iii) we apply our approach to MapReduce, locking, and e-commerce services.

7.2.1.2. Cloud Capacity Planning and Elasticity.

Capacity management is a process used to manage the capacity of IT services and the IT infrastructure. Its primary goal is to ensure that IT resources (services, infrastructure) are right-sized to meet current and future requirements in a cost-effective and timely manner. In [34], we present a comprehensive overview of capacity planning and management for cloud computing. First, we state the problem of capacity management in the context of cloud computing from the point of view of several service providers. Second, we provide a brief discussion about *when* capacity planning should take place. Finally, we survey a number of methods for capacity planning and management proposed by both people from industry and researchers.

In his PhD [12], Simon Dupont proposes to extend the concept of elasticity to higher layers of the cloud, and more precisely to the SaaS level. He presents the new concept of *software elasticity* by defining the ability of the software to adapt, ideally in an autonomous way, to cope with workload changes and/or limitations of IaaS elasticity. This brings the consideration of Cloud elasticity in a multi-layer way through the adaptation of all kind of Cloud resources (software, virtual machines, physical machines). In [23], we introduce ElaScript, a DSL that offers Cloud administrators a simple and concise way to define complex elasticity-based reconfiguration plans. ElaScript is capable of dealing with both infrastructure and software elasticities, independently or together, in a coordinated way. We validate our approach by first showing the interest to have a DSL offering multiple levels of control for Cloud elasticity, and then by showing its integration with a realistic well-known application benchmark deployed in OpenStack and Grid'5000 infrastructure testbed.

7.2.1.3. Infrastructure.

Academic and industry experts are now advocating for going from large-centralized Cloud Computing infrastructures to smaller ones massively distributed at the edge of the network (aka., Fog and Edge Computing solutions). Among the obstacles to the adoption of this model is the development of a convenient and powerful IaaS system capable of managing a significant number of remote data-centers in a unified way.

In 2016, we achieved three major results in this context.

The first result is related to the economical viability of Fog/Edge Computing infrastructures that is often debated w-r-t large cloud computing data centers operated by US giants such as Amazon, Google .... To answer such a question, we conducted a specific study that goes beyond the state of the art of the current cost model of Distributed Cloud infrastructures. First, we provided a classification of the different ways of deploying Distributed Cloud platforms. Then, we proposed a versatile cost model that can help new actors evaluate the viability of deploying a Fog/Edge Computing offer. We illustrated the relevance of our proposal by instantiating it over three use-cases and comparing them according to similar computation capabilities provided by the Amazon solution. Such a study clearly showed that deploying a Distributed Cloud infrastructure makes sense for telcos as well as new actors willing to enter the game [29].

The second result is related to the preliminary revisions we made in OpenStack. The OpenStack software suite has become the de facto open-source solution to operate, supervise and use a Cloud Computing infrastructure. Our objective is to study to what extent current OpenStack mechanisms can handle massively distributed cloud infrastructures and to propose revisions/extensions of internal mechanisms when appropriate. The work we conducted this year focused on the Nova service of OpenStack.More precisely, we modified the code base in order to use a distributed key/value store instead of the centralized SQL backend. We conducted several experiments that validate the correct behavior and gives performance trends of our prototype through an emulation of several data-centers using Grid'5000 testbed. In addition to paving the way to the first large-scale and Internet-wide IaaS manager, we expect this work will attract a community of specialists from both distributed system and network areas to address the Fog/Edge Computing challenges within the OpenStack ecosystem [36], [27]. These and additional corresponding results have been presented in a more detailed manner as part of Jonathan Pastor's PhD thesis [14].

The third result is related to the data management in Fog/Edge Computing infrastructures. Our ultimate goal is to propose an Amazon-S3 like system, *i.e.*, a blob storage service, that can take into account Fog/Edge specifics. The study we achieved this year is preliminary. We first identified a list of properties a storage system should meet in this context. Second, we evaluated through performance analysis three "off-the-shelf" object store solutions, namely Rados, Cassandra and InterPlanetary File System (IPFS). In particular, we focused (i) on access times to push and get objects under different scenarios and (ii) on the amount of network traffic that is exchanged between the different sites during such operations. We also evaluated how the network latencies influence the access times and how the systems behave in case of network partitioning. Experiments have been conducted using the Yahoo Cloud System Benchmark (YCSB) on top of the Grid'5000 testbed. We showed that among the three tested solutions IPFS fills most of the criteria expected for a Fog/Edge computing infrastructure. [33], [32]

7.2.2. Renewable energy

With the emergence of the Future Internet and the dawning of new IT models such as cloud computing, the usage of data centers (DC), and consequently their power consumption, increase dramatically. Besides the ecological impact, the energy consumption is a predominant criterion for DC providers since it determines the daily cost of their infrastructure. As a consequence, power management becomes one of the main challenges for DC infrastructures and more generally for large-scale distributed systems. We have design the EpoCloud prototype, from hardware to middleware layers. This prototype aims at optimizing the energy consumption of mono-site Cloud DCs connected to the regular electrical grid and to renewable-energy sources. [17]

7.2.2.1. Green Energy awareness in SaaS Application.

With the proliferation of Cloud computing, data centers have to urgently face energy consumption issues. Although recent efforts such as the integration of renewable energy to data centers or energy efficient techniques in (virtual) machines contribute to the reduction of carbon footprint, creating green energy awareness around *Interactive Cloud Applications* by smartly using the presence of green energy has not been yet addressed. By *awareness*, we mean the inherited capability of SaaS applications to dynamically adapt with the availability of green energy and to reduce energy consumption while green energy is scarce or absent. In [25], we present two application controllers based on different metrics (e.g., availability of green energy, response time, user experience level). Based on extensive experiments with a real application benchmark and

workloads in Grid'5000, results suggest that providers revenue can be increased as high as 64%, while 13% brown energy can be reduced without deprovisioning any physical or virtual resources at IaaS layer and 17 fold increment of performance can be guaranteed.

ASPI Project-Team

6. New Results

6.1. Central limit theorem for adaptive multilevel splitting

Participants: Frédéric Cérou, Arnaud Guyader, Mathias Rousset.

This is a collaboration with Bernard Delyon (université de Rennes 1).

In this work, we consider the adaptive multilevel splitting algorithm as a Fleming–Viot particle system: the particles are indexed by levels instead of time, and the associated states are given by first entrance into level sets, in a similar fashion as in [38]. A rigorous proof of a central limit theorem has been obtained in [24] for Fleming–Viot particle systems. The application to AMS (adaptive multilevel splitting) algorithm is in preparation.

6.2. An efficient algorithm for video super-resolution based on a sequential model

Participant: Patrick Héas.

This is a collaboration with Angélique Drémeau (ENSTA Bretagne, Brest) and Cédric Herzet (EPI FLUMINANCE, Inria Rennes–Bretagne Atlantique)

In [16], we propose a novel procedure for video super–resolution, that is the recovery of a sequence of high–resolution images from its low-resolution counterpart. Our approach is based on a "sequential" model (i.e., each high-resolution frame is supposed to be a displaced version of the preceding one) and considers the use of sparsity-enforcing priors. Both the recovery of the high–resolution images and the motion fields relating them is tackled. This leads to a large–dimensional, non-convex and non-smooth problem. We propose an algorithmic framework to address the latter. Our approach relies on fast gradient evaluation methods and modern optimization techniques for non-differentiable/non-convex problems. Unlike some other previous works, we show that there exists a provably-convergent method with a complexity linear in the problem dimensions. We assess the proposed optimization method on several video benchmarks and emphasize its good performance with respect to the state of the art.

6.3. Low–rank approximation and dynamic mode decomposition

Participant: Patrick Héas.

This is a collaboration with Cédric Herzet (EPI FLUMINANCE, Inria Rennes–Bretagne Atlantique)

Dynamic mode decomposition (DMD) has emerged as a powerful tool for analyzing the dynamics of non-linear systems from experimental datasets. Recently, several attempts have extended DMD to the context of low-rank approximations. This low-rank extension takes the form of a non–convex optimization problem. To the best of our knowledge, only sub–optimal algorithms have been proposed in the literature to compute the solution of this problem. In [26], we prove that there exists a closed-form optimal solution to this problem and design an effective algorithm to compute it based on singular value decomposition (SVD). Based on this solution, we then propose efficient procedures for reduced-order modeling and for the identification of the low-rank DMD modes and amplitudes. Experiments illustrates the gain in performance of the proposed algorithm compared to state-of-the-art techniques.

6.4. Model reduction from partial observations

Participant: Patrick Héas.

This is a collaboration with Angélique Drémeau (ENSTA Bretagne, Brest) and Cédric Herzet (EPI FLUMINANCE, Inria Rennes–Bretagne Atlantique)

In [25], we deal with model order reduction of parametric partial differential equations (PPDE). We consider the specific setup where the solutions of the PPDE are only observed through a partial observation operator and address the task of finding a good approximation subspace of the solution manifold. We provide and study several tools to tackle this problem. We first identify the best worst–case performance achievable in this setup and propose simple procedures to approximate this optimal solution. We then provide, in a simplified setup, a theoretical analysis relating the achievable reduction performance to the choice of the observation operator and the prior knowledge available on the solution manifold.

In [22], we focus on reduced modeling of dynamical systems, in an analogous partial observation setup. Assuming prior knowledge available, we provide a unified reduction framework based on an a posteriori characterisation of the uncertainties on the solution manifold. Relying on sequential Monte Carlo (SMC) samples, we provide a closed-form approximation of solutions to the problem of choosing an optimal Galerkin projection or an optimal low–rank linear approximation. Numerical results obtained for a standard geophysical model show the gain brought by exploiting this posterior information for building a reduced model.

6.5. Combining analog method and ensemble data assimilation

Participants: Thi Tuyet Trang Chau, François Le Gland, Valérie Monbet.

This is a collaboration with Pierre Ailliot (université de Bretagne Occidentale, Brest), Ronan Fablet and Pierre Tandéo (Télé´com Bretagne, Brest), Anne Cuzol (université de Bretagne Sud, Vannes) and Bernard Chapron (IFREMER, Brest).

Nowadays, ocean and atmosphere sciences face a deluge of data from spatial observations, in situ monitoring as well as numerical simulations. The availability of these different data sources offer new opportunities, still largely underexploited, to improve the understanding, modeling and reconstruction of geophysical dynamics. The classical way to reconstruct the space–time variations of a geophysical system from observations relies on data assimilation methods using multiple runs of the known dynamical model. This classical framework may have severe limitations including its computational cost, the lack of adequacy of the model with observed data, modeling uncertainties. In [60], we explore an alternative approach and develop a fully data—driven framework, which combines machine learning and statistical sampling to simulate the dynamics of complex system. As a proof concept, we address the assimilation of the chaotic Lorenz–63 model and imputation of missing data in multisite wind and rain time series. We demonstrate that a nonparametric sampler from a catalog of historical datasets, namely local linear regression, combined with a classical stochastic data assimilation scheme, the ensemble Kalman filter and the particular filter, reach state–of–the–art performances, without online evaluations of the physical model. The use of local regression instead of analog sampler allows to improve the performance of the filters.

6.6. Classification trees, functional data, applications in biology

Participants: Valérie Monbet, Audrey Poterie.

This is a collaboration with Jean–François Dupuy (INSA Rennes) and Laurent Rouvière (université de Haute Bretagne, Rennes).

Classification and discriminant analysis methods have grown in depths during the past 20 years. Fisher linear discriminant analysis (LDA) is the basic but standard approach. As the structure and dimension of the data becomes more complex in a wide range of applications, such as functional data, there is a need for more flexible nonparametric classification and discriminant analysis tools, especially when the ratio of learning sample size to number of covariates is low and the covariates are highly correlated and the covariance matrix is highly degenerated or when the large number of covariates are generally weak in predicting the class labels. For some data such as spectrometry data, only some parts of the observed curves are discriminant leading to groups of variables.

We proposed a classification tree based on groups of variables. Like usual tree-based methods, the algorithm partitions the feature space into M regions, by recursively performing binary splits. The main difference is that each split is based on groups of variables and the boundary between both classes is the hyperplane which minimizes the Bayes risk in the set generated by the selected group of variables. We demonstrate on several toy examples and real spectrometry data that the performances of the proposed tree groups algorithm are at least as good as the one of the standard CART algorithm and group Lasso logistic regression.

CAIRN Project-Team

7. New Results

7.1. Reconfigurable Architecture Design

7.1.1. Dynamic Reconfiguration Support in FPGA

Participants: Olivier Sentieys, Christophe Huriaux.

Almost since the creation of the first SRAM-based FPGAs there has been a desire to explore the benefits of partially reconfiguring a portion of an FPGA at run-time while the remainder of design functionality continues to operate uninterrupted. Currently, the use of partial reconfiguration imposes significant limitations on the FPGA design: reconfiguration regions must be constrained to certain shapes and sizes and, in many cases, bitstreams must be precompiled before application execution depending on the precise region of the placement in the fabric. We developed an FPGA architecture that allows for seamless translation of partially-reconfigurable regions, even if the relative placement of fixed-function blocks within the region is changed.

In [4], we proposed a design flow for generating compressed configuration bitstreams abstracted from their final position on the logic fabric, the Virtual Bit-Streams (VBS). Those configurations can then be decoded and finalized in real-time and at run-time by a dedicated reconfiguration controller to be placed at a given physical location. The VPR (Versatile Place and Route) framework was expanded to include bitstream generation features. The configuration stream format was proposed along with its associated decoding architecture. We analyzed the compression induced by our coding method and proved that compression ratios of at least $2.5\times$ can be achieved on the 20 largest MCNC benchmarks. The introduction of clustering which aggregates multiple routing resources together showed compression ratio up to a factor of $10\times$, at the cost of a more complex decoding step at runtime.

The emergence of 2.5D and 3D packaging technologies enables the integration of FPGA dice into more complex systems. Both heterogeneous manycore designs, which include an FPGA layer, and interposer-based multi-FPGA systems support the inclusion of reconfigurable hardware in 3D-stacked integrated circuits. In these architectures, the communication between FPGA dice or between FPGA and fixed-function layers often takes place through dedicated communication interfaces spread over the FPGA logic fabric, as opposed to an I/O ring around the fabric. In [39], we investigate the effect of organizing FPGA fabric I/O into coarse-grained interface blocks distributed throughout the FPGA fabric. Specifically, we consider the quality of results for the placement and routing phases of the FPGA physical design flow. We evaluate the routing of I/O signals of large applications through dedicated interface blocks at various granularities in the logic fabric, and study its implications on the critical path delay of routed designs. We show that the impact of such I/O routing is limited and can improve chip routability and circuit delay in many cases.

7.1.2. Hardware Accelerated Simulation of Heterogeneous Platforms

Participant: François Charot.

When considering designing heterogeneous multi-core platforms, the number of possible design combinations leads to a huge design space, with subtle trade-offs and design interactions. To reason about what design is best for a given target application requires detailed simulation of many different possible solutions. Simulation frameworks exist (such as gem5) and are commonly used to carry out these simulations. Unfortunately, these are purely software-based approaches and they do not allow a real exploration of the design space. Moreover, they do not really support highly heterogeneous multi-core architectures. These limitations motivate the study of the use of hardware to accelerate the simulation, and in particular of FPGA components. In this context, we are currently investigating the possibility of building hardware accelerated simulators using the HAsim simulation infrastructure, jointly developed by MIT and Intel. HAsim is a FPGA-accelerated simulator that is able to simulate a multicore with a high-detailed pipeline, cache hierarchy and detailed on-chip network on a single FPGA. We work on integrating a model of the RISC-V instruction set architecture in the HAsim infrastructure. This work is done with the perspective of studying hardware accelerated simulation of heterogeneous multicore architectures mixing RISC-V cores and hardware accelerators.

7.1.3. Optical Interconnections for 3D Multiprocessor Architectures

Participants: Jiating Luo, Ashraf El-Antably, Pham Van Dung, Cédric Killian, Daniel Chillet, Olivier Sentieys.

To address the issue of interconnection bottleneck in multiprocessor on a single chip, we study how an Optical Network-on-Chip (ONoC) can leverage 3D technology by stacking a specific photonics die. The objectives of this study target: i) the definition of a generic architecture including both electrical and optical components, ii) the interface between electrical and optical domains, iii) the definition of strategies (communication protocol) to manage this communication medium, and iv) new techniques to manage and reduce the power consumption of optical communications. The first point is required to ensure that electrical and optical components can be used together to define a global architecture. Indeed, optical components are generally larger than electrical components, so a trade-off must be found between the size of optical and electrical parts. For example, if the need in terms of communications is high, several waveguides and wavelengths must be necessary, and can lead to an optical area larger than the footprint of a single processor. In this case, a solution is to connect (through the optical NoC) clusters of processors rather than each single processor. For the second point, we study how the interface can be designed to take applications needs into account. From the different possible interface designs, we extract a high-level performance model of optical communications from losses induced by all optical components to efficiently manage Laser parameters. Then, the third point concerns the definition of high-level mechanisms which can handle the allocation of the communication medium for each data transfer between tasks. This part consists in defining the protocol of wavelength allocation. Indeed, the optical wavelengths are a shared resource between all the electrical computing clusters and are allocated at run time according to application needs and quality of service. The last point concerns the definition of techniques allowing to reduce the power consumption of on-chip optical communications. The power of each Laser can be dynamically tuned in the optical/electrical interface at run time for a given targeted bit-error-rate. Due to the relatively high power consumption of such integrated Laser, we study how to define adequate policies able to adapt the laser power to the signal losses.

We are currently designing an Optical-Network-Interface (ONI) to connect one processor, or a cluster of several processors, to the optical communication medium. This interface, constrained by the 10 Gb/s data-rate of the Lasers, integrates Error Correcting Codes and a communication manager. This manager can select, at run-time, the communication mode to use depending on timing or power constraints. Indeed, as the use of ECC is based on redundant bits, it increases the transmission time, but saves power for a given Bit Error Rate (BER). Moreover, our ONI allows for data to be sent using several wavelengths in parallel, hence increasing transmission bandwidth.

However, multiple signals sharing simultaneously a waveguide can lead to inter-channel crosstalk noise. This problem impacts the Signal to Noise Ratio (SNR) of the optical signal, which leads to an increase in the Bit Error Rate (BER) at the receiver side. In [40], [59], we proposed a Wavelength Allocation (WA) method allowing to search for performance and energy trade-offs based on application constraints. We showed that for a 16-core WDM ring-based ONoC architecture using 12 wavelengths, more than 100,000 allocation solutions exist and only 51 are on a Pareto front giving a tradeoff between execution time and energy per bit (derived from the BER). The optimized solutions reached reduce the execution time by 37% or the energy from 7,6fJ/bit to 4,4fJ/bit.

7.1.4. Communication-Based Power Modelling for Heterogeneous Multiprocessor Architectures

Participants: Baptiste Roux, Olivier Sentieys, Steven Derrien.

Programming heterogeneous multiprocessor architectures is a real challenge dealing with a huge design space. Computer-aided design and development tools try to circumvent this issue by simplifying instantiation mechanisms. However, energy consumption is not well supported in most of these tools due to the difficulty to obtain fast and accurate power estimation. To this aim, in [46] we proposed and validated a power model for such platforms. The methodology is based on micro-benchmarking to estimate the model parameters. The energy model mainly relies on the energy overheads induced by communications between processors in a

parallel application. Power modelling and micro-benchmarks are validated using a Zynq-based heterogeneous architecture showing the accuracy of the model for several tested synthetic applications.

7.1.5. Arithmetic Operators for Cryptography and Fault-Tolerance

Participants: Arnaud Tisserand, Emmanuel Casseau, Pierre Guilloux, Karim Bigou, Gabriel Gallin, Audrey Lucas, Franck Bucheron, Jérémie Métairie.

Arithmetic Operators for Fast and Secure Cryptography.

Our paper [21], published in IEEE Transactions on Computers, extends our fast RNS modular inversion for finite fields arithmetic published at CHES 2013 conference. It is based on the binary version of the plus-minus Euclidean algorithm. In the context of elliptic curve cryptography (*i.e.* 160–550 bits finite fields), it significantly speeds-up modular inversions. In this extension, we propose an improved version based on both radix 2 and radix 3. This new algorithm leads to 30 % speed-up for a maximal area overhead about 4 % on Virtex 5 FPGAs. This work was done in the ANR PAVOIS project.

Our paper [32], presented at ARITH-23, presents an hybrid representation of large integers, or prime field elements, combining both positional and residue number systems (RNS). Our *hybrid position-residues* (HPR) number system mixes a high-radix positional representation and digits represented in RNS. RNS offers an important source of parallelism for addition, subtraction and multiplication operations. But, due to its non-positional property, it makes comparisons and modular reductions more costly than in a positional number system. HPR offers various trade-offs between internal parallelism and the efficiency of operations requiring position information. Our current application domain is asymmetric cryptography where HPR significantly reduces the cost of some modular operations compared to state-of-the-art RNS solutions. This work was done in the ANR PAVOIS project.

An ASIC circuit has been implemented in the 65nm ST CMOS technology and sent to fabrication in June 2016 (chip delivery is expected for January 2017). The implemented cryptoprocessor was designed for 256-bit prime finite fields elements and generic curves. It embeds: 1 multiplier, 1 adder and 1 inversion units for field-level computations. Various algorithms for scalar multiplication primitives can be programmed in software for curve-level computations. It was designed to evaluate algorithmic and arithmetic protections against side channel attacks (there is no hardware protection embedded in this ASIC version). This work was done in the ANR PAVOIS project.

In the HAH project, funded by CominLabs and Lebesgue Labex, we study hardware implementation of cryptoprocessors for hyperelliptic curves. The poster [61] presents the current state of the project for FPGA implementations.

Arithmetic Operators for Fault-Tolerance.

Various methods have been proposed for fault detection and fault tolerance in digital integrated circuits. In the case of *arithmetic circuits*, the selection of an efficient method depends on several elements: type of operation, type(s) of operand(s), computation algorithms, internal representations of numbers, optimizations at architecture and circuit levels, and acceptable accuracy level (i.e. mathematical error) of the result(s) including both rounding errors and errors due to the faults. High-level mathematical models are not sufficient to capture the effect of faults in arithmetic circuits. Simulation of intensive fault scenarios in all components of the arithmetic circuit (data-path, control, gates with important fan-out such as some partial products generation in large multipliers, etc.) is widely used. But cycle accurate and bit accurate software simulations at gate level are too slow for large circuits and numerous fault scenarios. *FPGA emulation* is a popular method to speed-up fault simulation.

We are developing an hardware-software platform dedicated to fault emulation for ASIC arithmetic circuits. The platform is based on a parallel cluster of Zynq FPGA cards and a Linux server. Various arithmetic circuits and fault models will be demonstrated in the context of digital signal and image processing. Our paper [57], presented at Compas, describes the very first version of our platform. This platform has also been presented in a poster at GDR SoC-SiP [58] and in a Demo Night at DASIP [56]. This work was done in the ANR ARDyT and Reliasic projects.

7.1.6. Adaptive Overclocking, Error Correction, and Voltage Over-Scaling for Error-Resilient Applications

Participants: Rengarajan Ragavan, Benjamin Barrois, Cédric Killian, Olivier Sentieys.

Error detection and correction based on double-sampling is used as common technique to handle timing errors while scaling V_{dd} for energy efficiency. Implementation and advantages of double-sampling technique in FPGAs are simpler and significant compared to the conventional highly pipelined processors due to the higher flexibility of the reconfigurable architectures. It is common practice to insert shadow flipflop in the critical paths of the design, which will fail while scaling down the supply voltage, or to correct timing errors while over clocking the datapaths. Overclocking, and error detection and correction capabilities of these methods are limited due to the fixed speculation window used by these methods. In [44], we presented a Dynamic Speculation Window in double-sampling for timing error detection and correction in FPGAs. The proposed method employs online slack measurement and conventional shadow flipflop approach to adaptively overclock the design and also to detect and correct timing errors due to temperature and other variability effects. We demonstrated this method in the Xilinx VC707 Virtex 7 FPGA for various benchmarks. We achieved maximum of 71% overclocking for unsigned 32-bit multiplier with the area overhead of 1.9% LUTs and 1.7% FFs.

Voltage scaling has been used as a prominent technique to improve energy efficiency in digital systems, scaling down supply voltage effects in quadratic reduction in energy consumption of the system. Reducing supply voltage induces timing errors in the system that are corrected through additional error detection and correction circuits. In [43], we proposed voltage over-scaling based approximate operators for applications that can tolerate errors. We characterized the basic arithmetic operators using different operating triads (combination of supply voltage, body-biasing scheme and clock frequency) to generate models for approximate operators. Error-resilient applications can be mapped with the generated approximate operator models to achieve optimum trade-off between energy efficiency and error margin. Based on the dynamic speculation technique, best possible operating triad is chosen at runtime based on the user definable error tolerance margin of the application. In our experiments in 28nm FDSOI, we achieved maximum energy efficiency of 89% for basic operators like 8-bit and 16-bit adders at the cost of 20% Bit Error Rate (ratio of faulty bits over total bits) by operating them in near-threshold regime.

7.2. Compilation and Synthesis for Reconfigurable Platform

7.2.1. Adaptive dynamic compilation for low power embedded systems

Participants: Steven Derrien, Simon Rokicki.

Dynamic binary translation (DBT) consists in translating – at runtime – a program written for a given instruction set to another instruction set. Dynamic Translation was initially proposed as a means to enable code portability between different instruction sets and can be implemented in software or hardware. DBT is also used to improve the energy efficiency of high performance processors, as an alternative to out-of-order microarchitectures. In this context, DBT is used to uncover instruction level parallelism (ILP) in the binary program, and then target an energy efficient wide issue VLIW architecture. This approach is used in Transmeta Crusoe [75] and NVidia Denver [68] processors. Since DBT operates at runtime, its execution time is directly perceptible by the user, hence severely constrained. As a matter of fact, this overhead has often been reported to have a huge impact on actual performance, and is considered as being the main weakness of DBT based solutions. This is particularly true when targeting a VLIW processor: the quality of the generated code depends on efficient scheduling; unfortunately scheduling is known to be the most time-consuming component of a JIT compiler or DBT. Improving the responsiveness of such DBT systems is therefore a key research challenge. This is however made very difficult by the lack of open research tools or platform to experiment with such platforms. In this work, we have been addressing these two issues by developing an open hardware/software platform supporting DBT. The platform was designed using HLS tools and validated on a FPGA board. The DBT uses RISC-V as host ISA, and can target varying issue width VLIW architectures. Our platform uses custom hardware accelerators to improve the reactivity of our optimizing DBT flow. Our results show that, compared to a software implementation, our approach offers speed-up by 8× while consuming 18× less energy.

7.2.2. Leveraging Power Spectral Density for Scalable System-Level Accuracy Evaluation

Participants: Benjamin Barrois, Olivier Sentieys.

The choice of fixed-point word-lengths critically impacts the system performance by impacting the quality of computation, its energy, speed and area. Making a good choice of fixed-point word-length generally requires solving an NP-hard problem by exploring a vast search space. Therefore, the entire fixed-point refinement process becomes critically dependent on evaluating the effects of accuracy degradation. In [30], a novel technique for the system-level evaluation of fixed-point systems, which is more scalable and that renders better accuracy, was proposed. This technique makes use of the information hidden in the power-spectral density of quantization noises. It is shown to be very effective in systems consisting of more than one frequency sensitive components. Compared to state-of-the-art hierarchical methods that are agnostic to the quantization noise spectrum, we show that the proposed approach is $5\times$ to $500\times$ more accurate on some representative signal processing kernels.

7.2.3. Approximate Computing

Participants: Benjamin Barrois, Olivier Sentieys.

Many applications are error-resilient, allowing for the introduction of approximations in the calculations, as long as a certain accuracy target is met. Traditionally, fixed-point arithmetic is used to relax accuracy, by optimizing the bit-width. This arithmetic leads to important benefits in terms of delay, power and area. Lately, several hardware approximate operators were invented, seeking the same performance benefits. However, a fair comparison between the usage of this new class of operators and classical fixed-point arithmetic with careful truncation or rounding, has never been performed. In [31], we first compare approximate and fixed-point arithmetic operators in terms of power, area and delay, as well as in terms of induced error, using many state-of-the-art metrics and by emphasizing the issue of data sizing. To perform this analysis, we developed a design exploration framework, APXPERF, which guarantees that all operators are compared using the same operating conditions. Moreover, operators are compared in several classical real-life applications leveraging relevant metrics. In [31], we show that considering a large set of parameters, existing approximate adders and multipliers tend to be dominated by truncated or rounded fixed-point ones. For a given accuracy level and when considering the whole computation data-path, fixed-point operators are several orders of magnitude more accurate while spending less energy to execute the application. A conclusion of this study is that the entropy of careful sizing is always lower than approximate operators, since it require significantly less bits to be processed in the data-path and stored. Approximated data therefore always contain on average a greater amount of costly erroneous, useless information.

7.2.4. Real-Time Scheduling of Reconfigurable Battery-Powered Multi-Core Platforms

Participants: Daniel Chillet, Aymen Gammoudi.

Reconfigurable real-time embedded systems are constantly increasingly used in applications like autonomous robots or sensor networks. Since they are powered by batteries, these systems have to be energy-aware, to adapt to their environment and to satisfy real-time constraints. For energy harvesting systems, regular recharges of battery can be estimated, and by including this parameter in the operating system, it is then possible to develop strategy able to ensure the best execution of the application until the next recharge. In this context, operating system services must control the execution of tasks to meet the application constraints. Our objective concerns the proposition of a new real-time scheduling strategy that considers execution constraints such as the deadline of tasks and the energy.

To address this issue, we first focus on mono-processor scheduling [38] and propose to classify the tasks that have similar periods (or WCETs) in packs and to manage the execution parameters of these packs. For each reconfiguration scenario, parameter modifications are performed on packs/tasks to meet the real-time and energy constraints. Compared to previous work, task delaying is significantly improved in [36]. Furthermore, we also develop a strategy for multi-cores systems considering the dependencies between tasks [37] by adding the cost of communication between cores.

7.2.5. Optimization of loop kernels using software and memory information

Participant: Angeliki Kritikakou.

Current compilers cannot generate code that can compete with hand-tuned code in efficiency, even for a simple kernel like matrix–matrix multiplication (MMM). A key step in program optimization is the estimation of optimal values for parameters such as tile sizes and number of levels of tiling. The scheduling parameter values selection is a very difficult and time-consuming task, since parameter values depend on each other; this is why they are found by using searching methods and empirical techniques. To overcome this problem, the scheduling sub-problems must be optimized together, as one problem and not separately. In [24], an MMM methodology is presented where the optimum scheduling parameters are found by decreasing the search space theoretically, while the major scheduling sub-problems are addressed together as one problem and not separately according to the hardware architecture parameters and input size; for different hardware architecture parameters and/or input sizes, a different implementation is produced. This is achieved by fully exploiting the software characteristics (e.g., data reuse) and hardware architecture parameters (e.g., data caches sizes and associativities), giving high-quality solutions and a smaller search space. This methodology refers to a wide range of CPU and GPU architectures.

The size required to store an array is crucial for an embedded system, as it affects the memory size, the energy per memory access and the overall system cost. Existing techniques for finding the minimum number of resources required to store an array are less efficient for codes with large loops and not regularly occurring memory accesses. They have to approximate the accessed parts of the array leading to overestimation of the required resources. Otherwise their exploration time is increased with an increase over the number of the different accessed parts of the array. In [25], we propose a methodology to compute the minimum resources required for storing an array which keeps the exploration time low and provides a near-optimal result for regularly and non-regularly occurring memory accesses and overlapping writes and reads.

7.2.6. Adaptive Software Control to Increase Resource Utilization in Mixed-Critical Systems

Participant: Angeliki Kritikakou.

Automotive embedded systems need to cope with antagonist requirements: on the one hand, the users and market pressure push car manufacturers to integrate more and more services that go far beyond the control of the car itself. On the other hand, recent standardization efforts in the safety domain has led to the development of the ISO 26262 norm that defines means and requirements to ensure the safe operation of automotive embedded systems. In particular, it led to the definition of ASIL (Automotive Safety and Integrity Levels), i.e., it formally defines several criticality levels. Handling the increased complexity of new services makes new architectures, such as multi or many-cores, appealing choices for the car industry. Yet, these architectures provide a very low level of timing predictability due to shared resources, which goes in contradiction with timing guarantees required by ISO 26262. For highest criticality level tasks, Worst-Case Execution Time analysis (WCET) is required to guarantee that timing constraints are respected. The WCET analyzers consider the worst-case scenario: whenever a critical task accesses a shared resource in a multi/many-core platform, a WCET analyzer considers that all cores use the same resource concurrently. To improve the system performance, we proposed in a earlier work an approach where a critical task can be run in parallel with less critical tasks, as long as the real-time constraints are met. When no further interferences can be tolerated, the proposed run-time control in [54] suspends the low critical tasks until the termination of the critical task. In an automotive context, the approach can be translated as a highly critical partition, namely a classic AUTOSAR one, that runs on one dedicated core, with several cores running less critical Adaptive AUTOSAR application(s). We briefly describe in [54] the design of our proven-correct approach. Our strategy is based on a graph grammar to formally model the critical task as a set of control flow graphs on which a safe partial WCET analysis is applied and used at run-time to control the safe execution of the critical task.

CELTIQUE Project-Team

4. New Results

4.1. Monitoring attacker knowledge with information flow analysis

Participants: Thomas Jensen, Frédéric Besson.

Motivated by the problem of stateless web tracking (fingerprinting) we have investigated a novel approach to hybrid information flow monitoring by tracking the knowledge that an attacker can learn about secrets during a program execution. We have proposed a general framework for combining static and dynamic information flow analysis, based on a precise representation of attacker knowledge. This hybrid analysis computes a precise description of what an attacker learns about the initial configuration (and in particular the secret part of it) by observing a specific output. An interesting feature of this knowledge-based information flow analysis is that it can be used to improve other information flow control mechanisms, such as no-sensitive upgrade. The whole framework is accompanied by a formalisation of the theory in the Coq proof assistant [18].

4.2. Semantic analysis of functional specifications of system software

Participants: Thomas Jensen, Oana Andreescu, Pauline Bolignano.

We have developed a static analysis for correlating input and output values in functional specifications, written in a functional, strongly typed, high-level specification formalism developed by the SME Prove & Run. In the context of interactive formal verification of complex systems, much effort is spent on proving the preservation of the system invariants. However, most operations have a localized effect on the system. Identifying correlations (in particular equalities) between input and output can substantially ease the proof burden for the programmer. Our correlation analysis is a flow-sensitive interprocedural analysis that handles arrays, structures and variant data types, and which computes a conservative approximation of the equality between sub-structures of input and of output fragments [27]. In a separate strand of work, we have used abstraction-based techniques for structuring and simplifying the proof of simulation between a high-level and a low-level specification of memory management algorithms in a hypervisor [22]. Both strands of work was carried out and validated on system software (a micro-kernel and a hypervisor) developed using the formal approach defined by Prove & Run.

4.3. Certified Static Analyses

4.3.1. Certified Semantics and Analyses for JavaScript

Participants: Martin Bodin, Gurvan Cabon, Thomas Jensen, Alan Schmitt.

We have continued our work on the certification of the semantics of JavaScript and of analyses for JavaScript on three different fronts.

First, on the language side, we have developed at tool in collaboration with Arthur Charguéraud (Inria Saclay) and Thomas Wood (Imperial College) to interactively explore the specification of JavaScript. More precisely, we have written a compiler for a subset of OCaml to a subset of JavaScript that generates an interpreter that can be executed step by step, inspecting both the state of the interpreted program but also the state of the interpreter. We have used this compiler on the JavaScript interpreter extracted from our Coq semantics of JavaScript. The resulting tool is available here and a demo can be run here. The tool has been presented to the Ecma TC39 committee in charge of standardizing JavaScript. We are currently identifying the improvements required to make it useful for the standardization process.

Second, Bodin, Schmitt, and Jensen have designed an abstract domain based on separation logic to faithfully abstract JavaScript heaps. This domain is able to capture interlinked dynamic and extensible objects, a central feature of the JavaScript memory model. In addition, we have introduced the notion of *membranes* that let us correctly define abstractions in a way that is compatible both with separation logic and abstract interpretation. As an extension of last year's work [32], this approach is globally correct as soon as each rule is independently proven correct. This result illustrates the robustness of our approach to define certified abstract semantics based on pretty-big-step semantics. This work has not yet been published.

Third, Cabon and Schmitt are developing a framework to automatically derive an information-flow tracking semantics from a pretty-big-step semantics. We have manually shown the approach works for complex examples, and are currently proving it in Coq. This work is submitted for publication.

4.3.2. Certified Analyses for C and lower-level programs

Participants: Sandrine Blazy, David Pichardie, Alix Trieu.

We have continued our work on the static analyzer Verasco [37], based on abstract interpretation and operating over most of the ISO C 1999 language (excluding recursion and dynamic allocation). Verasco establishes the absence of run-time errors in the analyzed programs. It enjoys a modular architecture that supports the extensible combination of multiple abstract domains. We have extended the memory abstract domain (that takes as argument any standard numerical abstract domain), so that it finely tracks properties about memory contents, taking into account union types, pointer arithmetic and type casts [19]. This memory domain is implemented and verified inside the Coq proof assistant with respect to the CompCert compiler memory model.

Motivated by applications to security and high efficiency, we are reusing the Verasco static analyzer and the CompCert compiler in order to design a lightweight and automated methodology for validating on low-level intermediate representations the results of a source-level static analysis. Our methodology relies on two main ingredients: a relative-safety checker, an instance of a relational verifier which proves that a program is safer than another, and a transformation of programs into defensive form which verifies the analysis results at runtime.

4.4. Certified Compilation

Participants: Sandrine Blazy, Frédéric Besson, Pierre Wilke, Alexandre Dang.

The CompCert C compiler provides the formal guarantee that the observable behaviour of the compiled code improves on the observable behaviour of the source code. A first limitation of this guarantee is that if the source code goes wrong, i.e. does not have a well-defined behaviour, any compiled code is compliant. Another limitation is that CompCert 's notion of observable behaviour is restricted to IO events.

Over the past years, we have developed the semantics theory so that unlike CompCert but like Gcc, the binary representation of pointers can be manipulated much like integers and where memory is a finite resource. We have now a formally verified C compiler, CompCertS, which is essentially the CompCert compiler, albeit with a stronger formal guarantee. The semantics preservation theorem applies to a wider class of existing C programs and, therefore, their compiled version benefits from the formal guarantee of CompCertS. CompCertS preserves not only the observable behaviour of programs but also ensures that the memory consumption is preserved by the compiler. As a result, we have the formal guarantee that the compiled code requires no more memory than the source code. This ensures that the absence of stack-overflows is preserved by compilation.

The whole proof of CompCertS represents a significant proof-effort and the details can be found in Pierre Wilke's PhD thesis [39].

CompCertS also implements the Portable Software Fault Isolation approach pioneered by Kroll *et al.* [38]. The advantage of CompCertS is that the masking operation of pointers has a defined semantics and can therefore be directly reasoned about.

4.5. Mechanical Verification of SSA-based Compilation Techniques

Participants: Delphine Demange, Yon Fernandez de Retana, David Pichardie.

We have continued our work on the mechanical verification of SSA-based compilation techniques [30], [31], [36].

A crucial phase for efficient machine code generation is the destruction of a middle-end SSA-like IR. To this end, we have studied a variant of SSA, namely the Conventional SSA form, which simplifies the destruction back to non-SSA code (i.e. at the exit point of the middle-end). This had long remained a difficult problem, even in a non-verified environment. We formally defined and proved the properties of the generation of Conventional SSA. Finally, we implemented and proved correct a coalescing destruction of the Conventional SSA form, à la Boissinot et al. [33], where variables can be coalesced according to a refined notion of interference. Our CSSA-based, coalescing destruction allows us to coalesce more than 99% of introduced copies, on average, and leads to encouraging results concerning spilling and reloading during post-SSA allocation. This work has been published in [24].

4.6. Semantics for shared-memory concurrency

Participants: Gurvan Cabon, David Cachera, David Pichardie.

Modern multicore processor architectures and compilers of shared-memory concurrent programming languages provide only weak memory consistency guarantees. A *memory model* specifies which write action can be seen by a read action between concurrent threads.

In a previous work on the Java memory model [35], we defined in an axiomatic style, a memory model where we embed the reorderings of memory accesses directly in the semantics, so that formalizing optimizations and their correctness proof is easier.

This year, following a similar approach, we have studied the RMO (Relaxed- Memory Order) model. More precisely, we defined a new multibuffer operational semantics with write and read buffers. We also introduced an intermediate semantics inspired from Boudol et al. [34], where actions are reordered within a single pipeline. Finally, another model formalizes the reordering semantics in an axiomatic way. We fully proved the equivalence between the first two models and present a methodology for the remaining part. This work has been published in an international workshop [23].

4.7. Static analysis of functional programs using tree automata and term rewriting

Participant: Thomas Genet.

We develop a specific theory and the related tools for analyzing programs whose semantics is defined using term rewriting systems. The analysis principle is based on regular approximations of infinite sets of terms reachable by rewriting. Regular tree languages are (possibly) infinite languages which can be finitely represented using tree automata. To over-approximate sets of reachable terms, the tools we develop use the Tree Automata Completion (TAC) algorithm to compute a tree automaton recognizing a superset of all reachable terms. This over-approximation is then used to prove properties on the program by showing that some “bad” terms, encoding dangerous or problematic configurations, are not in the superset and thus not reachable. This is a specific form of, so-called, Regular Tree Model Checking. In [16], we have shown two results. The first result is a precision result guaranteeing that, for most of term rewriting systems known to have a regular set of reachable terms, TAC always compute it in an exact way. The second result shows that tree automata completion can be applied to functional programs to over-approximate their image. In particular, we have shown that tree automata completion computes a safe over-approximation of the image of any first-order, purely functional, complete and terminating program. Now, our first next objective is to demonstrate the accuracy of those regular approximations to perform lightweight formal verification of functional programs. The second objective is to lift those results to higher-order purely functional programs.

CIDRE Project-Team

7. New Results

7.1. Intrusion Detection

7.1.1. Intrusion Detection in Distributed Systems

Alert Correlation: In large systems, multiple (host and network) Intrusion Detection Systems (IDS) and many sensors are usually deployed. They continuously and independently generate notifications (event's observations, warnings and alerts). To cope with this amount of collected data, alert correlation systems have to be designed. An alert correlation system aims at exploiting the known relationships between some elements that appear in the flow of low level notifications to generate high semantic meta-alerts. The main goal is to reduce the number of alerts returned to the security administrator and to allow a higher level analysis of the situation. However, producing correlation rules is a highly difficult operation, as it requires both the knowledge of an attacker, and the knowledge of the functionalities of all IDSes involved in the detection process. In the context of the PhD of Erwan Godefroy [1], we focus on the transformation process that allows to translate the description of a complex attack scenario into correlation rules and its assessment. We show that, once a human expert has provided an action tree derived from an attack tree, a fully automated transformation process can generate exhaustive correlation rules that would be tedious and error prone to enumerate by hand.

Long lived attack campaigns known as Advanced Persistent Threats (APTs) have emerged as a serious security risk. These attack campaigns are customised for their target and performed step by step during months on end. The major difficulty in detecting an APT is keeping track of the different steps logged over months of monitoring and linking them. In [11], we describe TerminAPTor, an APT detector which highlights links between the traces left by attackers in the monitored system during the different stages of an attack campaign. TerminAPTor tackles this challenge by resorting to Information Flow Tracking (IFT). Our main contribution is showing that IFT can be used to highlight APTs. Additionally, we describe a generic representation of APTs and validate our IFT-based APT detector.

Inferring the normal behavior of an application: In [29], [6], [41], we propose an approach to detect intrusions that affect the behavior of distributed applications. To determine whether an observed behavior is normal or not (occurence of an attack), we rely on a model of normal behavior. This model has been built during an initial training phase (machine learning approach). During this preliminary phase, the application is executed several times in a safe environment. The gathered traces (sequences of actions) are used to generate an automaton that characterizes all these acceptable behaviors. To reduce the size of the automaton and to be able to accept more general behaviors that are close to the observed traces, the automaton is transformed. These transformations may lead to introduce unacceptable behaviors. Our current work aims at identifying the possible errors tolerated by the compacted automaton.

This approach is particularly appealing to detect intrusions in industrial control systems since these systems exhibit well-defined behaviors at different levels: network level (network communication patterns, protocol specifications, etc.), control level (continue and discrete process control laws), or even the state of the local resources (memory or CPU). Industrial control systems (ICS) can be subject to highly sophisticated attacks which may lead the process towards critical states. Due to the particular context of ICS, protection mechanisms are not always practical, nor sufficient. On the other hand, developing a process-aware intrusion detection solution with satisfactory alert characterization remains an open problem. In [20], we focus on process-aware attacks detection in sequential control systems. We build on results from runtime verification and specification mining to automatically infer and monitor process specifications. Such specifications are represented by sets of temporal safety properties over states and events corresponding to sensors and actuators. The properties are then synthesized as monitors which report violations on execution traces. We develop an efficient specification mining algorithm and use filtering rules to handle the large number of mined properties. Furthermore, we introduce the notion of activity and discuss its relevance to both specification mining and attack detection

in the context of sequential control systems. The proposed approach is evaluated in a hardware-in-the-loop setting subject to targeted process-aware attacks. Overall, due to the explicit handling of process variables, the solution provides a better characterization of the alerts and a more meaningful understanding of false positives.

7.1.2. Illegal Information Flow Detection

Our research work on intrusion detection based on information flow has been initiated in 2002. This research work has resulted in Blare, a framework for Intrusion Detection Systems [0], including KBlare, an implementation as a Linux Security Module (LSM), JBlare, an implementation for the Java Virtual Machine (JVM), and AndroBlare, for Android applications.

Illegal Information Flow in Web-browser: In the context of the CominLabs SECCLOUD project, we were interested in implementing our approach to detect illegal information flow in web-browser. We have proposed a new secure information flow control model specifically designed for JavaScript [28]. In our approach, we augment the standard symbol table with a mechanism that replaces the reference address for secret values based on the current execution stack. This mechanism also ensures that the secret is stored in a dedicated memory location thereby protecting the secret from any unintended leakage or modification by a malicious JavaScript. This work on detection of illegal information flow in JavaScript has received the best paper award at the 9th International Conference on Security of Information and Networks (SIN 2016) [28].

Later Deepak Subramanian has improved this approach and optimized the computation time required to determine the legacy of information flows. An approach which begins with a learning phase allows to increase the accuracy of the proposed solution. Information about the modified variables are kept in memory to perform a more accurate analysis of the indirect information flows. This self-correcting information flow control model for a web-browser is described in [27].

Information Leaks: Qualitative information flow aims at detecting information leaks, whereas the emerging quantitative techniques target the estimation of information leaks. Quantifying information flow in the presence of low inputs is challenging, since the traditional techniques of approximating and counting the reachable states of a program no longer suffice. In [32], we propose an automated quantitative information flow analysis for imperative deterministic programs with low inputs. The approach relies on a novel abstract domain, the cardinal abstraction, in order to compute a precise upper-bound over the maximum leakage of batch-job programs. We prove the soundness of the cardinal abstract domain by relying on the framework of abstract interpretation. We also prove its precision with respect to a flow-sensitive type system for the two-point security lattice.

More generally, for his research activities during his PhD thesis, Mounir Assaf has received the 2016 thesis prize awarded by the GDR GPL (Engineering Programming and Software).

Characterizing Android Malwares: Android has become the world's most popular mobile operating system, and consequently the most popular target for unscrupulous developers. These developers seek to make money by taking advantage of Android users who customise their devices with various applications, which are the main malware infection vector. Indeed, the most likely way a user executes a repackaged application is by downloading a seemingly harmless application from a store and executing it. Such an application may have been modified by an attacker in order to add malicious pieces of code.

To fight repackaged applications containing malicious code, most official application marketplaces have implemented security analysis tools that try to detect and remove malware. Countermeasures adopted by the attackers to bypass these new controls can be divided into two main approaches: avoiding static analysis and avoiding dynamic analysis [39]. A static analysis of an application consists of analysing its code and its resources without executing it. Conversely, dynamic analysis stands for any kind of analysis that requires executing the application in order to observe its actions.

The Kharon project [19] goes a step further from classical dynamic analysis of malware (http://kharon.gforge.inria.fr). Funded by the Labex CominLabs and involving partners of Centrale-Supélec, Inria and INSA Centre Val de Loire, this project aims to capture a compact and comprehensive

[0] http://www.blare-ids.org

representation of malware. To achieve such a goal we have developed tools to monitor operating systems' information flows induced by the execution of a marked application. We support the idea that the best way to understand malware impact is to observe it in its normal execution environment i.e., a real smartphone. Additionally, the main challenge is to be able to trigger malicious behaviours even if the malware tries to escape dynamic analysis.

In this context, we have developed an original solution that mainly consists of 'helping the malware to execute'. In other words we slightly modify the bytecode of the infected application in order to defeat the protection against dynamic analysis and we execute the suspicious code in its most favourable execution conditions. Thus, our software helps us understand malware's objectives and the consequences on the health of a user's device. In particular, we use a global control flow graph (CFG) to exhibit an execution path to reach specific parts of code [42].

To achieve stealthiness when attacking a mobile device, an effective approach is the use of a covert channel built by two colluding applications to locally exchange data. Since this process is tightly coupled with the used hiding method, its detection is a challenging task, also worsened by the very low transmission rates. Using general indicators such as the energy consumed by the device, we propose in [5] an approach to detect the hidden data exchange between colluding applications and show its feasibility and effectiveness through different experimental results.

Our main research direction and challenge is to develop new and original protections against malicious applications that try to defeat classical dynamic analysis.

7.1.3. Intrusion Detection in Low-Level Software Components

In order to protect the IDS itself, we have initiated different research activities in the domain of hardware security. Our goal is to use co-design software/hardware approaches against traditional software attacks. In a bilateral research project with HP Inc Research Labs, we investigate how dedicated hardware could be used to monitor the whole software stack (from the firmware to the user-mode applications). In the CominLabs HardBlare project, we study the use of a dedicated co-processor to enforce Dynamic Information Flow Control on the main CPU. Finally, in the context of the PhD thesis of Thomas Lethan (ANSSI), we investigate the use of formal methods to evaluate the security guarantees provided by hardware platforms, which combine different CPUs, chipsets and memories. Over time, hardware designs have constantly grown in complexity and modern platforms involve multiple interconnected hardware components. During the last decade, several vulnerability disclosures have proven that trust in hardware can be misplaced. In [21], [37], we give a formal definition of Hardware-based Security Enforcement (HSE) mechanisms, a class of security enforcement mechanisms such that a software component relies on the underlying hardware platform to enforce a security policy. We then model a subset of a x86-based hardware platform specifications and we prove the soundness of a realistic HSE mechanism within this model using Coq, a proof assistant system.

The HardBlare project proposes a software/hardware co-design methodology to ensure that security properties are preserved all along the execution of the system but also during files storage. It is based on the Dynamic Information Flow Tracking (DIFT) that generally consists in attaching tags to denote the type of information that are saved or generated within the system. These tags are then propagated when the system evolves and information flow control is performed in order to guarantee the safe execution and storage within the system monitored by security policies [43].

In [30] we introduce an efficient approach for DIFT (Dynamic Information Flow Tracking) implementations on reconfigurable chips. Existing solutions are either hardly portable or bring unsatisfactory time overheads. This work presents an innovative implementation for DIFT on reconfigurable SoCs such as Xilinx Zynq devices.

In [7], we detail a hardware-assisted approach for information flow tracking implemented on reconfigurable chips. Current solutions are either time-consuming or hardly portable (modifications of both software/hardware layers). This work takes benefits from debug components included in ARMv7 processors to retrieve details on instructions committed by the CPU. First results in terms of silicon area and time overheads are also given.

7.1.4. Vizualization

The large quantities of alerts generated by intrusion detection systems (IDS) make very difficult to distinguish on a network real threats from noise. To help solving this problem, we propose VEGAS [12], an alerts visualization and classification tool that allows first line security operators to group alerts visually based on their principal component analysis (PCA) representation. VEGAS is included in a workflow in such a way that once a set of similar alerts has been collected and diagnosed, a filter is generated that redirects forthcoming similar alerts to other security analysts that are specifically in charge of this set of alerts, in effect reducing the flow of raw undiagnosed alerts.

Our research on visualization of security events has lead to two proofs-of-concept (See ELVIS and VEGAS softwares). We are currently pursuing business opportunities on this topic. Indeed SplitSec is a soon to be founded startup developing tools to help security experts to better manage and understand security data. Scalable analysis solutions and data visualisations adapted for security are combined into powerful tools for incident response. Christopher Humphries is a technology transfer engineer employed by Inria to build these tools based on promising research prototypes.

7.2. Privacy

7.2.1. Image Encryption

More and more users prefer to share their photos through image-sharing platforms of social networks than using e-mail or personal webpages. Since the provider of the image-sharing platform can clearly know the contents of any published images, the users have to trust the provider to respect their privacy or has to encrypt their images. In the context of the PhD of Kun He [18], [17], [16], we have proposed an IND-CPA image encryption algorithm that preserve the image format after encryption, and we have shown that our encryption algorithm can be used on several widely used image-sharing platforms such as Flickr, Pinterest, Google+ and Twitter.

7.2.2. Fingerprinting

Active fingerprinting schemes were originally invented to deter malicious users from illegally releasing an item, such as a movie or an image. To achieve this, each time an item is released, a different fingerprint is embedded in it. In the context of the PhD of Julien Lolive, we have defined the first privacy-preserving asymmetric fingerprinting protocol based on Tardos codes [2]. This protocol is optimal with respect to traitor tracing. We also formally proved that our protocol achieves the properties of correctness, anti-framing, traitor tracing, as well as buyer- and item-unlinkability.

7.3. Communication and Synchronization in Distributed Systems

7.3.1. Routing Protocol for Tactical Mobile Ad Hoc Networks

In the context of the PhD thesis of Florian Grandhomme, we propose new secure and efficient algorithms and protocols to provide inter-domain routing in the context of tactical mobile ad hoc network. The proposed protocol has to handle context modification due to the mobility of Mobile Ad hoc NETwork (MANET), that is to say split of a MANET, merge of two or more MANET, and also handle heterogeneity of technology and infrastructure. The solution has to be independent from the underlying intra-domain routing protocol and from the infrastructure: wired or wireless, fixed or mobile. This work is done in cooperation with DGA-MI.

New generation military equipment, soldiers and vehicles, use wireless technology to communicate on the battlefield. During missions, they form a MANET. Since the battlefield includes coalition, each group may communicate with another group, and inter-MANET communication may be established. Inter-MANET (or inter-domain MANET) communication should allow communication, but maintain a control on the exchanged information. Several protocols have been proposed in order to handle inter-domain routing for tactical MANETs. In [14], [33], we describe and compare three solutions. Based on this analysis, we propose some preconizations to design Inter-domain protocols for MANET.

In [15], we present a coalition context and describe the functional hypothesis we used. Then, we propose a protocol that would fit such a network and conduct experimentation that tend to show that our proposition is quite efficient.

7.3.2. Communication and Synchronization Primitives

Use of Primitives to Limit Equivocation: We consider the approximate consensus problem in a partially connected network of n nodes where at most f nodes may suffer from Byzantine faults. In [22], we study under which conditions this problem can be solved using an iterative algorithm. A Byzantine node can equivocate: it may provide different values to its neighbors. To restrict the possibilities of equivocation, the 3-partial multicast primitive is considered. When a (correct or faulty) node uses this communication primitive, it provides necessarily the same value to the two identified receivers. Based on this communication primitive, a novel condition called f-resilient is proposed and proved to be necessary and sufficient to solve the approximate Byzantine consensus problem in a synchronous network.

The Test&Set Problem: In [35], we present a solution to the well-known problem of synchronization in a distributed asynchronous system prone to process crashes. This problem is also known as the Test&Set problem. The Test&Set is a distributed synchronization protocol that, when invoked by a set of processes, returns a unique winning process. This unique process is then allowed to use, for instance, a shared resource. Recently many advances in implementing Test&Set objects have been achieved, however all of them uniquely target the shared memory model. In this paper we propose an implementation of a Test&Set object for a message passing distributed system. This implementation can be invoked by any number $n \leq N$ of processes where N is the total number of processes in the system. We show in this paper, using a Markov model, that our implementation has an expected step complexity in O(log n) and we give an explicit formula for the distribution of the number of steps needed to solve the problem.

7.3.3. Dependability in Cloud Storage

The quantity of data in the world is steadily increasing bringing challenges to storage system providers to find ways to handle data efficiently in terms of dependability and in a cost-effectively manner. We have been interested in cloud storage which is a growing trend in data storage solution. For instance, the International Data Corporation (IDC) predicts that by 2020, nearly 40% of the data in the world will be stored or processed in a cloud. The thesis of Pierre Obame [3] addressed challenges around data access latency and dependability in cloud storage. We proposed Mistore, a distributed storage system that we designed to ensure data availability, durability, low access latency by leveraging the Digital Subscriber Line (xDSL) infrastructure of an Internet Service Provider (ISP). Mistore uses the available storage resources of a large number of home gateways, Points of Presence, and datacenters for content storage and caching facilities. Mistore also targets data consistency by providing multiple types of data consistency criteria and a versioning system. We also considered the data security and confidentiality in the context of storage systems applying data deduplication which is becoming one of the most popular data technologies to reduce the storage cost and we design a data deduplication method that is secure against malicious clients while remaining efficient in terms of network bandwidth and storage space savings.

7.3.4. Decentralized Cryptocurrency Systems

Decentralized cryptocurrency systems offer a medium of exchange secured by cryptography, without the need of a centralized banking authority. Among others, Bitcoin is considered as the most mature one [10]. Its popularity lies on the introduction of the concept of the blockchain, a public distributed ledger shared by all participants of the system. Double spending attacks and blockchain forks are two main issues in blockchain-based protocols. The first one refers to the ability of an adversary to use the very same bitcoin more than once, while blockchain forks cause transient inconsistencies in the blockchain. In [9], we show through probabilistic analysis that the reliability of recent solutions that exclusively rely on a particular type of Bitcoin actors, called miners, to guarantee the consistency of Bitcoin operations, drastically decreases with the size of the blockchain.

Some recent works have proposed to improve upon Bitcoin weaknesses. In [31], we analyze of one of these recent works, and shows through an analytical performance evaluation that new Bitcoin improvements are still needed.

7.3.5. Large Scale Systems

Population Protocol: the computational model of population protocols is a formalism that allows the analysis of properties emerging from simple and pairwise interactions among a very large number of anonymous finite-state agents. Significant work has been done so far to determine which problems are solvable in this model and at which cost in terms of states used by the protocols and time needed to converge. The problem tackled in [23] is the population proportion problem: each agent starts independently from each other in one of two states, say A or B, and the objective is for each agent to determine the proportion of agents that initially started in state A, assuming that each agent only uses a finite set of state, and does not know the number n of agents. We propose a solution which guarantees with any high probability that after O(log n) interactions any agent outputs with a precision given in advance, the proportion of agents that start in state A. The population proportion problem is a generalization of both the majority and counting problems, and thus our solution solves both problems. We show that our solution is optimal in time and space. Simulation results illustrate our theoretical analysis.

Propagation Time of a Rumor: the context of this work is the well studied dissemination of information in large scale distributed networks through pairwise interactions. This problem, originally called rumor mongering, and then rumor spreading has mainly been investigated in the synchronous model. This model relies on the assumption that all the nodes of the network act in synchrony, that is, at each round of the protocol, each node is allowed to contact a random neighbor. In [24], we drop this assumption under the argument that it is not realistic in large scale systems. We thus consider the asynchronous variant, where at time unit, a single node interacts with a randomly chosen neighbor. We perform a thorough study of the total number of interactions needed for all the nodes of the network to discover the rumor.

Distributed Stream Processing Systems: shuffle grouping is a technique used by stream processing frameworks to share input load among parallel instances of stateless operators. With shuffle grouping each tuple of a stream can be assigned to any available operator instance, independently from any previous assignment. A common approach to implement shuffle grouping is to adopt a Round-Robin policy, a simple solution that fares well as long as the tuple execution time is almost the same for all the tuples. However, such an assumption rarely holds in real cases where execution time strongly depends on tuple content. As a consequence, parallel stateless operators within stream processing applications may experience unpredictable unbalance that, in the end, causes undesirable increase in tuple completion times. In [25], [26] we propose Online Shuffle Grouping (OSG), a novel approach to shuffle grouping aimed at reducing the overall tuple completion time. OSG estimates the execution time of each tuple, enabling a proactive and online scheduling of input load to the target operator instances. Sketches are used to efficiently store the otherwise large amount of information required to schedule incoming load. We provide a probabilistic analysis and illustrate, through both simulations and a running prototype, its impact on stream processing applications.

Load shedding is a technique employed by stream processing systems to handle unpredictable spikes in the input load whenever available computing resources are not adequately provisioned. A load shedder drops tuples to keep the input load below a critical threshold and thus avoid unbounded queuing and system trashing. In [38] we propose Load-Aware Shedding (LAS), a novel load shedding solution that, unlike previous works, does not rely neither on a pre-defined cost model nor on any assumption on the tuple execution duration. Leveraging sketches, LAS efficiently builds and maintains at runtime a cost model to estimate the execution duration of each tuple with small error bounds. This estimation enables a proactive load shedding of the input stream at any operator that aims at limiting queuing latencies while dropping as few tuples as possible. We provide a theoretical analysis. Furthermore, through an extensive practical evaluation based on simulations and a prototype, we evaluate its impact on stream processing applications, which validate the robustness and accuracy of LAS.

DIONYSOS Project-Team

7. New Results

7.1. Performance Evaluation of Call Centers

Participant: Pierre L'Ecuyer.

We develop research activities around the analysis and design of call centers, from a performance perspective. The effective management of call centers is a challenging task mainly because managers are consistently facing considerable uncertainty.

One aspect studied in [23] is the development of stochastic models for the daily arrival rate in a call center. Models in which the busyness factors are independent across periods, or in which a common busyness factor applies to all periods, have been studied previously. But they are not sufficiently realistic. We examine alternative models for which the busyness factors have some form of dependence across periods.

We also carry out in [14] large-scale data-based investigation of service times in a call center with many heterogeneous agents and multiple call types to investigate the validity of traditionally used standard Erlang queueing models, based on independent and identically distributed exponential random variables. Our study provides empirical support to the theoretical research that goes beyond standard modelling assumptions in service systems.

In [56], we consider a stochastic staffing problem with uncertain arrival rates. The objective is to minimize the total cost of agents under some chance constraints, defined over the randomness of the service level in a given time period. We present a method that combines simulation, mixed integer programming, and cut generation to solve this problem. In [84], we consider a particular staffing problem with probabilistic constraints in an emergency call center. We propose an algorithm to solve the problem, and validate it with a simulation model based on real data from the 911 emergency call center of Montreal, Canada.

We are also interested in predicting the waiting time of customers upon their arrival in some service system such as a call center or emergency service. In [86], we propose two new predictors that are very simple to implement and can be used in multiskill settings. They are based on the waiting times of previous customers of the same class. In our simulation experiments, these new predictors are very competitive with the optimal ones for a simple queue, and for multiskill centers they perform better than other predictors of comparable simplicity.

7.2. Analytic models

Participants: Gerardo Rubino, Bruno Sericola.

Sojourn times in Markovian models. In [98], we discuss different issues related to the time a Markov chain spends in a part of its state space. This is relevant in many application areas including those interesting Dionysos, namely, in the performance and dependability analysis of complex systems. For instance, in dependability, the reliability of a system subject to failures and repairs of its components, is, in terms of a discrete-space model of it, the probability that it remains in the subset of operational or up states during the whole time interval $[0, t]$. In performance, the occupancy factor of some server is the probability that, in steady state, the model belongs to the subset of states where the server is busy. This book chapter reviews some past work done by the authors on this topic (see our book [111] for a synthesis of these works), and add some new insights on the properties of these sojourn times.

Queuing systems in equilibrium. In the late 70s, Leonard Kleinrock proposed a metric able to capture the tradeoff between the work done by a system and its cost, or, in terms of queueing systems, between throughput and mean response time. The new metric was called *power* and among its properties, it satisfies a nice one informally called "keep the pipe full", specifying that the operation point of many queues that maximizes their power also leads to a mean backlog equal to exactly one customer. Last year [110] we explored what happens with this metric when we consider Jackson queuing networks. After showing that the same property holds for them, we showed that the power metric has some drawbacks, mainly when considering multiserver queues and networks of queues. We then proposed a new metric that we called *effectiveness*, identical to power when there is a single queue with a single server, but different otherwise, that avoids these drawbacks. We analyze it and, in particular, we showed that the same "keep the pipe full" holds for it. In the keynote [34] we presented these ideas together with some new results (for example, the analysis of G-queues from this point of view).

For other analytical-oriented work, see [72] for new applications of queueing theory used at the Markovian level, and [72] for applications of stochastic analysis to general problems where performance and dependability are simultaneously taken into account in the same model.

7.3. Performance Evaluation of Distributed Systems

Participants: Bruno Sericola, Yann Busnel, Yves Mocquard.

Detection of distributed deny of service attacks. A Deny of Service (DoS) attack tries to progressively take down an Internet resource by flooding this resource with more requests than it is capable to handle. A Distributed Deny of Service (DDoS) attack is a DoS attack triggered by thousands of machines that have been infected by a malicious software, with as immediate consequence the total shut down of targeted web resources (*e.g.*, e-commerce websites). A solution to detect and to mitigate DDoS attacks it to monitor network traffic at routers and to look for highly frequent signatures that might suggest ongoing attacks. A recent strategy followed by the attackers is to hide their massive flow of requests over a multitude of routes, so that locally, these flows do not appear as frequent, while globally they represent a significant portion of the network traffic. The term "iceberg" has been recently introduced to describe such an attack as only a very small part of the iceberg can be observed from each single router. The approach adopted to defend against such new attacks is to rely on multiple routers that locally monitor their network traffic, and upon detection of potential icebergs, inform a monitoring server that aggregates all the monitored information to accurately detect icebergs [41]. Now to prevent the server from being overloaded by all the monitored information, routers continuously keep track of the c (among n) most recent high flows (modeled as items) prior to sending them to the server, and throw away all the items that appear with a small probability. Parameter c is dimensioned so that the frequency at which all the routers send their c last frequent items is low enough to enable the server to aggregate all of them and to trigger a DDoS alarm when needed. This amounts to compute the time needed to collect c distinct items among n frequent ones. A thorough analysis of the time needed to collect c distinct items appears in [10].

Stream Processing Systems. Stream processing systems are today gaining momentum as tools to perform analytics on continuous data streams. Their ability to produce analysis results with sub-second latencies, coupled with their scalability, makes them the preferred choice for many big data companies.

A stream processing application is commonly modeled as a direct acyclic graph where data operators, represented by nodes, are interconnected by streams of tuples containing data to be analyzed, the directed edges (the arcs). Scalability is usually attained at the deployment phase where each data operator can be parallelized using multiple instances, each of which will handle a subset of the tuples conveyed by the operators' ingoing stream. Balancing the load among the instances of a parallel operator is important as it yields to better resource utilization and thus larger throughputs and reduced tuple processing latencies. We have proposed a new key grouping technique targeted toward applications working on input streams characterized by a skewed value distribution [80]. Our solution is based on the observation that when the values used to perform the grouping have skewed frequencies, the few most frequent values (the *heavy hitters*) drive the load distribution, while the remaining largest fraction of the values (the *sparse items*) appear so rarely in the stream that the relative impact of each of them on the global load balance is negligible. We have shown, through a theoretical analysis, that our solution provides on average near-optimal mappings using sub-linear spaces in the number of tuples

read from the input stream in the learning phase and the support (value domain) of the tuples. In particular this analysis presents new results regarding the expected error made on the estimation of the frequency of heavy hitters.

Load shedding is a technique employed by stream processing systems to handle unpredictable spikes in the input load whenever available computing resources are not adequately provisioned. A load shedder drops tuples to keep the input load below a critical threshold and thus avoid unbounded queuing and system trashing. In [102] and [79] we propose Load-Aware Shedding (LAS), a novel load shedding solution that, unlike previous works, does not rely neither on a pre-defined cost model nor on any assumption on the tuple execution duration. Leveraging sketches, LAS efficiently builds and maintains at runtime a cost model to estimate the execution duration of each tuple with small error bounds. This estimation enables a proactive load shedding of the input stream at any operator that aims at limiting queuing latencies while dropping as few tuples as possible. We provide a theoretical analysis proving that LAS is an (ε, δ)-approximation of the optimal online load shedder. Furthermore, through an extensive practical evaluation based on simulations and a prototype, we evaluate its impact on stream processing applications, which validate the robustness and accuracy of LAS.

Shuffle grouping is a technique used by stream processing frameworks to share input load among parallel instances of stateless operators. With shuffle grouping each tuple of a stream can be assigned to any available operator instance, independently from any previous assignment. A common approach to implement shuffle grouping is to adopt a Round-Robin policy, a simple solution that fares well as long as the tuple execution time is almost the same for all the tu-ples. However, such an assumption rarely holds in real cases where execution time strongly depends on tuple content. As a consequence, parallel stateless operators within stream processing applications may experience unpredictable unbalance that, in the end, causes undesirable increase in tuple completion times. In [77] we propose Online Shuffle Grouping (OSG), a novel approach to shuffle grouping aimed at reducing the overall tuple completion time. OSG estimates the execution time of each tuple, enabling a proactive and online scheduling of input load to the target operator instances. Sketches are used to efficiently store the otherwise large amount of information required to schedule incoming load. We provide a probabilistic analysis and illustrate, through both simulations and a running prototype, its impact on stream processing applications.

Estimating the frequency of any piece of information in large-scale distributed data streams became of utmost importance in the last decade (*e.g.*, in the context of network monitoring, big data, *etc.*). If some elegant solutions have been proposed recently, their approximation is computed from the inception of the stream. In a runtime distributed context, one would prefer to gather information only about the recent past. This may be led by the need to save resources or by the fact that recent information is more relevant. In [78], we consider the *sliding window* model and propose two different (on-line) algorithms that approximate the items frequency in the active window. More precisely, we determine a (ε, δ)-additive-approximation meaning that the error is greater than ε only with probability δ. These solutions use a very small amount of memory with respect to the size N of the window and the number n of distinct items of the stream, namely, $O(\frac{1}{\varepsilon}\log\frac{1}{\delta}(\log N + \log n))$ and $O(\frac{1}{\tau\varepsilon}\log\frac{1}{\delta}(\log N + \log n))$ bits of space, where τ is a parameter limiting memory usage. We also provide their distributed variant, *i.e.*, considering the *sliding window functional monitoring* model, with a communication cost of $O(\frac{k}{\varepsilon^2}\log\frac{1}{\delta}\log N)$ bits per window (where k is the number of nodes). We compared the proposed algorithms to each other and also to the state of the art through extensive experiments on synthetic traces and real data sets that validate the robustness and accuracy of our algorithms.

Randomized Message-Passing Test-and-Set. In [101], we have presented a solution to the well-known Test&Set operation in an asynchronous system prone to process crashes. Test&Set is a synchronization operation that, when invoked by a set of processes, returns yes to a unique process and returns no to all the others. Recently, many advances in implementing Test&Set objects have been achieved. However, all of them target the shared memory model. In this paper we propose an implementation of a Test&Set object in the message passing model. This implementation can be invoked by any number $p \leq n$ of processes where n is the total number of processes in the system. It has an expected individual step complexity in $O(\log p)$ against an oblivious adversary, and an expected individual message complexity in $O(n)$. The proposed Test&Set object is built atop a new basic building block, called selector, that allows to select a winning group among two

groups of processes. We propose a message- passing implementation of the selector whose step complexity is constant. We are not aware of any other implementation of the Test&Set operation in the message passing model.

Throughput Prediction in Cellular Networks Downlink data rates can vary significantly in cellular networks, with a potentially non-negligible effect on the user experience. Content providers address this problem by using different representations (*e.g.*, picture resolution, video resolution and rate) of the same content and switch among these based on measurements collected during the connection. If it were possible to know the achievable data rate before the connection establishment, content providers could choose the most appropriate representation from the very beginning. We have conducted a measurement campaign involving 60 users connected to a production network in France, to determine whether it is possible to predict the achievable data rate using measurements collected, before establishing the connection to the content provider, on the operator's network and on the mobile node. We show that it is indeed possible to exploit these measurements to predict, with a reasonable accuracy, the achievable data rate [81].

Population Protocol Model. The computational model of population protocols, introduced by Angluin and his colleagues in 2006, is a formalism that allows the analysis of properties emerging from simple and pairwise interactions among a very large number of anonymous finite-state agents. Significant work has been done so far to determine which problems are solvable in this model and at which cost in terms of states used by the protocols and time needed to converge. The problem tackled in [74] is the population proportion problem: each agent starts independently from each other in one of two states, say A or B, and the objective is for each agent to determine the proportion of agents that initially started in state A, assuming that each agent only uses a finite set of state, and does not know the number n of agents. We propose a solution which guarantees with any high probability that after $O(\log n)$ interactions any agent outputs with a precision given in advance, the proportion of agents that start in state A. The population proportion problem is a generalization of both the majority and counting problems, and thus our solution solves both problems. We show that our solution is optimal in time and space. Simulation results illustrate our theoretical analysis.

The context of [75] is the well studied dissemination of information in large scale distributed networks through pairwise interactions. This problem, originally called "rumor mongering", and then "rumor spreading", has mainly been investigated in the synchronous model. This model relies on the assumption that all the nodes of the network act in synchrony, that is, at each round of the protocol, each node is allowed to contact a random neighbor. In this paper, we drop this assumption under the argument that it is not realistic in large scale systems. We thus consider the asynchronous variant, where at time unit, a single node interacts with a randomly chosen neighbor. We perform a thorough study of T_n, the total number of interactions needed for all the n nodes of the network to discover the rumor. While most of the existing results involve huge constants that do not allow for comparing different protocols, we prove that in a complete graph of size $n \geq 2$, the probability that $T_n > k$ for all $k \geq 1$ is less than $(1 + 2k(n-2)^2/n)(1-2/n)^{k-1}$. We also study the behavior of the complementary distribution of T_n at point $cE(T_n)$ when n tends to infinity, in function of c. This paper received the Best Student Paper Award from the 15th IEEE Symposium on Network Computing and Applications (IEEE NCA 2016).

Bitcoin. Decentralized cryptocurrency systems offer a medium of exchange secured by cryptography, without the need of a centralized banking authority. Among others, Bitcoin is considered as the most mature one. Its popularity lies on the introduction of the concept of the blockchain, a public distributed ledger shared by all participants of the system. Double spending attacks and blockchain forks are two main issues in blockchain-based protocols. The first one refers to the ability of an adversary to use the very same bitcoin more than once, while blockchain forks cause transient inconsistencies in the blockchain. We show in [43], [89], [42] through probabilistic analysis that the reliability of recent solutions that exclusively rely on a particular type of Bitcoin actors, called miners, to guarantee the consistency of Bitcoin operations, drastically decreases with the size of the blockchain.

7.4. Future networks and architectures

Participants: Adlen Ksentini, Bruno Sericola, Yassine Hadjadj-Aoul, Jean-Michel Sanner, Hamza Ben Ammar.

SDN and NFV. Network Function Visualization (NFV) and Software Defined Network (SDN) currently play a key role to transform the network architecture from hardware-based to software-based.

SDN is in the process of revolutionizing the way of managing networks by providing a new way to support current and future services. However, by relocating the control functionality in a remote entity, the measurements' accuracy of the resources' utilization becomes more difficult, which complicates the decision making. Although there are previous works focusing on the problem of network management and measurement in SDN networks, only a few proposed solutions have taken into consideration the trade-off existing between statistics' polling frequency (i.e. generated overhead), and the accuracy of monitoring results (i.e. optimized resourcesâ allocation). In [62], we proposed a new approach calculating accurately the bandwidth utilization while adapting the polling frequency according to ports/switches activity. The emulations' results under Mininet clearly demonstrate the effectiveness of the proposed solution, which proved to be scalable compared to classical approaches. The controllers' placement is another important concern that emerged recently to solve the scalability and the reliability issues of SDN networks. The placement efficiency is influenced by both network operators (NO) strategy and the supported service requirements, which makes more complex the decision-making process. In particular, the need to support QoS-constrained services may lead NO to guide the controllers' placement in a way to ensure services efficiency while optimizing the underlying infrastructure. In [82] and [66], we proposed a model for the placement of network controllers, and we formulated a general optimization problem. To provide more flexibility and to avoid time-prohibitive calculations, we proposed a hierarchical clustering strategy for the controllers' placement allowing to minimize the number of network controllers while reducing the potential disparity of burden between the different controllers. Besides, the algorithms' structure makes it easy to act on other network parameters to improve the reliability of the SDN network. In [107], we proposed an improvement of such algorithms, by considering an evolutionary solution based on a genetic technique with an ad hoc cross-over operator designed to solve a mono-objective controller placement problem.

To connect the VNFs hosted in the same Data Center (DC) or across multiple DCs, virtual switches are required. Besides forwarding functions, virtual switches can be configured to mirror traffics for network management needs. Among the existing virtual switch solutions, Open vSwitch (OVS) is the most known and used. OVS is open source, and included in most of the existing Linux distributions. However, OVS performance in terms of throughput for smaller packets is very smaller than of line rate of the interface. To overcome this limitation, OVS was ported to Data Plane Development Kit (DPDK), namely OVDK. The latter achieves an impressive line rate throughput across physical interfaces. In [83], we presented the result of OVDK performance test when flow and port mirroring are activated, which was not tested so far. The performance test focuses on two parameters, throughput and latency in OVDK, allowing to validate the use of OVDK for flow forwarding and network management in the envisioned virtualized network architecture.

Mobile cloud. To cope with the tremendous growth in mobile data traffic on one hand, and the modest average revenue per user on the other hand, mobile operators have been exploring network virtualization and cloud computing technologies to build cost-efficient and elastic mobile networks and to have them offered as a cloud service. In such cloud-based mobile networks, ensuring service resilience is an important challenge to tackle. Indeed, high availability and service reliability are important requirements of carrier grade, but not necessarily intrinsic features of cloud computing. Building a system that requires the five nines reliability on a platform that may not always grant it is therefore a hurdle. Effectively, in carrier cloud, service resilience can be heavily impacted by a failure of any network function (NF) running on a virtual machine (VM). In [31], we introduce a framework, along with efficient and proactive restoration mechanisms, to ensure service resilience in carrier cloud. As restoration of a NF failure impacts a potential number of users, adequate network overload control mechanisms are also proposed. A mathematical model is developed to evaluate the performance of the proposed mechanisms. The obtained results are encouraging and demonstrate that the proposed mechanisms efficiently achieve their design goals.

Typically, maintaining a static pool of cloud resources to meet peak requirements with good service quality makes the cloud infrastructure costly. To cope with this, [58] proposes an approach that enables a cloud infrastructure to automatically and dynamically scale-up or scale-down resources of a virtualized environment aiming for efficient resource utilization and improved quality of experience (QoE) of the offered services. The QoE-aware approach ensures a truly elastic infrastructure, capable of handling sudden load surges while reducing resource and management costs. The paper also discusses the applicability of the proposed approach within the ETSI NFV MANO framework for cloud-based 5G mobile systems.

Video distribution. Due to the Internet usage evolution over these last years, the current IP-based architecture becomes heavier and less efficient for providing Internet services. In order to face this shortcoming, "Content Centric Networking" has been proposed. One of its important features is the use of in-network caching as a way of improving network performance and service scalability. However, in most of the existing CCN-based approaches several copies of the same content are present in the network, which reduce its efficiency. In [45], we proposed the "CLIque-based cooperative Caching" (CLIC) strategy, which basically consists in detecting cliques within the network topology to allocate more efficiently the content in the network. The main motivation of the proposed solution is to eliminate contents' redundancy between neighboring nodes while promoting the most popular contents. This approach guarantees a sufficient number of copies of popular files within the network while maximizing the number of distinct content items. We evaluate the proposed scheme through simulation. The results show significant improvements in terms of cache management and network performance.

In [59], we make the case for opening the telco CDN infrastructure to content providers by means of network function virtualization (NFV) and cloud technologies. We design and implement a CDN-as-a-Service architecture, where content providers can lease CDN resources on demand at regions where the ISP has presence. Using open northbound RESTful APIs, content providers can express performance requirements and demand specifications, which can be translated to an appropriate service placement on the underlying cloud substrate. To gain insight which can be applied to the design of such service placement mechanisms, we evaluate the capabilities of key enabling virtualization technologies by extensive testbed experiments.

Network design using new dependability metrics. When designing a network taking into account its capabilities face to possible failures to its components, the basic theoretical framework is classical network reliability, where the system under study is represented by a graph with perfect nodes and imperfect links randomly and independently failing. The corresponding connectivity-based metrics must then be evaluated in order to quantify the robustness of the networking architecture. Recently, a new family of metrics, called diameter-constrained, have been proposed and analyzed by Dionysos' collaborators and members. In [53], we developed some elements for a factoring theory associated with these metrics. The paper is focused on the detection of irrelevant components, a key task when evaluating these types of quantities using factorization. The paper also includes a factoring algorithm, which is an up-to-date procedure exploiting all available results for implementing the pivoting idea (proved to be one of the most powerful methods in classical reliability analysis).

In [54], we consider an homogeneous network (identical and independent components). In this context, if p is the probability that each of the components works, then any reliability metric is necessarily a polynomial in p, and computing these metrics can be reduced to counting problems (counting specific classes of paths or of cuts, for instance). In the paper, we quantify, in some sense, the "degree of difficulty" of these counting processes, and we identify the situations where they are "easy". The second contribution of the paper is to propose a fundamental problem from survivable network design, called the Network Utility Problem. The goal is to maximize network utility (defined as the opposite of the level of difficulty minus one), under a minimum edge-connectivity requirement.

Optical network design. Paper [65] presents a fast and accurate mathematical method to evaluate the blocking probability (the probability of a burst loss) in dynamic WDM networks without wavelength conversion (the present used technology). We assume that all links have the same number of wavelengths (the same capacity). The proposed model considers different traffic loads at each network connection (heterogeneous traffic). To take into account the wavelength continuity constraint, the method divides the network into a sequence of

networks where all the links have capacity 1. Every network in the sequence is evaluated separately using an analytical technique. Then, a procedure combines the results of these evaluations in a way that captures the dependencies that occur in the real system due to the competition for bandwidth between the different connections. The method efficiently achieves results very close to those obtained by simulation, but orders of magnitude faster, allowing the evaluation of the blocking probability of all users (connections) for mesh network topologies.

7.5. Network Economics

Participants: Bruno Tuffin, Pierre L'Ecuyer.

The general field of network economics, analyzing the relationships between all acts of the digital economy, has been an important subject for years in the team. The whole problem of network economics, from theory to practice, describing all issues and challenges, is described in our book published in 2014 [109].

Network neutrality. Most of our activity has been devoted to the vivid network neutrality debate, going beyond the traditional for or against neutrality. We especially responded to the public consultation on draft BEREC Guidelines on implementation of net neutrality rules held during Summer 2016.

Network neutrality is often advocated by content providers, stressing that side payments to Internet Service Providers would hinder innovation. However, we also observe some content provider actually paying those fees. In [20] we intend to explain such behaviors through economic modeling, illustrating how side payments can be a way for an incumbent content provider to prevent new competitors from entering the market. We investigate the conditions under which the incumbent can benefit from such a barrier-to-entry, and the consequences of that strategic behavior on the other actors: content providers, users, and the Internet Service Provider. We also describe how the Nash bargaining solution concept can be used to determine the side payment.

In [105], we explain how non neutrality may be pushed by big CPs to their benefits. Major content/service providers are publishing grades they give to ISPs about the quality of delivery of their content. The goal is to inform customers about the "best" ISPs. But this could be an incentive for, or even a pressure on, ISPs to differentiate service and provide a better quality to those big content providers in order to be more attractive. This fits the network neutrality debate, but instead of the traditional vision of ISPs pressing content providers, we face here the opposite situation, still possibly at the expense of small content providers though. We design in [105] a model describing the various actors and their strategies, analyzes it thanks to non-cooperative game theory, and quantifies the impact of those advertised grades with respect to the situation where no grade is published. We illustrate that a non-neutral behavior, differentiating traffic, is not leading to a desirable situation.

While neutrality is focusing on the behavior of ISPs, we claim that the debate should be generalized. Indeed, the reality of the Internet in the 2010s is that various actors contribute to the delivery of data, with sometimes contradictory objectives. We highlight in [19] the fact that neutrality principles can be bypassed in many ways without violating the rules currently evoked in the debate. For example via Content Delivery Networks (CDNs), which deliver content on behalf of content providers for a fee, or via search engines, which can hinder competition and innovation by affecting the visibility and accessibility of content. We therefore call for an extension of the net neutrality debate to all the actors involved in the Internet delivery chain. We particularly challenge the definition of net neutrality as it is generally discussed. Our goal is to initiate a relevant debate for net neutrality in an increasingly complex Internet ecosystem, and to provide examples of possible neutrality rules for different levels of the delivery chain, this level separation being inspired by the OSI layer model.

The impact of a revenue-oriented CDN is particularly investigated in [104] and [70]. Content Delivery Networks (CDN) have become key telecommunication actors. They contribute to improve significantly the quality of services delivering content to end users. However, their impact on the ecosystem (end-users, the network operators and the content providers) raises concerns about their "neutrality", and therefore the question of their inclusion in the network neutrality debate becomes relevant. We compare the outcome with that of a neutral behavior, and at investigating whether some regulation should be introduced. We present a

mathematical model and show that there exists a unique optimal revenue-maximizing policy for a CDN actor, in terms of dimensioning and allocation of its storage capacity, and depending on parameters such as prices for service/transport/storage. In addition, using the real traces, we compare the revenue-based policy with policies based on several fairness criteria. The CDN activity being potentially lucrative and not included in the neutrality debate, we analyze in [71] the revenue-optimal strategies and impact of a vertically integrated ISP-CDNs, which can sell those services to content providers. Our approach is based on an economic model of revenues and costs, and a multilevel game-theoretic formulation of the interactions among actors. Our model incorporates the possibility for the vertically-integrated ISP to partially offer CDN services to competitors in order to optimize the trade-off between CDN revenue (if fully offered) and competitive advantage on subscriptions at the ISP level (if not offered to competitors). Our results highlight two counterintuitive phenomena: an ISP may prefer an independent CDN over controlling (integrating) a CDN; and from the user point of view, vertical integration is preferable to an independent CDN or a no- CDN configuration. Hence, a regulator may want to elicit such CDN-ISP vertical integrations rather than prevent them.

Online platforms and search engines. Another set of key actors in the Internet economy is the online platforms and search engines. When a keyword-based search query is received by a search engine, a classified ads website, or an online retailer site, the platform has exponentially many choices in how to sort the search results. Two extreme rules are (a) to use a ranking based on estimated relevance only, which improves customer experience in the long run because of perceived quality, and (b) to use a ranking based only on the expected revenue to be generated immediately, which maximizes short-term revenue. Typically, these two objectives (and the corresponding rankings) differ. A key question then is what middle ground between them should be chosen.We introduce in [16] stochastic models that yield elegant solutions for this situation, and we propose effective solution methods to compute a ranking strategy that optimizes long-term revenues. This strategy has a very simple form and is easy to implement if the necessary data is available. It consists in ordering the output items by decreasing order of a score attributed to each. This score results from evaluating a simple function of the estimated relevance, the expected revenue of the link, and a real-valued parameter. We find the latter via simulation-based optimization, and its optimal value is related to the endogeneity of user activity in the platform as a function of the relevance offered to them.

The impact on other actors of search engines has led to the so-called search neutrality debate, as a parallel to the network neutrality debate. Search engines accused of biasing the ranking of their organic links to provide a competitive advantage to their own content. Based on the optimal ranking policy for a search engine obtained in [16], we investigate in [67] on an example whether non-neutrality impacts innovation. We illustrate that a revenue-oriented search engine may indeed deter innovation at the content level, hence the validity of the argument (without necessarily meaning that search engines should be regulated).

Sponsored auctions. Advertisement in dedicated webpage spaces or in search engines sponsored slots is usually sold using auctions, with a payment rule that is either per impression or per click. But advertisers can be both sensitive to being viewed (brand awareness effect) and being clicked (conversion into sales). In [33], [92], we generalize the auction mechanism by including both pricing components: the advertisers are charged when their ad is displayed, and pay an additional price if the ad is clicked. Applying the results for Vickrey-Clarke-Groves (VCG) auctions, we show how to compute payments to ensure incentive compatibility from advertisers as well as maximize the total value extracted from the advertisement slot(s). We provide tight upper bounds for the loss of efficiency due to applying only pay-per-click (or pay-per-view) pricing instead of our scheme. Those bounds depend on the joint distribution of advertisement visibility and population likelihood to click on ads, and can help identify situations where our mechanism yields significant improvements. We also describe how the commonly used generalized second price (GSP) auction can be extended to this context.

7.6. Monte Carlo

Participants: Bruno Tuffin, Gerardo Rubino, Pierre L'Ecuyer.

We maintain a research activity in different areas related to dependability, performability and vulnerability analysis of communication systems, using both the Monte Carlo and the Quasi-Monte Carlo approaches to

evaluate the relevant metrics. Monte Carlo (and Quasi-Monte Carlo) methods often represent the only tool able to solve complex problems of these types.

Rare event simulation. However, when the events of interest are rare, simulation requires a special attention, to accelerate the occurrence of the event and get unbiased estimators of the event of interest with a sufficiently small relative variance (see our book [108] for a global introduction to the field). This is the main problem in the area. Dionysos' work focuses then on dealing with the rare event situation, with a particular focus on dependability [40].

A non-negligible part of our activity on the application of rare event simulation was about the evaluation of static network reliability models. In a static network reliability model one typically assumes that the failures of the components of the network are independent. This simplifying assumption makes it possible to estimate the network reliability efficiently via specialized Monte Carlo algorithms. Hence, a natural question to consider is whether this independence assumption can be relaxed, while still attaining an elegant and tractable model that permits an efficient Monte Carlo algorithm for unreliability estimation. In [12], we provide one possible answer by considering a static network reliability model with dependent link failures, based on a Marshall-Olkin copula, which models the dependence via shocks that take down subsets of components at exponential times, and propose a collection of adapted versions of permutation Monte Carlo (PMC, a conditional Monte Carlo method), its refinement called the turnip method, and generalized splitting (GS) methods, to estimate very small unreliabilities accurately under this model. The PMC and turnip estimators have bounded relative error when the network topology is fixed while the link failure probabilities converge to zero, whereas GS does not have this property. But when the size of the network (or the number of shocks) increases, PMC and turnip eventually fail, whereas GS works nicely (empirically) for very large networks, with over 5000 shocks in our examples. In [73], we propose a methodology for calibrating a dependent failure model to compute the reliability in a telecommunication network, following a similar starting point (that is, using Marshall-Olkin copulas). In practice, this model is difficult to calibrate because it requires the estimation of a number of parameters that is exponential in the number of links. We formulate an optimization problem for calibrating a Marshall-Olkin copula model to attain given marginal failure probabilities for all links and the correlations between them. Using a geographic failure model, we calibrate various Marshall-Olkin copula models using our methodology, we simulate them, and we benchmark the reliabilities thus obtained. Our experiments show that considering the simultaneous failures of small and connected subsets of links is the key for obtaining a good approximation of reliability, confirming what it is suggested by the telecommunication literature.

A related problem is when links have random capacities and a certain target amount of flow must be carried from some source nodes to some destination nodes is considered in [47]. Each destination node has a fixed demand that must be satisfied and each source node has a given supply. The goal is to estimate the unreliability of the network, defined as the probability that given the realized link capacities, the network cannot carry the required amount of flow to meet the demand at all destination nodes. We adapt GS and PMC to this context. In [55], we explore other methods designed to reduce the variance of the estimators in this context. All of them are adaptations of methods originally developed to make reliability estimations on different network models. These methods are introduced together with a brief review of the algorithms on which they are based.

A new application of our previously designed zero-variance approximation importance sampling method has been developed in [76]: To accurately estimate the reliability of highly reliable rail systems and comply with contractual obligations, rail system suppliers such as ALSTOM require efficient reliability estimation techniques. While in our previous works, the studied graph models were dealing with failing links, we propose an adaptation of the algorithm to evaluate the reliability of real transport systems where nodes are the failing components. This is more representative of railway telecommunication system behavior. Robustness measures of the accuracy of the estimates, bounded or vanishing relative error properties, are discussed and results from a real network (Data Communication System used in automated train control system) showing bounded relative error property, are presented.

Random variable generation. Simulation requires the use of pseudo-random generators. In [18], we examine the requirements and the available methods and software to provide (or imitate) uniform random numbers in parallel computing environments. In this context, for the great majority of applications, independent streams

of random numbers are required, each being computed on a single processing element at a time. Sometimes, thousands or even millions of such streams are needed. We explain how they can be produced and managed. We devote particular attention to multiple streams for GPU devices.

Sampling from the Normal distribution truncated to some finite or semi-infinite interval is of particular interest for certain applications in Bayesian statistics, such as to perform exact posterior simulations for parameter inference. We study and compare in [46] various methods to generate such random variables, with special attention to the situation where the interval is far in the tail. The algorithms are implemented and available in Java, R, and MATLAB, and the software is freely available.

Quasi-Monte Carlo (QMC). Finally, we have continued our work on QMC methods. In [15], we review the Array-RQMC method, its variants, sorting strategies, and convergence results. We are interested in the convergence rate of measures of discrepancy of the states at a given step of the chain, as a function of the sample size, and also the convergence rate of the variance of the sample average of a (cost) function of the state at a given step, viewed as an estimator of the expected cost. We summarize known convergence rate results and show empirical results that suggest much better convergence rates than those that are proved. We also compare different types of multivariate sorts to match the chains with the RQMC points, including a sorting procedure based on a Hilbert curve.

The description of a new software tool and library named Lattice Builder, written in C++, that implements a variety of construction algorithms for good rank-1 lattice rules (a familiy of sequences used in QMC methods) is provided in [17]. The library is extensible, thanks to the decomposition of the algorithms into decoupled components, which makes it easy to implement new types of weights, new search domains, new figures of merit, etc.

7.7. Wireless Networks

Participants: Osama Arouk, Btissam Er-Rahmadi, Adlen Ksentini, Meriem Bouzouita, Pantelis Frangoudis, Yassine Hadjadj-Aoul, César Viho, Quang Pham, Gerardo Rubino.

We are continuing our activities around wireless and mobile networks, by focusing more on leveraging the current mobile and wireless architecture toward building the 5G systems.

Congestion control for M2M applications. Machine-to-Machine (M2M) communications are expected to be one of the major drivers for the future 5G network. It is expected that M2M will come up with substantial revenue growth for Mobile Network Operators (MNO), but they represent at the same time the most important challenge they are facing. For instance, a massive number of Machine-to-Machine (M2M) devices performs simultaneously Random Accesses (RA), which causes severe congestions and reduces the RA success probability. To control the Radio Access Network (RAN) overload and alleviate the congestion between M2M devices, 3GPP developed the Access Class Barring (ACB) procedure that depends on an access probability called the ACB factor. In [48][24], we first presented a simple fluid model of M2M devices' random access. This model is then used to derive an optimal regulator of the ACB factor based on nonlinear non affine control theory. The main advantages of the proposed approach are twofold. First, the proposal is fully compliant with the standard while it reduces significantly the computation and the signaling overheads. Second, it provides an efficient mean to regulate adaptively the ACB factor as it guarantees having an optimal number of M2M devices accessing concurrently to the RAN. The obtained results based on simulations show clearly the robustness of the proposed approach, and its superiority compared to existing proposals. However, such a model assumes a perfect knowledge about the number of M2M attempting the ACB and the RA, which is not possible in realistic use cases. For this reason, we proposed in [50] a system-agnostic controller, which computes the barring factor dynamically based only on the mismatch between the average number of M2M devices succedding in the RA and the optimal number of M2M which should succeed. We base our controlling algorithm in a Propostional Integral Derivative (PID)-based controller. Simulation results show that the algorithm outperforms the existing solutions by improving significantly the access success probability while minimizing radio resources' underutilization.

Different schemes were proposed in the literature to solve the congestion problem by regulating the M2M devices' opportunities of transmission. Nonetheless, as revealed in [51], these schemes turn out to be ineffective in case of heavily congested M2M networks. In fact, in such a condition, the unpredictable and increasingly accumulated number of devices cannot be blocked. This augments the risk of M2M devices' synchronized access, which may result in a congestion collapse. Consequently, we proposed, in [49], a methodology for a better estimation of the number of M2M devices attempting the access. We also proposed a novel implementation of the ACB process, which dynamically computes the ACB factor according to the network's overload conditions and includes a corrective action adapting the controller work, based on the mismatch existing between the computed and the targeted mean load. The simulation results show that the proposed algorithms allow improving considerably the estimation of the number of M2M devices' arrivals, while outperforming existing techniques.

In [32], we proposed a novel approach to deal with massive synchronous access attempts, tailored for both M2M delay-sensitive applications and energy constrained ones. The main idea behind the paper is to leverage crowd sourcing data, transmitted from the devices succeeding in the RACH procedure, to tune the access parameters, without requiring too complex techniques for the estimation of the number of attempts. Simulation results show that the proposed scheme achieves sub-optimal performance in the wireless resources' utilization while reducing significantly both the number of access attempts and the access latency for delay sensitive applications. This allows guaranteeing energy conservation.

In [44], we proposed two optimal solutions that use Device-to-Device (D2D) communications to lightweight the overhead of M2M devices on 5G networks. Each scheme has a specific objective, and aims to manage the communications between devices and eNodeBs to achieve its objective. The proposed solutions nominate the devices that should communicate using D2D communications and those that should directly communicate with eNodeBs. The first solution aims to reduce the energy consumption, whereas the second one aims to reduce the data transfer delay at the eNodeBs. The performance of the proposed schemes is evaluated via simulations and the obtained results demonstrate their feasibility and ability in achieving their design goals.

Network selection and optimization. With the explosion of mobile data traffic, the Fixed and Mobile Converged (FMC) network are being heavily required. Mobile devices have the capability of connecting simultaneously to different access networks in the FMC architecture. Access network selection becomes an issue when mobile devices are under coverage of different access networks, since a bad selection may lead to network congestion and degrade the QoE of users. In order to address this problem, in [91] we modeled and analyzed the interface selection procedure using control theory. Based on our model, we designed a controller which can send to mobile devices a network selection command calculated instantaneously for the access network selection.

Dynamic Adaptive Streaming over HTTP (DASH), with its different proprietary versions, is presently the most widely adopted technology for video delivery over the Internet. DASH offers significant advantages, enabling users to switch dynamically between different available video qualities responding to variations in the current network conditions during video playback. This is particularly interesting in wireless and mobile access networks, which present such variations in a hard to predict manner, but sometimes quite frequently. Moreover, mobile users of these networks share a common radio access link and, thus, a common bottleneck in case of congestion, which may cause user experience to degrade. In this context, the Mobile Edge Computing (MEC) emerging standard gives new opportunities to improve DASH performance, by moving IT and cloud computing capabilities down to the edge of the mobile network. In [69] and [103] we proposed a novel architecture for adaptive HTTP video streaming tailored to a MEC environment. The proposed architecture includes an adaptation algorithm running as a MEC service, aiming to relax network congestion while improving the Quality of Experience (QoE) for mobile users. Our mechanism is standards-compliant and compatible with receiver-driven adaptive video delivery algorithms, with which it cooperates in a transparent manner.

Low-rate wireless personal area networks (WPANs) (and also wireless sensor networks) suffer from many constraints. The IEEE 802.15.4 standard proposes the slotted CSMA/CA as a communication channel access mechanism with collision avoidance that takes into account the constraints of WPANs. In [22], we proposed

to introduce a data fragmentation mechanism into slotted CSMA/CA to improve a bandwidth utilisation. The novelty here is the use of the fragmentation mechanism to replace an acknowledgement frame after the transmission of the fragment and the remaining frame. The beacon frame is used to confirm the success transmission of a data fragment. To evaluate the performance of our proposition, we have developped a three dimension Markov chain which modelises the behaviour of the node using IEEE 802.15.4 with data fragmentation mechanism without using an ACK frame. The analytical results concerning the network throughput and the transmission success delay demonstrate the improvement of the bandwidth occupation.

Mobile networks' improvements. In [85], we introduced the concept of elastic bearer in Evolved Packet System (EPS), which allows the users to enhance on-demand the performance of certain applications and permits the network to efficiently manage the resource allocation taking into account the application type. In particular, the paper introduces a set of mechanisms to trigger and support bearer elasticity in EPS based on the Quality of Experience (QoE) perceived by users or based on feedback from Radio Access Network (RAN). Bearer elasticity can be attained through potential Packet Data Network/Serving Gateway (PDN/S-GW) relocation to eventually improve QoE within and beyond the mobile network operator premises. The paper also introduces a set of methods to identify and cope with a storm of requests for particular applications at densely populated areas.

One important objective of 5G mobile networks is to accommodate a diverse and ever-increasing number of user equipment (UEs). Coping with the massive signaling overhead expected from UEs is an important hurdle to tackle so as to achieve this objective. In [11], we devised an efficient tracking area list management (ETAM) framework that aims for finding optimal distributions of tracking areas (TAs) in the form of TA lists (TALs) and assigning them to UEs, with the objective of minimizing two conflicting metrics, namely paging overhead and tracking area update (TAU) overhead. ETAM incorporates an online part and an and offline one, in order to achieve its design goal. In the online part, two strategies were proposed to assign in real time, TALs to different UEs, while in the offline part, three solutions were proposed to optimally organize TAs into TALs. The performance of ETAM is evaluated via analysis and simulations, and the obtained results demonstrate its feasibility and ability in achieving its design goals, improving the network performance by minimizing the cost associated with paging and TAU.

QoE aware routing in wireless networks. This year we continued our research on QoE-based optimization routing for wireless mesh networks. First, we approximate PSQA models by explicit mathematical forms, which can be used to find the optimal or near to optimal routes. Next, the hardness of the problem is studied and decentralized algorithms are proposed. The quality of the solution, computational complexity of the proposed algorithm, and the fairness are the main concerns of this work. Several centralized approximation algorithms have been proposed in order to address the complexity and the quality of the published solutions. The results can be found in the following papers: [25],[94], [95] and [26]. However, these centralized algorithms are not appropriate in large-scale networks. Thus, a distributed algorithm is necessary as a complement of the existing centralized methods. This is currently studied at the team.

DIVERSE Project-Team

7. New Results

7.1. Results on Variability modeling and management

7.1.1. Feature Model Synthesis: Algorithms and Empirical Studies

We attack the problem of synthesising feature models by considering both configuration semantics and ontological semantics of a feature model. We define a generic synthesis procedure that computes the likely siblings or parent candidates for a given feature. We develop six heuristics for clustering and weighting the logical, syntactical and semantical relationships between feature names. We then perform an empirical evaluation on hundreds of feature models, coming from the SPLOT repository and Wikipedia. We provide evidence that a fully automated synthesis (i.e., without any user intervention) is likely to produce models far from the ground truths. As the role of the user is crucial, we empirically analyze the strengths and weaknesses of heuristics for computing ranking lists and different kinds of clusters. We show that a hybrid approach mixing logical and ontological techniques outperforms state-of-the-art solutions.

Numerous synthesis techniques and tools have been proposed, but only a few consider both configuration and ontological semantics of a feature model. We also boil down several feature model management operations to a synthesis problem. Our approach, the FAMILIAR environment, and empirical results support researchers and practitioners working on feature models. The synthesis problem is a core issue when reverse engineering, merging, slicing, or refactoring feature models. An article has been published in 2016 at Empirical Software Engineering journal, a major avenue for software engineering research [19].

7.1.2. Product Comparison Matrix

Product Comparison Matrices (PCMs) are widely used for documenting or comparing a set of products. PCMs are simple tabular data in which products are usually organized as rows, features as columns, while each cell define how a product implements the corresponding feature. We develop metamodeling and feature modeling techniques for formalizing PCMs. We perform numerous empirical experiments with users, tools, and data for validating our proposal. We also develop automated techniques to extract PCMs out of informal product descriptions, written in natural language. We establish a connection between PCMs and variability modeling formalism, which is of interest for the product line community. OpenCompare is a direct output of this research and is an important step towards the creation of a community around PCMs. We mined millions of Wikipedia tabular data together with end-users and developers to cross-validate our model-based approach [19]. We also mined data from BestBuy [17].

7.1.3. Machine Learning and Variability Testing

We propose the use of a machine learning approach to infer variability constraints from an oracle that is able to assess whether a given configuration is correct. We propose an automated procedure to randomly generate configurations, classify them according to the oracle, and synthesize cross-tree constraints. We validate our approach on a product-line video generator, using a simple computer vision algorithm as an oracle. We show that an interesting set of cross-tree constraint can be generated, with reasonable precision and recall. Our learning-based testing technique complements our initial effort in engineering an industrial video generator. The use of learning allows to significantly narrow the configuration space and discover complex constraints, hard to discover even for experts. We conduct a series of work in the computer vision domain to generate variants of videos, investigating the usefulness and effectiveness of variability techniques in novel areas. Our approach is novel and general: the same principles can be applied to other configurable systems [55].

7.1.4. Enumeration of All Feature Model Configurations

Feature models are widely used to encode the configurations of a software product line in terms of mandatory, optional and exclusive features as well as propositional constraints over the features. Numerous computationally expensive procedures have been developed to model check, test, configure, debug, or compute relevant information of feature models. We explore the possible improvement of relying on the enumeration of all configurations when performing automated analysis operations. We tackle the challenge of how to scale the existing enumeration techniques by relying on distributed computing. We show that the use of distributed computing techniques might offer practical solutions to previously unsolvable problems and opens new perspectives for the automated analysis of software product lines [40].

7.1.5. Software Unbundling

Unbundling is a phenomenon that consists of dividing an existing software artifact into smaller ones. It can happen for different reasons, one of them is the fact that applications tend to grow in functionalities and sometimes this can negatively influence the user experience. It can be seen as a way to produce different variants of an application. For example, mobile applications from well-known companies are being divided into simpler and more focused new ones. Despite its current importance, little is known or studied about unbundling or about how it relates to existing software engineering approaches, such as modularization. Consequently, recent cases point out that it has been performed unsystematically and arbitrarily. Our main goal is to present this novel and relevant concept and its underlying challenges in the light of software engineering, also exemplifying it with recent cases. We relate unbundling to standard software modularization, presenting the new motivations behind it, the resulting problems, and drawing perspectives for future support in the area [23].

7.1.6. Featured Model Types

By analogy with software product reuse, the ability to reuse (meta)models and model transformations is key to achieve better quality and productivity. To this end, various opportunistic reuse techniques have been developed, such as higher-order transformations, metamodel adaptation, and model types. However, in contrast to software product development that has moved to systematic reuse by adopting (model-driven) software product lines, we are not quite there yet for modelling languages, missing economies of scope and automation opportunities. Our vision is to transpose the product line paradigm at the metamodel level, where reusable assets are formed by metamodel and transformation fragments and "products" are reusable language building blocks (model types). We introduce featured model types to concisely model variability amongst metamodelling elements, enabling configuration, automated analysis, and derivation of tailored model types [53].

7.1.7. A Formal Modeling and Analysis Framework for SPL of Pre-emptive Real-time Systems

We present a formal analysis framework to analyze a family of platform products w.r.t. real-time properties. First, we propose an extension of the widely-used feature model, called Property Feature Model (PFM), that distinguishes features and properties explicitly Second, we present formal behavioral models of components of a real-time scheduling unit such that all real-time scheduling units implied by a PFM are automatically composed to be analyzed against the properties given by the PFM. We apply our approach to the verification of the schedulability of a family of scheduling units using the symbolic and statistical model checkers of Uppaal [44].

7.1.8. Exploration of Architectural Variants

In systems engineering, practitioners shall explore numerous architectural alternatives until choosing the most adequate variant. The decision-making process is most of the time a manual, time-consuming, and error-prone activity. The exploration and justification of architectural solutions is ad-hoc and mainly consists in a series of tries and errors on the modeling assets.

We report on an industrial case study in which we apply variability modeling techniques to automate the assessment and comparison of several candidate architectures (variants). We first describe how we can use a model-based approach such as the Common Variability Language (CVL) to specify the architectural variability. We show that the selection of an architectural variant is a multi-criteria decision problem in which there are numerous interactions (veto, favor, complementary) between criteria. We present a tooled process for exploring architectural variants integrating both CVL and the MYRIAD method for assessing and comparing variants based on an explicit preference model coming from the elicitation of stakeholders' concerns. This solution allows understanding differences among variants and their satisfactions with respect to criteria. Beyond variant selection automation improvement, this experiment results highlight that the approach improves rationality in the assessment and provides decision arguments when selecting the preferred variants. It is a joint work and collaboration with Thales [47].

7.1.9. A Complexity Tale: Web Configurators

Online configurators are basically everywhere. From physical goods (cars, clothes) to services (cloud solutions, insurances, etc.) such configurators have pervaded many areas of everyday life, in order to provide the customers products tailored to their needs. Being sometimes the only interfaces between product suppliers and consumers, much care has been devoted to the HCI aspects of configurators, aiming at offering an enjoyable buying experience. However, at the backend, the management of numerous and complex configuration options results from ad-hoc process rather than a systematic variability-aware engineering approach. We present our experience in analysing web configurators and formalising configuration options in terms of feature models or product configuration matrices. We also consider behavioural issues and perspectives on their architectural design [32].

7.2. Results on Software Language Engineering

7.2.1. Safe Model Polymorphism for Flexible Modeling

Domain-Specific Languages (DSLs) are increasingly used by domain experts to handle various concerns in systems and software development. To support this trend, the Model-Driven Engineering (MDE) community has developed advanced techniques for designing new DSLs. However, the widespread use of independently developed, and constantly evolving DSLs is hampered by the rigidity imposed to the language users by the DSLs and their tooling, e.g., for manipulating a model through various similar DSLs or successive versions of a given DSL. In [24] we propose a disciplined approach that leverages type groups' polymorphism to provide an advanced type system for manipulating models, in a polymorphic way, through different DSL interfaces. A DSL interface, aka. model type, specifies a set of features, or services, available on the model it types, and subtyping relations among these model types define the safe substitutions. This type system complements the Melange language workbench and is seamlessly integrated into the Eclipse Modeling Framework (EMF), hence providing structural interoperability and compatibility of models between EMF-based tools. We illustrate the validity and practicability of our approach by bridging safe interoperability between different semantic and syntactic variation points of a finite-state machine (FSM) language, as well as between successive versions of the Unified Modeling Language (UML).

7.2.2. Execution Framework for Model Debugging

The development and evolution of an advanced modeling environment for a Domain-Specific Modeling Language (DSML) is a tedious task, which becomes recurrent with the increasing number of DSMLs involved in the development and management of complex software-intensive systems. Recent efforts in language workbenches result in advanced frameworks that automatically provide syntactic tooling such as advanced editors. However, defining the execution semantics of languages and their tooling remains mostly hand crafted. Similarly to editors that share code completion or syntax highlighting, the development of advanced debuggers, animators, and others execution analysis tools shares common facilities, which should be reused among various DSMLs. In [37] we present the execution framework offered by the GEMOC studio, an Eclipse-based language and modeling workbench. The framework provides a generic interface to plug in different execution

engines associated to their specific metalanguages used to define the discrete-event operational semantics of DSMLs. It also integrates generic runtime services that are shared among the approaches used to implement the execution semantics, such as graphical animation or omniscient debugging.

7.2.3. Variability Management in Language Families

The use of domain-specific languages (DSLs) has become a successful technique in the development of complex systems. Nevertheless, the construction of this type of languages is time-consuming and requires highly-specialized knowledge and skills. An emerging practice to facilitate this task is to enable reuse through the definition of language modules which can be later put together to build up new DSLs. In [29], we report on an effort for organizing the literature on language product line engineering. More precisely, we propose a definition for the life-cycle of language product lines, and we use it to analyze the capabilities of current approaches. In addition, we provide a mapping between each approach and the technological space it supports.

Still, the identification and definition of language modules are complex and error-prone activities, thus hindering the reuse exploitation when developing DSLs. In [50], [51], we propose a computer-aided approach to i) identify potential reuse in a set of legacy DSLs; and ii) capitalize such potential reuse by extracting a set of reusable language modules with well defined interfaces that facilitate their assembly. We validate our approach by using realistic DSLs coming out from industrial case studies and obtained from public GitHub repositories. We also developed a publicly available tool, namely Puzzle, that uses static analysis to facilitate the detection of specification clones in DSLs implemented under the executable metamodeling paradigm. Puzzle also enables the extraction specification clones as reusable language modules that can be later used to build up new DSLs.

7.2.4. A Tool-Supported Approach for Concurrent Execution of Heterogeneous Models

In the software and systems modeling community, research on domain-specific modeling languages (DSMLs) is focused on providing technologies for developing languages and tools that allow domain experts to develop system solutions efficiently. Unfortunately, the current lack of support for explicitly relating concepts expressed in different DSMLs makes it very difficult for software and system engineers to reason about information spread across models describing different system aspects. As a particular challenge, we investigate in [38] relationships between, possibly heterogeneous, behavioral models to support their concurrent execution. This is achieved by following a modular executable metamodeling approach for behavioral semantics understanding, reuse, variability and composability. This approach supports an explicit model of concurrency (MoCC) and domain-specific actions (DSA) with a well-defined protocol between them (incl., mapping, feedback and callback) reified through explicit domain-specific events (DSE). The protocol is then used to infer a relevant behavioral language interface for specifying coordination patterns to be applied on conforming executable models. All the tooling of the approach is gathered in the GEMOC studio, and outlined in the next section. Currently, the approach is experienced on a systems engineering language provided by Thales, named Capella.

7.2.5. Various Dimensions of Reuse

Reuse, enabled by modularity and interfaces, is one of the most important concepts in software engineering. This is evidenced by an increasingly large number of reusable artifacts, ranging from small units such as classes to larger, more sophisticated units such as components, services, frameworks, software product lines, and concerns. We give evidence in [43] that a canonical set of reuse interfaces has emerged over time: the variation, customization, and usage interfaces (VCU). A reusable artifact that provides all three interfaces reaches the highest potential of reuse, as it explicitly exposes how the artifact can be manipulated during the reuse process along these three dimensions. We demonstrate the wide applicability of the VCU interfaces along two axes: across abstraction layers of a system specification and across existing reuse techniques. The former is shown with the help of a comprehensive case study including reusable requirements, software, and hardware models for the authorization domain. The latter is shown with a discussion on how the VCU interfaces relate to existing reuse techniques.

7.2.6. Modeling for Sustainability

The complex problems that computational science addresses are more and more benefiting from the progress of computing facilities (e.g., simulators, librairies, accessible languages). Nevertheless, the actual solutions call for several improvements. Among those, we address the needs for leveraging on knowledge and expertise by focusing on Domain-Specific Modeling Languages application. In this work we explored, through concrete experiments, how the last DSML research help getting closer the problem and implementation spaces.

Various disciplines use models for different purposes. While engineering models, including software engineering models, are often developed to guide the construction of a non- existent system, scientific models, in contrast, are created to better understand a natural phenomenon (i.e., an already existing system). An engineering model may incorporate scientific models to build a system. Both engineering and scientific models have been used to support sustainability, but largely in a loosely-coupled fashion, independently developed and maintained from each other. Due to the inherent complex nature of sustainability that must balance trade-offs between social, environmental, and economic concerns, modeling challenges abound for both the scientific and engineering disciplines. In [39] we propose a vision that synergistically combines engineering and scientific models to enable broader engagement of society for addressing sustainability concerns, informed decision-making based on more accessible scientific models and data, and automated feed-back to the engineering models to support dynamic adaptation of sustainability systems. To support this vision, we identify a number of challenges to be addressed with particular emphasis on the socio-technical benefits of modeling.

As first experiments, we presented at the EclipseCon France, Europe and North America 2016, an approach to develop smart cyber physical systems in charge of managing the production, distribution and consumption of energies (e.g., water, electricity). The main objective is to enable a broader engagement of society, while supporting a more informed decision-making, possibly automatically, on the development and run-time adaptation of sustainability systems (e.g., smart grid, home automation, smart cities). We illustrate this approach through a system that allows farmers to simulate and optimize their water consumption by combining the model of a farming system together with agronomical models (e.g., vegetable and animal lifecycle) and open data (e.g., climate series). To do so, we use Model Driven Engineering (MDE) and Domain Specific Languages (DSL) to develop such systems driven by scientific models that define the context (e.g., environment, social and economy), and model experiencing environments to engage general public and policy makers.

7.2.7. Formal Specification of a Packet Filtering Language Using the K Framework

Many project-specific languages, including in particular filtering languages, are defined using non-formal specifications written in natural languages. This leads to ambiguities and errors in the specification of those languages. In [46] we report on an industrial experiment on using a tool-supported language specification framework (K) for the formal specification of the syntax and semantics of a filtering language having a complexity similar to those of real-life projects. This experimentation aims at estimating, in a specific industrial setting, the difficulty and benefits of formally specifying a packet filtering language using a tool-supported formal approach.

7.2.8. Correct-by-construction model driven engineering composition operators

Model composition is a crucial activity in Model Driven Engineering both to reuse validated and verified model elements and to handle separately the various aspects in a complex system and then weave them while preserving their properties. Many research activities target this compositional validation and verification (V & V) strategy: allow the independent assessment of components and minimize the residual V & V activities at assembly time. However, there is a continuous and increasing need for the definition of new composition operators that allow the reconciliation of existing models to build new systems according to various requirements. These ones are usually built from scratch and must be systematically verified to assess that they preserve the properties of the assembled elements. This verification is usually tedious but is mandatory to avoid verifying the composite system for each use of the operators. Our work addresses these issues, we first target the use of proof assistants for specifying and verifying compositional verification frameworks relying on formal verification techniques instead of testing and proofreading. Then, using a divide

and conquer approach, we focus on the development of elementary composition operators that are easy to verify and can be used to further define complex composition operators. In our approach [27], proofs for the complex operators are then obtained by assembling the proofs of the basic operators. To illustrate our proposal, we use the Coq proof assistant to formalize the language-independent elementary composition operators Union and Substitution and the proof that the conformance of models with respect to metamodels is preserved during composition. We show that more sophisticated composition operators that share parts of the implementation and have several properties in common (especially: aspect oriented modeling composition approach, invasive software composition, and package merge) can then be built from the basic ones, and that the proof of conformance preservation can also be built from the proofs of basic operators.

7.2.9. Engineering Modeling Languages

The DiverSE project-team is deeply involved in transferring research knowledge into education. In particular, one book in English have been published in 2016 as a textbook [59]. The book cover the broad scope of MDE, and are based on the experience of the project-team members.

7.3. Results on Heterogeneous and dynamic software architectures

We have selected three main contributions : two are in the field of runtime management, while the third one is in the field of non-functionnal software testing.

7.3.1. Precise and efficient resource management using models@runtime

Contribution. We have developed an efficient monitoring framework to quickly spot an abnormal resource consumption within a complex application. In these papers [25], we have proposed an optimistic adaptive monitoring system to determine the faulty components of an application. Suspected components are finely analyzed by the monitoring system, but only when required. Unsuspected components are left untouched and execute normally.

Originality. Current solutions that perform permanent and extensive monitoring to detect anomalies induce high overhead on the system, and can, by themselves, make the system unstable. Our system performs localized just-in-time monitoring that decreases the accumulated overhead of the monitoring system. Through our evaluation, we show that our technique correctly detects faulty components, while reducing overhead by 92.98 on average%.

Impact. Beyond the scientific originality of this work, the main impacts of this novel approach approach to monitor software component performance has been to (i) reinforce DIVERSE's visibility in the academic and industrial communities on software components and (ii) to create several research tracks that are currently explored in different projects of the team (HEADS and B-com PhD thesis). This work has been integrated within the Kevoree platform.

7.3.2. Dynamic web application using models@runtime

Contribution. We have developed a component-based platform supporting the development of dynamically adaptable single Web page applications. An important part of this contribution lies in the possibility to dynamically move code from the server to the client side allowing a great flexibility in the performance management. This contribution [56] is based on a models@runtime approach and has been implemented in our open source KevoreeJS platform.

Originality. Current solutions to create single Web page application are limited to a static code repartition between clients and server, thus limiting the flexibility at runtime.

Impact. Beyond the scientific originality of this work, the main impacts of this novel approach to monitor software component performance has been to (i) reinforce DIVERSE's visibility in the open-source community, (ii) to start several research tracks that are currently explored in different projects of the team (HEADS, STAMP, GRevis). This platforms is modular, one of the component has a monthly download count greater than 100k [0]).

[0] https://www.npmjs.com/package/npmi

7.3.3. Testing non-functional behavior of compiler and code generator

Contribution. We have developed NOTICE [36], [35], a component-based framework for non-functional testing of compilers through the monitoring of generated code in a controlled sand-boxing environment. In this work, we have proposed an automatic way of testing non-functional properties of compilers, while optimizing the generated application with respect to a set of specific non-functional properties (CPU, memory usage, energy consumption, *etc.*).

Originality. Compiler users generally apply different optimizations to generate efficient code with respect to specific non-functional properties such as energy consumption, execution time, etc. However, due to the huge number of optimizations provided by modern compilers, finding the best optimization sequence for a specific objective and a given program is more and more challenging.

Impact. Beyond the scientific originality of this work, the main impact of this novel approach is to enable the auto-tuning of compilers according to user requirements and to construct optimizations that yield to performance results that are better than standard optimization levels.

7.3.4. Automatic Microbenchmark Generation to Prevent Dead Code Elimination and Constant Folding

Contribution. Microbenchmarking consists of evaluating, in isolation, the performance of small code segments that play a critical role in large applications. The accuracy of a microbenchmark depends on two critical tasks: wrap the code segment into a payload that faithfully recreates the execution conditions that occur in the large application; build a scaffold that runs the payload a large number of times to get a statistical estimate of the execution time. While recent frameworks such as the Java Microbenchmark Harness (JMH) take care of the scaffold challenge, developers have very limited support to build a correct payload. This year, we focus on the automatic generation of pay-loads, starting from a code segment selected in a large application [54]. In particular, we aim at preventing two of the most common mistakes made in microbenchmarks: dead code elimination and constant folding. Since a microbench-mark is such a small program, if not designed carefully, it will be *over-optimized* by the JIT and result in distorted time measures. Our technique hence automatically extracts the segment into a compilable payload and generates additional code to prevent the risks of *over-optimization*. The whole approach is embedded in a tool called AutoJMH, which generates payloads for JMH scaffolds. We validate the capabilities AutoJMH, showing that the tool is able to process a large percentage of segments in real programs. We also show that AutoJMH can match the quality of payloads handwritten by performance experts and outperform those written by professional Java developers without experience un microbenchmarking.

7.3.5. Collaborations

This year, we had a close and fruitful collaboration with the industrial partners that are involved in the HEADS and Occiware projects, in particular an active interaction with the Tellu company in Norway in the Heads context [49]. Tellu relies on Kevoree and KevoreeJS to build their health management systems. They will be also a active member the new Stamp project led by DIVERSE. We can cite also an active collaboration with Orange Labs through Kevin Corre's joint PhD thesis. Another joint industrial (CIFRE) PhD started in September 2016, and we are also partner in a new starting FUI project. Finally, DIVERSE collaborates with the B-COM IRT (https://b-com.com/en), as one permanent member has a researcher position of one day per week at B-COM and a new joint PhD started in September [52].

At the academic level we collaborate actively with the Spiral team at Inria Lille (several joint projects), the Tacoma team (with two co-advised PhD students), the Myriad team (1 co-advised PhD student) and we have started two collaborations with the ASAP team.

7.4. Results on Diverse Implementations for Resilience

Diversity is acknowledged as a crucial element for resilience, sustainability and increased wealth in many domains such as sociology, economy and ecology. Yet, despite the large body of theoretical and experimental

science that emphasizes the need to conserve high levels of diversity in complex systems, the limited amount of diversity in software-intensive systems is a major issue. This is particularly critical as these systems integrate multiple concerns, are connected to the physical world through multiple sensors, run eternally and are open to other services and to users. Here we present our latest observational and technical results about (i) new approaches to increase diversity in software systems, and (ii) software testing to assess the validity of software.

7.4.1. Software diversification

A main achievement in our investigations of software diversity, is a large scale analysis of browser fingerprints [45]. Browser fingerprinting consists in collecting information about a user's browser and its execution environment. A distinctive feature of these fingerprints is that they are unique and can be used to track users. We show that innovations in HTML5 provide access to highly discriminating attributes, notably with the use of the Canvas API which relies on multiple layers of the user's system. In addition, we show that browser fingerprinting is as effective on mobile devices as it is on desktops and laptops, albeit for radically different reasons due to their more constrained hardware and software environments. We also evaluate how browser fingerprinting could stop being a threat to user privacy if some technological evolutions continue (e.g., disappearance of plugins) or are embraced by browser vendors (e.g., standard HTTP headers).

As for automatic diversification of programs, we have had a strong focus on runtime transformations. Online Genetic Improvement embeds the ability to evolve and adapt inside a target software system enabling it to improve at runtime without any external dependencies or human intervention. We recently developed a general purpose tool enabling Online Genetic Improvement in software systems running on the java virtual machine. This tool, dubbed ECSELR, is embedded inside extant software systems at runtime, enabling such systems to autonomously generate diverse variants [31]. We have also worked on diversification against just-in-Time (JIT) Spraying: a technique that embeds return-oriented programming (ROP) gadgets in arithmetic or logical instructions as immediate offsets. We introduce libmask, a JIT compiler extension that transforms constants into global variables and marks the memory area for these global variables as read only. Hence, any constant is referred to by a memory address making exploitation of arithmetic and logical instructions more difficult. Then, these memory addresses are randomized to further harden the security [42].

7.4.2. Software testing

Our work in the area of software testing focuses on tailoring the testing tools (analysis, generation, oracle, etc.) to specific domains and purposes. This allows us to consider domain specific knowledge (e.g., architectural patterns for GUI implementation) in order to increase the relevance and the efficiency of testing. The main results of this year are about test case refactoring and testing code generators.

Software developers design test suites to verify that software meets its expected behaviors. Yet, many dynamic analysis techniques are performed on the exploitation of execution traces from test cases. In practice, one test case may imply various behaviors. However, the execution of a test case only yields one trace, which can hide the others. We have developed a new technique of test code refactoring, which splits a test case into small test fragments that cover a simpler part of the control flow to provide better support for dynamic analysis. This technique can effectively improve the execution traces of the test suite: exception contracts are better verified via applying this refactoring to original test suites [30].

Finding the smallest set of valid test configurations that ensure sufficient coverage of the system's feature interactions is essential, especially when the execution of test configurations is costly or time-consuming. However, this problem is NP-hard in general and approximation algorithms have often been used to address it in practice. We explore an approach based on constraint programming to increase the effectiveness of configuration testing while keeping the number of configurations as low as possible. For 79% of 224 feature models, our technique generated up to 60% fewer test configurations than the competitor tools [26].

The intensive use of generative programming techniques provides an elegant engineering solution to deal with the heterogeneity of platforms and technological stacks. Yet, producing correct and efficient code generator is complex and error-prone. We describe a practical approach based on a runtime monitoring infrastructure to automatically check the potential inefficient code generators. We evaluate our approach by analyzing the

performance of Haxe, a popular high-level programming language that involves a set of cross-platform code generators. The results show that our approach is able to detect some performance inconsistencies that reveal real issues in Haxe code generators [36], [35]

Graphical User Interfaces (GUIs) intensively rely on event-driven programming: widgets send GUI events, which capture users' interactions, to dedicated objects called *controllers*. Controllers implement several *GUI listeners* that handle these events to produce GUI commands. We study to what extent the number of GUI commands that a GUI listener can produce has an impact on the code quality. We then identify a new type of design smell, called *Blob listener* that characterizes GUI listeners that can produce more than two GUI commands. We propose a systematic static code analysis procedure that searches for *Blob Listener* instances that we implement in *InspectorGuidget* [48].

DYLISS Project-Team

7. New Results

7.1. Data integration and pre-processing with semantic-based technologies

Participants: Meziane Aite, Marie Chevallier, Olivier Dameron, Aurélie Evrard, Clémence Frioux, Xavier Garnier, Jeanne Got, François Moreews, Yann Rivault, Anne Siegel, Pierre Vignet, Denis Tagu, Camille Trottier.

Integration and query of biological datasets with Semantic Web technologies. The purpose of this work is to obtain quick answers to biological questions demanding currently hours of manual search in several spreadsheet results files. We introduce an integration and interogation framework using an RDF model and the SPARQL query language. It allows biologists to transparently integrate and query their data without any a priori proficiency about RDF and SPARQL. [*O. Dameron, A. Evrard, X. Garnier*] [37], [45]

Handling the heterogeneity of genomic and metabolic networks data within flexible workflows with the PADMet toolbox A main challenge of the era of fast and massive genome sequencing is to transform sequences into biological knowledge. The high diversity of input files and tools required to run any metabolic networks reconstruction protocol represents an important drawback: it appears very difficult to ensure that input files agree among them. Such a heterogeneity produces loss of information during the use of the protocols and generates uncertainty in the final metabolic model. Here we introduce the PADMet-toolbox which allows conciliating genomic and metabolic network information. The toolbox centralizes all this information in a new graph-based format: PADMet (PortAble Database for Metabolism) and provides methods to import, update and export information. For the sake of illustration, the toolbox was used to create a workflow, named AuReMe, aiming to produce high-quality genome-scale metabolic networks and eventually input files to feed most platforms involved in metabolic network analyses. We applied this approach to two exotic organisms and our results evidenced the need of combining approaches and reconciling information to obtain a functional metabolic network to produce biomass. [*M. Chevallier, M. Aite, C. Frioux, J. Got, A. Siegel, C. Trottier, P. Vignet*] [34]

PEPS: a platform for supporting studies in pharmaco-epidemiology using medico-administrative databases We showed that Semantic Web technologies are technically adapted for representing patients' data from medico-administrative databases as RDF and querying them using SPARQL. We also demonstrated that this approach is relevant as it supports the combination of patients' data with hierarchical knowledge in order to address the problem of reconciling precise patients data with more general query criteria. [*O. Dameron, Y. Rivault*] [33], [31], [30]

Telemedicine : ontology-based reasoning and data integration We have developed a system based on a formal ontology that integrates the alert information and the patient data extracted from the electronic health record in order to better classify the importance of alerts. A pilot study was conducted on atrial fibrillation alerts. The results suggest that this approach has the potential to significantly reduce the alert burden in telecardiology. The methods may be extended to other types of connected devices. We also worked on a telemedicine application for monitoring patients with chronic diseases. We proposed an architecture supporting data exchange in the context of multiple chronic diseases [*O. Dameron*] [26], [25], [18]

7.2. Data and knowledge integration based on combinatorial optimization

Participants: Marie Chevallier, Damien Eveillard, Jeanne Got, Julie Laniau, François Moreews, Jacques Nicolas, Anne Siegel.

Packing graphs with ASP for landscape simulation This study is part of a more general research track on graph compression, a fundamental issue for the analysis of biological networks that we address with Answer Set Programming (ASP)modelling. The general issue is to cover a given graph by a set of subgraphs. The IJCAI paper describes an application to crop allocation for generating realistic landscapes. The aim is to cover optimally a bare landscape, represented by its plot graph, with spatial patterns describing local arrangements of crops. This problem belongs to the hard class of graph packing problems. The approach provides a compact solution to the basic problem and at the same time allows extensions such as a flexible integration of expert knowledge. Particular attention is paid to the treatment of symmetries, especially due to sub-graph isomorphism issues. Experiments were conducted on a database of simulated and real landscapes. Currently, our program can process graphs of medium size, a size that enables studies on real agricultural practices. [*J. Nicolas*] [29]

Deciphering transcriptional regulations coordinating the response to environmental changes We introduce a method that extracts from a transcriptional regulatory network determined from a set of predicted transcription factors (TF) and binding site (BS) a subnetwork explaining a given set of observed co-expressions, highlighting those TFs and BSs most likely involved in the co-regulation. The method solves an optimization problem on a graph to select confident paths within the given transcriptional regulatory network joining a putative common regulator with two co-expressed genes via regulatory cascades. It provides a useful modeling scheme for deciphering the regulatory mechanisms that underly the phenotypical response of an organism to environmental challenges and can be used as a reliable tool for further research on genome scale transcriptional regulation studies. [*M. Chevallier, D. Eveillard, A. Siegel*] [13]

Putative bacterial interactions from metagenomic knowledge with an integrative systems ecology approach. Our software tool *shogen* [62] was used to decipher functional roles within a consortium of five mining bacteria through the integration of genomic and metabolic knowledge at genome scale. We first reconstructed a global metabolic network. Next, using a parsimony assumption, we deciphered sets of genes, called Sets from Genome Segments (SGS), that (i) are close on their respective genomes, (ii) take an active part in metabolic pathways and (iii) whose associated metabolic reactions are also closely connected within metabolic networks. The use of SGS (shogen) pinpoints a functional compartmentalization among the investigated species and exhibits putative bacterial interactions necessary for promoting these pathways. [*M. Chevallier, D. Eveillard, A. Siegel*] [15]

Molecular alterations induced by a high-fat high-fiber diet in porcine adipose tissues: variations according to the anatomical fat location Our methods based on the integration of metabolic and regulatory regulations [61] were combined to statistical approaches and applied to the understanding of fatty acid metabolism in porcs and chicken. The analyses evidenced that a high-fat high-fiber diet depressed glucose and lipid anabolic molecular pathways, thus counteracting adipose tissue expansion. Interaction effects between dietary intake of fiber and lipids on gene expression may modulate innate immunity and inflammation, a response which is of interest with regard to chronic inflammation and its adverse effects on health and performance. [*F. Moreews, A. Siegel*] [20]

7.3. Systems biology

Participants: Jérémie Bourdon, Jean Coquet, Victorien Delannée, Jacques Nicolas, Anne Siegel, Nathalie Théret, Pierre Vignet.

Representation of symbolic dynamical systems generated by a substitution. Iterated morphisms are combinatorial processes which are related to several classes of dynamical systems appearing in several fields of computer sciences and mathematics: numeration, ergodic theory, discrete geometry. They may be associated to fractal sets called "Rauzy fractals" whose topological properties are linked to the properties of the underlying dynamical system. We have introduced a generic algorithm framework to check such topological properties within a complete family of iterated morphism. This makes efficient the verification of conjectures on several families of substitutions related to multi-dimensional continued fraction algorithms. [*A. Siegel*] [14]

Identification of logical models for signaling pathways. Logical models of signaling pathways are a promising way of building effective in silico functional models of a cell. The automated learning of Boolean logic models describing signaling pathways can be achieved by training to phosphoproteomics data. In our work, combinatorial optimization methods based on recent logic programming paradigm allow to enumerate, and discriminate the family of logical models explaining data. Together, these approaches enable a robust understanding of the system response. The results are implemented in the *caspo* software. The main weakness of ASP-based learning algorithm is that they focus on the comparison of two time-points and assumes that the system has reached an early steady state. We have generalized such a learning procedure in order to discriminate Boolean networks according to their transient dynamics. To that goal, we exhibit a necessary condition that must be satisfied by a Boolean network dynamics to be consistent with a discretized time series trace. [*A. Siegel*] [23], [28]

Model of the Delayed Translation of Cyclin B Maternal mRNA After Sea Urchin Fertilization. An extended model of the numerical model introduced in [74] was developed to have a better understanding of the role of cyclin B in protein synthesis within minutes after fertilization of sea urchin eggs. The model confirms that regulation of cyclin B biosynthesis is an example of a select protein whose translation is controlled by pathways that are distinct from housekeeping proteins, even though both involve the same cap-dependent initia- tion pathway. Therefore, this model should help provide insight to the signaling utilized for the biosynthesis of cyclin B and other select proteins. [*J. Bourdon, A. Siegel*] [24]

Deciphering pathways involved in TGF-β signalling network. TGF-β is a multifunctional cytokine that regulates mammalian development, differentiation, and homeostasis. As a growth inhibitor of epithelial, endothelial, and hematopoietic cells, TGF-β is a potent anticancer agent in normal tissue. At the opposite TGF-β acts as a promoter of tumor by inducing the hallmarks of the cancer. Consequently targeting the deleterious effects of TGF-β without affecting its physiological role is the common goal of therapeutic strategies. While several strategies based on blocking TGF-β antibodies or small inhibitors of TGF-β receptors have been investigated, they did not take into account the impact of the (extracellular matrix) ECM remodeling that regulates TGF-β bioavailability and the complexity of TGF-β-dependent signaling pathways which regulate both physiological and pathological processes depending on context. In accordance with this, we recently demonstrated the benefical anti-tumor effect of the interplay between TGF-β signaling and the CD103 integrin pathway. At the opposite we have previously demonstrated that the disintegrin ADAMTS1 promotes TGF-β activation in chronic liver disease and we recently characterized interaction with inhibitor peptide to block such effects, using in silico approach. Importantly, we need to take into account a system-wide view and develop predictive models for therapeutic benefit. In that context we demonstrated that the ratio of TGFBR2 to TGFBR1 receptors concentrations can be used to discriminate between metastable regimes of TGF-β signaling model and predic the tumor cell aggressiveness [*N. Théret*][27], [16], [21].

7.4. Sequence and structure annotation

Participants: Aymeric Antoine-Lorquin, Catherine Belleannée, François Coste, Jacques Nicolas.

Detection of mutated primers on metagenomics sequences to detect more species. In targeted metagenomics, an initial task is the detection in each sequence of the primers used for amplifying the targeted region. The selected sequences are then trimmed and clustered in order to inventory species present in the sample. Common pratices consist in retaining only the sequences with perfect primers (i.e. non-mutated by sequencing error). In the context of a study characterizing the biodiversity of tropical soils in unicellular eukaryotes, we have implemented the search for mutated primers, using the grammatical pattern matching tool Logol, and shown that retrieving sequences with mutated primers has a significant impact on targeted metagenomics results, as it makes possible to detect more species (7% additional OTUs in our study) [*A. Antoine-Lorquin, C. Belleannée*] [32], [11].

VIRALpro: a tool to identify viral capsid and tail sequences. Not only sequence data continues to outpace annotation information, but the problem is further exacerbated when organisms are underrepresented in the annotation databases. This is the case with non human-pathogenic viruses which occur frequently in metagenomic projects. Thus there is a need for tools capable of detecting and classifying viral sequences.

Based on machine learning techniques, we have proposed a new effective tool for identifying capsid and tail protein sequences, which are the cornerstones toward viral sequence annotation and viral genome classification. The software and corresponding web server are publicly available as part of the SCRATCH suite. [*F. Coste, C. Galiez*] [19]

Learning substitutable context-free grammars to model protein families. In the first experiments on learning substitutable context-free grammars to model protein families, an identified bottleneck for larger scale experimentation was parsing time. We have implemented a new parsing strategy enabling to handle efficiently the ambiguity of 'gap loops', enabling a factor 20 speedup in practice. We have also begun to investigate the inference of more expressive classes, said contextually substitutable, and have proposed a refined graph approach to learn smaller contextually substitutable grammars from smaller training samples in the framework that we have initiated with ReGLiS. [*F. Coste*] [43], [35]

How to measure the topological quality of protein grammars? To assess the quality of grammars modelling protein families, one is interested in their performances to predict new members of the families, classically measured on the basis of recall and precision in the machine learning framework, but also by their modelling power, which is more difficult to evaluate. We propose here to address this later point by measuring the consistency of grammar's parse trees with 3D structures of proteins, when they are available, by the introduction of a set of measures based on respective internal distances. [*F. Coste*] [36]

Tutorial chapter: Learning the Language of Biological Sequences. Learning the language of biological sequences is an appealing challenge for the grammatical inference research field. While some first successes have already been recorded, such as the inference of profile hidden Markov models or stochastic context-free grammars which are now part of the classical bioinformatics toolbox, it is still a source of open and nice inspirational problems for grammatical inference, enabling us to confront our ideas to real fundamental applications. As an introduction to this field, we survey here the main ideas and concepts behind the approaches developed in pattern/motif discovery and grammatical inference to characterize successfully the biological sequences with their specificities. [*F. Coste*] [40]

FLUMINANCE Project-Team

6. New Results

6.1. Fluid motion estimation

6.1.1. Stochastic uncertainty models for motion estimation

Participants: Shengze Cai, Etienne Mémin, Musaab Khalid Osman Mohammed.

The objective consists here in relying on a stochastic transport formulation to propose a luminance conservation assumption dedicated to the measurement of large-scale fluid flows velocity. This formulation has the great advantage to incorporate from the beginning an uncertainty on the motion measurement. This uncertainty modeled as a possibly inhomogeneous random field uncorrelated in time can be estimated jointly to the motion estimates. Such a formulation, besides providing estimates of the velocity field and of its associated uncertainties, allows us to naturally define a linear multiresolution scale-space framework. It provides also a reinterpretation, in terms of uncertainty, of classical regularization functionals proposed in the context of motion estimation. This estimator, which extend a local motion estimator previously proposed in the team, has shown to improve significantly the results of the corresponding deterministic estimator. This kind method is assessed in the context of river hydrologics applications through a collaboration with an Irstea Lyon research group (HHLY). This study is performed within the PhD thesis of Musaab Mohammed.

6.1.2. Development of an image-based measurement method for large-scale characterization of indoor airflows

Participants: Dominique Heitz, Etienne Mémin, Romain Schuster.

The goal is to design a new image-based flow measurement method for large-scale industrial applications. From this point of view, providing in situ measurement technique requires the development of precise models relating the large-scale flow observations to the velocity, appropriate large-scale regularization strategies, and adapted seeding and lighting systems, like Hellium Filled Soap Bubles (HFSB) and led ramp lighting.This work conducted within the PhD of Romain Schuster in collaboration with the compagny ITGA has started in february 2016. The first step has been to evaluate the performances of a stochastic uncertainty motion estimator when using large scale scalar images, like those obtained when seeding a flow with smoke.

6.1.3. 3D flows reconstruction from image data

Participants: Dominique Heitz, Cédric Herzet.

Our work focuses on the design of new tools for the estimation of 3D turbulent flow motion in the experimental setup of Tomo-PIV. This task includes both the study of physically-sound models on the observations and the fluid motion, and the design of low-complexity and accurate estimation algorithms.

This year, we keep on our investigation on the problem of efficient volume reconstruction. Our work takes place within the context of some modern optimization techniques. First, we focussed our attention on the family of proximal and splitting methods and showed that the standard techniques commonly adopted in the TomoPIV literature can be seen as particular cases of such methodologies. Recasting standard methodologies in a more general framework allowed us to propose extensions of the latter: i) we showed that the parcimony characterizing the sought volume can be accounted for without increasing the complexity of the algorithms (e.g., by including simple thresholding operations); ii) we emphasized that the speed of convergence of the standard reconstruction algorithms can be improved by using Nesterov's acceleration schemes; iii) we also proposed a totally novel way of reconstructing the volume by using the so-called "alternating direction of multipliers method" (ADMM). In 2016, this work has led to the publication of a contribution in the international journal IOP Measurement Science and Technology.

On top of this work, we also focussed on another crucial step of the volume reconstruction problem, namely the pruning of the model. The pruning task consists in identifying some positions in the volume of interest which cannot contains any particle. Removing this position from the problem can then potentially allow for a dramatic dimensionality reduction. This year, we provide a methodological answer to this problem through the prism of the so-called "screening" techniques which have been proposed in the community of machine learning. In 2016, this work led to the publication of one contribution in the proceedings of the international conference on acoustics, speech and signal processing (ICASSP'16).

6.1.4. Sparse-representation algorithms

Participant: Cédric Herzet.

The paradigm of sparse representations is a rather new concept which turns out to be central in many domains of signal processing. In particular, in the field of fluid motion estimation, sparse representation appears to be potentially useful at several levels: i) it provides a relevant model for the characterization of the velocity field in some scenarios; ii) it plays a crucial role in the recovery of volumes of particles in the 3D Tomo-PIV problem.

Unfortunately, the standard sparse representation problem is known to be NP hard. Therefore, heuristic procedures have to be devised to access to the solution of this problem. Among the popular methods available in the literature, one can mention orthogonal matching pursuit (OMP), orthogonal least squares (OLS) and the family of procedures based on the minimization of sparsity inducing norms. In order to assess and improve the performance of these algorithms, theoretical works have been undertaken in order to understand under which conditions these procedures can succeed in recovering the "true" sparse vector.

This year, we contributed to this research axis by deriving conditions of success for the algorithms mentioned above when the amplitudes of the nonzero coefficients in the sparse vector obey some decay. In a TomoPIV context, this decay corresponds to the fact that not all the particles in the fluid diffuse the same quantity of light (notably because of illumination or radius variation). In particular, we show that the standard coherence-based guarantees for OMP/OLS can be relaxed by an amount which depends on the decay of the nonzero coefficients. In 2016, our work has led to the publication of one paper in the journal IEEE Transactions on Information Theory.

We also investigated a new methodology to take sparsity into account into variational assimilation problems. We focussed on the problem of estimating of scalar transported by an unknown velocity field, when only low-resolution observations of the scalar are supposed to be available. The goal is to reconstruct both a high-resolution version of the scalar and the velocity field, assuming that these quantities admit a sparse decomposition in some proper frames. The associated optimization problem typically involves millions of variables and thus requires dedicated optimization procedures to be tractable. In 2016, we proposed a new assimilation scheme combining state-of-the-art optimization techniques (forward-backward propagation, ADMM, Attouch's procedure) to address this problem. Our algorithm is provably convergent while exhibiting a complexity per iteration evolving linearly with the problem's dimensions. This contribution has led to a journal publication in SIAM Journal on Imaging Science.

6.2. Tracking, Data assimilation and model-data coupling

6.2.1. Stochastic fluid flow dynamics under uncertainty

Participants: Pierre Derian, Etienne Mémin, Valentin Resseguier.

In this research axis we aim at devising Eulerian expressions for the description of fluid flow evolution laws under uncertainties. Such an uncertainty is modeled through the introduction of a random term that allows taking into account large-scale approximations or truncation effects performed within the dynamics analytical constitution steps. This includes for instance the modeling of unresolved scales interaction in large eddies simulation (LES) or in Reynolds average numerical simulation (RANS), but also uncertainties attached to non-uniform grid discretization. This model is mainly based on a stochastic version of the Reynolds transport theorem. Within this framework various simple expressions of the drift component can be exhibited for

different models of the random field carrying the uncertainties we have on the flow. We aim at using such a formalization within image-based data assimilation framework and to derive appropriate stochastic versions of geophysical flow dynamical modeling. This formalization has been published in the journal Geophysical and Astrophysical Fluid Dynamics [10]. Numerical simulation on divergence free wavelets basis of 3D viscous Taylor-Green vortex and Crow instability have been performed within a collaboration with Souleymane Kadri-Harouna. Besides, we explore in the context of Valentin Resseguier's PhD the extension of such framework to oceanic models and to satellite image data assimilation. This PhD thesis takes place within a fruitful collaboration with Bertrand Chapron (CERSAT/IFREMER). This year we have more deeply explored several uncertainty representations of classical geophysical models for ocean and atmosphere. This study have led to very promising stochastic representation for the Quasi Geostophic approximation (QG) with noises of different energy.

6.2.2. Free surface flows reconstruction and tracking

Participants: Dominique Heitz, Etienne Mémin.

We investigated the combined use of a Kinect depth sensor and of a stochastic data assimilation method to recover free-surface flows. More generally, we proposed a particle filter method to reconstruct the complete state of free-surface flows from a sequence of depth images only. The data assimilation scheme introduced accounts for model and observations errors. We evaluated the developed approach on two numerical test cases: a collapse of a water column as a toy-example and a flow in an suddenly expanding flume as a more realistic flow. The robustness of the method to simulated depth data quality and also to initial conditions was considered. We illustrated the interest of using two observations instead of one observation into the correction step. Then, the performance of the Kinect sensor to capture temporal sequences of depth observations was investigated. Finally, the efficiency of the algorithm was qualified for a wave in a real rectangular flat bottom tank. It was shown that for basic initial conditions, the particle filter rapidly and remarkably reconstructed velocity and height of the free surface flow based on noisy measures of the elevation

6.2.3. Optimal control techniques for the coupling of large scale dynamical systems and image data

Participants: Pranav Chandramouli, Dominique Heitz, Etienne Mémin.

In this axis of work we are exploring the use of optimal control techniques for the coupling of Large Eddies Simulation (LES) techniques and 2D image data. The objective is to reconstruct a 3D flow from a set of simultaneous time resolved 2D image sequences visualizing the flow on a set of 2D plans enlightened with laser sheets. This approach will be experimented on shear layer flows and on wake flows generated on the wind tunnel of Irstea Rennes. Within this study we wish also to explore techniques to enrich large-scale dynamical models by the introduction of uncertainty terms or through the definition of subgrid models from the image data. This research theme is related to the issue of turbulence characterization from image sequences. Instead of predefined turbulence models, we aim here at tuning from the data the value of coefficients involved in traditional LES subgrid models. The longer-term goal is to learn empirical subgrid models directly from image data. An accurate modeling of this term is essential for Large Eddies Simulation as it models all the non resolved motion scales and their interactions with the large scales.

We have pursued the first investigations on a 4DVar assimilation technique, integrating PIV data and Direct Numerical Simulation (DNS), to reconstruct two-dimensional turbulent flows. The problem we are dealing with consists in recovering a flow obeying Navier-Stokes equations, given some noisy and possibly incomplete PIV measurements of the flow. By modifying the initial and inflow conditions of the system, the proposed method reconstructs the flow on the basis of a DNS model and noisy measurements. The technique has been evaluated in the wake of a circular cylinder. It denoises the measurements and increases the spatiotemporal resolution of PIV time series. These results have been recently published in the Journal of Computational Physics [7]. Along the same line of studies the 3D case is ongoing. The goal consists here to reconstruct a 3D flow from a set of simultaneous time resolved 2D images of planar sections of the 3D volume. This work has been mainly conducted within the PhD of Cordelia Robinson. The development of the variational assimilation code has been initiated within a collaboration with A. Gronskis, S. Laizé (lecturer, Imperial College, UK) and

Eric Lamballais (institut P' Poitiers). A High Reynolds number simulation of the wake behind a cylinder has been recently performed within this collaboration. The 4DVar assimilation technique based on the numerical code Incompact3D is now implemented. We are currently trying to reconstruct a 3D turbulent flow from dual plane velocity observations. The control of subgrid parameterizations will be the main objective of the PhD of Pranav Chandramouli that is just starting.

6.2.4. Ensemble variational data assimilation of large scale fluid flow dynamics with uncertainty

Participant: Etienne Mémin.

This study is focused on the coupling of a large scale representation of the flow dynamics built from the location uncertainty principle with image data of finer resolution. The velocity field at large scales is described as a regular smooth component whereas the complement component is a highly oscillating random velocity field defined on the image grid but living at all the scales. Following this route we have assessed the performance of an ensemble variational assimilation technique with direct image data observation. Preliminary encouraging results have been obtained for simulation under uncertainty of 1D and 2D shallow water models.

6.2.5. Reduced-order models for flows representation from image data

Participants: Mamadou Diallo, Cédric Herzet, Etienne Mémin, Valentin Resseguier.

During the PhD thesis of Valentin Ressguier we proposed a new decomposition of the fluid velocity in terms of a large-scale continuous component with respect to time and a small-scale non continuous random component. Within this general framework, an uncertainty based representation of the Reynolds transport theorem and Navier-Stokes equations can be derived, based on physical conservation laws. This physically relevant stochastic model has been applied in the context of the POD-Galerkin method. The pertinence of this reduced order model has been successfully assessed on several wake flows. This study has been published in two conference papers and one journal article.

On the other hand, we investigated the problem of reduced-model construction from partial observations. In this line of search, our contribution was twofold. We first proposed a Bayesian framework for the construction of reduced-order models from image data. Our framework enables to account for any prior information on the system to reduce and takes the uncertainties on the parameters of the model into account. Interestingly, the proposed approach reduces to some well-known model-reduction techniques when the observations are not partial (i.e., the observation operator can be inverted). Second, we provided a theoretical analysis of our methodology in a simplified context (namely, the observations are supposed to be noiseless linear combinations of the state of the system). This result provides worst-case guarantees on the reconstruction performance which can be achieved by a reduced model built from the data. These contributions have led to the publications of one contribution in the proceedings of the international conference on acoustics, speech and signal processing (ICASSP'16). A journal version of these contributions has been submitted.

6.3. Analysis and modeling of turbulent flows

6.3.1. Singular and regular solutions to the Navier-Stokes equations (NSE) and relative turbulent models

Participant: Roger Lewandowski.

The common thread of this work is the problem set by J. Leray in 1934 : does a regular solution of the Navier-Stokes equations (NSE) with a smooth initial data develop a singularity in finite time, what is the precise structure of a global weak solution to the Navier-Stokes equations, and are we able to prove any uniqueness result of such a solution. This is a very hard problem for which there is for the moment no answer. Nevertheless, this question leads us to reconsider the theory of Leray for the study of the Navier-Stokes equations in the whole space with an additional eddy viscosity term that models the Reynolds stress in the context of large-scale flow modelling. It appears that Leray's theory cannot be generalized turnkey for this problem, so that things must be reconsidered from the beginning. This problem is approached by a regularization process using

mollifiers, and particular attention must paid to the eddy viscosity term. For this regularized problem and when the eddy viscosity has enough regularity, we have able to prove the existence of a global unique solution that is of class C^∞ in time and space and that satisfies the energy balance. Moreover, when the eddy viscosity is of compact support in space, uniformly in time, we recently shown that this solution converges to a turbulent solution to the corresponding Navier-Stokes equations when, the regularizing parameter goes to 0. These results are described in a paper that will be soon submitted to the journal Archive for Rational Mechanics and Analysis (ARMA).

In the same direction, we also finalized a paper in collaboration with L. Berselli (Univ. Pisa, Italy) about the well known Bardina's turbulent model. In this problem, we consider the Helmholtz filter usually used within the framework of Large Eddy Simulation. We carry out a similar analysis, by showing in particular that no singularity occurs for Bardina's model.

Another study in collaboration with B. Pinier, P. Chandramouli and E. Memin has been undertaken. This work takes place within the context of the PhD work of B. Pinier. We considered the standard turbulent models involving the Navier-Stokes equations with an eddy viscosity that depends on the Turbulent Kinetic Energy (TKE), coupled with an addition equation for the TKE. The problem holds in a 3D bounded domain, with the Manning law at the boundary for the velocity. We have modeled a flux condition at the boundary for the TKE. We prove that with these boundary conditions, the resulting problem has a distributional solution. Then a serie of numerical tests is performed in a parallelepiped with a non trivial bottom, showing the accuracy of the model in comparison with a direct numerical simulation of the Navier-Stokes equations.

6.3.2. Turbulence similarity theory for the modeling of Ocean Atmosphere interface

Participants: Roger Lewandowski, Etienne Mémin, Benoit Pinier.

The Ocean Atmosphere interface plays a major role in climate dynamics. This interaction takes place in a thin turbulent layer. To date no sastifying universal models for the coupling of atmospheric and oceanic models exists. In practice this coupling is realized through empirically derived interaction bulks. In this study, corresponding to the PhD thesis of Benoit Pinier, we aim at exploring similarity theory to identify universal mean profile of velocity and temperature within the mixture layer. The goal of this work consists in exhibiting eddy viscosity models within the primitive equations. We will also explore the links between those eddy viscocity models and the subgrid tensor derived from the uncertainty framework studied in the Fluminance group. In that prospect, we have started to study the impact of the introduction of a random modeling of the friction velocity on the classical wall law expression.

6.3.3. Hot-wire anemometry at low velocities

Participant: Dominique Heitz.

A new dynamical calibration technique has been developed for hot-wire probes. The technique permits, in a short time range, the combined calibration of velocity, temperature and direction calibration of single and multiple hot-wire probes. The calibration and measurements uncertainties were modeled, simulated and controlled, in order to reduce their estimated values. Based on a market study the french patent application has been extended this year to a Patent Cooperation Treaty (PCT) application.

6.3.4. Numerical and experimental image and flow database

Participants: Pranav Chandramouli, Dominique Heitz.

The goal was to design a database for the evaluation of the different techniques developed in the Fluminance group. The first challenge was to enlarge a database mainly based on two-dimensional flows, with three-dimensional turbulent flows. Synthetic image sequences based on homogeneous isotropic turbulence and on circular cylinder wake have been provided. These images have been completed with time resolved Particle Image Velocimetry measurements in wake and mixing layers flows. This database provides different realistic conditions to analyse the performance of the methods: time steps between images, level of noise, Reynolds number, large-scale images. The second challenge was to carried out orthogonal dual plane time resolved stereoscopic PIV measurements in turbulent flows. The diagnostic employed two orthogonal and synchronized

stereoscopic PIV measurements to provide the three velocity components in planes perpendicular and parallel to the streamwise flow direction. These temporally resolved planar slices observations will be used in 4DVar assimilation technique, integrating Direct Numerical Simulation (DNS) and Large Eddies Simulation (LES), to reconstruct three-dimensional turbulent flows. This reconstruction will be conducted within the PhD of Pranav Chandramouli. The third challenge was to carried out a time resolved tomoPIV experiments in a turbulent wake flow. These temporally resolved volumic observations will be used to assess the algorithms developped in the PhD of Ioana Barbu and in the postdoc of Kai Berger. Then this data will be used in 4DVar assimilation technique to reconstruct three-dimensional turbulent flows. This reconstruction will be conducted within the PhD of Cordelia Robinson.

6.4. Visual servoing approach for fluid flow control

6.4.1. Closed-loop control of a spatially developing shear layer

Participants: Christophe Collewet, Johan Carlier.

This study aims at controling one of the prototypical flow configurations encountered in fluid mechanics: the spatially developing turbulent shear layer occuring between two parallel incident streams with different velocities. Our goal is to maintain the shear-layer in a desired state and thus to reject upstream perturbations. As in all our previous works in flow control, we propose a vision-based approach to control this flow. We investigate the use of an optimal control based on a reduced linearized state space model of the Navier-Stokes equations. A steady desired state was first considered leading to a linear time-invariant system. The main problem consists to maintain the flow in his desired state in presence of unknown perturbation. Different strategies have been evaluated for different types of actuators and different cost functions. Even if our control law is based on a linearized approach, its efficiency has been validated on a realistic numerical Navier-Stokes 3D solver. This work has been submitted to the 20th World Congress of the International Federation of Automatic Control (IFAC).

6.5. Reactive transport

6.5.1. Reactive transport in porous media

Participant: Jocelyne Erhel.

In many environmental applications, transport of solutes is coupled with chemical reactions, either kinetic or at equilibrium. These reactions involve not only solutes, but also sorbed species and minerals. The mathematical model is a coupled set of nonlinear partial algebraic differential equations. A classical approach is to discretize first in space then in time. Since the problem is rather stiff, explicit time discretization suffers from a drastic CFL-like condition. On the other hand, implicit schemes allow large timesteps during some periods of simulation. Implicit Euler scheme is often used for monotonicity properties. The Jacobian is computed from the transport operator and the chemical operator. We have designed such a global approach and implemented it in our software GRT3D. We have done numerical experiments on the benchmark MoMaS.

Publications: 2 conferences and one journal article [15], [20], [21]

Grant: H2MNO4

6.5.2. Reactive transport in fractured-porous media

Participants: Yvan Crenner, Benjamin Delfino, Jean-Raynald de Dreuzy, Jocelyne Erhel.

Even in small numbers, fractures must be carefully considered for the geological disposal of radioactive waste. They critically enhance diffusivity, speed up solute transport, extend mixing fronts and, in turn, modify the physicochemical conditions of reactivity around possible storage sites. Numerous studies addressing various applications (e.g. radioactive waste storage, CO2 sequestration, geothermal storage, hydrothermal alteration) have shown that fractures cannot be simply integrated within an equivalent porous medium. Our objective is to develop a reactive transport model based on the separation of the fracture and matrix domains, with diffusion conditions differing between the fracture and in the matrix, appropriate flow-rock interactions at equilibrium in the matrix and fracture-matrix exchange conditions at their interface.

This year, we developed a numerical model for a chemical system with several minerals, which is representative of a storage site.

Publications: 2 conferences [28], [27]

Grant: ANDRA

6.6. Linear solvers

6.6.1. Sparse linear solvers

Participants: Jocelyne Erhel, David Imberti.

Sparse linear systems arise in computational science and engineering. The goal is to reduce the memory requirements and the computational cost, by means of high performance computing algorithms. We introduce a new variation on s-step GMRES in order to improve its stability, reduce the number of iterations necessary to ensure convergence, and thereby improve parallel performance. In doing so, we develop a new block variant that allows us to express the stability difficulties in s-step GMRES more fully.

Grants and projects: EXA2CT 8.2.1 , EoCoE 8.2.2 , C2S@EXA 8.1.7

Publications: 3 conférences [22], [23], [39]

GENSCALE Project-Team

7. New Results

7.1. HTS data processing

7.1.1. Providing end-user solutions, example from the Colib'read on galaxy project

Participants: Claire Lemaitre, Camille Marchet, Pierre Peterlongo.

Colib'read tools suite uses optimized reference-free algorithms for various analyses of NGS datasets, such as variant calling or read set comparisons. To facilitate data analysis and tools dissemination, we developed Galaxy tools and tool shed repositories. The galaxy package, facilitates the analyzis of raw NGS data for a broad range of life scientists [16].

7.1.2. Assembly of Streptococcus Bacteria

Participant: Dominique Lavenier.

With the microbiological and bacteriological group of the Rennes hospital, we design a new strategy to assemble the genomes of 40 *Streptococcus* bacteria. Each strain has been sequenced and independently assembled using different assembly tools. For a specific strain, a merge of the contigs is done using the MIX software. This step allows the number of contigs to be significantly reduced, resulting in a better final assembly compared to each individual assembly. The comparison with other known *Streptococcus* genomes indicates where phages are located in the genome [20].

7.1.3. Data-mining applied to GWAS

Participants: Pham Hoang Sun, Dominique Lavenier.

Identifying variant combination association with disease is a bioinformatics challenge. This problem can be solved by discriminative pattern mining that uses statistical functions to evaluate the significance of individual biological patterns. There is a wide range of such measures. However, selecting an appropriate measure as well as a suitable threshold in some specific practical situations is a difficult task. We propose to use the skypattern technique which allows combinations of measures to be used to evaluate the importance of variant combinations without having to select a given measure and a fixed threshold. Experiments on several real variant datasets demonstrate that the skypattern method effectively identifies the risk variant combinations related to diseases [28].

7.1.4. Variant detection in transcriptomic data

Participant: Camille Marchet.

We defined a method to identify, quantify and annotate SNPs (Single Nucleotide Polymorphisms) using RNA-seq reads only. Organisms with a poor quality or no reference genome can take benefit of this approach, as well as studies where not enough material is available for sequencing from one individual, where samples can be pooled. The method relies on motifs discovery and post-treatment in de Bruijn graphs built from the reads. It can be used for any species to annotate SNPs and predict their impact on proteins as well as test their association to a phenotype of interest. The approach has been validated using well known human RNA-seq data. Results have been compared with state of the art approaches for variant calling. We showed that the methods perform similarly in terms of precision and recall. Then we focused on the main target of the study, namely the non-model species. We finally validated experimentally the predictions of our method [18].

7.1.5. Faster de Bruijn graph compaction

Participant: Antoine Limasset.

We developed a new algorithm, called BCALM2, for the compaction of de Bruijn graphs. BCALM2 is a parallel algorithm based on minimizer repartition of sequences. This repartition allows the compaction of extremely large graphs with moderate memory usage and time. The compaction of a human sequencing graph can be done in 1 hour with only 3GB of memory and huge genomes, such as the pine and white spruce ones (more than 20Gbp each), can be handled using our approach on a regular server (2 days and 40GB of memory). Those results argue that BCALM2 is one order of magnitude more efficient than available approaches and can tackle the assembly bottleneck of constructing a compacted de Bruijn graph [14].

7.1.6. Scaffolding

Participants: Rumen Andonov, Sebastien François, Dominique Lavenier.

We developed a method for solving genome scaffolding as a problem of finding a long simple path in a graph defined by the contigs that satisfies additional constraints encoding the insert-size information. Then we solved the resulting mixed integer linear program to optimality using the Gurobi solver. We tested our algorithm on several chloroplast genomes and showed that it outperforms other widely-used assembly solvers by the accuracy of the results [25].

7.2. Sequence comparison

7.2.1. Metagenomics datasets comparison

Participants: Gaetan Benoit, Dominique Lavenier, Claire Lemaitre, Pierre Peterlongo.

We developed a new method, called Simka, to compare simultaneously numerous large metagenomics datasets. The method computes pairwise distances based on the amount of shared k-mers between datasets. The method scales to a large number of datasets thanks to an efficient kmer-counting step that processes all datasets simultaneoulsy. Additionnally, several distance definitions were implemented and compared, including some originating from the ecological domain. The method is currently applied to the TARA oceans project (more than 2000 datasets) which aims at comparing worldwide sea water samples (ANR HydrGen project) [12].

7.2.2. Read similarity detection

Participants: Camille Marchet, Antoine Limasset, Pierre Peterlongo.

Retrieving similar reads inside or between read sets is a fundamental task either for algorithmic reasons or for analyses of biological data. This task is easy in small datasets, but becomes particularly hard when applied to millions or billions of reads. In [24] we used a straightforward indexing structure that scales to billions of elements. We proposed two direct applications in genomics and metagenomics. These applications consist in either approximating the number of similar reads between dataset(s) or to simply retrieve these similar reads. They can be applied on distinct read sets or on a read set against itself.

7.3. Parallelism

7.3.1. Processing-in-Memory

Participants: Charles Deltel, Dominique Lavenier.

The concept of PIM (Processor In Memory) aims to dispatch the computer power near the data. Together with the UPMEM company (http://www.upmem.com/), which is currently developing a DRAM memory enhanced with computing units, we investigate the parallelization of two bioinformatics algorithms for this new type of memory: sequence alignment and mapping [34] [33]. The first results show that blast-like algorithms or mapping algorithms can highly benefit from such memory and speed-up of more than 25 can be achieved [26].

7.3.2. GPU for graph algorithms

Participants: Rumen Andonov, Dominique Lavenier.

We describe three algorithms and their associated GPU implementations for two types of shortest path problems. These implementations target computations on graphs with up to millions of vertices and executions on GPU clusters. The first two algorithms solve the All-Pairs Shortest Path (APSP) problem. The first of these two algorithms allows computations on graphs with negative edges while the second trades this ability for better parallel scaling properties and improved memory access. The third algorithm solves the Single-Pair Shortest Path (SPSP) query problem. Our implementations efficiently exploit the computational power of 256 GPUs simultaneously. All shortest paths of a million vertex graph can be computed in 6 minutes and shortest path queries on the same graph are answered in a quarter of a millisecond. These implementations proved to be orders of magnitude faster than existing parallel approaches[30].

7.4. Data representation

7.4.1. Computational pan-genomics: status, promises and challenges

Participant: Pierre Peterlongo.

We took part to the Computational Pan-Genomics Consortium producing a “white paper” dedicated to computational pan-genomic. A pan-genome is a representation of the union of the genomes of closely related individuals (eg from a same species). Computational pan-genomics is a new sub-area of research in computational biology. In [19], we generalized existing definitions and we examined already available approaches to construct and use pan-genomes, discussed the potential benefits of future technologies and methodologies and reviewed open challenges from the vantage point of the above-mentioned biological disciplines.

7.4.2. Mapping reads on graphs

Participants: Pierre Peterlongo, Antoine Limasset.

Many published genome sequences remain in the state of a large set of contigs. Each contig describes the sequence found along some path of the assembly graph, however, the set of contigs does not record all the sequence information contained in that graph. Although many subsequent analyses can be performed with the set of contigs, one may ask whether mapping reads on the contigs is as informative as mapping them on the paths of the assembly graph.

In [17], we proposed a formal definition of mapping a sequence on a de Bruijn graph, we analysed the problem complexity, and we provided a practical solution. The proposed tool can map millions of reads per CPU hour on a de Bruijn graph built from a large set of human genomic reads. Results show that up to 22 % more reads can be mapped on the graph but not on the contig set.

7.5. Applications

7.5.1. Study of the rapeseed genome structure

Participants: Sebastien Letort, Pierre Peterlongo, Dominique Lavenier, Claire Lemaitre, Fabrice Legeai.

In collaboration with IGEPP (Institut de Génétique, Environnement et Protection des Plantes), INRA, and through two national projects, PIA Rapsodyn and France-Génomique Polysuccess, we are involved in the genome analysis of several rapeseed varieties. The Rapsodyn project has the ambition to insure long-term competitiveness of the rapeseed production through improvement of the oil yield and reduction of nitrogen inputs during the crop cycle. Rapeseed varieties must thus be selected from genotypes that favor low nitrogen input. DiscoSnp++ is here used to locate new variants among the large panel of rapeseed varieties which have been sequenced during the project.

The PolySuccess project aims to answer the following question: how a polyploid, such as the oilseed rape plant, becomes a new species? Oilseed rape (Brassica napus) being a natural hybrid between B.rapa and B.oleracea, different genomes of these three species have been sequenced to study their structures. The Minia assembly pipeline provides a fast way to generate contigs that are used for studying gene specificities.

7.5.2. GATB Production Pipeline

Participants: Patrick Durand, Charles Deltel.

The entire set of libraries and tools related to the GATB Software have been introduced within a professional environment to support high-quality C++ developments. It relies on the use of technology platforms available at Inria: OpenStack and Jenkins. Considering the latter, we have setup more than 50 Jenkins tasks to automate the entire software development based on GATB: C++ code compiling and testing, documentation creation, packaging and preparation of official releases, mirroring on public Github repositories. Code compilation and tests are done on Linux and MacOSX VMs. https://ci.inria.fr/gatb-core/

7.5.3. Variant predictions in the pea genome

Participant: Pierre Peterlongo.

Progress in genetics and breeding in pea suffered from the limited availability of molecular resources. SNP markers that can be identified through affordable sequencing processes without the need for prior genome reduction or a reference genome allow the discovery of thousands of molecular markers.

We have been involved with IGEPP (Institut de Génétique, Environnement et Protection des Plantes, INRA) in the application of the discoSnp++ tool, discovering SNPs on HiSeq whole genome sequencing of four pea lines. Validation of a subset of predicted SNPs showed that almost all generated SNPs are highly designable and that most (95 %) deliver highly qualitative genotyping result [13].

7.5.4. Analysis of insect pest genomes

Participant: Fabrice Legeai.

Within a large international network of biologists, GenScale has contributed to various projects for identifying important components involved in the adaptation of major agricultural pests to their environment. We provided the assemblies, the annotations and the comparisons of various insects genomes [29]. Following specific agreement or policy, these results are available for browsing and consulting to a restricted consortium or a large community through the BioInformatics platform for Agro-ecosystems Arthropods (http://bipaa.genouest.org/is). In particular, this year our work helped to identify aphid genes involved in the adaptation to their favorite plant [15], or genes that are differentially expressed between leaf- and root-feeding phylloxera [21]. Furthermore, in order to help scientists to consult and cross genomics and postgenomics data, we are developping AskOmics, an integration and interrogation software for (linked) biological data, within a strong partnership, with Dyliss and GenOuest [36], [27].

HYBRID Project-Team

7. New Results

7.1. Virtual Reality and 3D Interaction

7.1.1. Perception in Virtual Environments

With the increasing demand in consumer VR applications, the need to understand how users perceive the virtual environment and their virtual self (avatar) is becoming more and more important. In particular, with the potential of virtual reality to alter and control avatars in different ways, the user representation in the virtual world does not always necessarily match the user body structure. Besides, the study of how the users perceive their surrounding environment (e.g. depth perception) is another active field of research in VR.

The role of interaction in virtual embodiment: Effects of the virtual hand representation

Participants: Ferran Argelaguet and Anatole Lécuyer

First, we have studied how people appropriate their virtual hand representation when interacting in virtual environments [14]. In order to answer this question, we conducted an experiment studying the sense of embodiment when interacting with three different virtual hand representations (see Figure 2), each one providing a different degree of visual realism but keeping the same control mechanism. The main experimental task was a Pick-and-Place task in which participants had to grasp a virtual cube and place it to an indicated position while avoiding an obstacle (brick, barbed wire or fire). Results show that the sense of agency is stronger for less realistic virtual hands which also provide less mismatch between the participant's actions and the animation of the virtual hand. In contrast, the sense of ownership is increased for the human virtual hand which provides a direct mapping between the degrees of freedom of the real and virtual hand.

This work was done in collaboration with MimeTIC team.

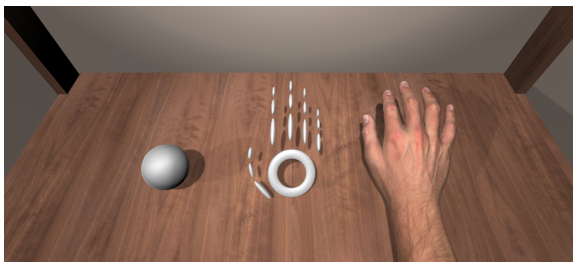


Figure 2. Evaluated virtual hand representations: abstract (left), iconic (center) and realistic virtual hands (right). Each virtual hand had its own visual feedback when the grasping operation was triggered.

Wow! I Have Six Fingers!": Would You Accept Structural Changes of Your Hand in VR?

Participants: Ferran Argelaguet and Anatole Lécuyer

In a different context, we have explored how users would accept as their own a six-digit realistic virtual hand [6]. By measuring participants' senses of ownership (i.e., the impression that the virtual hand is actually our own hand) and agency (i.e., the impression to be able to control the actions of the virtual hand), we somehow evaluate the possibility of creating a Six-Finger Illusion in VR. We measured these two dimensions of virtual embodiment in a virtual reality experiment where participants performed two tasks successively: (1) a self-manipulation task inducing visuomotor feedback, where participants mimicked finger movements presented in the virtual scene and (2) a visuotactile task inspired by Rubber Hand Illusion protocols, where an experimenter stroked the hand of the user with a brush (see Figure 3). The real and virtual brushes were synchronously stroking the participants' real and virtual hand, and in the case when the virtual brush was stroking the additional virtual digit, the real ring finger was also synchronously stroked to provide consistent tactile stimulation and elicit a sense of embodiment. Results of the experiment show that participants did experience high levels of ownership and agency of the six-digit virtual hand as a whole. These results bring preliminary insights about how avatar with structural differences can affect the senses of ownership and agency experienced by users in VR.

This work was done in collaboration with MimeTIC team.

Figure 3. The virtual six-finger hand and the participant's hand are synchronously stimulated using a virtual and a real brush respectively.

CAVE Size Matters: Effects of Screen Distance and Parallax on Distance Estimation in Large Immersive Display Setups

Participants: Ferran Argelaguet and Anatole Lécuyer

When walking within a CAVE-like system, accommodation distance, parallax, and angular resolution vary according to the distance between the user and the projection walls, which can alter spatial perception. As these systems get bigger, there is a need to assess the main factors influencing spatial perception in order to better design immersive projection systems and virtual reality applications. In this work, we performed two experiments that analyze distance perception when considering the distance toward the projection screens and parallax as main factors. Both experiments were conducted in a large immersive projection system with up to 10-meter interaction space. The first experiment showed that both the screen distance and parallax

have a strong asymmetric effect on distance judgments. We observed increased underestimation for positive parallax conditions and slight distance overestimation for negative and zero parallax conditions. The second experiment further analyzed the factors contributing to these effects and confirmed the observed effects of the first experiment with a high-resolution projection setup providing twice the angular resolution and improved accommodative stimuli. In conclusion, our results suggest that space is the most important characteristic for distance perception, optimally requiring about 6- to 7-meter distance around the user, and virtual objects with high demands on accurate spatial perception should be displayed at zero or negative parallax [3].

This work was done in collaboration with MimeTIC team and the University of Hamburg.

7.1.2. 3D User Interfaces

GiAnt: stereoscopic-compliant multi-scale navigation in VEs

Participants: Ferran Argelaguet

Navigation in multi-scale virtual environments (MSVE) requires the adjustment of the navigation parameters to ensure optimal navigation experiences at each level of scale (see Figure 4). In particular, in immersive stereoscopic systems, e.g. when performing zoom-in and zoom-out operations, the navigation speed and the stereoscopic rendering parameters have to be adjusted accordingly. Although this adjustment can be done manually by the user, it can be complex, tedious and strongly depends on the virtual environment. We have proposed GiAnt (GIant/ANT) [15], a new multi-scale navigation technique which automatically and seamlessly adjusts the navigation speed and the scale factor of the virtual environment based on the user's perceived navigation speed. The adjustment ensures an almost-constant perceived navigation speed while avoiding diplopia effects or diminished depth perception due to improper stereoscopic rendering configurations. The results from the conducted user evaluation shows that GiAnt is an efficient multi-scale navigation which minimizes the changes of the scale factor of the virtual environment compared to state-of-the-art multi-scale navigation techniques.

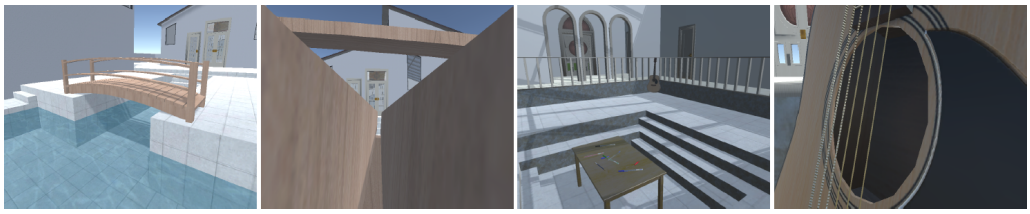


Figure 4. Multi-scale navigation sequence requiring the adaptation of the camera speed and the stereoscopic rendering parameters (e.g. parallax). GiAnt ensures that the navigation speed and the scale factor of the virtual environment are adjusted ensuring a comfortable navigation experience.

Enjoying 360° Vision with the FlyVIZ

Participants: Florian Nouviale, Maud Marchal and Anatole Lécuyer

FlyVIZ is a novel concept of wearable display device which enables to extend the human field-of-view up to 360°. With the FlyVIZ users can enjoy an artificial omnidirectional vision and see "with eyes behind their back"! We propose a novel version of our approach called the FlyVIZ v2. It is based on affordable and on the shelf components. For image acquisition, the FlyVIZ v2 relies on an iPhone4S smart-phone combined with a GoPano lens that contains a curved mirror enabling the capture of video with 360° horizontal field-of-view. For image transformation, we developed a dedicated software for iPhone that processes the video stream and transforms it into a real-time meaningful representation for the user. The "FlyVIZ_v2" was demonstrated at the ACM SIGGRAPH Emerging Technologies (2016).

Figure 5. (Left) Overview of the system. (Middle) 360° panoramic image displayed in the HMD when walking in a corridor. (Right) User grabbing an object located outside his natural field-of-view.

D3PART: A new Model for Redistribution and Plasticity of 3D User Interfaces

Participants: Jérémy Lacoche and Bruno Arnaldi

D3PART (Dynamic 3D Plastic And Redistribuable Technology) is a new model that we introduced to handle redistribution for 3D user interfaces. Redistribution consists in changing the components distribution of an interactive system across different dimensions such as platform, display and user. We extended previous plasticity models with redistribution capabilities, which lets developers create applications where 3D content and interaction tasks can be automatically redistributed across the different dimensions at runtime [21].

This work was done in collaboration with b<>com, ENIB and Telecom Bretagne.

Integration concept and model of Industry Foundation Classes (IFC) for interactive virtual environments

Participants: Anne-Solène Dris, Valérie Gouranton and Bruno Arnaldi

We defined a concept of Building Information Modeling (BIM) in combination with an integration model in order to enable interaction in Virtual Environments (see Figure 6). Such model, rich of information could be used to increase the level of abstraction of the interaction process. We proposed to explore and define how to create a BIM to ensure interoperability with the Industry Foundation Classes (IFC) model. The IFC model provides a definition of building objects, geometry, relation between objects, and other attributes such as layers, systems, link to planning, construction method, materials, domain (HVAC, Electrical, Architectural, Structure...) and quantities. The interoperability will enrich the virtual environment with the aim of creating an informed and interactive virtual environments, thus reducing the costs of applications' development. We defined a BIM modeling methodology extending the IFC interoperability to the interactive virtual environment [19].

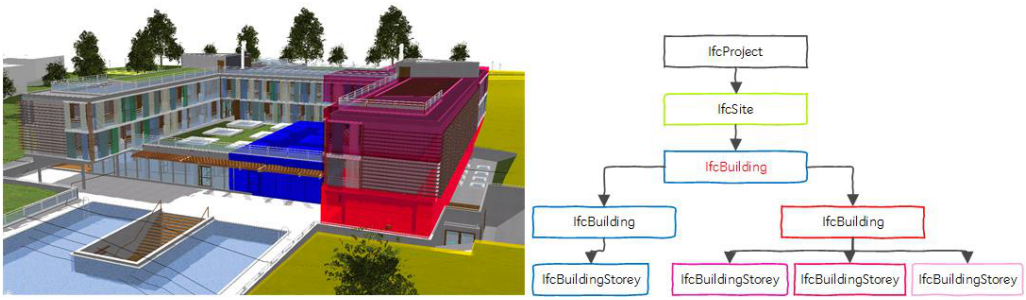


Figure 6. Interaction in virtual environments related to construction area based on BIM and a model of Industry Foundation Classes (IFC).

7.1.3. Virtual Archaeology

Digital and handcrafting processes applied to sound-studies of archaeological bone flutes

Participants: Jean-Baptiste Barreau, Ronan Gaugne, Bruno Arnaldi and Valérie Gouranton.

Bone flutes make use of a naturally hollow raw-material. As nature does not produce duplicates, each bone has its own inner cavity, and thus its own sound-potential. This morphological variation implies acoustical specificities, thus making it impossible to handcraft a true and exact sound-replica in another bone. This phenomenon has been observed in a handcrafting context and has led us to conduct two series of experiments (the first one using a handcrafting process, the second one using a 3D process) in order to investigate its exact influence on acoustics as well as on sound-interpretation based on replicas. The comparison of the results has shed light upon epistemological and methodological issues that have yet to be fully understood. This work contributes to assessing the application of digitization, 3D printing and handcrafting to flute-like sound instruments studied in the field of archaeomusicology [26].

This work was done in collaboration with MimeTIC team, ARTeHis, LBBE and Atelier El Block.

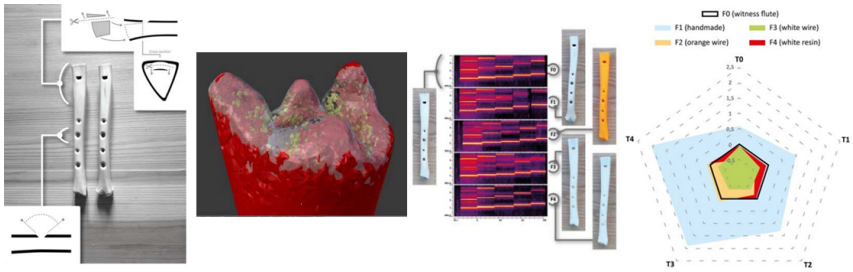


Figure 7. Sound-studies of archaeological bone flutes: (a) control flute (left) and its replica (right) both made out of goat's tibias, (b) 3D sculpted patch (transparent gray) on Blender (based on the geometry of the cloud), (c) Diagrams analysis , (d) The sound proximity of each replica comparing to the control flute, for each finger hole (numeric scale in semi-tones).

Internal 3D Printing of Intricate Structures

Participants: Ronan Gaugne, Valérie Gouranton and Bruno Arnaldi.

Additive technologies are increasingly used in Cultural Heritage process, for example in order to reproduce, complete, study or exhibit artefacts. 3D copies are based on digitization techniques such as laser scan or photogrammetry. In this case, the 3D copy remains limited to the external surface of objects. Medical images based digitization such as MRI or CT scan are also increasingly used in CH as they provide information on the internal structure of archaeological material. Different previous works illustrated the interest of combining 3D printing and CT scan in order to extract concealed artefacts from larger archaeological material. The method was based on 3D segmentation techniques within volume data obtained by CT scan to isolate nested objects. This approach was useful to perform a digital extraction, but in some case it is also interesting to observe the internal spatial organization of an intricate object in order to understand its production process. We propose a method for the representation of a complex internal structure based on a combination of CT scan and emerging 3D printing techniques mixing colored and transparent parts [25], [11]. This method was successfully applied to visualize the interior of a funeral urn and is currently applied on a set of tools agglomerated in a gangue of corrosion (see Figure 8).

This work was done in collaboration with Inrap and Image ET.

Figure 8. Front and bottom views of our 3D printed urn.

7.2. Physically-Based Simulation and Multisensory Feedback

7.2.1. Physically-based Simulation

Real-time tracking of deformable targets in 3D ultrasound sequences

Participants: Maud Marchal

Soft-tissue motion tracking is an active research area that consists in providing accurate evaluation about the location of anatomical structures. To do so, ultrasound imaging is often used since it is non-invasive, real-time and portable. Thus, several ultrasound tracking approaches have been developed in order to estimate soft tissue displacements that are caused by physiological motions and manipulations by medical tools. These methods have gained significant interest for image-guided therapies such as radio-frequency ablation or high-intensity focused ultrasound. In our work, we present a real-time approach that allows tracking deformable structures in 3D ultrasound sequences [8]. Our method consists in obtaining the target displacements by combining robust dense motion estimation and mechanical model simulation. We performed an evaluation of our method through simulated data, phantom data, and real-data. Results demonstrate that this novel approach has the advantage of providing correct motion estimation regarding different ultrasound shortcomings including speckle noise, large shadows and ultrasound gain variation. Furthermore, we show the good performance of our method with respect to state-of-the-art techniques by testing on the 3D databases provided by MICCAI CLUST'14 and CLUST'15 challenges.

This work was done in collaboration with LAGADIC team and b<>com.

7.2.2. 3D Haptic Interaction

DesktopGlove: a Multi-finger Force Feedback Interface Separating Degrees of Freedom Between Hands

Participants: Merwan Achibet and Maud Marchal

In virtual environments, interacting directly with our hands and fingers greatly contributes to immersion, especially when force feedback is provided for simulating the touch of virtual objects. Yet, common haptic interfaces are unfit for multi-finger manipulation and only costly and cumbersome grounded exoskeletons do provide all the efforts expected from object manipulation. To make multi-finger haptic interaction more accessible, we have proposed to combine two affordable haptic interfaces into a bimanual setup named DesktopGlove. With this approach, each hand is in charge of different components of object manipulation: one commands the global motion of a virtual hand while the other controls its fingers for grasping (see Figure 9). In addition, each hand is subjected to forces that relate to its own degrees of freedom so that users perceive a variety of haptic effects through both of them. Our results show that (1) users are able to

integrate the separated degrees of freedom of DesktopGlove to efficiently control a virtual hand in a posing task, (2) DesktopGlove shows overall better performance than a traditional data glove and is preferred by users, and (3) users considered the separated haptic feedback realistic and accurate for manipulating objects in virtual environments [12].

This work was done in collaboration with MJOLNIR team.

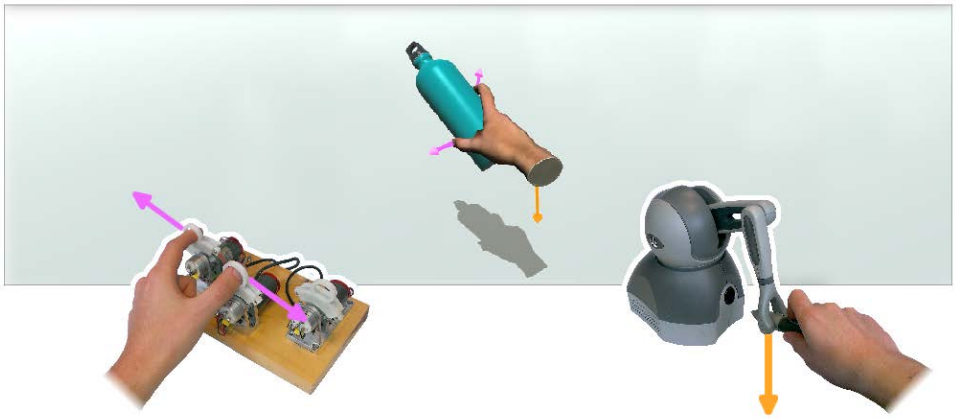


Figure 9. DesktopGlove separates the control of one virtual hand between both user's hands: a common haptic arm handles the global motion and a custom multi-finger interface controls the virtual fingers. The force feedback is split between both interfaces so that each hand is exposed to forces that relate to its own frame of reference.

ElasticArm: leveraging passive haptic feedback in virtual environments

Participants: Merwan Achibet, Adrien Girard, Anatole Lécuyer and Maud Marchal

Haptic feedback is known to improve 3D interaction in virtual environments but current haptic interfaces remain complex and tailored to desktop interaction. In [2], we describe an alternative approach called "Elastic-Arm" for incorporating haptic feedback in immersive virtual environments in a simple and cost-effective way. The Elastic-Arm is based on a body-mounted elastic armature that links the user's hand to his body and generates a progressive egocentric force when extending the arm. A variety of designs can be proposed with multiple links attached to various locations on the body in order to simulate different haptic properties and sensations such as different levels of stiffness, weight lifting, bimanual interaction, etc. Our passive haptic approach can be combined with various 3D interaction techniques and we illustrate the possibilities offered by the Elastic-Arm through several use cases based on well-known techniques such as the Bubble technique, redirected touching, and pseudo-haptics. A user study was conducted which showed the effectiveness of our pseudo-haptic technique as well as the general appreciation of the Elastic-Arm. We believe that the Elastic-Arm could be used in various VR applications which call for mobile haptic feedback or human-scale haptic sensations.

Vision-based adaptive assistance and haptic guidance for safe wheelchair corridor following

Participants: Maud Marchal

In case of motor impairments, steering a wheelchair can become a hazardous task. Joystick jerks induced by uncontrolled motions may lead to wall collisions when a user steers a wheelchair along a corridor. In [7] we introduce a low-cost assistive and guidance system for indoor corridor navigation in a wheelchair, which uses purely visual information, and which is capable of providing automatic trajectory correction and haptic guidance in order to avoid wall collisions. A visual servoing approach to autonomous corridor following

serves as the backbone to this system. The algorithm employs natural image features which can be robustly extracted in real time. This algorithm is then fused with manual joystick input from the user so that progressive assistance and trajectory correction can be activated as soon as the user is in danger of collision. A force feedback in conjunction with the assistance is provided on the joystick in order to guide the user out of his dangerous trajectory. This ensures intuitive guidance and minimal interference from the trajectory correction system. In addition to being a low-cost approach, it can be seen that the proposed solution does not require an a-priori environment model. Experiments on a robotised wheelchair equipped with a monocular camera prove the capability of the system to adaptively guide and assist a user navigating in a corridor.

This work was done in collaboration with LAGADIC team.

7.2.3. Tactile Interaction at Fingertips

The fingertips are one of the most important and sensitive parts of our body. They are the first stimulated areas of the hand when we interact with our environment. Providing haptic feedback to the fingertips in virtual reality could, thus, drastically improve perception and interaction with virtual environments. Within this context, we proposed two contributions for tactile feedback and haptic interaction at the fingertips.

The Haptip

Participants: Adrien Girard, Yoren Gaffary, Anatole Lécuyer and Maud Marchal

In [5], we present a modular approach called HapTip to display such haptic sensations at the level of the fingertips. This approach relies on a wearable and compact haptic device able to simulate 2 Degree of Freedom (DoF) shear forces on the fingertip with a displacement range of 2 mm. Several modules can be added and used jointly in order to address multi-finger and/or bimanual scenarios in virtual environments. For that purpose, we introduce several haptic rendering techniques to cover different cases of 3D interaction, such as touching a rough virtual surface, or feeling the inertia or weight of a virtual object. In order to illustrate the possibilities offered by HapTip, we provide four use cases focused on touching or grasping virtual objects (see Figure 10). To validate the efficiency of our approach, we also conducted experiments to assess the tactile perception obtained with HapTip. Our results show that participants can successfully discriminate the directions of the 2 DoF stimulation of our haptic device. We found also that participants could well perceive different weights of virtual objects simulated using two HapTip devices. We believe that HapTip could be used in numerous applications in virtual reality for which 3D manipulation and tactile sensations are often crucial, such as in virtual prototyping or virtual training.

Figure 10. Illustrative use cases of our approach HapTip: the user can get in contact and tap a virtual bottle, touch a surface and feel its texture, and heft an object and feel its weight.

This work was done in collaboration with CEA List.

Studying one and two-finger perception of tactile directional cues

Participants: Yoren Gaffary, Anatole Lécuyer and Maud Marchal

In [20], we study the perception of tactile directional cues by one or two fingers, using either the index, middle, or ring finger, or any of their combination. Therefore, we use tactile devices able to stretch the skin of the fingertips in 2 DOF along four directions: horizontal, vertical, and the two diagonals. We measure the recognition rate in each direction, as well as the subjective preference, depending on the (couple of) finger(s) stimulated (see Figure 11). Our results show first that using the index and/or middle finger performs significantly better than using the ring finger on both qualitative and quantitative measures. The results when comparing one versus two-finger configurations are more contrasted. The recognition rate of the diagonals is higher when using one finger than two, whereas two fingers enable a better perception of the horizontal direction. These results pave the way to other studies on one versus two-finger perception, and raise methodological considerations for the design of multi-finger tactile devices.

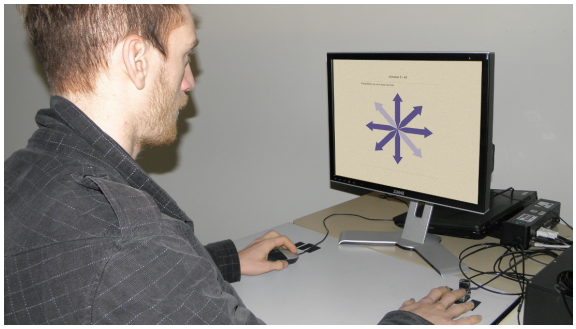


Figure 11. Experimental setup of our study on the perception of tactile directional cues by one or two fingers: the participant reports the direction of the stimulus he just perceived on the fingertip of his middle finger.

This work was done in collaboration with CEA List, IRMAR and Agrocampus Ouest.

7.3. Collaborative Virtual Environments

7.3.1. Acting in Collaborative Virtual Environments

VR Rehearsals for Acting with Visual Effects

Participants: Rozenn Bouville, Valérie Gouranton and Bruno Arnaldi,

We studied the use of Virtual Reality for movie actors rehearsals of VFX-enhanced scenes. The impediment behind VFX scenes is that actors must be filmed in front of monochromatic green or blue screens with hardly any cue to the digital scenery that is supposed to surround them. The problem is worsens when the scene includes interaction with digital partners. The actors must pretend they are sharing the set with imaginary creatures when they are, in fact, on their own on an empty set. To support actors in this complicated task, we introduced the use of VR for acting rehearsals not only to immerse actors in the digital scenery but to provide them with advanced features for rehearsing their play. Indeed, our approach combines a fully interactive environment with a dynamic scenario feature to allow actors to become familiar with the virtual elements while rehearsing dialogue and action at their own speed. The interactive and creative rehearsals enabled by the system can be either single-user or multiuser. Moreover, thanks to the wide range of supported platforms, VR rehearsals can take place either onset or offset. We conducted a preliminary study to assess whether VR training can replace classical training (see Figure 12). The results show that VR-trained actors deliver a performance just as good as ordinarily trained actors. Moreover, all the subjects in our experiment preferred VR training to classic training [17].

Synthesis and Simulation of Collaborative Surgical Process Models

Figure 12. The use of VR for acting rehearsal enables actors to rehearse being immersed in the virtual scenery before being shot on a green and empty set.

Participants: Guillaume Claude, Valérie Gouranton and Bruno Arnaldi

The use of Virtual Reality for surgical training has been mostly focused on technical surgical skills. We proposed a novel approach by focusing on the procedural aspects [4]. Our system relies on a specific workflow, which enbables to generate a model of the procedure based on real case surgery observations made in the operating room (see Figure 13). In addition, in the context of the project S3PM we then proposed an innovative workflow to integrate the generic model of the procedure (generated from the real-case surgery observation) as a scenario model in the VR training system (see Figure 14). We described how the generic procedure model could be generated, as well as its integration in the virtual environment [18].

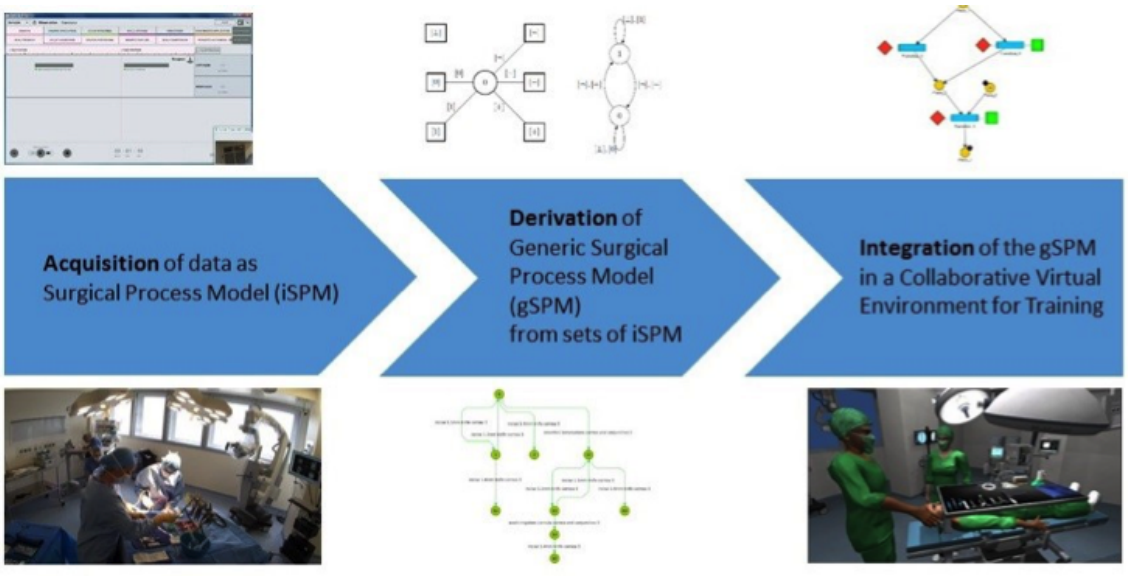


Figure 13. Collaborative Virtual Environments for Training in Surgical Procedures, based on observations during real surgeries. Observation data is integrated into a system providing a Generalised Surgical Process Model (gSPM) of the procedure. This Model is integrated as the scenario of the Virtual Environment.

This work was done in collaboration with HYCOMES team and LTSI Inserm Medicis.

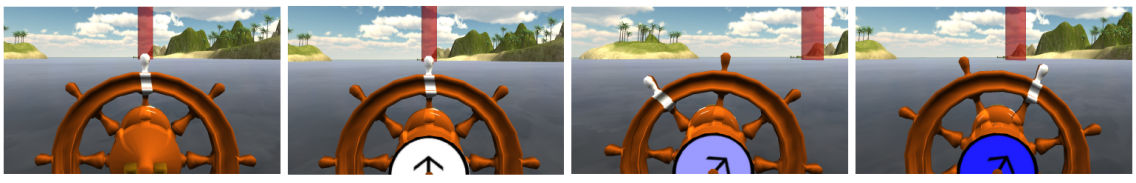
Figure 14. Virtual replica of a real operating room of Rennes hospital (CHU Rennes) in the Immersia CAVE-like setup (IRISA/Inria Rennes).

7.3.2. Awareness for Collaboration in Virtual Environments

Take-Over Control Paradigms in Collaborative Virtual Environments for Training

Participants: Gwendal Le Moulec, Ferran Argelaguet, Anatole Lécuyer and Valérie Gouranton

We studied the notion of Take-Over Control in Collaborative Virtual Environments for Training (CVET). The Take-Over Control represents the transfer (the take over) of the interaction control of an object between two or more users. This paradigm is particularly useful for training scenarios, in which the interaction control could be continuously exchanged between the trainee and the trainer, e.g. the latter guiding and correcting the trainee's actions. We proposed a formalization of the Take-Over Control followed by an illustration focusing in a use-case of collaborative maritime navigation. In the presented use-case, the trainee has to avoid an under-water obstacle with the help of a trainer who has additional information about the obstacle. The use-case allows to highlight the different elements a Take-Over Control situation should enforce, such as user's awareness. Different Take-Over Control techniques were provided and evaluated focusing on the transfer exchange mechanism and the visual feedback (see Figure 15). The results show that participants preferred the Take-Over Control technique which maximized the user awareness [24].



(a) *Initial state*
No feedback condition

(b) *Initial state*
$w_1 = 1,\ w_2 = 0$

(c) *In progress*
$w_1 = 0.5,\ w_2 = 0.5$

(d) *Completed*
$w_1 = 0,\ w_2 = 1$

Figure 15. Our illustrative use case inspired by maritime navigation for Take-Over Control during training in a collaborative virtual environment. The user was instructed to steer a boat towards a semi-transparent red column (target destination) by controlling the heading of the boat. A white handle indicated the rotation angle of the boat(a). The sequence (b,c,d) shows the evolution of the contribution of the trainer on the steering angle, from no control to full control.

Vishnu: Virtual Immersive Support for HelpiNg Users: An Interaction Paradigm for Collaborative Remote Guiding in Mixed Reality

Participants: Morgan Le Chénéchal, Valérie Gouranton and Bruno Arnaldi

Increasing networking performances as well as the emergence of Mixed Reality (MR) technologies make possible providing advanced interfaces to improve remote collaboration. We presented a novel interaction paradigm called Vishnu that aims to ease collaborative remote guiding. We focus on collaborative remote maintenance as an illustrative use case. It relies on an expert immersed in Virtual Reality (VR) in the remote workspace of a local agent helped through an Augmented Reality (AR) interface. The main idea of the Vishnu paradigm is to provide the local agent with two additional virtual arms controlled by the remote expert who can use them as interactive guidance tools. Many challenges come with this: collocation, inverse kinematics (IK), the perception of the remote collaborator and gestures coordination. Vishnu aims to enhance the maintenance procedure thanks to a remote expert who can show to the local agent the exact gestures and actions to perform (see Figure 16). Our pilot user study shows that it may decrease the cognitive load compared to a usual approach based on the mapping of 2D and de-localized informations, and it could be used by agents in order to perform specific procedures without needing to have an available local expert [22].

This work was done in collaboration with b<>com and Telecom Bretagne.

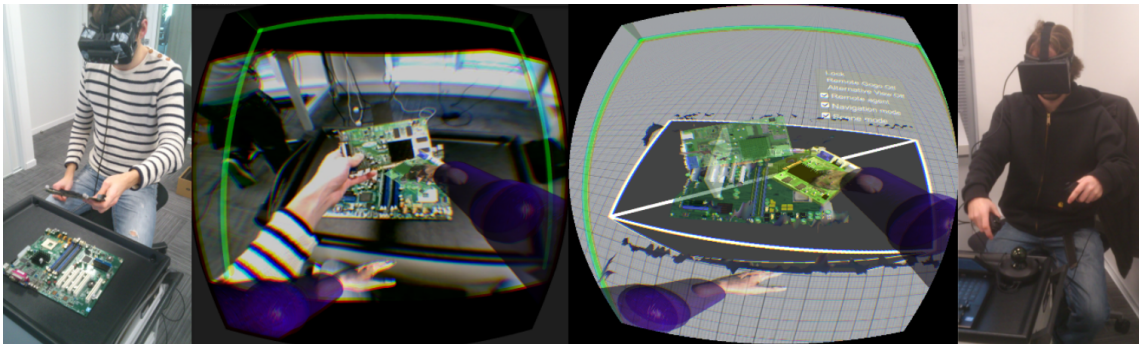


Figure 16. Illustration of the Vishnu approach: system and viewpoints of the agent (left) and the expert (right) in a motherboard assembly scenario.

When the Giant meets the Ant: An Asymmetric Approach for Collaborative and Concurrent Object Manipulation in a Multi-Scale Environment

Participants: Morgan Le Chénéchal, Jérémy Lacoche, Valérie Gouranton and Bruno Arnaldi

We proposed a novel approach to enable two or more users to manipulate an object collaboratively. Our goal is to benefits from the wide variety of todays VR devices. Our solution is based on an asymmetric collaboration pattern at different scales in which users benefit from suited points of views and interaction techniques according to their device setups. Each user application is adapted thanks to plasticity mechanisms. Our system provides an efficient way to co-manipulate an object within irregular and narrow courses, taking advantages of asymmetric roles in synchronous collaboration (see Figure 17). Moreover, it aims to provide a way to maximize the filling of the courses while the object moves on its path [23],[35].

This work was done in collaboration with b<>com and Telecom Bretagne.

7.4. Brain-Computer Interfaces

7.4.1. Contribution to a Reference Book on BCI

We have largely contributed to a reference book on BCI released in 2016 in French and English, and co-edited by Fabien Lotte, Maureen Clerc and Laurent Bougrain for ISTE (French version [36] [37]) and Wiley

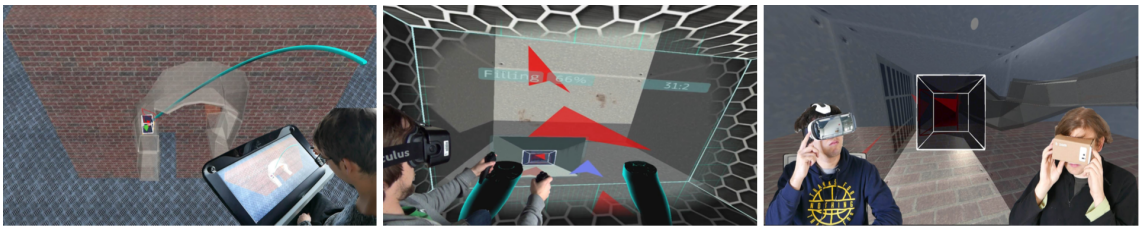


Figure 17. When the Giant meets the Ant: Collaborative manipulation of a virtual object (here, a cube) based on an asymmetric setting between two users who can be helped by two additional users. (a) The first participant has a global view of the scene and moves the object with a 3D bent ray. (b) The second user is placed inside the object and precisely rotates and scales it. (c) Two additional roles can be added. The first one helps to scale the object using a third person view of it. The other one is a spectator who switches between the other participants' viewpoints and helps them with oral communication.

(English version [39] [40]) publishers. This book provides keys for understanding and designing these multi-disciplinary interfaces, which require many fields of expertise such as neuroscience, statistics, informatics and psychology. This work corresponds to four different book chapters, all published in both French and English, which are presented hereafter.

Book chapter on BCI and videogames

Participants: Anatole Lécuyer

Videos games are often cited as a very promising field of applications for brain-computer interfaces. In a first chapter [30] [31], we described state of the art in the field of video games played "with the mind". In particular, we considered the results of the OpenViBE2 project: one of the most important research projects in this area. We presented a selection of prototypes developed during this OpenViBE2 project which is illustrative of the state of the art in this field and of the use of BCIs in video games, such as based on imagining a motion of the left and right hands to score goals, or in another example using the P300 cerebral potential to destroy spaceships in a remake of well-known Japanese game.

Book chapter on BCI softwares

Participants: Jussi Lindgren and Anatole Lécuyer

In a second chapter [28] [29], we described OpenViBE and other software platforms used to study the subject. The chapter gave an overview of such platforms. We described how the software components of the platforms reflect typical signal acquisition and signal processing stages used in BCI. Finally, we presented a high-level account of differences between major BCI platforms and gave a few pieces of advice and recommendation regarding BCI platform selection.

Book chapter on BCI and HCI

Participants: Andéol Evain, Ferran Argelaguet and Anatole Lécuyer

In a third chapter [34], we focused on the link between BCI and Human-Computer Interaction (HCI), and studied how HCI concepts can apply to BCIs. First, we presented an overview of the main concepts of HCI. We then studied the main characteristics of BCIs related to these concepts. This chapter also discussed the choice of cerebral patterns to use, depending on the interaction task and the use context. Finally, we presented the most promising new interaction paradigms for interaction with BCIs.

This work was done in collaboration with MJOLNIR team.

Book chapter on Neurofeedback

Participants: Lorraine Perronnet and Anatole Lécuyer

We proposed a fourth chapter called Brain training with Neurofeedback [33] [32]. We first defined the concept of neurofeedback (NF) and gave an overall view of the current status in this domain. Then we described the design of a NF training program and the typical course of a NF session, as well as the learning mechanisms underlying NF. We retraced the history of NF, explaining the origin of its questionable reputation and providing a foothold for understanding the diversity of existing approaches. We also discussed how the fields of NF and BCIs might potentially overlap in future with the development of "restorative" BCIs. Finally, we presented a few applications of NF and summarized the state of research of some of its major clinical applications.

This work was done in collaboration with VISAGES team.

7.4.2. BCI Methods and Techniques

Do the Stimuli of a BCI Have to be the Same as the Ones Used for Training it?

Participants: Andéol Evain, Ferran Argelaguet and Anatole Lécuyer

Does the stimulation used during the training on an SSVEP-based BCI have to be similar to that of the end use? We conducted an experiment in which we recorded six-channel EEG data from 12 subjects in various conditions of distance between targets , and of difference in color between targets [10]. Our analysis revealed that the stimulation configuration used for training which leads to the best classification accuracy is not always the one which is closest to the end use configuration. We found that the distance between targets during training is of little influence if the end use targets are close to each other, but that training at far distance can lead to a better accuracy for far distance end use. Additionally, an interaction effect is observed between training and testing color: while training with monochrome targets leads to good performance only when the test context involves monochrome targets as well, a classifier trained on colored targets can be efficient for both colored and monochrome targets. In a nutshell, in the context of SSVEP-based BCI, training using distant targets of different colors seems to lead to the best and more robust performance in all end use contexts.

This work was done in collaboration with MJOLNIR team.

A Novel Fusion Approach Combining Brain and Gaze Inputs for Target Selection

Participants: Andéol Evain, Ferran Argelaguet and Anatole Lécuyer

Gaze-based interfaces and Brain-Computer Interfaces (BCIs) allow for hands-free human-computer interaction. We investigated the combination of gaze and BCIs. We proposed a novel selection technique for 2D target acquisition based on input fusion [9]. This new approach combines the probabilistic models for each input, in order to better estimate the intent of the user. We evaluated its performance against the existing gaze and brain-computer interaction techniques. Twelve participants took part in our study, in which they had to search and select 2D targets with each of the evaluated techniques (see Figure 18). Our fusion-based hybrid interaction technique was found to be more reliable than the previous gaze and BCI hybrid interaction techniques for 10 participants over 12, while being 29% faster on average. However, similarly to what has been observed in hybrid gaze-and-speech interaction, gaze-only interaction technique still provides the best performance. Our results should encourage the use of input fusion, as opposed to sequential interaction, in order to design better hybrid interfaces.

This work was done in collaboration with MJOLNIR team.

7.4.3. BCI User Experience and Neurofeedback

Influence of Error Rate on Frustration of BCI Users

Participants: Andéol Evain, Ferran Argelaguet and Anatole Lécuyer

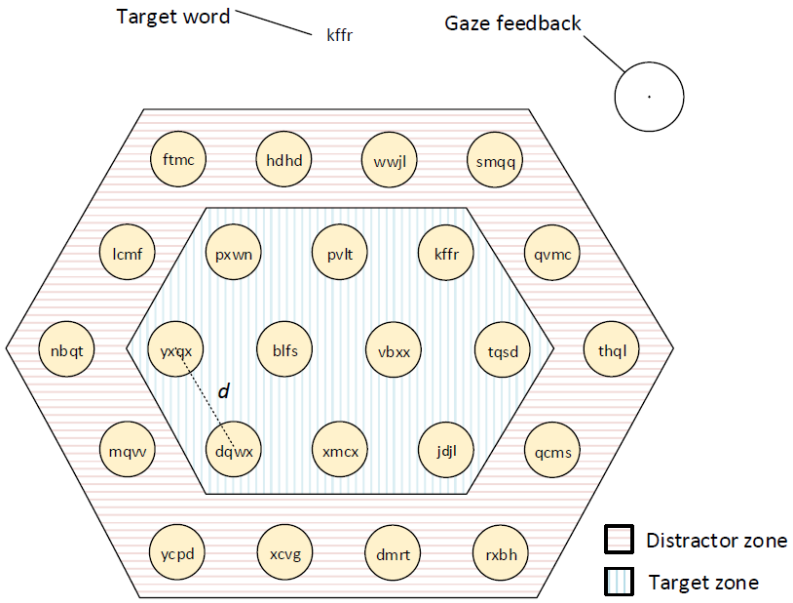


Figure 18. Experimental task combining gaze and brain inputs. The user has to look for the goal word displayed at the top of the screen, then, the user has to select the target with the exact same word. The detected gaze position is displayed under the form of a circle and a central point (visual feedback). For all trials the size of the targets remained constant, and only the length of the target word and the separation (d) between targets varied. The targets at the outer circle were distractors in which the target word was never placed.

Brain-Computer Interfaces (BCIs) are still much less reliable than other input devices. The error rates of BCIs range from 5% up to 60%. We assessed the subjective frustration, motivation, and fatigue of BCI users, when confronted to different levels of error rate [27]. We conducted a BCI experiment in which the error rate was artificially controlled (see Figure 19). Our results first show that a prolonged use of BCI significantly increases the perceived fatigue, and induces a drop in motivation. We also found that user frustration increases with the error rate of the system but this increase does not seem critical for small differences of error rate. Thus, for future BCIs, we advise to favor user comfort over accuracy when the potential gain of accuracy remains small.

This work was done in collaboration with MJOLNIR team.

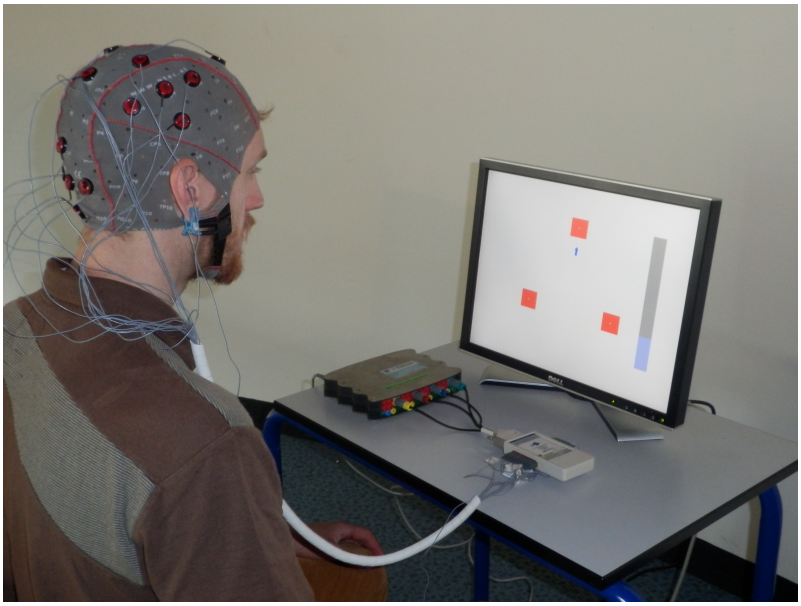


Figure 19. Experimental setup: a participant is using an SSVEP-based BCI which error rate is artificially controlled.

Design of an Experimental Platform for Hybrid EEG-fMRI Neurofeedback Studies

Participants: Marsel Mano, Lorraine Perronnet and Anatole Lécuyer

During a neurofeedback (NF) experiment one or more brain activity measuring technologies are used to estimate the changes of the acquired neural signals that reflect the changes of the subject's brain activity in real-time. There exist a variety of NF research applications that use only one type of neural signals (i.e. uni-modal) like EEG or fMRI, but there are very few NF researches that use two or more neural signals (i.e. multi-modal). We have developed a hybrid EEG-fMRI platform for bi-modal NF experiments, as part of the project Hemisfer. Our system is based on the integration and the synchronization of an MR-compatible EEG and fMRI acquisition subsystems. The EEG signals are acquired with a 64 channel MR-compatible solution from Brain Products and the MR imaging is performed on a 3T Verio Siemens scanner (VB17) with a 12-ch head coil. We have developed two real-time pipelines for EEG and fMRI that handle all the necessary signal processing, the Joint NF module that calculates and fuses the NF and a visualize module that displays the NF to the subject. The control and the synchronization of both subsystems with each-other and with the experimental protocol is handled by the NF Control. Our platform showed very good real-time performance with various pre-processing, filtering, and NF estimation and visualization methods. The entire fMRI process

from acquisition to NF takes always less than 200ms, well below the TR of regular EPI sequences (2s). The same process for EEG, with NF update cycles varying 2-5Hz, is done in virtually real time (50Hz).

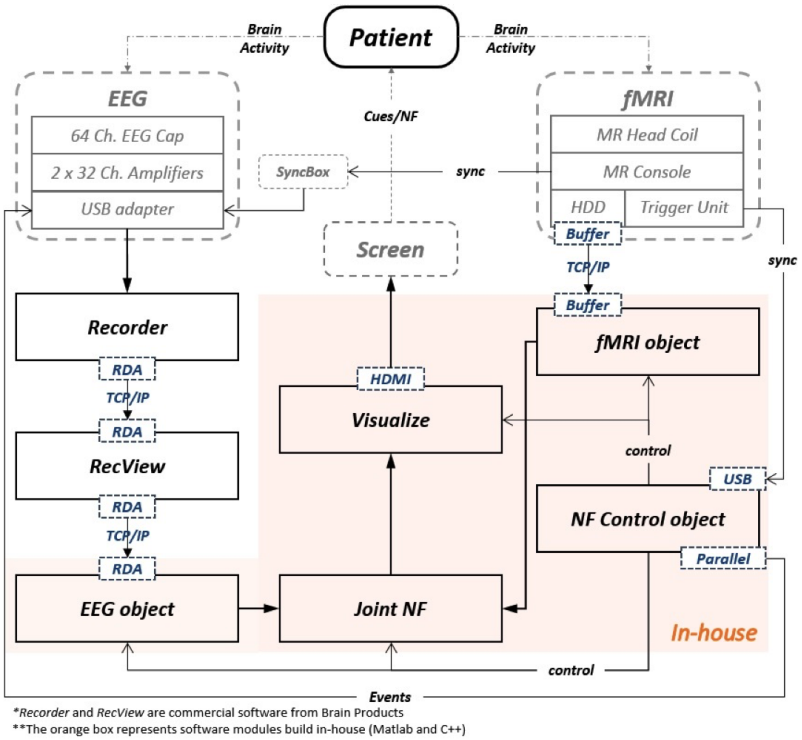


Figure 20. Architecture of our hybrid EEG-fMRI neurofeedback platform.

This work was done in collaboration with VISAGES team and presented as poster at OHBM 2016.

Unimodal versus Bimodal EEG-fMRI Neurofeedback

Participants: Lorraine Perronnet, Anatole Lécuyer and Marsel Mano

In the context of the HEMISFER project, we proposed a simultaneous EEG-fMRI experimental protocol in which 10 healthy participants performed a motor-imagery task in unimodal and bimodal neurofeedback conditions. With this protocol we were able to compare for the first time the effects of unimodal EEG-neurofeedback and fMRI-neurofeedback versus bimodal EEG-fMRI-neurofeedback by looking both at EEG and fMRI activations. We also introduced a new feedback metaphor for bimodal EEG-fMRI-neurofeedback that integrates both EEG and fMRI signal in a single bi-dimensional feedback (a ball moving in 2D). Such a feedback is intended to relieve the cognitive load of the subject by presenting the bimodal neurofeedback task as a single regulation task instead of two. Additionally, this integrated feedback metaphor gives flexibility on defining a bimodal neurofeedback target. Participants were able to regulate activity in their motor regions in all neurofeedback conditions. Moreover, motor activations as revealed by offline fMRI analysis were stronger during EEG-fMRI-neurofeedback than during EEG-neurofeedback. This result suggests that EEG-fMRI-neurofeedback could be more specific or more engaging than EEG-neurofeedback. Our results also suggest that during EEG-fMRI-neurofeedback, participants tended to regulate more the modality that was harder to control. Taken together our results shed light on the specific mechanisms of bimodal EEG-fMRI-neurofeedback and on its added-value as compared to unimodal EEG-neurofeedback and fMRI-neurofeedback.

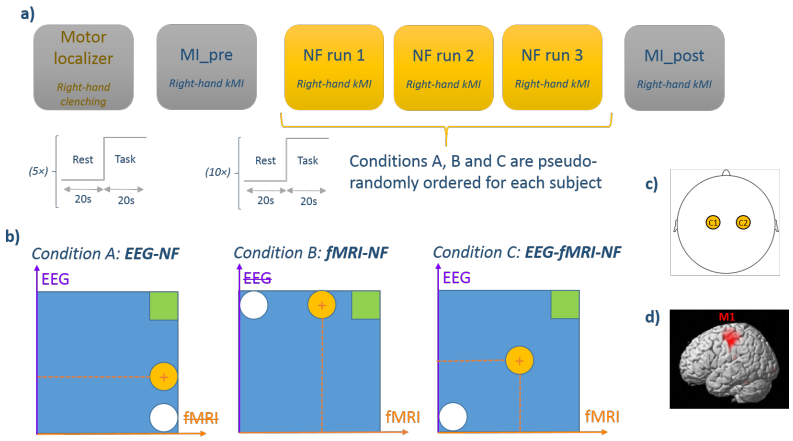


Figure 21. Experimental procedure for comparing unimodal versus bimodal EEG-fMRI neurofeedback.

This work was done in collaboration with VISAGES team and presented as poster at OHBM 2016. Experiments were conducted at NEURINFO platform from University of Rennes 1.

HYCOMES Project-Team

7. New Results

7.1. Structural Analysis of Multi-Mode DAEs

Differential Algebraic Equation (DAE) systems constitute the mathematical model supporting physical modeling languages such as Modelica or Simscape. Unlike Ordinary Differential Equations, or ODEs, they exhibit subtle issues because of their implicit *latent equations* and related *differentiation index*. Multi-mode DAE (mDAE) systems are much harder to deal with, not only because of their mode-dependent dynamics, but essentially because of the events and resets occurring at mode transitions. Unfortunately, the large literature devoted to the numerical analysis of DAEs do not cover the multi-mode case. It typically says nothing about mode changes. This lack of foundations cause numerous difficulties to the existing modeling tools. Some models are well handled, others are not, with no clear boundary between the two classes. In [11], we develop a comprehensive mathematical approach to the *structural analysis* of mDAE systems which properly extends the usual analysis of DAE systems. We define a constructive semantics based on nonstandard analysis and show how to produce execution schemes in a systematic way. This work has been accepted for presentation at the HSCC 2017 conference [19] in April 2017.

7.2. Decoupling Abstractions

In [10], we investigated decoupling abstractions, by which we seek to simulate (i.e. abstract) a given system of ordinary differential equations (ODEs) by another system that features completely independent (i.e. uncoupled) sub-systems, which can be considered as separate systems in their own right. Beyond a purely mathematical interest as a tool for the qualitative analysis of ODEs, decoupling can be applied to verification problems arising in the fields of control and hybrid systems. Existing verification technology often scales poorly with dimension. Thus, reducing a verification problem to a number of independent verification problems for systems of smaller dimension may enable one to prove properties that are otherwise seen as too difficult. We show an interesting correspondence between Darboux polynomials and decoupling simulating abstractions of systems of polynomial ODEs and give a constructive procedure for automatically computing the latter.

7.3. Formal Verification of the ACAS X System

The *Next-Generation Airborne Collision Avoidance System* (ACAS X) is intended to be installed on all large aircraft to give advice to pilots and prevent mid-air collisions with other aircraft. It is currently being developed by the Federal Aviation Administration (FAA). In [6], we determine the geometric configurations under which the advice given by ACAS X is safe under a precise set of assumptions and formally verify these configurations using hybrid systems theorem proving techniques. We consider subsequent advisories and show how to adapt our formal verification to take them into account. We examine the current version of the real ACAS X system and discuss some cases where our safety theorem conflicts with the actual advisory given by that version, demonstrating how formal hybrid systems proving approaches are helping to ensure the safety of ACAS X. Our approach is general and could also be used to identify unsafe advice issued by other collision avoidance systems or confirm their safety.

7.4. Chattering-Free Simulation

Chattering is a fundamental phenomenon that is unique to hybrid systems, due to the complex interaction between discrete dynamics (in the form of discrete transitions) and continuous dynamics (in the form of time). In practice, simulating chattering hybrid systems is challenging in that simulation effectively halts near the chattering time point, as an infinite number of discrete transitions would need to be simulated. In [7],

formal conditions are provided for when the simulated models of hybrid systems display chattering behavior, and methods are proposed for avoiding chattering "on the fly" in runtime. We utilize dynamical behavior analysis to derive conditions for detecting chattering without enumeration of modes. We also present a new iterative algorithm to allow for solutions to be carried past the chattering point, and we show by a prototypical implementation how to generate the equivalent chattering-free dynamics internally by the simulator in the main simulation loop.

I4S Project-Team

7. New Results

7.1. Outdoor InfraRed Thermography

7.1.1. Autonomous software architecture standardized for infrared and environmental SHM : Cloud2IR

Participants: Antoine Crinière, Jean Dumoulin, Laurent Mevel.

Cloud2IR is an autonomous software architecture, allowing multi-sensor connection (i.e. Infrared Thermography), dedicated to the long term monitoring of infrastructures. Past experimentations have shown the need as well as usefulness of such system. The system has been developed in order to cut down software integration time which facilitates the system adaptation to each experiment specificity. That is why we propose a bi-headed architecture. A specialized part, it represents the sensor specific development as well as their drivers and their different fixed configurations. In our case, as infrared camera are slightly different than other kind of sensors, the system implement in addition an RTSP server which can be used to set up the FOV as well as other measurement parameter considerations and a generic part, which can be seen as the data management side. This last part can be seen as the first embryo of a future generic framework dedicated to the data management of local multisensors (DaMaLoS). It is able to aggregate any sensor data, type or size and automatically encapsulate them in various generic data format as HDF5 or cloud data as OGC SWE standard. This whole part is also responsible of the acquisition scenario the local storage management and the network management through SFTP or SOAP for OGC Web services. Cloud2IR has been deployed on field since more than one year at the SenseCity outdoor test bed and several month at the Inria test bed, both located in France. The system aggregates various sensors as infrared camera, a GPS, multiple pyranometers, a weather station and a proprietary access to the SenseCity data viewer.[40][41]

7.1.2. GPU Improved quantitative analysis of Longterm Infrared-Thermography Data

Participants: Antoine Crinière, Jean Dumoulin, Laurent Mevel.

Since the past decade, infrared thermography coupled with inverse models based on 1d thermal quadrupoles have shown their usefulness in civil engineering by first showing their ability to assess the quantitative non destructive testing of concrete repaired by bonded CFRP plate over a wide area (i.e. repaired or reinforced concrete beams). On the other hand early implementations of long terms monitoring methods based on such approach have given their first results over a whole bridge deck. The experimental method, allow us to have the apparent surface temperature field evolution with time for a wide area divided in pixels. Knowing this specificity, the procedure aims to apply an independent model to each pixel in order to retrieve physical properties map. Such treatment can have a high computational cost. We propose various improvement of our procedure based on GPGPU paradigm in order to shorten the computational time. This study will detail an experimental procedure able to assess the long term thermal monitoring of a bridge deck over days and to draw properties maps of the inner structure. [28]

7.1.3. Infrared thermography for cultural heritage monitoring

Participant: Jean Dumoulin.

Radiation theory helps us to introduce infrared thermography. Infrared thermography is first presented in its passive mode and followed by considerations on active mode. Some processing analysis approaches are described. They belong to signal and image processing domain or to heat transfer domain. Illustration of results obtained with such analysis approaches are described on two experiments carried out in quasi laboratory conditions. Then, a case case study of the monitoring of the Viaduct Basento in Potenza (Southern Italy) is presented. Two features make fascinating this case study. The first one regards the fact that Viaduct Basento is probably the most important and visionary architectural work of the famous structural engineer Sergio Musmeci. The second aspect concerns the application, almost unique in the scientific literature, of an integrated diagnosis approaches combining a wide set of electromagnetic sensing technologies combined with advanced civil engineering analysis methodologies and tools.[44] [45] [22] [23]

7.2. Smarts roads and R5G

7.2.1. Positive surface temperature pavement

Participants: Jean Dumoulin, Nicolas Le Touz.

The mobility during winter season in France mainly relies on the use of de-icers, with an amount ranging from two hundreds thousands tons up to two millions tons for the roads only. Besides the economic impact, there are many concerns on their environmental and infrastructure, both on roads and on airports. In such context and in the framework of the R5G (5th Generation Road) project driven by IFSTTAR, investigations were carried out on the way to modify the infrastructure to maintain pavement surface at a temperature above water freezing point. Two distinct approaches, that can could be combined, were selected. The first one consisted in having a heated fluid circulating in a porous layer within an asphalt concrete pavement sample. The second one specifically relied on the use of paraffin phase change materials (PCM) in cement concrete pavement ones. Experiments on enhanced pavement samples were conducted in a climatic chamber to simulate winter conditions for several continuous days, including wind and precipitations, and monitored by infrared thermography. [24]

7.3. Methods for building performance assessment

7.3.1. Building performance assessment

Participants: Jordan Brouns, Jean Dumoulin, Alexandre Nassiopoulos, Nicolas Le Touz.

Accurate building performance assessment is necessary for the design of efficient energy retrofit operations and to foster the development of energy performance contracts. An important barrier however is that simulation tools fail to accurately predict the actual energy consumption. Two methodology are adressed, first combining thermal sensor output and inverse algorithms to determine the key parameters of a multizone thermal model [15] then assessing wall thermal resistance estimation using infrared thermography and microwave coupling [38][34][43]

7.4. System identification

7.4.1. Variance estimation of modal parameters from subspace-based system identification

Participants: Michael Doehler, Laurent Mevel.

This work has been carried out in collaboration with Philippe Mellinger (former PhD student with Dassault Aviation, now CEA).

An important step in the operational modal analysis of a structure is to infer on its dynamic behavior through its modal parameters. When output-only data is available, i.e. measured responses of the structure, frequencies, damping ratios and mode shapes can be identified assuming that ambient sources like wind or traffic excite the system sufficiently. When also input data is available, i.e. signals used to excite the structure, input/output identification algorithms are used. The use of input information usually provides better modal estimates in a desired frequency range. When identifying the modal parameters from noisy measurement data, the information on their uncertainty is most relevant. In this work, new variance computation schemes for modal parameters are developed for four subspace algorithms, including output-only and input/output methods, as well as data-driven and covariance-driven methods. For the input/output methods, the known inputs are considered as realizations of a stochastic process. Based on Monte Carlo validations, the quality of identification, accuracy of variance estimations and sensor noise robustness are discussed. Finally these algorithms are applied on real measured data obtained during vibrations tests of an aircraft. [19] [37]

7.4.2. Bayesian parameter estimation for parameter varying systems using interacting Kalman filters

Participants: Antoine Crinière, Laurent Mevel, Jean Dumoulin.

Method based on the use of Bayesian modal parameter recursive estimation based on a particular Kalman filter algorithm with decoupled distributions for mass and stiffness. Particular Kalman filtering is a combination of two widely used Bayesian estimation methods working together: the particle filter (also called sequential Monte Carlo samplings) and the Kalman filter. Usual system identification techniques for civil and mechanical structures assume the availability of large set of data derived from a stationary quasi steady structure. On the opposite, several scenarios involve time varying structures. For example, due to interaction with aerodynamics in aeronautics, some critical parameter may have to be monitored, for instability monitoring (leading possibly to flutter) of in flight data due to fuel consumption and speed change. This relates to the monitoring of time varying structural parameters such as frequencies and damping ratios. The main idea of a particular Kalman filter is to consider stochastic particles evolving in the parameter space. For each particle, a corresponding linear state is recursively estimated by applying a Kalman filter to the mechanical system, whose modal parameters are driven by the evolution of this time-varying particle. In order to provide fast and convincing results for large time varying structure, such as an airplane, the execution time of the method has to be improved. Within the Cloud2sm ADT a GPGPU implementation of the algorithm have been developed, now a post-doctoral position have been obtained to improve the algorithm reliability.[29]

7.4.3. Stability of the Kalman filter for continuous time output error systems

Participant: Qinghua Zhang.

This work has been carried out in collaboration with Boyi Ni (SAP Labs China).

The stability of the Kalman filter is usually ensured by the uniform complete controllability *regarding the process noise* and the uniform complete observability of linear time varying systems. This work studies the case of continuous time *output error* systems, in which the process noise is totally absent. The classical stability analysis assuming the controllability regarding the process noise is thus not applicable. It is shown in this work that the uniform complete observability *alone* is sufficient to ensure the asymptotic stability of the Kalman filter applied to time varying *output error* systems, regardless of the stability of the considered systems themselves. The exponential or polynomial convergence of the Kalman filter is then further analyzed for particular cases of stable or unstable output error systems. The results of this work have been published in [20].

7.4.4. Parameter uncertainties quantification for finite element based subspace fitting approaches

Participants: Guillaume Gautier, Laurent Mevel, Michael Doehler.

This work has been carried out in collaboration with Jean-Mathieu Mencik and Roger Serra (INSA Centre Val de Loire).

We address the issue of quantifying uncertainty bounds when updating the finite element model of a mechanical structure from measurement data. The problem arises as to assess the validity of the parameters identification and the accuracy of the results obtained. A covariance estimation procedure is proposed about the updated parameters of a finite element model, which propagates the data-related covariance to the parameters by considering a first-order sensitivity analysis. In particular, this propagation is performed through each iteration step of the updating minimization problem, by taking into account the covariance between the updated parameters and the data-related quantities. Numerical simulations on a beam show the feasibility and the effectiveness of the method. [31]

7.4.5. Embedded subspace-based modal analysis and uncertainty quantification

Participants: Vincent Le Cam, Michael Doehler, Mathieu Le Pen, Ivan Guéguen, Laurent Mevel.

Operational modal analysis is an important step in many methods for vibration-based structural health monitoring. These methods provide the modal parameters (frequencies, damping ratios and mode shapes) of the structure and can be used for monitoring over time. For a continuous monitoring the excitation of a structure is usually ambient, thus unknown and assumed to be noise. Hence, all estimates from the vibration measurements are realizations of random variables with inherent uncertainty due to unknown excitation, measurement noise and finite data length. Estimating the standard deviation of the modal parameters on the same dataset offers significant information on the accuracy and reliability of the modal parameter estimates. However, computational and memory usage of such algorithms are heavy even on standard PC systems in Matlab, where reasonable computational power is provided. In this work, we examine an implementation of the covariance-driven stochastic subspace identification on the wireless sensor platform PEGASE, where computational power and memory are limited. Special care is taken for computational efficiency and low memory usage for an on-board implementation, where all numerical operations are optimized. The approach is validated from an engineering point of view in all its steps, using simulations and field data from a highway road sign structure. [33]

7.5. Damage diagnosis

7.5.1. Estimation of distributed and lumped ohmic losses in electrical cables

Participants: Nassif Berrabah, Qinghua Zhang.

This work has been carried out in the framework of a CIFRE PhD project in collaboration with EDF R&D.

Cables play an important role in modern engineering systems, from power transmission to data communication. In order to ensure reliable and cost-efficient operations, as well as a high level of performance, efficient tools are needed to assess and monitor cables. Hard faults are well handled by existing techniques, whereas soft fault diagnosis still represents an important challenge for current researches. This work focuses on the detection, localization, and estimation of resistive soft fault in electrical cables from reflectometry measurements. A method for the computation of the distributed resistance profile along the cable under test has been developed. Both experimental and simulation results confirm its effectiveness, as reported in the conference paper [26]. A patent based on this work has been registered at INPI (see Section 10.1.4.1).

7.5.2. Fault detection, isolation and quantification from Gaussian residuals

Participants: Michael Doehler, Laurent Mevel, Qinghua Zhang.

Despite the general acknowledgment in the Fault Detection and Isolation (FDI) literature that FDI are typically accomplished in two steps, namely residual generation and residual evaluation, the second step is by far less studied than the first one. This work investigates the residual evaluation method based on the local approach to change detection and on statistical tests. The local approach has the remarkable ability of transforming quite general residuals with unknown or non Gaussian probability distributions into a standard Gaussian framework, thanks to a central limit theorem. In this work, the ability of the local approach for fault quantification is exhibited, whereas previously it was only presented for fault detection and isolation. The numerical computation of statistical tests in the Gaussian framework is also revisited to improve numerical efficiency. An example of vibration-based structural damage diagnosis is presented to motivate the study and to illustrate the performance of the proposed method. [17]

7.5.3. Performance of damage detection in dependence of sample length and measurement noise

Participants: Saeid Allahdadian, Michael Doehler, Laurent Mevel.

In this work the effects of measuring noise and number of samples is studied on the stochastic subspace damage detection (SSDD) technique. In previous studies, the effect of these practical parameters was examined on simulated measurements from a model of a real structure. In this study, these effects are formulated for the expected damage index evaluated from a Chi-square distributed value. Several theorems that describe the effects are proposed and proved. These theorems are used to develop a guideline to serve the user of the SSDD method to face these effects. [25]

7.5.4. Statistical damage localization with stochastic load vectors

Participants: Md Delwar Hossain Bhuyan, Michael Doehler, Laurent Mevel.

The Stochastic Dynamic Damage Locating Vector (SDDLV) method is an output-only damage localization method based on both a Finite Element (FE) model of the structure and modal parameters estimated from output-only measurements in the damage and reference states of the system. A vector is obtained in the null space of the changes in the transfer matrix computed in both states and then applied as a load vector to the model. The damage location is related to this stress where it is close to zero. In previous works an important theoretical limitation was that the number of modes used in the computation of the transfer function could not be higher than the number of sensors located on the structure. It would be nonetheless desirable not to discard information from the identification procedure. In this work, the SDDLV method has been extended with a joint statistical approach for multiple mode sets, overcoming this restriction on the number of modes. The new approach is validated in a numerical application, where the outcomes for multiple mode sets are compared with a single mode set. From these results, it can be seen that the success rate of finding the correct damage localization is increased when using multiple mode sets instead of a single mode set. [27]

7.5.5. Classification of vibration-based damage localization methods

Participant: Michael Doehler.

This work, issued from the COST Action TU1402, is in collaboration with M.P. Limongelli (Politecnico Milan), E. Chatzi (ETH Zürich), G. Lombaert and E. Reynders (both KU Leuven).

After a brief review of vibration based damage identification methods, three different algorithms for damage identification are applied to the case of the benchmark Z24 bridge. Data-driven as well as model-based methods are discussed, including input-output algorithms for taking into account the effect of environmental and/or operational sources on the variability of damage features. A further class of data-driven methods that use finite element information is finally introduced as a possible future development. [35]

7.5.6. Structural system reliability updating with subspace-based damage detection information

Participant: Michael Doehler.

This work is in collaboration with S. Thöns (DTU).

Damage detection systems and algorithms (DDS and DDA) provide information of the structural system integrity in contrast to e.g. local information by inspections or non-destructive testing techniques. However, the potential of utilizing DDS information for the structural integrity assessment and prognosis is hardly exploited nor treated in scientific literature up to now. In order to utilize the information provided by DDS for the structural performance, usually high computational efforts for the pre-determination of DDS reliability are required. In this work, an approach for the DDS performance modelling is introduced building upon the non-destructive testing reliability which applies to structural systems and DDS containing a strategy to overcome the high computational efforts for the pre-determination of the DDS reliability. This approach takes basis in the subspace-based damage detection method and builds upon mathematical properties of the damage detection algorithm. Computational efficiency is gained by calculating the probability of damage indication directly without necessitating a pre-determination for all damage states. The developed approach is applied to a static, dynamic, deterioration and reliability structural system model, demonstrating the potentials for utilizing DDS for risk reduction. [30]

7.5.7. Structural system model updating based on different sensor types

Participants: Dominique Siegert, Xavier Chapeleau, Ivan Guéguen.

Detecting and quantifying early structural damages using deterministic and probabilistic model updating techniques can be achieved by local information in a form of optical strain measurement. The strategy consists in updating physical parameters associated to damages, such as Young's modulus, in order to minimize the gap between the numerical strain obtained from finite element solves and the strain sensor outputs. Generally, the damage estimation is an ill-posed inverse problem, and hence requires regularization. Herein, three model updating techniques are considered involving different type of regularization: classical Tikhonov regularization, constitutive relation error based updating method and Bayesian approach [21]. This work follows an experimental campaign carried out on a post tensioned concrete beam with the aim of investigating the possibility to detect early warning signs of deterioration based on static and/or dynamic tests. Responses of a beam were measured by an extensive set of instruments consisting of accelerometers, inclinometers, displacement transducers, strain gauges and optical fibers. [18].

IPSO Project-Team

4. New Results

4.1. List of results

4.1.1. Landau damping in Sobolev spaces for the Vlasov-HMF model

In [25], the authors consider the Vlasov-HMF (Hamiltonian Mean-Field) model. They consider solutions starting in a small Sobolev neighborhood of a spatially homogeneous state satisfying a linearized stability criterion (Penrose criterion). They prove that these solutions exhibit a scattering behavior to a modified state, which implies a nonlinear Landau damping effect with polynomial rate of damping.

4.1.2. Fast Weak-Kam Integrators for separable Hamiltonian systems

In [4], the authors consider a numerical scheme for Hamilton-Jacobi equations based on a direct discretization of the Lax-Oleinik semi-group. They prove that this method is convergent with respect to the time and space stepsizes provided the solution is Lipschitz, and give an error estimate. Moreover, They prove that the numerical scheme is a *geometric integrator* satisfying a discrete weak-KAM theorem which allows to control its long time behavior. Taking advantage of a fast algorithm for computing min–plus convolutions based on the decomposition of the function into concave and convex parts, they show that the numerical scheme can be implemented in a very efficient way.

4.1.3. The weakly nonlinear large-box limit of the 2D cubic nonlinear Schrödinger equation

In [23], the authors consider the cubic nonlinear Schrödinger (NLS) equation set on a two dimensional box of size L with periodic boundary conditions. By taking the large box limit $L \to \infty$ in the weakly nonlinear regime (characterized by smallness in the critical space), we derive a new equation set on $\mathbb{R}^2$ that approximates the dynamics of the frequency modes. This nonlinear equation turns out to be Hamiltonian and enjoys interesting symmetries, such as its invariance under Fourier transform, as well as several families of explicit solutions. A large part of this work is devoted to a rigorous approximation result that allows to project the long-time dynamics of the limit equation into that of the cubic NLS equation on a box of finite size.

4.1.4. An asymptotic preserving scheme for the relativistic Vlasov–Maxwell equations in the classical limit

In [13], the authors consider the relativistic Vlasov–Maxwell (RVM) equations in the limit when the light velocity c goes to infinity. In this regime, the RVM system converges towards the Vlasov–Poisson system and the aim of this work is to construct asymptotic preserving numerical schemes that are robust with respect to this limit. A number of numerical simulations are conducted in order to investigate the performances of our numerical scheme both in the relativistic as well as in the classical limit regime. In addition, they derive the dispersion relation of the Weibel instability for the continuous and the discretized problem.

4.1.5. Free Vibrations of Axisymmetric Shells: Parabolic and Elliptic cases

In [41], approximate eigenpairs (quasimodes) of axisymmetric thin elastic domains with laterally clamped boundary conditions (Lamé system) are determined by an asymptotic analysis as the thickness (2ε) tends to zero. The departing point is the Koiter shell model that we reduce by asymptotic analysis to a scalar model that depends on two parameters: the angular frequency k and the half-thickness ε. Optimizing k for each chosen ε, we find power laws for k in function of ε that provide the smallest eigenvalues of the scalar reductions. Corresponding eigenpairs generate quasimodes for the 3D Lamé system by means of several reconstruction operators, including boundary layer terms. Numerical experiments demonstrate that in many cases the constructed eigenpair corresponds to the first eigenpair of the Lamé system.

Geometrical conditions are necessary to this approach: The Gaussian curvature has to be nonnegative and the azimuthal curvature has to dominate the meridian curvature in any point of the midsurface. In this case, the first eigenvector admits progressively larger oscillation in the angular variable as ε tends to 0. Its angular frequency exhibits a power law relation of the form $k = \gamma\varepsilon^{-\beta}$ with $\beta = \frac{1}{4}$ in the parabolic case (cylinders and trimmed cones), and the various βs $\frac{2}{5}$, $\frac{3}{7}$, and $\frac{1}{3}$ in the elliptic case. For these cases where the mathematical analysis is applicable, numerical examples that illustrate the theoretical results are presented.

4.1.6. High frequency oscillations of first eigenmodes in axisymmetric shells as the thickness tends to zero

In [30], the lowest eigenmode of thin axisymmetric shells is investigated for two physical models (acoustics and elasticity) as the shell thickness (2ε) tends to zero. Using a novel asymptotic expansion we determine the behavior of the eigenvalue $\lambda(\varepsilon)$ and the eigenvector angular frequency $k(\varepsilon)$ for shells with Dirichlet boundary conditions along the lateral boundary, and natural boundary conditions on the other parts.

First, the scalar Laplace operator for acoustics is addressed, for which $k(\varepsilon)$ is always zero. In contrast to it, for the Lamé system of linear elasticity several different types of shells are defined, characterized by their geometry, for which $k(\varepsilon)$ tends to infinity as ε tends to zero. For two families of shells: cylinders and elliptical barrels we explicitly provide $\lambda(\varepsilon)$ and $k(\varepsilon)$ and demonstrate by numerical examples the different behavior as ε tends to zero.

4.1.7. On numerical Landau damping for splitting methods applied to the Vlasov-HMF model

In [24], we consider time discretizations of the Vlasov-HMF (Hamiltonian Mean-Field) equation based on splitting methods between the linear and non-linear parts. We consider solutions starting in a small Sobolev neighborhood of a spatially homogeneous state satisfying a linearized stability criterion (Penrose criterion). We prove that the numerical solutions exhibit a scattering behavior to a modified state, which implies a nonlinear Landau damping effect with polynomial rate of damping. Moreover, we prove that the modified state is close to the continuous one and provide error estimates with respect to the time stepsize.

4.1.8. High-order Hamiltonian splitting for Vlasov-Poisson equations

In [5], we consider the Vlasov-Poisson equation in a Hamiltonian framework and derive new time splitting methods based on the decomposition of the Hamiltonian functional between the kinetic and electric energy. Assuming smoothness of the solutions, we study the order conditions of such methods. It appears that these conditions are of Runge-Kutta-Nyström type. In the one dimensional case, the order conditions can be further simplified, and efficient methods of order 6 with a reduced number of stages can be constructed. In the general case, high-order methods can also be constructed using explicit computations of commutators. Numerical results are performed and show the benefit of using high-order splitting schemes in that context. Complete and self-contained proofs of convergence results and rigorous error estimates are also given.

4.1.9. Uniformly accurate exponential-type integrators for Klein-Gordon equations with asymptotic convergence to classical splitting schemes in the nonlinear schrödinger limit

In [34], we introduce efficient and robust exponential-type integrators for Klein-Gordon equations which resolve the solution in the relativistic regime as well as in the highly-oscillatory non-relativistic regime without any step-size restriction under the same regularity assumptions on the initial data required for the integration of the corresponding nonlinear Schrödinger limit system. In contrast to previous works we do not employ any asymptotic or multiscale expansion of the solution. This allows us to derive uniform convergent schemes under far weaker regularity assumptions on the exact solution. In addition, the newly derived first- and second-order exponential-type integrators converge to the classical Lie, respectively, Strang splitting in the nonlinear Schrödinger limit.

4.1.10. Convergence of a normalized gradient algorithm for computing ground states

In [45], we consider the approximation of the ground state of the one-dimensional cubic nonlinear Schrödinger equation by a normalized gradient algorithm combined with linearly implicit time integrator, and finite difference space approximation. We show that this method, also called *imaginary time evolution method* in the physics literature, is locally convergent, and we provide error estimates: for an initial data in a neighborhood of the ground state, the algorithm converges exponentially towards a modified soliton that is a space discretization of the exact soliton, with error estimates depending on the discretization parameters.

4.1.11. Improved error estimates for splitting methods applied to highly-oscillatory nonlinear Schrödinger equations

In [8], we analyse the error behavior of operator splitting methods for highly-oscillatory differential equations. The scope of applications includes time-dependent nonlinear Schrödinger equations, where the evolution operator associated with the principal linear part is highly-oscillatory and periodic in time. In a first step, a known convergence result for the second-order Strang splitting method applied to the cubic Schrödinger equation is adapted to a wider class of nonlinearities. In a second step, the dependence of the global error on the decisive parameter $0 < \varepsilon < 1$, defining the length of the period, is examined. The main result states that, compared to established error estimates, the Strang splitting method is more accurate by a factor ε, provided that the time stepsize is chosen as an integer fraction of the period. This improved error behavior over a time interval of fixed length, which is independent of the period, is due to an averaging effect. The extension of the convergence result to higher-order splitting methods and numerical illustrations complement the investigations.

4.1.12. Solving highly-oscillatory NLS with SAM: numerical efficiency and geometric properties

In [7], we present the Stroboscopic Averaging Method (SAM), which aims at numerically solving highly-oscillatory differential equations. More specifically, we first apply SAM to the Schrödinger equation on the 1-dimensional torus and on the real line with harmonic potential, with the aim of assessing its efficiency: as compared to the well-established standard splitting schemes, the stiffer the problem is, the larger the speed-up grows (up to a factor 100 in our tests). The geometric properties of SAM are also explored: on very long time intervals, symmetric implementations of the method show a very good preservation of the mass invariant and of the energy. In a second series of experiments on 2-dimensional equations, we demonstrate the ability of SAM to capture qualitatively the long-time evolution of the solution (without spurring high oscillations).

4.1.13. Highly-oscillatory evolution equations with non-resonant frequencies: averaging and numerics

In [40], we are concerned with the application of the recently introduced multi-revolution composition methods, on the one hand, and two-scale methods, on the other hand, to a class of highly-oscillatory evolution equations with multiple frequencies. The main idea relies on a well-balanced reformulation of the problem as an equivalent mono-frequency equation which allows for the use of the two aforementioned techniques.

4.1.14. A formal series approach to the Center Manifold theorem

In [35], we consider near-equilibrium systems of ordinary differential equations with explicit separation of the slow and stable manifolds. Formal B-series like those previously used to analyze highly-oscillatory systems or to construct modified equations are employed here to construct expansions of the change of variables, the center invariant manifold and the reduced model. The new approach may be seen as a process of reduction to a normal form, with the main advantage, as compared to the standard view conveyed by the celebrated center manifold theorem, that it is possible to recover the complete solution at any time through an explicit change of variables.

4.1.15. Uniformly accurate time-splitting methods for the semi-classical Schrödinger equation, Part II: Numerical analysis

This article [39] is second part of a twofold paper devoted to the construction of numerical methods which remain insensitive to the smallness of the semiclassical parameter for the Schrödinger equation in the semiclassical limit. Here, we specifically analyse the convergence behavior of the first-order splitting introduced in Part I, for a linear equation with smooth potential. Our main result is a proof of uniform accuracy.

4.1.16. Averaging of highly-oscillatory transport equations

In [38], we develop a new strategy aimed at obtaining high-order asymptotic models for transport equations with highly-oscillatory solutions. The technique relies upon recent developments averaging theory for ordinary differential equations, in particular normal form expansions in the vanishing parameter. Noteworthy, the result we state here also allows for the complete recovery of the exact solution from the asymptotic model. This is done by solving a companion transport equation that stems naturally from the change of variables underlying high-order averaging. Eventually, we apply our technique to the Vlasov equation with external electric and magnetic fields. Both constant and non-constant magnetic fields are envisaged, and asymptotic models already documented in the literature and re-derived using our methodology. In addition, it is shown how to obtain new high-order asymptotic models.

4.1.17. Asymptotic preserving and time diminishing schemes for rarefied gas dynamic

In [11], we introduce a new class of numerical schemes for rarefied gas dynamic problems described by collisional kinetic equations. The idea consists in reformulating the problem using a micro-macro decomposition and successively in solving the microscopic part by using asymptotically stable Monte Carlo methods. We consider two types of decompositions, the first leading to the Euler system of gas dynamics while the second to the Navier-Stokes equations for the macroscopic part. In addition, the particle method which solves the microscopic part is designed in such a way that the global scheme becomes computationally less expensive as the solution approaches the equilibrium state as opposite to standard methods for kinetic equations which computational cost increases with the number of interactions. At the same time, the statistical error due to the particle part of the solution decreases as the system approach the equilibrium state. This causes the method to degenerate to the sole solution of the macroscopic hydrodynamic equations (Euler or Navier-Stokes) in the limit of infinite number of collisions. In a last part, we will show the behaviors of this new approach in comparisons to standard Monte Carlo techniques for solving the kinetic equation by testing it on different problems which typically arise in rarefied gas dynamic simulations.

4.1.18. Asymptotic Preserving scheme for a kinetic model describing incompressible fluids

The kinetic theory of fluid turbulence modeling developed by Degond and Lemou (2002) is considered for further study, analysis and simulation. Starting with the Boltzmann like equation representation for turbulence modeling, a relaxation type collision term is introduced for isotropic turbulence. In order to describe some important turbulence phenomenology, the relaxation time incorporates a dependency on the turbulent microscopic energy and this makes difficult the construction of efficient numerical methods. To investigate this problem, we focus in this work [17] on a multi-dimensional prototype model and first propose an appropriate change of frame that makes the numerical study simpler. Then, a numerical strategy to tackle the stiff relaxation source term is introduced in the spirit of Asymptotic Preserving Schemes. Numerical tests are performed in a one-dimensional framework on the basis of the developed strategy to confirm its efficiency.

4.1.19. Numerical schemes for kinetic equations in the diffusion and anomalous diffusion limits. Part I: the case of heavy-tailed equilibrium

In [15], we propose some numerical schemes for linear kinetic equations in the diffusion and anomalous diffusion limit. When the equilibrium distribution function is a Maxwellian distribution, it is well known that for an appropriate time scale, the small mean free path limit gives rise to a diffusion type equation. However, when a heavy-tailed distribution is considered, another time scale is required and the small mean free path limit leads to a fractional anomalous diffusion equation. Our aim is to develop numerical schemes for the

original kinetic model which works for the different regimes, without being restricted by stability conditions of standard explicit time integrators. First, we propose some numerical schemes for the diffusion asymptotics; then, their extension to the anomalous diffusion limit is studied. In this case, it is crucial to capture the effect of the large velocities of the heavy-tailed equilibrium, so that some important transformations of the schemes derived for the diffusion asymptotics are needed. As a result, we obtain numerical schemes which enjoy the Asymptotic Preserving property in the anomalous diffusion limit, that is: they do not suffer from the restriction on the time step and they degenerate towards the fractional diffusion limit when the mean free path goes to zero. We also numerically investigate the uniform accuracy and construct a class of numerical schemes satisfying this property. Finally, the efficiency of the different numerical schemes is shown through numerical experiments.

4.1.20. Numerical schemes for kinetic equations in the anomalous diffusion limit. Part II: degenerate collision frequency

In [14], which is the continuation of [15], we propose numerical schemes for linear kinetic equation which are able to deal with the fractional diffusion limit. When the collision frequency degenerates for small velocities it is known that for an appropriate time scale, the small mean free path limit leads to an anomalous diffusion equation. From a numerical point of view, this degeneracy gives rise to an additional stiffness that must be treated in a suitable way to avoid a prohibitive computational cost. Our aim is therefore to construct a class of numerical schemes which are able to undertake these stiffness. This means that the numerical schemes are able to capture the effect of small velocities in the small mean free path limit with a fixed set of numerical parameters. Various numerical tests are performed to illustrate the efficiency of our methods in this context.

4.1.21. Multiscale schemes for the BGK-Vlasov-Poisson system in the quasi-neutral and fluid limits. Stability analysis and first order schemes

In [12], we deal with the development and the analysis of asymptotic stable and consistent schemes in the joint quasi-neutral and fluid limits for the collisional Vlasov-Poisson system. In these limits, the classical explicit schemes suffer from time step restrictions due to the small plasma period and Knudsen number. To solve this problem, we propose a new scheme stable for choices of time steps independent from the small scales dynamics and with comparable computational cost with respect to standard explicit schemes. In addition, this scheme reduces automatically to consistent discretizations of the underlying asymptotic systems. In this first work on this subject, we propose a first order in time scheme and we perform a relative linear stability analysis to deal with such problems. The framework we propose permits to extend this approach to high order schemes in the next future. We finally show the capability of the method in dealing with small scales through numerical experiments.

4.1.22. Uniformly accurate forward semi-Lagrangian methods for highly oscillatory Vlasov-Poisson equations.

In [16], we deal with the numerical simulation of a Vlasov-Poisson equation modeling charged particles in a beam submitted to a highly oscillatory external electric field. A numerical scheme is constructed for this model. This scheme is uniformly accurate with respect to the size of the fast time oscillations of the solution, which means that no time step refinement is required to simulate the problem. The scheme combines the forward semi-Lagrangian method with a class of Uniformly Accurate (UA) time integrators to solve the characteristics. These UA time integrators are derived by means of a two-scale formulation of the characteristics, with the introduction of an additional periodic variable. Numerical experiments are done to show the efficiency of the proposed methods compared to conventional approaches.

4.1.23. Multi-scale methods for the solution of the radiative transfer equation

Various methods have been developed and tested over the years to solve the radiative transfer equation (RTE) with different results and trade-offs. Although the RTE is extensively used, the approximate diffusion equation is sometimes preferred, particularly in optically thick media, due to the lower computational requirements. Recently, multi-scale models, namely the domain decomposition methods, the micro-macro model and the

hybrid transport- diffusion model, have been proposed as an alternative to the RTE. In domain decomposition methods, the domain is split into two subdomains, namely a mesoscopic subdomain where the RTE is solved and a macroscopic subdomain where the diffusion equation is solved. In the micro-macro and hybrid transport-diffusion models, the radiation intensity is decomposed into a macroscopic component and a mesoscopic one. In both cases, the aim is to reduce the computational requirements, while maintaining the accuracy, or to improve the accuracy for similar computational requirements. In [10], these multi-scale methods are described, and the application of the micro-macro and hybrid transport-diffusion models to a three- dimensional transient problem is reported. It is shown that when the diffusion approximation is accurate, but not over the entire domain, the multi-scale methods may improve the solution accuracy in comparison with the solution of the RTE. The order of accuracy of the numerical schemes and the radiative properties of the medium play a key role in the performance of the multi-scale methods.

4.1.24. Nonlinear Geometric Optics method based multi-scale numerical schemes for highly-oscillatory transport equations

In [42], we introduce a new numerical strategy to solve a class of oscillatory transport PDE models which is able to capture accurately the solutions without numerically resolving the high frequency oscillations in both space and time. Such PDE models arise in semiclassical modeling of quantum dynamics with band-crossings, and other highly oscillatory waves. Our first main idea is to use the nonlinear geometric optics ansatz, which builds the oscillatory phase into an independent variable. We then choose suitable initial data, based on the Chapman-Enskog expansion, for the new model. For a scalar model, we prove that so constructed model will have certain smoothness, and consequently, for a first order approximation scheme we prove uniform error estimates independent of the (possibly small) wave length. The method is extended to systems arising from a semiclassical model for surface hopping, a non-adiabatic quantum dynamic phenomenon. Numerous numerical examples demonstrate that the method has the desired properties.

4.1.25. Asymptotic Preserving numerical schemes for multiscale parabolic problems

In [18], we consider a class of multiscale parabolic problems with diffusion coefficients oscillating in space at a possibly small scale ε. Numerical homogenization methods are popular for such problems, because they capture efficiently the asymptotic behaviour as ε goes to 0, without using a dramatically fine spatial discretization at the scale of the fast oscillations. However, known such homogenization schemes are in general not accurate for both the highly oscillatory regime ($\varepsilon << 1$) and the non oscillatory regime ($\varepsilon \approx 1$). In this paper, we introduce an Asymptotic Preserving method based on an exact micro-macro decomposition of the solution which remains consistent for both regimes.

4.1.26. Uniformly accurate numerical schemes for the nonlinear dirac equation in the nonrelativistic limit regime

In [47], we apply the two-scale formulation approach to propose uniformly accurate (UA) schemes for solving the nonlinear Dirac equation in the nonrelativistic limit regime. The nonlinear Dirac equation involves two small scales ε and ε^2 with $\varepsilon \to 0$ in the nonrelativistic limit regime. The small parameter causes high oscillations in time which bring severe numerical burden for classical numerical methods. We present a suitable two-scale formulation as a general strategy to tackle a class of highly oscillatory problems involving the two small scales ε and ε^2. A numerical scheme with uniform (with respect to $\varepsilon \in [0, 1]$) second order accuracy in time and a spectral accuracy in space are proposed. Numerical experiments are done to confirm the UA property.

4.1.27. Semiclassical Sobolev constants for the electro-magnetic Robin Laplacian

In [26], we deal with the asymptotic analysis of the optimal Sobolev constants in the semiclassical limit and in any dimension. We combine semiclassical arguments and concentration-compactness estimates to tackle the case when an electromagnetic field is added as well as a smooth boundary carrying a Robin condition. As a byproduct of the semiclassical strategy, we also get exponentially weighted localization estimates of the minimizers.

4.1.28. On the MIT bag model: self-adjointness and non-relativistic limit

This paper [32] is devoted to the mathematical investigation of the MIT bag model, that is the Dirac operator on a smooth and bounded domain with certain boundary conditions. We prove that the operator is self-adjoint and, when the mass goes to infinity, we provide spectral asymptotic results.

4.1.29. Global behavior of N competing species with strong diffusion: diffusion leads to exclusion

It is known that the competitive exclusion principle holds for a large kind of models involving several species competing for a single resource in an homogeneous environment. Various works indicate that the coexistence is possible in an heterogeneous environment. We propose in [6] a spatially heterogeneous system modeling the competition of several species for a single resource. If spatial movements are fast enough, we show that our system can be well approximated by a spatially homogeneous system, called aggregated model, which can be explicitly computed. Moreover, we show that if the competitive exclusion principle holds for the aggregated model, it holds for the spatially heterogeneous model too.

4.1.30. Extended Rearrangement inequalities and applications to some quantitative stability results

In [28], we prove a new functional inequality of Hardy-Littlewood type for generalized rearrangements of functions. We then show how this inequality provides *quantitative* stability results of steady states to evolution systems that essentially preserve the rearrangements and some suitable energy functional, under minimal regularity assumptions on the perturbations. In particular, this inequality yields a *quantitative* stability result of a large class of steady state solutions to the Vlasov-Poisson systems, and more precisely we derive a quantitative control of the L^1 norm of the perturbation by the relative Hamiltonian (the energy functional) and rearrangements. A general non linear stability result has been obtained recently by Lemou, Méhats and Raphaël (2012) in the gravitational context, however the proof relied in a crucial way on compactness arguments which by construction provides no quantitative control of the perturbation. Our functional inequality is also applied to the context of 2D-Euler system and also provides quantitative stability results of a large class of steady-states to this system in a natural energy space.

4.1.31. Mate Finding, Sexual Spore Production, and the Spread of Fungal Plant Parasites

Sexual reproduction and dispersal are often coupled in organisms mixing sexual and asexual reproduction, such as fungi. The aim of this study [27] is to evaluate the impact of mate limitation on the spreading speed of fungal plant parasites. Starting from a simple model with two coupled partial differential equations, we take advantage of the fact that we are interested in the dynamics over large spatial and temporal scales to reduce the model to a single equation. We obtain a simple expression for speed of spread, accounting for both sexual and asexual reproduction. Taking Black Sigatoka disease of banana plants as a case study, the model prediction is in close agreement with the actual spreading speed (100 km per year), whereas a similar model without mate limitation predicts a wave speed one order of magnitude greater. We discuss the implications of these results to control parasites in which sexual reproduction and dispersal are intrinsically coupled.

4.1.32. Dimension reduction for rotating Bose-Einstein condensates with anisotropic confinement

In [29], we consider the three-dimensional time-dependent Gross-Pitaevskii equation arising in the description of rotating Bose-Einstein condensates and study the corresponding scaling limit of strongly anisotropic confinement potentials. The resulting effective equations in one or two spatial dimensions, respectively, are rigorously obtained as special cases of an averaged three dimensional limit model. In the particular case where the rotation axis is not parallel to the strongly confining direction the resulting limiting model(s) include a negative, and thus, purely repulsive quadratic potential, which is not present in the original equation and which can be seen as an effective centrifugal force counteracting the confinement.

4.1.33. Averaging of nonlinear Schrödinger equations with strong magnetic confinement

In [46], we consider the dynamics of nonlinear Schrödinger equations with strong constant magnetic fields. In an asymptotic scaling limit the system exhibits a purely magnetic confinement, based on the spectral properties of the Landau Hamiltonian. Using an averaging technique we derive an associated effective description via an averaged model of nonlinear Schrödinger type. In a special case this also yields a derivation of the LLL equation.

4.1.34. The Interaction Picture method for solving the generalized nonlinear Schrödinger equation in optics

The interaction picture (IP) method is a very promising alternative to Split-Step methods for solving certain type of partial differential equations such as the nonlinear Schrödinger equation used in the simulation of wave propagation in optical fibers. The method exhibits interesting convergence properties and is likely to provide more accurate numerical results than cost comparable Split-Step methods such as the Symmetric Split-Step method. In [1] we investigate in detail the numerical properties of the IP method and carry out a precise comparison between the IP method and the Symmetric Split-Step method.

4.1.35. Diffusion limit for the radiative transfer equation perturbed by a Markovian process

In [21], we study the stochastic diffusive limit of a kinetic radiative transfer equation, which is non-linear, involving a small parameter and perturbed by a smooth random term. Under an appropriate scaling for the small parameter, using a generalization of the perturbed test-functions method, we show the convergence in law to a stochastic non-linear fluid limit.

4.1.36. Estimate for P_tD for the stochastic Burgers equation

In [20], we consider the Burgers equation on $H = L^2(0,1)$ perturbed by white noise and the corresponding transition semigroup P_t. We prove a new formula for $P_tD\phi$ (where $\phi : H \to \mathbb{R}$ is bounded and Borel) which depends on ϕ but not on its derivative. Then we deduce some new consequences for the invariant measure ν of P_t as its Fomin differentiability and an integration by parts formula which generalises the classical one for gaussian measures.

4.1.37. Degenerate Parabolic Stochastic Partial Differential Equations: Quasilinear case

In [22], we study the Cauchy problem for a quasilinear degenerate parabolic stochastic partial differential equation driven by a cylindrical Wiener process. In particular, we adapt the notion of kinetic formulation and kinetic solution and develop a well-posedness theory that includes also an L^1-contraction property. In comparison to the first-order case (Debussche and Vovelle, 2010) and to the semilinear degenerate parabolic case (Hofmanová, 2013), the present result contains two new ingredients: a generalized Itô formula that permits a rigorous derivation of the kinetic formulation even in the case of weak solutions of certain nondegenerate approximations and a direct proof of strong convergence of these approximations to the desired kinetic solution of the degenerate problem.

4.1.38. An integral inequality for the invariant measure of a stochastic reaction-diffusion equation

In [19], we consider a reaction-diffusion equation perturbed by noise (not necessarily white). We prove an integral inequality for the invariant measure ν of a stochastic reaction-diffusion equation. Then we discuss some consequences as an integration by parts formula which extends to ν a basic identity of the Malliavin Calculus. Finally, we prove the existence of a surface measure for a ball and a half-space of H.

4.1.39. Large deviations for the two-dimensional stochastic Navier-Stokes equation with vanishing noise correlation

In [36], we are dealing with the validity of a large deviation principle for the two-dimensional Navier-Stokes equation, with periodic boundary conditions, perturbed by a Gaussian random forcing. We are here interested in the regime where both the strength of the noise and its correlation are vanishing, on a length scale ε and

$\delta(\varepsilon)$, respectively, with $0 < \varepsilon, \delta(\varepsilon) << 1$. Depending on the relationship between ε and $\delta(\varepsilon)$ we will prove the validity of the large deviation principle in different functional spaces.

4.1.40. Quasilinear generalized parabolic Anderson model

In [33], we provide a local in time well-posedness result for a quasilinear generalized parabolic Anderson model in dimension two $\partial_t u = a(u)\Delta u + g(u)\xi$. The key idea of our approach is a simple transformation of the equation which allows to treat the problem as a semilinear problem. The analysis is done within the setting of paracontrolled calculus.

4.1.41. The Schrödinger equation with spatial white noise potential

In [44], we consider the linear and nonlinear Schrödinger equation with a spatial white noise as a potential in dimension 2. We prove existence and uniqueness of solutions thanks to a change of unknown used by Hairer and Labbé (2015) and conserved quantities.

KERDATA Project-Team

7. New Results

7.1. Convergence of HPC and Big Data

7.1.1. Transactional storage

Participants: Pierre Matri, Alexandru Costan, Gabriel Antoniu.

Concurrent Big Data applications often require high-performance storage, as well as ACID (Atomicity, Consistency, Isolation, Durability) transaction support. Although blobs (binary large objects) are an increasingly popular model for addressing the storage needs of such applications, state-of-the-art blob storage systems typically offer no transaction semantics. This demands users to coordinate access to data carefully in order to avoid race conditions, inconsistent writes, overwrites and other problems that cause erratic behavior. We argue there is a gap between existing storage solutions and application requirements, which limits the design of transaction-oriented applications.

Týr is the first blob storage system to provide built-in, multi-blob transactions, while retaining sequential consistency and high throughput under heavy access concurrency. Týr offers fine-grained random write access to data and in-place atomic operations.

Large-scale experiments on Microsoft Azure with a production application from CERN LHC show Týr throughput outperforms state-of-the-art solutions by more than 75 %.

Collaboration. *This work was done in collaboration with María Pérez, UPM, Spain.*

7.1.2. Big Data on HPC

Participants: Orçun Yildiz, Shadi Ibrahim, Gabriel Antoniu.

Over the last decade, Map-Reduce has stood as the most powerful Big Data processing model. Map-Reduce model is now used by many companies and research labs to facilitate large-scale data analysis. With the growing needs of users and size of data, commodity-based infrastructure (most commonly used as of now) will strain under the heavy weight of Big Data. On the other hand, HPC systems offer a rich set of opportunities for Big Data processing.

As first steps towards Big Data processing on HPC systems, several research efforts have been devoted to understand Map-Reduce performance on these systems. Yet, the impact of the specific features of HPC environments have not been fully investigated, yet.

We conducted an experimental campaign to provide a clearer understanding of Map-Reduce performance on HPC systems. We use Spark, a widely adopted Map-Reduce framework, and representative Big Data workloads on Grid'5000 testbed to evaluate how the latency, contention and file system's configuration can influence the application performance.

7.1.3. Energy vs. performance trade-offs

Participants: Mohammed-Yacine Taleb, Shadi Ibrahim, Gabriel Antoniu.

Most large popular web applications, like Facebook and Twitter, have been relying on large amounts of in-memory storage to cache data and provide a low response time. As the memory capacity of clusters and clouds increases, it becomes possible to keep most of the data in the main memory.

This motivates the introduction of in-memory storage systems. While prior work has focused on how to exploit the low latency of in-memory access at scale, there is still little knowledge regarding the energy efficiency of in-memory storage systems. This is unfortunate, as it is known that main memory is a major energy bottleneck in many computing systems. For instance, DRAM consumes up to 40 % of a server's power.

By means of experimental evaluation, we have studied the performance and energy-efficiency of RAMCloud — a well-known in-memory storage system. We demonstrated that although RAMCloud is scalable for read-only applications, it exhibits non-proportional power consumption. We also found that the current replication scheme implemented in RAMCloud limits the performance and results in high energy consumption. Surprisingly enough, we also showed that replication can even play a negative role in crash-recovery.

Collaboration. *This work was carried out in collaboration with Toni Cortes (BSC, Spain).*

7.2. Efficient I/O and communication for Extreme-scale HPC systems

7.2.1. Adaptive performance-constrained in situ visualisation

Participant: Lokman Rahmani.

While many parallel visualization tools now provide in situ visualization capabilities, the trend has been to feed such tools with large amounts of unprocessed output data and let them render everything at the highest possible resolution. This leads to an increased run time of simulations that still have to complete within a fixed-length job allocation.

We have been working on tackling the challenge of enabling in situ visualization under performance constraints. Our approach shuffles data across processes according to their contents and filters out part of them. Thereby, the visualization pipeline is only fed with a reorganized subset of the data produced by the simulation.

Our framework, as presented in [22], leverages fast, generic evaluation procedures to score blocks of data, using information theory, statistics, and linear algebra. It monitors its own performance and dynamically adapts to achieve appropriate visual fidelity within predefined performance constraints. Experiments on the Blue Waters supercomputer with the CM1 simulation show that our approach enables a 5-time speedup with respect to the initial visualization pipeline, and is able to meet performance constraints.

Collaboration. *This was was carried out with the collaboration of Matthieu Dorier, ANL, USA.*

7.2.2. Dragonfly

Participants: Nathanaël Cheriere, Shadi Ibrahim, Gabriel Antoniu.

High-radix direct network topologies such as Dragonfly have been proposed for Petascale and Exascale supercomputers. It has been shown that they ensure fast interconnections and reduce the cost of the network compared to traditional network topologies. However, current algorithms for communication do not consider the topology and thus waste numerous opportunities of optimization for performance.

In our studies, we exploit the strength of the Dragonfly with topology-aware algorithms for AllGather and Scatter operations. We analyze existing algorithms, then propose derived algorithms, that we evaluate using CODES, an event-driven simulator.

As expected, making AllGather algorithms topology-aware does improve the performance and reduces the link utilization. However, simulations of various Scatter algorithms show surprising results, and point out the important role played by hardware for the efficiency of the algorithms. In particular, the knowledge of the number and size of input-output buffers in routers can be exploited to accelerate the Scatter operation by a factor up to 2 times.

Collaboration. *This work was done in collaboration with Matthieu Dorier and Rob Ross, ANL, USA.*

7.2.3. Interference between HPC jobs

Participants: Orçun Yildiz, Shadi Ibrahim, Gabriel Antoniu.

As we move toward the Exascale era, performance variability in HPC systems remains a challenge. I/O interference, a major cause of this variability, is becoming more important every day with the growing number of concurrent applications that share larger machines. Earlier research efforts on mitigating I/O interference focus on a single potential cause of interference (e.g., the network). Yet the root causes of I/O interference can be diverse.

In [27], we conducted an extensive experimental campaign to explore the various root causes of I/O interference in HPC storage systems. We used micro-benchmarks on the Grid'5000 testbed to evaluate how I/O interference is influenced by the applications' access pattern, the network components, the file system's configuration, and the backend storage devices.

Our studies revealed that in many situations interference is a result of a bad flow control in the I/O path, rather than being caused by some single bottleneck in one of its components. We further show that interference-free behavior is not necessarily a sign of optimal performance. To the best of our knowledge, our work provides the first deep insight into the role of each of the potential root causes of interference and their interplay. Our findings can help developers and platform owners improve I/O performance and motivate further research addressing the problem across all components of the I/O stack.

Collaboration. *This work was done in collaboration with Matthieu Dorier and Rob Ross, ANL, USA.*

7.3. Workflow on clouds

7.3.1. Managing hot metadata for scientific workflows on multisite clouds

Participants: Luis Eduardo Pineda Morales, Alexandru Costan, Gabriel Antoniu.

Large-scale scientific applications are often expressed as workflows that help defining data dependencies between their different components. Such workflows may incur huge storage and computation requirements, so that they need to be processed in multiple (cloud-federated) datacenters. A major challenge in such multisite clouds is the long latency of the network links between datacenters, that limits the performance of multisite applications. Moreover, it has been shown that poor metadata handling can further impact the efficiency of computing systems. Many efforts have been done to improve metadata management; however, most of them concern only single-site, HPC systems to date.

In [26], we assert that some workflow metadata are more frequently accessed than other, and thus should be handled with higher priority during the workflow's execution. We call them *hot metadata*. We present a hybrid decentralized/distributed model for handling hot metadata in *multisite* architectures. We couple our model with a scientific workflow management system (SWfMS) to validate and tune its applicability to various real-life scientific scenarios. We show that efficient management of hot metadata improves the performance of SWfMS, reducing the workflow execution time up to 50 % for highly parallel jobs by enabling timely data provisioning and avoiding unnecessary *cold* metadata operations.

7.3.2. Probabilistic optimizations for resource provisioning of cloud workflows

Participants: Chi Zhou, Shadi Ibrahim.

In many data-intensive applications, data management routines can be represented as workflows, where tasks are organized according to data and computation dependencies. Recently, the optimal provisioning of resources (e.g., VMs) for workflows running in the cloud has attracted a lot of attention. Most resource provisioning solutions overlook the important factor of cloud dynamics, e.g., the fluctuation of I/O, network performance, and system failures. In our experiments on the Amazon EC2 cloud, these issues significantly impact resource allocation quality. Therefore, we study how cloud dynamics should be incorporated into the resource provisioning process.

Our approach models cloud dynamics as time-dependent random variables (e.g., a probability distribution of workflow execution times) and performs probabilistic optimizations for resource provisioning problems using those random variables as optimization input. This solution yields more effective resource provisioning for cloud workflows. However, it involves heavy computation effort due to the complex structures of workflows and the large number of probability calculations.

To overcome this problem, we develop a three-stage pruning process, which simplifies workflow structure and reduces probability evaluation overhead. We have also implemented our techniques in a runtime library, which allows users to integrate our techniques into their existing resource provisioning methods. Experiments on two common resource provisioning problems show that probabilistic solutions can improve the performance by 51 % —70 % compared with state-of-the-art, static solutions.

Collaboration. *This work was done in collaboration with Bingsheng He NUS, Singapore.*

7.3.3. A taxonomy and survey of scientific computing in the cloud

Participants: Chi Zhou, Shadi Ibrahim.

Cloud computing has evolved as a popular computing infrastructure for many applications. With (big) data acquiring a crucial role in eScience, efforts have been made recently to develop and deploy scientific applications efficiently on the unprecedentedly scalable cloud infrastructures.

In [29], we review recent efforts in developing and deploying scientific computing applications in the cloud. In particular, we introduce a taxonomy specifically designed for scientific computing in the cloud, and further review the taxonomy with four major kinds of science applications, including life sciences, physics sciences, social and humanities sciences, and climate and earth sciences.

Due to the large data size in most scientific applications, the performance of I/O operations can greatly affect the overall performance of the applications. As a consequence, the dynamic I/O performance of the cloud has made resource provisioning an important and complex problem for scientific applications in the cloud.

We present our efforts on improving the resource provisioning efficiency and effectiveness of scientific applications in the cloud. Finally, we present the open problems for developing the next-generation eScience applications and systems in the cloud and give our conclusions.

Collaboration. *This work was done in collaboration with Bingsheng He NUS, Singapore.*

7.4. Fault tolerant data processing

7.4.1. Fast recovery

Participants: Orçun Yildiz, Shadi Ibrahim, Gabriel Antoniu.

Hadoop has emerged as a prominent tool for Big Data processing in large-scale clouds. Failures are inevitable in large-scale systems, especially in shared environments. Consequently, Hadoop was designed with hardware failures in mind. In particular, Hadoop handles machine failures by re-executing all the tasks of the failed machine. Unfortunately, the efforts to handle failures are entirely entrusted to the core of Hadoop and hidden from Hadoop schedulers. This may prevent Hadoop schedulers from meeting their objectives (e.g., fairness, job priority, performance) and can significantly impact the performance of the applications.

In our previous work, we addressed this issue through the design and implementation of a new scheduling strategy called Chronos. Chronos is conductive to improving the performance of Map-Reduce applications by enabling an early action upon failure detection. Chronos tries to launch recovery tasks immediately by preempting tasks belonging to low priority jobs, thus avoiding to wait until slots are freed.

In [20], we further investigated the potential benefit of launching local recovery tasks by implementing and evaluating Chronos*. To this end, we slightly changed the smart slot allocation strategy of Chronos into aggressive slot allocation strategy. With Chronos, recovery tasks with higher priority would preempt the selected tasks with less priority. With Chronos*, we also allow recovery tasks to preempt the selected tasks with the same priority (e.g., recovery tasks belonging to the same job with selected tasks). The experimental results indicate that Chronos* results in 100 % locality execution for recovery tasks thanks to its aggressive slot allocation strategy. Moreover, Chronos* improves the completion time of the jobs by up to 17 %.

7.4.2. Dynamic replica placement

Participants: Pierre Matri, Alexandru Costan, Gabriel Antoniu.

Large-scale applications are ever-increasingly geo-distributed. Maintaining the highest possible *data locality* is crucial to ensure high performance of such applications. Dynamic replication addresses this problem by dynamically creating replicas of frequently accessed data close to the clients. This data is often stored in decentralized storage systems such as Dynamo or Voldemort, which offer support for *mutable data*.

However, existing approaches to dynamic replication for such mutable data remain centralized, thus incompatible with these systems. We introduce a write-enabled dynamic replication scheme that leverages the decentralized architecture of such storage systems. We propose an algorithm enabling clients to locate tentatively the closest data replica without prior request to any metadata node. Large-scale experiments show a read latency decrease of up to 42% compared to other state-of-the-art, caching-based solutions.

Collaboration. *This work was done in collaboration with María Pérez, UPM, Spain.*

7.5. Advanced data management on clouds

7.5.1. Benchmarking Spark and Flink

Participants: Ovidiu-Cristian Marcu, Alexandru Costan, Gabriel Antoniu.

Spark and Flink are two Apache-hosted data analytics frameworks that represent the state of the art in modern in-memory Map-Reduce processing. They facilitate the development of multi-step data pipelines using directly acyclic graph (DAG) patterns. In the framework of our BigStorage project, we performed a comparative study [23] which evaluates the performance of Spark versus Flink. The objective is to identify and explain the impact of the different architectural choices and the parameter configurations on the perceived end-to-end performance.

Based on empirical evidences, the study points out that in Big Data processing there is not a single framework for all data types, sizes and job patterns and emphasize a set of design choices that play an important role in the behaviour of a Big Data framework: memory management, pipelined execution, optimizations and parameter configuration easiness. What raises our attention is that a streaming engine (i.e., Flink) delivers in many benchmarks better performance than a batch-based engine (i.e., Spark), showing that a more general Big Data architecture (treating batches as finite sets of streamed data) is plausible and may subsume both streaming and batching use cases.

Collaboration. *This work was done in collaboration with María Pérez, UPM, Spain.*

7.5.2. Geo-distributed graph processing

Participants: Chi Zhou, Shadi Ibrahim.

Graph processing is an emerging model adopted by a wide range of applications to easily parallelize the computations over graph data. Partitioning graph processing workloads to multiple machines is an important task for reducing the communication cost and improving the performance of graph processing jobs. Recently, many real-world applications store their data on multiple geographically distributed datacenters (DCs) to ensure flexible and low-latency services. Due to the limited Wide Area Network (WAN) bandwidths and the network heterogeneity of the geo-distributed DCs, existing graph partitioning methods need to be redesigned to improve the performance of graph processing jobs in geo-distributed DCs.

To address the above challenges, we propose a heterogeneity-aware graph partitioning method named G-Cut, which aims at minimizing the runtime of graph processing jobs in geo-distributed DCs while satisfying the WAN usage budget. G-Cut is a two-stage graph partitioning method. In the traffic-aware graph partitioning stage, we adopt the one-pass edge assignment to place edges into different partitions while minimizing the inter-DC data traffic size. In the network-aware partition refinement stage, we map the partitions obtained in the first stage onto different DCs in order to minimize the inter-DC data transfer time. We evaluate the effectiveness and efficiency of G-Cut using real-world graphs and the evaluation results show that G-Cut can achieve both lower WAN usage and shorter inter-DC data transfer time compared to state-of-the-art graph partitioning methods.

Collaboration. *This work was done in collaboration with Bingsheng He NUS, Singapore.*

7.5.3. Fairness and scheduling

Participants: Orçun Yildiz, Shadi Ibrahim, Gabriel Antoniu.

Recently, Map-Reduce and its open-source implementation Hadoop have emerged as prevalent tools for big data analysis in the cloud. Fair resource allocation in-between jobs and users is an important issue, especially in multi-tenant environments such as clouds. Several scheduling policies have been developed to preserve fairness in multi-tenant Hadoop clusters. At the core of these schedulers, simple (non-) preemptive approaches are employed to free resources for tasks belonging to jobs with less share. For example, Hadoop Fair Scheduler is equipped with two approaches: wait and kill. While wait may introduce a serious violation in fairness, kill may result in a huge waste of resources. Yet, recently some work have introduced preemption approach in shared Hadoop clusters.

To this end, we closely examine three approaches including wait, kill and preemption when Hadoop Fair Scheduler is employed for ensuring fair execution between multiple concurrent jobs. We perform extensive experiments to assess the impact of these approaches on performance and resource utilization while ensuring fairness. Our experimental results bring out the differences between these approaches and illustrate that these approaches are only sub-optimal for different workloads and cluster configurations: the efficiency of achieving fairness and the overall performance varies with the workload composition, resource availability and the cost of the adopted preemption technique.

7.5.4. Stragglers in Map-Reduce

Participants: Tien-Dat Phan, Shadi Ibrahim.

Big Data systems (e.g., Map-Reduce, Hadoop, Spark) rely increasingly on speculative execution to mask slow tasks also known as stragglers because a job's execution time is dominated by the slowest task instance. Big Data systems typically identify stragglers and speculatively run copies of those tasks with the expectation a copy may complete faster to shorten job execution times.

There is a rich body of recent results on straggler mitigation in Map-Reduce. However, the majority of these do not consider the problem of accurately detecting stragglers. Instead, they adopt a particular straggler detection approach and then study its effectiveness in terms of performance, e.g., reduction in job completion time, or its efficiency, e.g., extra resource usage.

In this work, we consider a complete framework for straggler detection and mitigation. We start with a set of metrics that can be used to characterizes and detect stragglers such as Precision, Recall, Detection Latency, Undetected Time and Fake Positive. We then develop an architectural model by which these metrics can be linked to measures of performance including execution time and system energy overheads.

We further conduct a series of experiments to demonstrate which metrics and approaches are more effective in detecting stragglers and are also predictive of effectiveness in terms of performance and energy efficiency. For example, our results indicate that the default Hadoop straggler detector could be made more effective. In certain cases, precision is low and only 65 % of those detected are actual stragglers and recall, i.e., the proportion of stragglers which are actually detected, is also relatively low at 56 %. For the same case, the hierarchical approach (i.e., a green-driven detector based on the default one) achieves a precision of 98 % and a recall of 33 %.

Further, these increases in precision can be used to achieve lower execution time and energy consumption, and thus higher performance and energy efficiency. Compared to the default Hadoop mechanism, energy consumption is reduced by almost 30 %. These results demonstrate how our framework can offer useful insights and be applied in practical settings to characterize and design new straggler detection mechanisms for Map-Reduce systems.

Collaboration. *This work was carried out in collaboration with Guillaume Aupy and Padma Raghavan whilst they were affiliated with Vanderbilt University, USA.*

LACODAM Team

7. New Results

7.1. Introduction

In this section, we organize our contributions over three main research topics:

- Mining different kinds of patterns, from 7.2 to 7.9
- Data mining and decision support with ASP, from 7.10 to 7.13
- Model-based diagnosis, from 7.13 to 7.15 .

7.2. Customer Purchase Signatures: a New Model in Grocery Retail Context

Participants: Clément Gautrais, Peggy Cellier [Lis], Thomas Guyet, René Quiniou, Alexandre Termier.

In the retail context, there is an increasing need for understanding individual customer behavior in order to personalize marketing actions. We propose the novel concept of customer signature, that identifies a set of important products that the customer refills regularly. Both the set of products and the refilling time periods give new insights on the customer behavior. Our approach is inspired by methods from the domain of sequence segmentation, thus benefiting from efficient exact and approximate algorithms. Experiments on a real massive retail dataset show the interest of the signatures for understanding individual customers (under submission to PAKDD 2017 conference).

This new model is used to detect and explain customer defection in a grocery retail context from the evolution of each customer basket content. It therefore provides actionable knowledge for the retailer at an individual scale. In addition, this model is able to identify customers that are likely to defect in the future months [16].

7.3. Discriminant Chronicles for Care Pathway Analysis

Participants: Yann Dauxais, Thomas Guyet, David Gross-Amblard [Druid], André Happe [Brest University Hospital].

A care pathway is a sequence of events (drugs deliveries, hospitalisation, etc) extracted from medical databases (see section 4.3 for details). In some studies, each patient is labeled by a class (*e.g.* died or not died). This information can be taken into account for the discriminant analysis of care pathways. This year, our objective was to extract discriminant patterns from a dataset of care pathways that can discriminate patients on their labels. To this end we introduced the new task of discriminant chronicle mining. Conceptually, a chronicle is a graph whose vertices are events and edges represent quantitative time constraints between events. We also proposed DCM, an algorithm dedicated to discriminant chronicles mining. This algorithm is based on rule learning methods to extract the temporal constraints. Computational performances and discriminant power of extracted chronicles are evaluated on artificial and real data.

The paper describing this work has been accepted in the french national conference on data mining (EGC 2017) [4] and is nominated for the best paper award.

7.4. Identifying Genetic Variant Combinations using Skypatterns

Participants: Alexandre Termier, Hoang-Son Pham [Genscale], Dominique Lavenier [Genscale].

Identifying variant combination association with disease is a bioinformatics challenge. This problem can be solved by discriminative pattern mining that uses a statistical function to evaluate the significance of individual biological patterns. There is a wide range of such measures. However, selecting an appropriate measure as well as a suitable threshold in some specific practical situations is a difficult task. In this work, we propose to use the skypattern technique which enables using combinations of measures to evaluate the importance of variant combinations without having to select a given measure and a fixed threshold (Pareto frontier). Experiments on several real variant datasets demonstrates that the skypattern method effectively identifies the risk variant combinations related to diseases [13].

7.5. Steady Patterns

Participants: Alexandre Termier, Willy Ugarte [UGA Grenoble], Miguel Santana [STMicroelectronics].

Skypatterns are an elegant answer to the pattern explosion issue, when a set of measures can be provided. Skypatterns for all possible measure combinations can be explored thanks to recent work on the skypattern cube. However, this leads to too many skypatterns, where it is difficult to quickly identify which ones are more important. First, we introduce a new notion of pattern steadiness [14] which measures the conservation of the skypattern property across the skypattern cube, allowing to see which are the "most universal" skypatterns. Then, we extended this notion to partitions of the dataset, and show in our experiments that this both allows to discover especially stable skypatterns, and identify interesting differences between the partitions.

7.6. Dense Bag-of-Temporal-SIFT-Words for Time Series Classification

Participants: Adeline Bailly [IRISA/Obelix], Laetitia Chapel [IRISA/Obelix], Thomas Guyet, Simon Malinowski [LinkMedia], Romain Tavenard [IRISA/Obelix].

The SIFT framework has shown to be effective in the image classification context. In [15], we designed a Bag-of-Words approach based on an adaptation of this framework to time series classification. It relies on two steps: SIFT-based features are first extracted and quantized into words; histograms of occurrences of each word are then fed into a classifier. In this work, we investigate techniques to improve the performance of Bag-of-Temporal-SIFT-Words: dense extraction of keypoints and different normalizations of Bag-of-Words histograms. Extensive experiments show that our method significantly outperforms nearly all tested standalone baseline classifiers on UCR datasets.

7.7. Comparing Symbolic and Statistical Classifiers on Energy Consumption Data

Participant: Benjamin Négrevergne.

During his Inria Carnot postdoc, Benjamin Négrevergne aimed at testing various data mining and machine learning methods on energy consumption data from the Energiency startup. Two symbolic methods developed in Lacodam were evaluated: QTempIntMiner and discriminant chronicle mining. While QTempIntMiner was shown to be ill-adapted in this setting, discriminant chronicle mining gave promising results. These results were evaluated in collaboration with our industrial partner. We also shown the interest of other methods: Hidden Markov Models and Gaussian processes. An internal report has been written to relate the results.

7.8. Detecting Strategic Moves in HearthStone Matches

Participants: Boris Doux [M1 intern], Clément Gautrais, Benjamin Negrevergne.

In this work, we demonstrate how to extract strategic knowledge from gaming data collected among players of the popular video game HearthStone. Our methodology is as follows. First we train a series of classifiers to predict the outcome of the game during a match, then we demonstrate how to spot key strategic events by tracking sudden changes in the classifier prediction. This methodology is applied to a large collection of HeathStone matches that we have collected from top ranked European players. Expert analysis shows that the events identified with this approach are both important and easy to interpret with the corresponding data [12].

7.9. Towards Visualizing Hidden Structures

Participants: Rémy Dautriche [STMicroelectronics], Alexandre Termier, Renaud Blanch [UGA Grenoble], Miguel Santana [STMicroelectronics].

There is an increasing need to quickly understand the contents of log data. A wide range of patterns can be computed and provide valuable information: for example existence of repeated sequences of events or periodic behaviors. However pattern mining techniques often produce many patterns that have to be examined one by one, which is time consuming for experts. On the other hand, visualization techniques are easier to understand, but cannot provide the in-depth understanding provided by pattern mining approaches. Our contribution is to propose a novel visual analytics method that allows to immediately visualize hidden structures such as repeated sets/sequences and periodicity, allowing to quickly gain a deep understanding of the log [3].

7.10. Knowledge-based Sequence Mining with ASP

Participants: Thomas Guyet, René Quiniou, Torsten Schaub.

We have introduced a framework for knowledge-based sequence mining, based on Answer Set Programming (ASP) [10], [5]. We begin by modeling the basic task and refine it in the sequel in several ways. First, we show how easily condensed patterns can be extracted by modular extensions of the basic approach. Second, we illustrate how ASP's preference handling capacities can be exploited for mining patterns of interest. In doing so, we demonstrate the ease of incorporating knowledge into the ASP-based mining process. To assess the trade-off in effectiveness, we provide an empirical study comparing our approach with a related sequence mining mechanism.

7.11. Packing Graphs with ASP for Landscape Simulation

Participants: Thomas Guyet, Yves Moinard, Jacques Nicolas [Dyliss], René Quiniou.

This work [6] describes an application of Answer Set Programming (ASP) to crop allocation for generating realistic landscapes. The task is to optimally cover a bare landscape, represented by its plot graph, with spatial patterns describing local arrangements of crops. This problem belongs to the hard class of graph packing problems and is modeled in the framework of ASP. The approach provides a compact and elegant solution to the basic problem and at the same time allows extensions such as a flexible integration of expert knowledge. Particular attention is paid to the treatment of symmetries, especially due to sub-graph isomorphism issues. Experiments were conducted on a database of simulated and real landscapes. Currently, the approach can process graphs of medium size, a size that enables studies on real agricultural practices.

7.12. Care Pathway Analysis with ASP Sequence Mining

Participants: Ahmed Samet, Benjamin Négrevergne, Thomas Guyet.

This line of work aims at applying our ASP encoding for sequential pattern mining to care pathway analysis (see section 4.3 for applicative objectives). This year, we proposed an approach of meaningful rare sequential pattern mining based on the declarative programming paradigm of Answer Set Programming (ASP). The setting of rare sequential pattern mining is introduced. To cope with the huge amount of meaningless rare patterns, our ASP approach provides an easy manner to encode expert constraints on expected patterns. Encodings are presented and quantitatively compared to a procedural baseline. An application on care pathways analysis illustrates the qualitative interest of expert constraints encoding.

This work has been submitted to the PAKDD 2017 conference.

7.13. ASP and Diagnosis

Participants: Christine Largouët, Laurence Rozé.

A new approach for performing diagnosis with ASP has been explored. The system is described by automata and implemented in an ASP program whose task is to find trajectories compatible with observations. The experimentation is carried out on benchmarks already used for the diagnosis problem using SAT. These benchmarks consider different levels of difficulty and number of faults (from one to twenty) and three types of observations: timestamped observations, totally ordered observations and partially ordered observations. The results were good both for dated and for totally ordered sequences of observations, whereas work needs to be still improved for the partial ordered observation case.

7.14. Searching for Cost-Optimized Strategies. Application to Temporal Planning and Agricultural System

Participants: Christine Largouët, Marie-Odile Cordier.

We consider a system modeled as a set of interacting components evolving along time according to explicit timing constraints. The decision making problem consists in selecting and organizing actions in order to reach a goal state in a limited time and in an optimal manner, assuming actions have a cost. We propose to reformulate the planning problem in terms of model-checking and controller synthesis such that the state to reach is expressed using a temporal logic. We have chosen to represent each agent using the formalism of Priced Timed Game Automata (PTGA) and a set of knowledge. PTGA is an extension of Timed Automata that allows the representation of cost on actions and the definition of a goal (to reach or to avoid). A first paper describes two algorithms designed to address the planning problem on a network of agents and proposes a practical implementation using model-checking tools that shows promising results on an agricultural application: a grassland based dairy production system [9]. Another paper describes the expressivity of this approach on the classical Transport Domain which is extended in order to include timing constraints, cost values and uncontrollable actions. This work has been implemented and performances evaluated on benchmarks [8].

7.15. Integrating Socio-Economic Drivers in an Explicit-Time, Qualitative Fisheries Model

Participant: Christine Largouët.

EcoMata is an explicit-time, qualitative modelling tool for assessing the ecosystem impacts of fishing and evaluating options for fishery management. The model is being developed further by integrating simple socio-economic drivers in the fishery system. Specifically, we have introduced a new module of automata that describes the profits associated to a specific fishing intensity and specific timing. This new module allows the evaluation of management strategies that are economically viable. The approach is illustrated on a coral-reef fishery in the Pacific that has been the focus of previous modelling work. [7].

LAGADIC Project-Team

7. New Results

7.1. Visual Perception

7.1.1. Micro/nano Manipulation

Participants: Le Cui, Eric Marchand.

Le Cui's Ph.D. [15] ended with a contribution related to visual tracking and estimation of the 3D pose of a micro/nano-object. It is indeed a key issue in the development of automated manipulation tasks using visual feedback. The 3D pose of the micro object can be estimated based on a template matching algorithm. Nevertheless, a key challenge for visual tracking in a scanning electron microscope (SEM) was the difficulty to observe the motion along the depth direction. We then proposed a template-based hybrid visual tracking scheme that uses luminance information to estimate the object displacement on x-y plane and uses defocus information to estimate object depth [54].

7.1.2. 3D Localization for Space Debris Removal

Participants: Aurélien Yol, Eric Marchand, François Chaumette.

This study is realized in the scope of the FP7 Removedebris project (see Section 9.3.1.1) [27]. We compared two vision-based navigation methods for tracking space debris in a low Earth orbit environment. The proposed approaches rely on a frame to frame model-based tracking in order to obtain the complete 3D pose of the camera with respect to the target [2]. The proposed algorithms robustly combine points of interest and edge features, as well as color-based features if needed. Experimental results have been obtained demonstrating the robustness of the approaches on synthetic image sequences simulating a CubeSat satellite orbiting the Earth [75].

7.1.3. 3D Localization for Airplane Landing

Participants: Noël Mériaux, François Chaumette, Patrick Rives, Eric Marchand.

This study is realized in the scope of the ANR VisioLand project (see Section 9.2.2). In a first step, we have considered and adapted our model-based tracker [2] to localize the aircraft with respect to the airport surroundings. Satisfactory results have been obtained from real image sequences provided by Airbus. In a second step, we are now considering to perform this localization from a set of keyframe images corresponding to the landing trajectory.

7.1.4. Scene Registration based on Planar Patches

Participants: Renato José Martins, Eduardo Fernandez Moral, Patrick Rives.

Image registration has been a major problem in computer vision over the past decades. It implies searching an image in a database of previously acquired images to find one (or several) that fulfill some degree of similarity, e.g. an image of the same scene from a similar viewpoint. This problem is interesting in mobile robotics for topological mapping, re-localization, loop closure and object identification. Scene registration can be seen as a generalization of the above problem where the representation to match is not necessarily defined by a single image (i.e. the information may come from different images and/or sensors), attempting to exploit all information available to pursue higher performance and flexibility. This problem is ubiquitous in robot localization and navigation. We propose a probabilistic framework to improve the accuracy and efficiency of a previous solution for structure registration based on planar representation. Our solution consists of matching graphs where the nodes represent planar patches and the edges describe geometric relationships. The maximum likelihood estimation of the registration is estimated by computing the graph similarity from a series of geometric properties (areas, angles, proximity, etc.) to maximize the global consistency of the graph. Our technique has been validated on different RGB-D sequences, both perspective and spherical [26].

7.1.5. Direct RGB-D Registration

Participants: Renato José Martins, Eduardo Fernandez Moral, Patrick Rives.

Dense direct RGB-D registration methods are widely used in tasks ranging from localisation and tracking to 3D scene reconstruction [7]. This work addresses a peculiar aspect which drastically limits the applicability of direct registration, namely the weakness of the convergence domain. In general, registration is performed only between close frames (small displacements), since dense registration tasks are particularly sensible to the local convexity of the cost error function. The main contribution of this work is an adaptive RGB-D error cost function that has a larger convergence domain and a faster convergence in both simulated and real data [67], [68]. This formulation employs the relative condition number metric to update the weighting of the RGB and depth costs. This approach is performed within a multi-resolution framework, where an efficient pixel selection for both RGB and ICP costs reduces the computational cost whilst preserving the precision. The formulation results in a larger region of attraction and faster convergence than classical RGB, ICP and RGB-D costs. Experiments was conducted using real sequences of indoor and outdoor images using perspective and spherical RGB-D sensors. Significant improvements was denoted in terms of the convergence stability and the speed of convergence in comparison with state-of-the-art methods.

7.1.6. Online localization and mapping for UAVs

Participants: Muhammad Usman, Paolo Robuffo Giordano, Eric Marchand.

Localization and mapping in unknown environments is still an open problem, in particular for what concerns UAVs because of the typical limited memory and processing power available onboard. In order to provide our quadrotor UAVs with high autonomy, we started studying how to exploit onboard cameras for an accurate (but fast) localization and mapping in unknown indoor environments. We chose to base both processes on the newly available Semi-Direct Visual Odometry (SVO) library (http://rpg.ifi.uzh.ch/software) which has gained considerable attention over the last years in the robotics community. The idea is to exploit dense images (i.e., with little image pre-processing) for obtaining an incremental update of the camera pose which, when integrated over time, can provide the camera localization (pose) w.r.t. the initial frame. In order to reduce drifts during motion, a concurrent mapping thread is also used for comparing the current view with a set of keyframes (taken at regular steps during motion) which constitute a "map" of the environment. We have started porting the SVO library to our UAVs and the preliminary results showed good performance of the localization accuracy against the Vicon ground truth. We are now planning to close the loop and base the UAV flight on the reconstructed pose from the SVO algorithm.

7.1.7. Reflectance and Illumination Estimation for Realistic Augmented Reality

Participants: Salma Jiddi, Eric Marchand.

The acquisition of surface material properties and lighting conditions is a fundamental step for photo-realistic Augmented Reality. Human visual cues remain sensitive to the global coherence within a computer-generated image. Absence or bad rendered virtual shadows, unconsidered specular reflections and/or occlusions, confused color perception such as an exuberantly bright virtual object are all elements which may not help an AR user interact and commit to a target application. In this work, we studied a new method for the estimation of the diffuse and specular reflectance properties of an indoor real static scene. Using an RGB-D sensor, we further estimate the 3D position of light sources responsible for specular phenomena and propose a novel photometry-based classification for all the 3D points. The resulting algorithm allows convincing AR results such as realistic virtual shadows as well as proper illumination and specularity occlusions [60].

7.1.8. Optimal Active Sensing Control

Participants: Paolo Salaris, Riccardo Spica, Paolo Robuffo Giordano.

This study concerns the problem of active sensing control. The objective is to improve the estimation accuracy of an observer by determining the inputs of the system that maximize the amount of information gathered by the outputs. In [9] this problem has been solved within the Structure from Motion (SfM) framework for 3D structure estimation problems, i.e. a point, a sphere and a cylinder, in the particular case where the observability property is instantaneously guaranteed. The optimal estimation strategy is hence given in terms of the instantaneous velocity direction of the camera velocities.

Recently, we have extended the optimal active sensing control to the case where the observability property is not instantaneously guaranteed. To simplify the analysis, we considered nonlinear differentially flat systems. Moreover, to quantify the richness of the acquired information the Observability Gramian (OG) has been used. We have hence defined a trajectory for the flat outputs of the system by using B-Spline curves and then, we have exploited an online gradient descent strategy to move the control points of such B-Spline in order to actively maximise the smallest eigenvalue of the OG over the whole fixed planning time horizon. While the system travels along its planned (optimized) trajectory, an Extended Kalman Filter (EKF) is used to estimate the system state. In order to keep memory of the past acquired sensory data for online re-planning, the OG is also computed on the past estimated state trajectories. This is then used for an online replanning of the optimal trajectory during the robot motion which is continuously refined by exploiting the estimated system state by the EKF. In order to show the effectiveness of our method we have considered a simple but significant case of a planar robot with a single range measurement. The simulation results show that, along the optimal path, the EKF converges faster and provides a more accurate estimate than along any other possible (non-optimal) paths. These results have been submitted to ICRA'2017.

7.2. Sensor-based Robot Control

7.2.1. Determining Singularity Configurations in IBVS

Participant: François Chaumette.

This theoretical study has been achieved through an informal collaboration with Sébastien Briot and Philippe Martinet from IRCCyN in Nantes, France. It concerns the determination of the singularity configurations of image-based visual servoing using tools from the mechanical engineering community and the concept of "hidden" robot. In a first step, we have revisited the welknown case of using three image points as visual feature, and then solved the general case of n image points [22]. The case of three image straight lines has also been solved for the first time [23].

7.2.2. Interval-based IBVS convergence domain computation

Participant: Vincent Drevelle.

This work aims to compute the set of camera poses from which IBVS will converge to the desired pose (that corresponds to the reference image). Starting from a (small) initial attraction domain of the desired pose (obtained using Lyapunov theory), we employ subpavings and guaranteed integration to iteratively increase the proven convergence domain, using a viability-based approach. Image-domain and pose-domain constraints are also enforced, like feature points visibility or workspace boundaries. First results have been obtained for a 3DOF line-scan camera IBVS case [56].

7.2.3. Visual Servoing of Humanoid Robots

Participants: Giovanni Claudio, Don Joven Agravante, Fabien Spindler, François Chaumette.

This study is realized in the scope of the BPI Romeo 2 and H2020 Comanoid projects (see Sections 9.2.7 and 9.3.1.2).

In a first step, we have considered classical kinematic visual servoing schemes for gaze control and manipulation tasks, such as can or box grasping. Two-hand manipulation has also been achieved using a master/slave approach [53], [81]. In a second step, we have designed the modeling of the visual features at the acceleration level to embed visual tasks and visual constraints in an existing QP controller [20][80]. Experimental results have been obtained on Romeo (see Section 6.9.4).

7.2.4. Model Predictive Visual Servoing

Participants: Nicolas Cazy, Paolo Robuffo Giordano, François Chaumette.

This study is realized in collaboration with Pierre-Brice Wieber, from Bipop group at Inria Rhône Alpes.

Model Predictive Control (MPC) is a powerful control framework able to take explicitly into account the presence of constraints in the controlled system (e.g., actuator saturations, sensor limitations, and so on). In this research activity, we studied the possibility of using MPC for tackling one of the most classical constraints of visual servoing applications, that is, the possibility to lose tracking of features because of occlusions, limited camera field of view, or imperfect image processing/tracking. The MPC framework depends upon the possibility to predict the future evolution of the controlled system over some time horizon, for correcting the current state of the modeled system whenever new information (e.g., new measurements) become available. We have also explored the possibility of applying these ideas in a multi-robot collaboration scenario where a UAV with a downfacing camera (with limited field of view) needs to provide localization services to a team of ground robots [13].

7.2.5. Model Predictive Control for Visual Servoing of a UAV

Participants: Bryan Penin, Riccardo Spica, François Chaumette, Paolo Robuffo Giordano.

Visual servoing is a welknown class of techniques meant to control the pose of a robot from visual input by considering an error function directly defined in the image (sensor) space. These techniques are particularly appealing since they do not require, in general, a full state reconstruction, thus granting more robustness and lower computational loads. However, because of the quadrotor underaction and inherent sensor limitations (mainly limited camera field of view), extending the classical visual servoing framework to the quadrotor flight control is not straightforward. For instance, for realizing a horizontal displacement the quadrotor needs to tilt in the desired direction. This tilting, however, will cause any downlooking camera to point in the opposite direction with, e.g., possible loss of feature tracking because of the limited camera field of view.

In order to cope with these difficulties and achieve a high-performance visual servoing of quadrotor UAVs, we are exploring the possibility of using techniques borrowed from Model-Predictive Control (MPC) for explicitly dealing with this kind of constraints during flight. Indeed, MPC is a class of (numerical) optimal control techniques able to explicitly take into account state and input constraints, as well as complex (and underactuated) nonlinear dynamics of the controlled system. In particular, the ability to predict, over some future time window, the behavior of the visual features on the image plane will allow the quadrotor to fly "blindly" for some limited phases, for then regaining tracking of any lost feature. This possibility will be crucial for allowing quick maneuvering guided by a direct visual feedback. We have started addressing the case of a simulated planar UAV as a representative case study, and we are now working towards an experimental validation with a real quadrotor UAV equipped with an onboard camera.

7.2.6. Visual-based shared control

Participants: Firas Abi Farraj, Nicolò Pedemonte, Paolo Robuffo Giordano.

This work concerns our activities in the context of the RoMaNS H2020 project (see Section 9.3.1.3 . Our main goal is to allow a human operator to be interfaced in an intuitive way with a two-arm system, one arm carrying a gripper (for grasping an object), and the other one carrying a camera for looking at the scene (gripper + object) and providing the needed visual feedback. The operator should be allowed to control the two-arm system in an easy way for letting the gripper approaching the target object, and she/he should also receive force cues informative of how feasible her/his commands are w.r.t. the constraints of the system (e.g., joint limits, singularities, limited camera fov, and so on).

We have started working on this topic by proposing a shared control architecture in which the operator could provide instantaneous velocity commands along four suitable task-space directions not interfering with the main task of keeping the gripper aligned towards the target object (this main task was automatically regulated). The operator was also receiving force cues informative of how much her/his commands were conflicting with the system constraints, in our case joint limits of both manipulators. Finally, the camera was always moving so as to keep both the gripper and the target object at two fixed locations on the image plane [46].

We have then extended this framework in two directions: first, by allowing the possibility of controlling a whole future trajectory for both arms (gripper+camera) while coping with the system constraints. The operator was then receiving an 'integral' force feedback along the whole planned trajectory: in this way, the operator's actions and the corresponding force cues were function of a planned trajectory (thus, carrying information over a future time window) that could be manipulated at runtime. Second, we studied how to integrate learning from demonstration into our framework by first using learning techniques for extracting statistical regularities of 'expert users' executing successful trajectories for the gripper towards the target object. Then, these learned trajectories were used for generating force cues able to guide novice users during their teleoperation task by the 'hands' of the expert users who demonstrated the trajectories in the first place. Both works have been submitted to ICRA'2017.

7.2.7. Direct Visual Servoing

Participants: Quentin Bateux, Eric Marchand.

In the direct visual servoing methods such as photometric framework, the images as a whole are used to define the control law. This can be opposed to the classical visual servoing approaches that relies on geometric features and where image processing algorithms that extract and track visual features are necessary. In [21], we proposed a generic framework to consider histograms as visual features. A histogram is an estimate of the probability distribution of a variable (for example the probability of occurrence in an intensity, color, or gradient orientation in an image). We demonstrated that the framework we proposed applies, but is not limited to, a wide set of histograms and allows the definition of efficient control laws.

Nevertheless, the main drawback for the direct visual servoing class of methods comparing to the classical geometric visual servoing methods is their comparatively limited convergence range. We then proposed in [48] a new direct visual servoing control law that relies on a particle filter to perform non-local and non-linear optimization in order to increase the convergence domain. To each particle considered we associate a virtual camera that predicts the image it should capture by using image transfer techniques. This new control law has been validated on a 6 DOF positioning task performed on our Gantry robot (see Section 6.9.1).

7.2.8. Audio-based Control

Participants: Aly Magassouba, François Chaumette.

This study is concerned with the application of sensor-based control approach to audio sensors. It is made in collaboration with Nancy Bertin from Panama group at Irisa and Inria Rennes-Bretagne Atlantique. Auditory features such as Interaural Time Difference (ITD), Interaural Level Difference (ILD), and sound energy have been modeled and integrated in various control schemes to control the motion of a mobile robot with two microphones onboard [66], [64]. Experiments with Romeo and Pepper (see Section 6.9.4) have also been achieved [65]. They show the robustness of closed loop sensor-based control with respect to coarse modeling and that explicit sound source localization is not a mandatory step for aural servoing.

7.3. Medical Robotics

7.3.1. Non-rigid Target Tracking in Ultrasound Images

Participants: Lucas Royer, Alexandre Krupa.

We pursued our work concerning the development of a real-time approach that allows tracking deformable soft tissue structures in 3D ultrasound sequences. In previous work we proposed a method which consists in estimating the target deformation by combining robust dense motion estimation and mechanical model simulation. This year we improved the robustness of our method to several image artefacts as the presence of large shadows, local illumination changes and image occlusions that occur due to the modification of the imaging gain and re-orientation of the ultrasound beam induced by probe motion. To achieve this, we proposed a new dissimilarity criterion between the current and reference images based on the Sum of Conditional Variance (SCV). Our new criterion, that we named Sum of Confident Conditional Variance (SCCV), consists in discriminating unconfident voxels thanks to the use of a pixel-wise quality measurement of the ultrasound images. This improved approach was experimentally validated on organic soft tissues and the obtained results were published in [40].

7.3.2. Optimization of Ultrasound Image Quality by Visual Servoing

Participants: Pierre Chatelain, Alexandre Krupa.

This study is realized in collaboration with Prof. Nassir Navab from the Technical University of Munich (TUM).

In previous work, we have developed ultrasound-based visual servoing methods to fulfill various tasks, such as compensating for physiological motion, maintaining the visibility of an anatomic target during ultrasound probe teleoperation, or tracking a surgical instrument. However, due to the specific nature of ultrasound images, guaranteeing a good image quality during the procedure remains a challenge. Therefore we pursued our study on the use of ultrasound confidence maps as a new modality for automatically positioning an ultrasound probe in order to improve the image quality. In addition to our visual servoing approach that optimizes the global quality of the image, this year we proposed a control fusion to optimize the acoustic window for a specific anatomical target which is tracked in the ultrasound images [50]. Recently, we extended our confidence-driven control to the out-of-plane motion of a 3D ultrasound probe and experimentally validated it on a human volunteer at TUM [14].

7.3.3. Visual Servoing using Shearlet Transform

Participants: Lesley-Ann Duflot, Alexandre Krupa.

In collaboration with the Femto-ST lab in Besançon, we proposed in a first-hand a solution to reduce the acquisition time of an Optical Coherence Tomography (OCT) 3D imaging scanner. This latter consists in sweeping a laser beam on a tissue sample of interest. To increase the frame rate of this imaging device we proposed to apply an optimal trajectory to the laser that covers entirely the image but without performing all the OCT measurements. The reconstruction of the missing data is then achieved by applying an updated Fast Iterative Soft-Thresholding Algorithm (FISTA) on a sparse representation of the image that is based on the shearlet transform [57]. In a second hand, we studied the feasibility of using the subsampled shearlet coefficients of an ultrasound image as the visual features of an image-based visual servoing. In a preliminary study we estimated numerically the interaction matrix that links the variation of the shearlet coarsest coefficients to the 6 degrees of freedom motion of the ultrasound probe and uses it in the visual servoing framework. The results obtained in cases of automatic probe positioning and phantom motion compensation demonstrated the efficiency of the shearlet-based features in terms of accuracy, repeatability, robustness and convergence behavior [59]. Then we proposed to consider a more efficient and adequate shearlet implementation that consists in a non-subsampled representation of the image. In this case the shearlet coefficients represent different images, focused on different singularities of the initial image, and we consider directly their pixel intensity values in the visual feature vector similarly to the photometry-based visual servoing approach. The modeling of the interaction matrix was analytically derived and experimental results demonstrated the reliability of the new method and its robustness to speckle noise [58].

7.3.4. 3D Steering of Flexible Needle by Ultrasound Visual Servoing

Participants: Jason Chevrie, Marie Babel, Alexandre Krupa.

The objective of this work is to provide robotic assistance during needle insertion procedures such as biopsy or ablation of localized tumor. In the past we only considered the control of the insertion and needle rotation along and around its main axis by the use of a duty-cycling control strategy. This latter consists in adapting online from visual feedback the orientation of a beveled-tip flexible needle during its insertion for controlling the needle curvature in 3D space that is induced by asymmetrical forces exerted on the bevel. However, such strategy limits the workspace of the needle tip. Therefore we proposed a new control method for flexible needle steering that combines direct base manipulation and needle tip based control. The direct base manipulation control is generated thanks to the use of a 3D model of a flexible beveled tip needle that gives the adequate motion of the needle base to obtain a given motion of the needle tip. This 3D model is based on virtual springs that characterize the needle mechanical interaction with soft tissue and is adapted online from visual tracking of the needle shape. From this model, a measure of the controllability of the needle tip degrees of freedom was proposed in order to mix the control between the direct base manipulation and the duty cycling technique

[51]. Preliminary results of an automatic needle tip positioning in a translucent gelatine phantom, observed by 2 orthogonal cameras, demonstrated the feasibility of the combination between direct base manipulation and needle tip control for reaching a desired target. This hybrid control allows better targeting capabilities in terms of larger needle workspace and reduced needle bending. In order to predict the trajectory of a needle during insertion under lateral motion of the tissue, we also improved our 3D model of the flexible needle to take into account the effect of the motion of the tissues on the needle shape. This was achieved thanks to the design of an algorithm based on an unscented Kalman filter that estimates the tissue motion. Results obtained from several needle insertions in a moving soft tissue phantom showed that our model gives good performance in terms of needle trajectory prediction. This model was also considered in a closed-loop control approach to allow automatic reaching of a target in case of tissue lateral displacement [52]. Future work will address the consideration of 3D ultrasound as visual feedback.

7.3.5. Enhancement of Ultrasound Elastography by Visual Servoing and Force Control

Participants: Pedro Alfonso Patlan Rosales, Alexandre Krupa.

Elastography imaging is performed by applying continuous stress variation on soft tissues in order to estimate a strain map of the observed tissues. It is obtained by estimating, from the RF (radio-frequency) signal along each scan line of the probe transducer, the echo time delays between pre- and post-compressed tissue. Usually, this continuous stress variation is performed manually by the user who manipulates the US probe and it results therefore in an user-dependent quality of the elastography image. To improve the ultrasound elastography imaging and provide quantitative measurement, we developed an assistant robotic palpation system that automatically moves a 2D ultrasound probe for optimizing ultrasound elastography [70]. The main originality of this work is the use of the elastography modality directly as input of the robot controller. Force measures are also considered in the probe control in order to automatically induce soft tissue deformation needed for real-time elastography imaging process.

7.4. Navigation of Mobile Robots

7.4.1. Visual Navigation from an Image Memory

Participants: Suman Raj Bista, Paolo Robuffo Giordano, François Chaumette.

This study is concerned with visual autonomous navigation in indoor environments. As in our previous works concerning navigation outdoors [4], the approach is based on a topological localization of the current image with respect to a set of keyframe images, but the visual features used for this localization as well as for the visual servoing are not composed of points of interest, but either on mutual information [71] following the idea proposed in [3], or straight lines that are more common indoors [38], or finally on a combination of points of interest and straight lines [11]. Satisfactory experimental results have been obtained using the Pioneer mobile robot (see Section 6.9.2).

7.4.2. Robot-Human Interactions during Locomotion

Participant: Julien Pettré.

In collaboration with the Gepetto team of Laas in Toulouse and the Mimetic group in Rennes, we have studied how humans avoid collision with a robot. Understanding how humans achieve such avoidance is crucial to better anticipate humans' reactions to the presence of a robot and to control the robot to adapt its trajectory accordingly. It is generally assumed that humans avoid a robot just like they avoid another human. In this work, we bring the empirical demonstration that humans actually set a specific strategy to avoid robots, and that, more precisely, they show a preference to give way to a robot which is on a collision course with them [41]. This results brings useful insight about human-robot interactions during locomotion, and provides useful guidelines to design reactive navigation techniques for mobile robots aimed at moving among humans.

7.4.3. Scene Mapping based on Intelligent Human/Robot Interactions

Participant: Patrick Rives.

For mobile robots to operate in compliance with human presence, interpreting the impact of human activities and responding constructively is a challenging goal. Towards this objective, mapping an environment allows robots to be deployed in diverse workspaces, marking this skill as a primary element in the integration of robots into human-populated environments. We proposed an effective approach for using human activity cues in order to enhance robot mapping and navigation and in particular in filtering noisy human detections, detecting passages, inferring space occupancy and allowing navigation within unexplored areas. Our contributions [36] are based on the development of intelligent interactions among conceptually different mapping levels, namely, the metric, social and semantic levels. Experiments, using the Hannibal platform (see Section 6.9.2), highlighted a number of strong dependences among these levels and the way in which they can be used to enhance individual performances and in turn the global robot operation.

7.4.4. Autonomous Social Navigation of a Wheelchair

Participants: Vishnu Karakkat Narayanan, Marie Babel.

This work is realized in collaboration with Anne Spalanzani (Chroma team - Inria Grenoble).

A key issue that hinders the adoption of assistive robotic technologies such as robotized wheelchair, in the real world, is that they need to operate in mostly human environments and among human crowds. Indeed intelligent wheelchairs need to be deployed in a human environment thereby making it essential for such robots to incorporate a sense of human-awareness. Simply put, humans are special objects that have to be perceived and acted on in a special manner by robots that interact with us humans. Thus one can define Human-aware Navigation as an intersection between human-robot interaction and robotic motion planning.

In this context we introduced a low-level velocity controller that could be employed by a social robot like a robotic wheelchair for approaching a group of interacting humans, in order to become a part of the interaction. Taking into account an interaction space that is created when at least two humans interact, a meeting point can be calculated where the robot should reach in order to equitably share space among the interacting group. We then proposed a sensor-based control law which uses the position and orientation of the humans with respect to the sensor as inputs, to reach the meeting point while respecting spatial social constraints [61]. Experiments using a mobile robot equipped with a single laser scanner, realized in collaboration with Ren Luo (Taiwan) within the Sampen Inria associated team, also proved the success of the algorithm in a noisy real world scenario [62].

In addition, a semi-autonomous framework for human-aware navigation in an intelligent wheelchair has been designed. A generalized linear control sharing framework was proposed that was able to progressively correct the user teleoperation in order to avoid obstacles and in order to avoid disturbance to humans. Meanwhile, we also proposed a Bayesian approach for user intention estimation. The formulation was partly inferred from the design of the controller for assisted doorway passing, wherein we hypothesized that predicting short term goals is sufficient for eliminating user intention uncertainty [16].

7.4.5. Semi-autonomous Control of a Wheelchair for Navigation Assistance

Participants: Louise Devigne, Vishnu Karakkat Narayanan, Marie Babel.

To address the wheelchair driving assistance issue, we proposed a unified shared control framework able to smoothly correct the trajectory of the electrical wheelchair [16]. The system integrates the manual control with sensor-based constraints by means of a dedicated optimization strategy. The resulting low-complex and low-cost embedded system is easily plugged onto on-the-shelf wheelchairs [55]. The robotic solution is currently under validation process with volunteering patients of Pôle Saint Hélier (France) who present different disabling neuro-pathologies preventing them to drive non-assisted wheelchairs.

Within the frame of ISI4NAVE associated team (see Section 9.4.1.2), this shared-control solution has been then coupled with first experimental biofeedback devices such as haptic devices. Preliminary tests have been conducted within the PAMELA facility at University College of London and within the rehabilitation center of Pôle Saint Hélier in Rennes (see Section 8.1.3). They involved regular wheelchair users as well as medical staff. We have demonstrated the ability of the framework to provide relevant assistance and now need to focus on methods to fine-tune parameters and customize/calibrate to the individual and evolving needs of each user.

7.5. Multi-robot and Crowd Motion Control

7.5.1. Advanced multi-robot control and estimation

Participant: Paolo Robuffo Giordano.

The challenge of coordinating the actions of multiple robots is inspired by the idea that proper coordination of many simple robots can lead to the fulfillment of arbitrarily complex tasks in a robust (to single robot failures) and highly flexible way. Teams of multi-robots can take advantage of their number to perform, for example, complex manipulation and assembly tasks, or to obtain rich spatial awareness by suitably distributing themselves in the environment. Within the scope of robotics, autonomous search and rescue, firefighting, exploration and intervention in dangerous or inaccessible areas are the most promising applications.

In the context of multi-robot (and multi-UAV) coordinated control, *connectivity* of the underlying graph is perhaps the most fundamental requirement in order to allow a group of robots accomplishing common goals by means of *decentralized* solutions. In fact, graph connectivity ensures the needed continuity in the data flow among all the robots in the group which, over time, makes it possible to share and distribute the needed information. We gave two contributions in this field: in the first one [35], we proposed a decentralized exploration strategy for a team of 3D agents able to guarantee exploration of a finite space in a finite amount of time while coping with the constraints of a connected sensing/communication graph for the robot group against sensing/communication constraints (limited range, occluded line-of-sight), and of obstacle and inter-robot collision avoidance. The strategy exploits a suitable state machine for assigning dynamic roles to the agents in the group for allowing completion of the exploration in finite time. Second, in [28] we studied how the choice of a leader agent in a leader-follower scenario could affect the performance of the group when tracking a desired formation (shape and gross motion). The proposed strategy allows selecting the "best leader" online as a function of the current group state (relative positions and velocities) and of the group topology (assumed connected). By cycling among several connected topologies during motion, we could show that our proposed leader selection algorithm provides the best performance among other possible choices (including the random one) while coping with the constraint of a connected (but possibly time-varying) topology.

These works were realized in collaboration with the robotics group at the Max Planck Institute for Biological Cybernetics, Tübingen, Germany, and the RIS group at Laas in Toulouse.

7.5.2. Rigidity-based methods for formation control

Participants: Fabrizio Schiano, Riccardo Spica, Andrea Peruffo, Paolo Robuffo Giordano.

Most multi-robot applications must rely on *relative sensing* among the robot pairs (rather than absolute/external sensing such as, e.g., GPS). For these systems, the concept of *rigidity* provides the correct framework for defining an appropriate sensing and communication topology architecture. Rigidity is a combinatorial theory for characterizing the "stiffness" or "flexibility" of structures formed by rigid bodies connected by flexible linkages or hinges. In a broader context, rigidity turns out to be an important architectural property of many multi-agent systems when a common inertial reference frame is unavailable. Applications that rely on sensor fusion for localization, exploration, mapping and cooperative tracking of a target, all can benefit from notions in rigidity theory. The concept of rigidity, therefore, provides the theoretical foundation for approaching decentralized solutions to the aforementioned problems using distance measurement sensors, and thus establishing an appropriate framework for relating system level architectural requirements to the sensing and communication capabilities of the system.

In the recent past, we have proposed a decentralized gradient-based rigidity maintenance action for a group of quadrotor UAVs [10]. By starting in a rigid configuration, the group of UAVs was able to estimate their relative position from sole relative distance measurements, and then use these estimated relative positions in a control action able to preserve rigidity of the whole formation despite presence of sensor limitations (maximum range and line-of-sight occlusions), possible collisions with obstacles and inter-robot collisions. This (rigidity-based) control/estimation framework has now been extended to the case of *bearing rigidity* for directed graphs: here, rather than distances the measurements are the 3D bearing vectors expressed in the local body-frame of each agent. The theory has been extended to the case of 3D agents evolving in $\mathbb{R}^3 \times \mathbb{S}^1$ by proposing a decentralized

bearing controller/localization algorithm that only requires one single distance measurement (among an arbitrary pair of agents) for a correct convergence [72]. The proposed algorithm ensures stabilization towards a desired bearing formation, and allows for the possibility of actuating the motion directions in the null-space of the bearing constraints (that is, collective translations in 3D, expansion/retraction, and coordinated rotation about a vertical axis).

The need of a single distance measurement (for fixing the formation scale) has also been relaxed in [73] where an *active* scale estimation scheme has been proposed for allowing the (distributed) estimation of the various inter-agent distances online by processing the measured bearings and the known agent ego-motion (body-frame linear and angular velocities). Finally, we have also proposed an extension of the "distance" rigidity maintenance controller proposed in [10] to the case of bearing measurements (and bearing rigidity), by considering the typical sensing constraints of onboard cameras, that is, limited range, limited field of view, of possible mutual occlusions when two or more agents lie on the same line-of-sight. This work has been experimentally validated with 5 quadrotor UAVs, and has been submitted to ICRA'2017.

These works were realized in collaboration with the RIS group at Laas, Toulouse, and with Technion, Israel.

7.5.3. Cooperative localization using interval analysis

Participants: Ide Flore Kenmogne Fokam, Vincent Drevelle.

In the context of multi-robot fleets, cooperative localization consists in gaining better position estimate through measurements and data exchange with neighboring robots. Positioning integrity (i.e., providing reliable position uncertainty information) is also a key point for mission-critical tasks, like collision avoidance. The goal of this work is to compute position uncertainty volumes for each robot of the fleet, using a decentralized method (i.e., using only local communication with the neighbors). The problem is addressed in a bounded-error framework, with interval analysis and constraint propagation methods. These methods enable to provide guaranteed position error bounds, assuming bounded-error measurements. They are not affected by over-convergence due to data incest, which makes them a well sound framework for decentralized estimation. Ongoing work focuses on position uncertainty domain computation in image-based UAV localization [63], and its extension to cooperative localization in a multi-UAV fleet.

7.5.4. Numerical Models of Local Interactions during Locomotion

Participants: Julien Bruneau, Panayiotis Charalambous, David Wolinski, Julien Pettré.

The numerical models of local interactions are core components of reactive navigation techniques (which allows a robot to avoid dynamic obstacles) and of microscopic crowd simulation algorithms (which allows to simulate a crowd motion as a collection of agent trajectories). We have pursued our efforts to design local models of interactions which capture humans pedestrian behavior, to simulate how they adapt their trajectory so as to perform interactions with their neighbors [12]. This year, our efforts were focused on the simulation of grouping behaviors [39], and mid-term strategies human set to perform energy-efficient sequences of successive avoidance adaptations [24]. These two situations deal with complex situations of interactions, where several interactions of different kinds need to be combined to compute agents trajectories. For example, when moving in groups, agents have to keep close to the other members of their group while they should not collide with them, as well as they should avoid collision with any other agent or obstacle out of this group.

We also revisited the foundation of velocity-based models of local interaction for collision avoidance. Using a velocity-based model, a collision-free motion is computed for one agent by extrapolating the future motion of neighbor agents with respect to their current position and velocity. From this information, each agent can deduce the set of velocities (called admissible velocities) that lead to a collision-free motion in the near future. The extrapolation is generally simply based on a linear extrapolation of the future position along the current velocity vector. This is simplistic as it assumes that the current velocity vector is representative of the future motion, while it is often false when, for instance, the agent is currently performing adaptations due to ongoing collision avoidance, or when the agent is following a curvy path. To improve the accuracy of motion prediction and the resulting simulation, we have introduced a probabilistic representation of future position, that can be

computed from a set of context elements such as the layout of the environment or the agents past motion [42]. We demonstrate in this work the high impact on the level of realism of resulting simulations. This work is implemented in the WarpDriver software (see Section 6.7).

Finally, we address applications of our simulators to the Computer Animation. Crowd simulation agents generally have a simplistic geometrical and kinematics models, typically, an oriented 2D circle moving on a flat surface. In Computer Animation, an animation of a crowd of 3D realistic characters can be computed on top of the agents simulation by computing their internal joints trajectories so as to perform walking motion along computed agents trajectories. However, the discrepancies between the 2D model of agents and 3D full body characters may result into residual collisions between character shapes. In this collaboration with the Mimetic team, we demonstrate that simple secondary animations for characters, such as local shoulder motions, can be efficiently triggered to camouflage those artefacts, with a very low computational overhead [29].

7.5.5. Motion Planning for Digital Characters

Participant: Julien Pettré.

Motion planning is an important component for agents and robot navigation and control, providing them the ability to perform geometrical reasoning over their environment to transform a high-level distant goal in their environment into a sequence of local motions and sub-goals to reach. This year, we have been involved into two collaborations dealing with motion planning. First collaboration was with the University of Utrecht in the Netherlands. We have proposed a method to evaluate and compare various environment decomposition techniques [74]. Environment decomposition is an important step to perform navigation planning in large static environments. Second collaboration was with the University of North Carolina in Chapel Hill (see Section 9.4.1.1). We have coupled a contact planner for virtual characters with ITOMP, a motion optimization technique to achieve complex motion in cluttered environment [69].

LINKMEDIA Project-Team

7. New Results

7.1. Unsupervised motif and knowledge discovery

7.1.1. Multimodal person discovery in TV broadcasts

Participants: Guillaume Gravier, Gabriel Sargent, Ronan Sicre.

Work in collaboration with Silvio J. Guimarães, Gabriel B. de Fonseca and Izabela Lyon Freire, PUC Minas, in the framework of the Inria Associate Team MOTIF.

Pursuing efforts initiated in 2015 in the framework of the MediaEval benchmark on Multimodal Person Discovery, we investigated graph-based approaches to name the persons on screen and speaking in TV broadcasts with no prior information, leveraging text overlays, speech transcripts as well as face and voice comparison. We adopted a graph-based representation of speaking faces and investigated two tag-propagation approaches to associate overlays co-occurring with some speaking faces to other visually or audiovisually similar speaking faces. Given a video, we first build a graph from the detected speaking faces (nodes) and their audiovisual similarities (edges). Each node is associated to its co-occurring overlays (tags) when they exist. Then, we consider two tag-propagation approaches, respectively based on a random walk strategy and on Kruskal's minimum spanning tree algorithm for node clustering [28].

7.1.2. Efficient similarity self-join for near-duplicate video detection

Participants: Laurent Amsaleg, Guillaume Gravier.

Work in collaboration with Henrique B. da Silva, Silvio J. Guimarães, Zenilto do Patrocino Jr., PUC Minas, and Arnaldo de A. Araújo, UFMG, in the framework of the Inria Associate Team MOTIF.

The huge amount of redundant multimedia data, like video, has become a problem in terms of both space and copyright. Usually, the methods for identifying near-duplicate videos are neither adequate nor scalable to find pairs of similar videos. Similarity self-join operation could be an alternative to solve this problem in which all similar pairs of elements from a video dataset are retrieved. Methods for similarity self-join however exhibit poor performance when applied to high-dimensional data. In [33], we propose a new approximate method to compute similarity self-join in sub-quadratic time in order to solve the near-duplicate video detection problem. Our strategy is based on clustering techniques to find out groups of videos which are similar to each other.

7.1.3. Recommendation systems with matrix factorization

Participants: Raghavendran Balu, Teddy Furon.

Matrix factorization is a prominent technique for approximate matrix reconstruction and noise reduction. Its common appeal is attributed to its space efficiency and its ability to generalize with missing information. For these reasons, matrix factorization is central to collaborative filtering systems. In the real world, such systems must deal with million of users and items, and they are highly dynamic as new users and new items are constantly added. Factorization techniques, however, have difficulties to cope with such a demanding environment. Whereas they are well understood with static data, their ability to efficiently cope with new and dynamic data is limited. Scaling to extremely large numbers of users and items is also problematic. In [10], we propose to use the count sketching technique for representing the latent factors with extreme compactness, facilitating scaling.

In [11], we discovered that sketching techniques implicitly provide differential privacy guarantees thanks to the inherent randomness of the data structure. Collaborative filtering is a popular technique for recommendation system due to its domain independence and reliance on user behavior data alone. But the possibility of identification of users based on these personal data raise privacy concerns. Differential privacy aims to minimize these identification risks by adding controlled noise with known characteristics. The addition of noise impacts the utility of the system and does not add any other value to the system other than enhanced privacy.

7.2. Multimedia content description and structuring

7.2.1. Hierarchical topic structuring

Participants: Guillaume Gravier, Pascale Sébillot.

In [37], we investigated the potential of a topical structure of text-like data that we recently proposed [55] in the context of summarization and anchor detection in video hyperlinking. This structure is produced by an algorithm that exploits temporal distributions of words through word burst analysis to generate a hierarchy of topically focused fragments. The obtained hierarchy aims at filtering out non-critical content, retaining only the salient information at various levels of detail. For the tasks we choose to evaluate the structure on, the lost of important information is highly damaging. We show that the structure can actually improve the results of summarization or at least maintain state-of-the-art results, while for anchor detection it leads us to the best precision in the context of the Search and Anchoring in Video Archives task at MediaEval. The experiments were carried on written text and a more challenging corpus containing automatic transcripts of TV shows.

7.2.2. Multimedia-inspired descriptors for time series classification

Participant: Simon Malinowski.

The SIFT framework has shown to be effective in the image classification context. Recently, we designed a bag-of-words approach based on an adaptation of this framework to time series classification. It relies on two steps: SIFT-based features are first extracted and quantized into words; histograms of occurrences of each word are then fed into a classifier. In [38], we investigated techniques to improve the performance of bag-of-temporal-SIFT-words: dense extraction of keypoints and different normalizations of Bag-of-Words histograms. Extensive experiments have shown that our method significantly outperforms nearly all tested standalone baseline classifiers on publicly available UCR datasets. In [23], we also investigate the use of convolutional neural networks (CNN) for time series classification. Such networks have been widely used in many domains like computer vision and speech recognition, but only a little for time series classification. We have designed a convolutional neural network that consists of two convolutional layers. One drawback with CNN is that they need a lot of training data to be efficient. We propose two ways to circumvent this problem: designing data-augmentation techniques and learning the network in a semi-supervised way using training time series from different datasets. These techniques are experimentally evaluated on a benchmark of time series datasets.

7.2.3. Early time series classification

Participant: Simon Malinowski.

In time series classification, two antagonist notions are at stake. On the one hand, in most cases, the sooner the time series is classified , the higher the reward. On the other hand, an early classification is more likely to be erroneous. Most of the early classification methods have been designed to take a decision as soon as a sufficient level of reliability is reached. However, in many applications, delaying the decision with no guarantee that the reliability threshold will be met in the future can be costly. Recently, a framework dedicated to optimizing the trade-off between classification accuracy and the cost of delaying the decision was proposed, together with an algorithm that decides online the optimal time instant to classify an incoming time series. On top of this framework , we have built in [29] two different early classification algorithms that optimize the trade-off between decision accuracy and the cost of delaying the decision. These algorithms are non-myopic in the sense that, even when classification is delayed, they can provide an estimate of when the optimal classification time is likely to occur. Our experiments on real datasets demonstrate that the proposed approaches are more robust than existing methods.

7.3. Content-based information retrieval

7.3.1. Bi-directional embeddings for cross-modal content matching

Participants: Guillaume Gravier, Christian Raymond, Vedran Vukotić.

Common approaches to problems involving multiple modalities (classification, retrieval, hyperlinking, etc.) are early fusion of the initial modalities and crossmodal translation from one modality to the other. Recently, deep neural networks, especially deep autoencoders, have proven promising both for crossmodal translation and for early fusion via multimodal embedding. In [31], we propose a flexible cross-modal deep neural network architecture for multimodal and crossmodal representation. By tying the weights of two deep neural networks, symmetry is enforced in central hidden layers thus yielding a multimodal representation space common to the two original representation spaces. The proposed architecture is evaluated in multimodal query expansion and multimodal retrieval tasks within the context of video hyperlinking. In [32], we extend the approach, focusing on the evaluation of a good single-modal continuous representations both for textual and for visual information. word2vec and paragraph vectors are evaluated for representing collections of words, such as parts of automatic transcripts and multiple visual concepts, while different deep convolutional neural networks are evaluated for directly embedding visual information, avoiding the creation of visual concepts. We evaluate methods for multimodal fusion and crossmodal translation, with different single-modal pairs, in the task of video hyperlinking.

7.3.2. Intrinsic dimensions in language information retrieval

Participant: Vincent Claveau.

Examining the properties of representation spaces for documents or words in information retrieval (IR) brings precious insights to help the retrieval process. Recently, several authors have studied the real dimensionality of the datasets, called intrinsic dimensionality, in specific parts of these spaces. In [34], we propose to revisit this notion through a coefficient called α in the specific case of IR and to study its use in IR tasks. More precisely, we show how to estimate α from IR similarities and to use it in representation spaces used for documents and words. Indeed, we prove that α may be used to characterize difficult queries. We moreover show that this intrinsic dimensionality notion, applied to words, can help to choose terms to use for query expansion.

7.3.3. Evaluation of distributional thesauri

Participants: Vincent Claveau, Ewa Kijak.

With the success of word embedding methods, all the fields of distributional semantics have experienced a renewed interest. Beside the famous word2vec, recent studies have presented efficient techniques to build distributional thesaurus, including our work on information retrieval (IR) tools and concepts to build a thesaurus [14]. In [13], we address the problem of the evaluation of such thesauri or embedding models. Several evaluation scenarii are considered: direct evaluation through reference lexicons and specially crafted datasets, and indirect evaluation through a third party tasks, namely lexical subsitution and Information Retrieval. Through several experiments, we first show that the recent techniques for building distributional thesaurus outperform the word2vec approach, whatever the evaluation scenario. We also highlight the differences between the evaluation scenarii, which may lead to very different conclusions when comparing distributional models. Last, we study the effect of some parameters of the distributional models on these various evaluation scenarii.

7.3.4. Scaling group testing similarity search

Participants: Laurent Amsaleg, Ahmet Iscen, Teddy Furon.

The large dimensionality of modern image feature vectors, up to thousands of dimensions, is challenging high dimensional indexing techniques. Traditional approaches fail at returning good quality results within a response time that is usable in practice. However, similarity search techniques inspired by the group testing framework have recently been proposed in an attempt to specifically defeat the curse of dimensionality. Yet, group testing does not scale and fails at indexing very large collections of images because its internal procedures analyze an excessively large fraction of the indexed data collection. In [16], we identify these difficulties and proposes extensions to the group testing framework for similarity searches that allow to handle larger collections of feature vectors. We demonstrate that it can return high quality results much faster compared to state-of-the-art group testing strategies when indexing truly high-dimensional features that are indeed hardly indexable with traditional indexing approaches.

We also discovered that group testing helps in enforcing security and privacy in identification. We detail a particular scheme based on embedding and group testing. Whereas embedding poorly protects the data when used alone, the group testing approach makes it much harder to reconstruct the data when combined with embedding. Even when curious server and user collude to disclose the secret parameters, they cannot accurately recover the data. Our approach reduces as well the complexity of the search and the required storage space. We show the interest of our work in a benchmark biometrics dataset [17], where we verify our theoretical analysis with real data.

7.3.5. Large-scale similarity search using matrix factorization

Participants: Ahmet Iscen, Teddy Furon.

Work in collaboration with Michael Rabbat, McGill University, Montréal.

We consider the image retrieval problem of finding the images in a dataset that are most similar to a query image. Our goal is to reduce the number of vector operations and memory for performing a search without sacrificing accuracy of the returned images. In [18], we adopt a group testing formulation and design the decoding architecture using either dictionary learning or eigendecomposition. The latter is a plausible option for small-to-medium sized problems with high-dimensional global image descriptors, whereas dictionary learning is applicable in large-scale scenarios. Experiments with standard image search benchmarks, including the Yahoo100M dataset comprising 100 million images, show that our method gives comparable (and sometimes better) accuracy compared to exhaustive search while requiring only 10 % of the vector operations and memory. Moreover, for the same search complexity, our method gives significantly better accuracy compared to approaches based on dimensionality reduction or locality sensitive hashing.

7.4. Linking, navigation and analytics

7.4.1. Opinion similarity and target extraction

Participants: Vincent Claveau, Grégoire Jadi.

Work in collaboration with Laura Monceaux and Béatrice Daille, LINA, Nantes.

In [19], we propose to evaluate the lexical similarity information provided by word representations against several opinion resources using traditional information retrieval tools. Word representation have been used to build and to extend opinion resources, such as lexicon and ontology, and their performance have been evaluated on sentiment analysis tasks. We question this method by measuring the correlation between the sentiment proximity provided by opinion resources and the semantic similarity provided by word representations using different correlation coefficients. We also compare the neighbors found in word representations and list of similar opinion words. Our results show that the proximity of words in state-of-the-art word representations is not very effective to build sentiment similarity.

In [20], we present the development of an opinion target extraction system in English and transpose it to the French language. In addition, we realize an analysis of the features and their effectiveness in English and French which suggest that it is possible to build an opinion target extraction system independant of the domain. Finally, we propose a comparative study of the errors of our systems in both English and French and propose several solutions to these problems.

7.4.2. Reinformation and fake detection in social networks

Participants: Vincent Claveau, Ewa Kijak, Cédric Maigrot.

Traditional media are increasingly present on social networks, but these usual sources of information are confronted with other sources called reinformation sources. These last ones sometimes tend to distort the information relayed to match their ideologies, rendering it partially or totally false. In [25], we conduct a study pursuing two goals: first, we present a corpus containing Facebook messages issued from both types of media sources; secondly, we propose some experiments in order to automatically detect reinformation messages. In particular, we investigate the influence of shallow features versus features more specifically describing the message content. We also developed a multi-modal hoax detection system composed of text, source, and image analysis [24]. As hoax can be very diverse, we want to analyze several modalities to better detect them. This system is applied in the context of the Verifying Multimedia Use task of MediaEval 2016. Experiments show the performance of each separated modality as well as their combination.

7.4.3. Multimodal video hyperlinking

Participants: Rémi Bois, Guillaume Gravier, Christian Raymond, Pascale Sébillot, Ronan Sicre, Vedran Vukotić.

Pursuing previous work on video hyperlinking and recent advances in multimodal content matching [32], we benchmarked a full video hyperlinking system in the framework of the TRECVid international benchmark [12]. The video hyperlinking task aims at proposing a set of video segments, called targets, to complement a query video segment defined as anchor. The 2016 edition of the task encouraged participants to use multiple modalities. In this context, we chose to submit four runs in order to assess the pros and cons of using two modalities instead of a single one and how crossmodality differs from multimodality in terms of relevance. The crossmodal run performed best and obtained the best precision at rank 5 among participants. In parallel, we also demonstrated that, in this framework, multimodal and crossmodal approaches offer significantly more diversity in the set of target proposed than classical information retrieval based approaches where all modalities are combined. We compared bidirectionnal multimodal embeddings [31] with multimodal LDA approaches as experimented last year in TRECVid [49]. The former offers more accurate matching, the latter exhibiting slighlty more diveristy.

7.4.4. User-centric evaluation of hyperlinked news content

Participants: Rémi Bois, Guillaume Gravier, Pascale Sébillot, Arnaud Touboulic.

Work in collaboration with Éric Jamet, Martin Ragot and Maxime Robert, CRPCC, Rennes.

Following our study of professional user needs in multimedia news analytics [15], we developed a prototype news analytics interface that facilitates the exploration of collections of multimedia documents by journalists. The application, based on standard web technology, enriches classical functionalities for this type of applications (e.g., keyword highlights, named entity detection, keyword search, etc.) with navigation-based functionalities. The latter exploit a graph-based organization of the collection, established from content-based similarity graphs on which community detection is performed along with basic link characterization. We performed usage tests on students in journalism and on journalists where each user was asked to write a synthesis article on a given topic. Preliminary results indicate that the graph-based navigation improves the completeness of the synthesis by exposing users to more content than with a standard search engine.

7.5. Miscellaneous

In parallel with mainstream research activities, LINKMEDIA has a number of contribtions in other domains based on the expertise of the team members.

7.5.1. Bidirectional GRUs in spoken dialog

Participants: Christian Raymond, Vedran Vukotić.

Recurrent neural networks recently became a very popular choice for spoken language understanding (SLU) problems. They however represent a big family of different architectures that can furthermore be combined to form more complex neural networks. In [30], we compare different recurrent networks, such as simple recurrent neural networks, long short-term memory networks, gated memory units and their bidirectional versions, on the popular ATIS dataset and on MEDIA, a more complex French dataset. Additionally, we propose a novel method where information about the presence of relevant word classes in the dialog history is combined with a bidirectional gated recurrent unit (GRU).

7.5.2. Kernel principal components analysis with extreme learning machines

Participant: Christian Raymond.

Work in collaboration with M'Sila University, Algeria.

Nowadays, wind power and precise forecasting are of great importance for the development of modern electrical grids. In [26], we investigate a prediction system for time series based on kernel principal component analysis (KPCA) and extreme learning machine (ELM). Comparison with standard dimensionality reduction techniques show that the reduction of the original input space affects positively the prediction output.

7.5.3. Pronunciation adaptation for spontaneous speech synthesis

Work in collaboration with Gwénolé Lecorvé and Damien Lolive, IRISA, Rennes.

In [36], we present a new pronunciation adaptation method which adapts canonical pronunciations to a spontaneous style. This is a key task in text-to-speech as those pronunciation variants bring expressiveness to synthetic speech, thus enabling new potential applications. The strength of the method is to solely rely on linguistic features and to consider a probabilistic machine learning framework, namely conditional random fields, to produce the adapted pronunciations.

7.6. Participation in benchmarking initiatives

- Video hyperlinking, TRECVid
- Search and anchoring, Mediaeval Multimedia International Benchmark
- Multimodal person discovery in broadcast TV, Mediaeval Multimedia International Benchmark
- DeFT 2015 text-mining challenge

MIMETIC Project-Team

7. New Results

7.1. Outline

In 2016, MimeTIC pursued its efforts in improving virtual human simulation by initiating new projects in this domain, such as the Inria PRE with CAIRN team, and recruiting Antonio Mucherino in Inria half-delegation. Our main goal is to provide more natural human motion in real-time applications, which is a transversal requirement in many of MimeTIC's research domains.

- In Biomechanics, being able to rapidly simulate plausible human motion enables to explore new approaches to provide real-time feedback to users in many application domains, such a rehabilitation, sports training, ergonomics and industrial training.
- In computer graphics, simulating natural motion either relies on heavy mechanical simulation and optimal control or adapting motion capture data. We wish to push dynamic simulation a step forward to propose new biomechanically-based simulation, such as actuating the virtual human with muscles instead of rotating servos. We also wish to simplify the process of retargeting motion capture data, which is a process still difficult to automatize. In both cases, we also promote the idea of understanding how human perception behaves when facing inaccurate simulation, in order to provide accurate simulations only when necessary.
- In virtual reality, real-time motion capture and simulation are essential when using head mounted display devices as users cannot perceive their own body during immersive experiences. Hence, simulating natural avatar motion and reacting efficiently to the user's actions are key points to ensure good Presence and Embodiment. MimeTIC is collaborating with other teams in VR, such as Hybrid, to address this complex pluridisciplinary question.
- In digital storytelling, interactive autonomous virtual characters lever the potentiality of proposing complex stories on social and human themes. More stories are now created with the goal of proposing several interactive storylines, which massively enhances the possibilities of interactive entertainment, computer games and digital applications. Projects in MimeTIC explore for instance how to provide a seamless control of the balance between the autonomy of characters and the unfolding of the story through the narrative discourse.

Hence, the organization of the results is reflecting these main challenges in motion analysis, virtual human simulation, interaction in VR, and digital storytelling.

7.2. Motion Analysis

In motion analysis, we continued designing new approaches to measure human performance in specific applications, such as clinical gait assessment, ergonomics and sports. We also developed an original approach to concurrently analyze and simulate human motion, by addressing the problem of redundancy in musculoskeletal models.

7.2.1. Clinical gait assessment based on Kinect data

Participant: Franck Multon.

In clinical gait analysis, we proposed a method to overcome the main limitations imposed by the low accuracy of the Kinect measurements in real medical exams. Indeed, inaccuracies in the 3D depth images lead to badly reconstructed poses and inaccurate gait event detection. In the latter case, confusion between the foot and the ground leads to inaccuracies in the foot-strike and toe-off event detection, which are essential information to get in a clinical exam. To tackle this problem we assumed that heel strike events could be indirectly estimated by searching for the extreme values of the distance between the knee joints along the walking longitudinal axis. As Kinect sensor may not accurately locate the knee joint, we used anthropometrical data to select a body point located at a constant height where the knee should be in the reference posture. Compared to previous works using a Kinect, heel strike events and gait cycles are more accurately estimated, which could improve global clinical gait analysis frameworks with such a sensor. Once these events are correctly detected, it is possible to define indexes that enable the clinician to have a rapid state of the quality of the gait. We therefore proposed a new method to assess gait asymmetry based on depth images, to decrease the impact of errors in the Kinect joint tracking system. It is based on the longitudinal spatial difference between lower-limb movements during the gait cycle. The movement of artificially impaired gaits was recorded using both a Kinect placed in front of the subject and a motion capture system. The proposed longitudinal index distinguished asymmetrical gait, while other symmetry indices based on spatiotemporal gait parameters failed using such Kinect skeleton measurements. This gait asymmetry index measured with a Kinect is low cost, easy to use and is a promising development for clinical gait analysis.

This method has been challenged with other classical approaches to assess gait asymmetry using either cheap Kinect data or Vicon data. We demonstrate the superiority of the approach when using Kinect data for which traditional approaches failed to accurately detect gait asymmetry. It has been validated on healthy subjects who were forced to walk with a 5cm sole placed below each foot alternatively [2].

This work has been done in collaboration with the MsKLab from Imperial College London, to design new gait asymmetry indexes that could be used in daily clinical analysis.

7.2.2. New automatic methods to assess motion in industrial contexts based on Kinect

Participants: Franck Multon, Pierre Plantard.

Recording human activity is a key point of many applications and fundamental works. Numerous sensors and systems have been proposed to measure positions, angles or accelerations of the user's body parts. Whatever the system is, one of the main challenge is to be able to automatically recognize and analyze the user's performance according to poor and noisy signals. Hence, recognizing and measuring human performance are important scientific challenges especially when using low-cost and noisy motion capture systems. MimeTIC has addressed the above problems in two main application domains. In this section, we detail the ergonomics application of such an approach.

Firstly, in ergonomics, we explored the use of low-cost motion capture systems (i.e., a Microsoft Kinect) to measure the 3D pose of a subject in natural environments, such as on a workstation, with many occlusions and inappropriate sensor placements. Predicting the potential accuracy of the measurement for such complex 3D poses and sensor placements is challenging with classical experimental setups. After having evaluated the actual accuracy of the pose reconstruction method delivered by the Kinect, we have identified that occlusions were a very important problem to solve in order to obtain reliable ergonomic assessments in real cluttered environments. To this end, we extended previous correction methods proposed by Hubert Shum (Northumbria University) which consist in identifying the reliable and unreliable parts of the Kinect skeleton data, and to replace unreliable ones by prior knowledge recorded in a database. In collaboration with Hubert Shum, we extended this approach to deal with long occlusions that occur in real manufacturing conditions. To this end we proposed a new data structure named Filtered Pose Graph to speed-up the process, and select example poses that improve the quality of the correction, especially ensuring continuity. We have demonstrated a significant increase of the quality of the correction, especially when large tracking errors occur with the Kinect system [16].

This method has been applied to a complete ergonomic process outputting RULA scores based on the reconstructed and corrected poses. We also demonstrated that it delivers new ergonomic information compared

to traditional approaches based on isolated pictures: it provides time spent above a given RULA score which is a valuable information to support decision in ergonomics [15]. We also challenged this method with a reference motion capture system in laboratory conditions. In order to evaluate the actual use in ergonomics, we also compared the ergonomic scores obtained with this automatic method to two experts' scores in real factories. The results show very good agreements between automatic and manual assessments, and have been published in Applied Ergonomics journal [25].

This work was partially funded by the Faurecia company through a Cifre convention.

7.2.3. Evaluation and analysis of sports gestures: application to tennis serve

Participants: Richard Kulpa, Marion Morel, Benoit Bideau, Pierre Touzard.

Following the previous studies we made on tennis serve, we were able to evaluate the link between performance and risk of injuries. To go further, we made new experiments to quantify the influence of fatigue on the performance of tennis serve, that is to say the kinematic, kinetic and performance changes that occur in the serve throughout a prolonged tennis match play [12], [13]. To this end, we recorded serves of several advanced tennis players with a motion capture system before, at mid-match, and after a 3-hour tennis match. Before and after each match, we also recorded electromyographic data of 8 upper limb muscles obtained during isometric maximal voluntary contraction. These experiments showed a decrease in mean power frequency values for several upper limb muscles that is an indicator of local muscular fatigue. Decreases in serve ball speed, ball impact height, maximal angular velocities and an increase in rating of perceived exertion were also observed between beginning and end of match. However, no change in timing of maximal angular velocities was observed. The consistency in timing of maximal angular velocities suggests that advanced tennis players are able to maintain the temporal pattern of their serve technique, in spite of the muscular fatigue development [12]. Moreover, we showed that passive shoulder internal rotation and total range of motion are significantly decreased during a 3-hour tennis match that is identified as an injury risk factor among tennis players [13].

Overall, automatically evaluating and quantifying the performance of a player is a complex task since the important motion features to analyze depend on the type of performed action. But above all, this complexity is due to the variability of morphologies and styles of both novices and experts (who perform the reference motions). Only based on a database of experts' motions and no additional knowledge, we propose an innovative 2-level DTW (Dynamic Time Warping) approach to temporally and spatially align the motions and extract the imperfections of the novice's performance for each joint. We applied our method on tennis serves and karate katas [22].

7.2.4. Interactions between walkers

Participants: Anne-Hélène Olivier, Armel Crétual, Julien Bruneau, Richard Kulpa, Sean Lynch, Laurentius Meerhoff, Julien Pettré.

Interaction between people, and especially local interaction between walkers, is a main research topic of MimeTIC. We propose experimental approaches using both real and virtual environments to study both perception and action aspects of the interaction. This year, we developed new experiments in our immersive platform. In the context of Sean Lynch's PhD on the visual perception of human motion during interactions in locomotor tasks, we designed a study to investigate whether local limb motion is required to successfully avoid a single dynamic obstacle or if global motion alone provides sufficient information (Figures 4 .a and 4 .b). Sixteen healthy subjects were immersed in a virtual environment that required navigating towards a target, whilst an obstacle crossed its path. Within the virtual environment, four occluding walls prevented the subject observing the complete environment at the initiation of movement, ensuring steady state was reached prior to obstacle interaction. The velocity and heading of the obstacle were programmed to result in a range of future crossing distance (varying from 0.1 to 1.2m) in front and behind the subject. The velocity and heading of the obstacle were fixed, and the subject used a joystick to control its orientation to avoid collision. Five obstacle appearances were presented in a randomized order; a full body (control condition), trunk- or legs- only (i.e., local motion only), and a cylinder or sphere representing the center of gravity (COG) (i.e., global motion only). No significant difference for obstacle appearance was found on number of collisions. However, in both

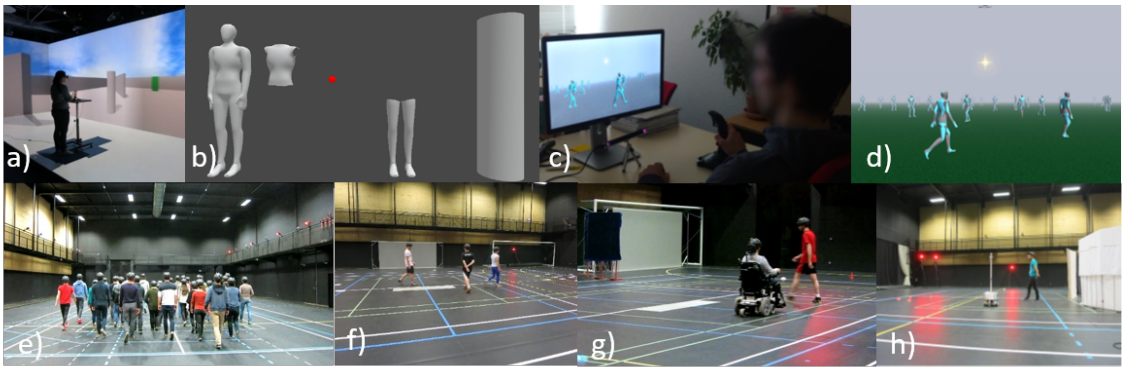


Figure 4. Experiments performed to investigate interactions between walkers.

global motion only conditions, subjects adopted alternative collision avoidance strategies compared to the full body control condition. Distance regulation and collision avoidance within daily activities may be principally regulated by global rather than local motion. Underlying mechanisms may differ accordingly to shape and size, however there is no impediment for successful completion of collision avoidance.

Second, we provide lot of efforts to investigate the complex case of multiple interactions while in previous studies we mainly focused on pairwise interactions. We developed a new experiment using an eye tracker to provide insight about the selection process of the interactions (Figures 4 .c and 4 .d). We proposed to study the human gaze during a navigation task in a crowded virtual environment. The characteristics of each virtual agent was known and controlled. Then, by recording the gaze activity, we are able to highlight the characteristics of each agent the participant was looking at. Results first showed a strong link between the fixated agents and the trajectory adaptations of the participants which means that participants looked at agents they are interacting with, which is an important result to validate the use of the eye tracker in such a situation. Concerning the characteristics of the fixated agents, results showed that human gaze, during navigation, is attracted by dangerous individuals: they were the ones presenting the higher risk of future collision with the participants. Future work is needed to evaluate the influence of other factors such as walking speed or the nature of agents trajectories. This year, we also performed an important experimental campaign including 80 participants to investigate collective behavior (Figure 4 .e). When people walk together in the street, they have to coordinate their own motion with the ones of their neighbors. From these local interactions, group motion emerges. The objective of this study was to understand how a collective behavior can emerge from these local interactions between individuals. Especially, the study aimed at identifying what is the neighborhood of a walker in a group from a perceptual point of view (who influences your motion). This work was performed in collaboration with William Warren (Brown University, Providence) and Cécile Appert-Rolland (CNRS, Orsay). Data analysis is still in process but from these results we hope to develop new knowledge on pedestrian behavior. These new results will help us to design new or improve existing crowd simulators based on local interactions. These simulators have important economic and societal roles. For example, they allow to validate the design of public places/building, which aims at hosting dense levels of public in perfectly safe conditions. The study of multiple interactions was also strengthened with the arrival of Laurentius Meerhoff as a post-doctoral student with a regional SAD funding in May 2016. Experiments involving 3 walkers were conducted (Figure 4 .f). We investigated how collision is avoided in small groups of people and whether people can successfully interact with the whole environment, or whether under some circumstance agents had to resort to sequential treatment. We proposed a method to detect whether the treatment was sequential or simultaneous and we showed the initial relative position between walkers strongly affects how interaction is engaged with.

Third, we started working on the interaction between a walker and a person on a motorized wheelchair (Figure 4 .g). This work was performed in collaboration with the Inria Lagadic team. The main objective was to design a control law that allows the wheelchair to automatically navigate in a crowded place without any collision. This is important for people who have difficulties to drive their wheelchair because of cognitive impairments. However, before reaching this objective, some steps are required to understand how walkers and persons on a wheelchair interact together. To this end, we developed a study where we recorded the trajectory of walkers and a person on a wheelchair in a collision avoidance and reaching scenario. Results will help to model such a control law for natural interactions.

Finally, we continue working on the interaction between a walker and a moving robot. This work was performed in collaboration with Philippe Souères and Christian Vassallo (LAAS, Toulouse). The development of Robotics accelerated these recent years, it is clear that robots and humans will share the same environment in a near future. In this context, understanding local interactions between humans and robots during locomotion tasks is important to steer robots among humans in a safe manner. Our work is a first step in this direction. Our goal is to describe how, during locomotion, humans avoid collision with a moving robot. We just published in Gait and Posture our results on collision avoidance between participants and a non-reactive robot (we wanted to avoid the effect of a complex loop by a robot reacting to participants' motion). Our objective was to determine whether the main characteristics of such interaction preserve the ones previously observed: accurate estimation of collision risk, anticipated and efficient adaptations. We observed that collision avoidance between a human and a robot has similarities with human-human interactions (estimation of collision risk, anticipation) but also leads to major differences [18]. Humans preferentially give way to the robot, even if this choice is not optimal with regard to motion adaptation to avoid the collision. In this new study, we considered the situation where the robot was reactive to the walker's motion (Figure 4 .h). First of all, it results that humans have a good understanding of the robot behavior and their reaction are smoother and faster with respect to the case with a non-collaborative robot. Second, humans adapt similarly to human-human study and the crossing order is respected in almost all cases. These results have strong similarities with the ones observed with two humans crossing each other.

7.2.5. Biomechanics for motion analysis-synthesis

Participants: Charles Pontonnier, Georges Dumont, Ana Lucia Cruz Ruiz, Antoine Muller, Diane Haering.

In the context of Ana Lucia Cruz Ruiz's PhD, whose goal is to define and evaluate muscle-based controllers for avatar animation, we developed an original control approach to reduce the redundancy of the musculoskeletal system for motion synthesis, based on the muscle synergy theory. For this purpose we ran an experimental campaign of overhead throwing motions. We recorded the muscle activity of 10 muscles of the arm and the motion of the subjects. Thanks to a synergy extraction algorithm, we extracted a reduced set of activation signals corresponding to the so called muscle synergies and used them as an input in a forward dynamics pipeline. Thanks to a two stage optimization method, we adapted the model's muscle parameters and the synergy signals to be as close as possible of the recorded motion. The results are compelling and ask for further developments [5]. We also proposed a classification about muscle-based controllers for animation that has been published in Computer Graphics Forum [6]. Ana Lucia defended her thesis on December 2nd, 2016.

We are also developing an analysis pipeline thanks to the work of Antoine Muller. This pipeline aims at using a modular and multiscale description of the human body to let users be able to analyse human motion. For now, the pipeline is able to assemble different biomechanical models in a convenient descriptive graph, to calibrate these models thanks to experimental data and to compute inverse dynamics to get joint torques from experimental motion capture data. We also investigated the capacity of motion-based methods to calibrate body segment inertial parameters in characterizing the part of the residuals due to the kinematical error into the dynamical one [23].

We also begin to work on the determination of maximal torque envelopes of the elbow thanks to the work of Diane Haering, Inria Post-doctoral fellow at MimeTIC. These results have a great potential of application i) to quantify the articular load during work tasks and ii) to help calibrating muscle parameters for musculoskeletal simulations. Preliminary results have been presented to an international biomechanics conference [21].

Finally, in collaboration with the CERAH (Centre d'étude et d'appareillage des handicapés, institut des invalides, Créteil, France), we proposed an identification-based method for knee prosthesis characteristics. The method is based on a forward dynamics framework enabling a matching between experimental data and model behavior [26].

7.3. Virtual Human Simulation

In addition to this last contribution on biomechanically-inspired character simulation, at the crossroad between motion analysis and simulation, we also explored two main directions for virtual human simulation in 2016. Firstly, with the arrival of Antonio Mucherino in the team, we pushed the idea of extending the idea of interaction meshes (introduced in 2010 by Taku Komura in Edinburgh) to model the constraints intrinsically associated with the motion. This approach requires developing new distance geometry algorithms in order to take time and rigid body constraints into account. Secondly, we continued to push the idea of using perceptual studies to efficiently adapt simulation in order to save computation time for less important details.

Julien Pettré moved to the Lagadic Inria team in March 2016. However we continue collaborating with him on crowd simulation problems, e.g., developing models related to interactions between pedestrians and designing perceptual studies to improve the realism of simulations.

7.3.1. Recent advances in discretizable distance geometry

Participants: Antonio Mucherino, Ludovic Hoyet, Franck Multon.

Since September 2016, Antonio Mucherino has a half-time Inria detachment in the MimeTIC team, in order to collaborate on exploring distance geometry-based problems in representing and editing human motion. In collaboration with various French and international partners, he has been working on the different facets of the discretization of the distance geometry. In 2016, he has mainly focused on the two following points. Firstly, since the discretization assumptions require the existence of a vertex ordering on the graph G which is used for representing a problem instance, he presented a new algorithm for the automatic detection of vertex orders that are also able to optimize a given set of objectives [7]. With the aim of making its exploration more efficient, the idea is to reduce in size the search space obtained with the discretization, while keeping in its interior the entire solution set. Secondly, he has started to investigate the possibility to extend the distance geometry (and its discretization) to a wider range of applications, by studying the overlaps between two different geometrical applications, arising in two different domains [3].

More related to the integration with the work in MimeTIC, we are currently exploring applying distance geometry approaches to other applications of interest for virtual human simulations, such as human motion editing and retargeting, and crowd simulations.

7.3.2. Perception of Secondary Motions in Crowd Scenarios

Participants: Ludovic Hoyet, Anne-Hélène Olivier, Richard Kulpa, Julien Pettré.

Creating plausible virtual character animations is of importance in topics researched in MimeTIC, especially for interactive applications where balancing realism and computational load is a requisite. Recently, we investigated how to improve realism of virtual crowd animations by exploring the effects of introducing secondary shoulder motions at the animation level. Typically, a crowd engine pipeline animates numerous moving characters according to a two-step process. First, a crowd simulator generates the characters' global 2D displacement trajectories in the environment, then an animation engine transforms these global trajectories into full body motions. This two-step decomposition is interesting for computational reasons, as crowd simulators raise quadratic complexity issues by nature. For the sake of simplicity, simulation models are often limited to 2D moving circles with 3 degrees of freedom (DoF), i.e., two translations and a rotation. The complete set of internal trajectories (30 to 60 DoF per character) is then considered at the animation step only, where characters are processed independently. This two-step process avoids combining the complexity of crowd simulators with the dimensionality of character kinematic models. However, it also leads to the notion of interactions between characters to be considered only at the simulation level, and to be lost at the animation level. Body animations are therefore not influenced by the presence of neighbours, only global trajectories are.

Final animations therefore often lead to residual collisions and/or characters walking as if they were alone, showing no sign to the influence of others.

In this work, we investigated the value of adding secondary motions on the perceived visual quality of crowd animations (i.e., perceived residual collisions and animation naturalness). We focused on adding shoulder motions to characters passing at close distances, and explored this question through two perceptual experiments. To understand the effects of shoulder motions on walking interactions, we first focused on understanding how these secondary motions affect how viewers perceive local interactions between two characters. We found that shoulder motions have strong positive effects on the visual quality of two-character animations, where such animations are perceived to be significantly more natural, and residual collisions become significantly less perceptible. Then we evaluated the benefits of displaying shoulder motions in the situation of crowded scenes, where shoulder motions are diluted into much more visually complex animations, and demonstrated positive effects on the animation naturalness. This increase of visual quality is obtained at a very low computational overhead, which demonstrates the relevance of the direction explored by our work. Our general conclusion is that adding secondary motions in character interactions has a significant impact on the visual quality of crowd animations, with a very light impact on the computational cost of the whole animation pipeline. Our results advance crowd animation techniques by enhancing the simulation of complex interactions between crowd characters with simple secondary motion triggering techniques.

These results were accepted and presented in SIGGRAPH 2016, the premier and most selective computer graphics scientific event, and published in ACM Transaction on Graphics [11].

7.4. Human Motions in VR

To carry-out natural and efficient interactions with a digital world, it is firstly necessary to recognize and evaluate the action of the user. We consequently initiated a collaboration with the Intuidoc IRISA team to adapt methods previously used in 2D gesture recognition to 3D motion. With the increasing use of head mounted display devices (especially cheap devices recently spread in the large public), the problem of avatar simulation and embodiment has become an important challenge. In this context, we initiated collaborative works with Hybrid to better understand embodiment and consequently imagine the future generation of avatars. Concurrently, we continued to explore the use of such technology in various application domains where human performance is a key point, such as ergonomics.

7.4.1. Motion recognition and classification

Participants: Franck Multon, Richard Kulpa, Yacine Boulahia.

Action recognition based on human skeleton structure represents nowadays a prospering research field. This is mainly due to the recent advances in terms of capture technologies and skeleton extraction algorithms. In this context, we observed that 3D skeleton-based actions share several properties with handwritten symbols since they both result from a human performance. We accordingly hypothesize that the action recognition problem can take advantage of trial and error approaches already carried out on handwritten patterns. Therefore, inspired by one of the most efficient and compact handwriting feature-set, we proposed a skeleton descriptor referred to as Handwriting-Inspired Features [20]. First of all, joint trajectories are preprocessed in order to handle the variability among actor's morphologies. Then we extract the HIF3D features from the processed joint locations according to a time partitioning scheme so as to additionally encode the temporal information over the sequence. Finally, we used Support Vector Machine (SVM) for classification. Evaluations conducted on two challenging datasets, namely HDM05 and UTKinect, testify the soundness of our approach as the obtained results outperform the state-of-the-art algorithms that rely on skeleton data.

This work has been carried-out in collaboration with the IRISA Intuidoc team, with Yacine Boulahia who is a co-supervised PhD student with Eric Anquetil.

7.4.2. Avatar Embodiment in Virtual Reality

Participant: Ludovic Hoyet.

With the massive development of virtual reality products investigated by major industrial companies (Google, Facebook, HTC, Sony, etc), there is a new need for understanding what makes users immersed in virtual environments, especially regarding their relation to their virtual representation (i.e., avatar). Amongst others, an important factor is for users to feel incarnated in their avatar, which is called *virtual embodiment*. As more and more technological limitations are now being unlocked, understanding such factors become important to lever new immersive applications, e.g., in education, ergonomics or entertainment.

In collaboration with the EPI Hybrid (Ferran Argelaguet and Anatole Lécuyer), we explore the capacity of avatars to convey such a sense of "virtual embodiment", i.e., the extent to which we accept an avatar to be our representation in the virtual environment. The question of embodiment originates from the famous Rubber Hand Illusion experiment of Botvinick and Cohen (1998). This experiment demonstrated that when participants are presented with a fake rubber hand positioned beside their real hidden hand, and that both hands are synchronously stroked by an experimenter, after some time participants consider their real hand to be positioned at the location of the fake rubber hand. Today, understanding how similar phenomena happen in virtual environments is crucial to provide a maximum immersion for users. For instance, previous work demonstrated that racial biases can be reduced when users are incarnated in virtual characters of a different race, or explored body weight perception by altering the morphology of the avatar. The innovative aspect of our contributions is that we explore this embodiment effect in terms of interactions of the user with the virtual environment.

So far, we explored how people appropriate avatars by evaluating how they accept different representations of their virtual hand in virtual environments. Using various representations ranging from simplistic to highly realistic when interacting in virtual environments [19], we demonstrated that the sense of ownership (i.e., the impression that the virtual hand is actually our own hand) is increased when displaying highly realistic hand representations, but that the sense of agency (i.e., the impression to be able to control the actions of the virtual hand) is stronger for less realistic representations. With the potential of VR to alter and control avatars in different ways, e.g., the user representation, we also explored how structural differences of the hand representation can influence embodiment through controlling a six-digit virtual hand [10]. We found that participants responded positively to the possibility of controlling the virtual hand despite the structural difference, and accepted it as their own to some extent. Overall, results from such experiments further our understanding of the capacity of avatars to elicit a sense of embodiment in the users, and help to design more immersive VR experiences.

7.4.3. VR and Ergonomics

Participants: Charles Pontonnier, Georges Dumont, Pierre Plantard, Franck Multon.

The use of virtual reality tools for ergonomics applications is a very important challenge in order to generalize the use of such devices for the design of workstations.

We proposed a framework for collaborative ergonomic design in virtual environments. The framework consists in defining design modes and metaphors that help the users (engineers, ergonomists, end-users) to find a good trade-off between their own design constraints that can be contradictory at some point. We evaluated the framework and concluded that the active user has to be carefully chosen with regard to the design specifications, since the active user is favouring systematically its own constraints. This work has been published in the Journal on Multimodal User Interfaces [14].

7.5. Digital Storytelling

A transversal research of MimeTIC is digital storytelling as it enables to analyse, capture, model and simulate scenarios involving several humans (real and/or virtual). In this context, it is important to propose annotation tools and languages being able to capture such scenarios and stylistic informations before being able to simulate new ones. Moreover, when living an immersive experience in VR the user may want to have a summarize of his experience, which goes beyond simply replaying the recorded motions. Narration techniques can be positively used to highlight key events and actions, with nonlinear storytelling and intelligent camera placement to convey the desired emotion. The research in this field in MimeTIC contributes to the creation of

complex stories on social and human themes. Such approaches are more and more required to create interactive storylines, which massively enhances the possibilities of interactive entertainment, training, computer games and digital applications.

7.5.1. Trip Synopsis: virtual camera control applied to route visualisation

Participant: Marc Christie.

Computerized route planning tools are widely used today by travelers all around the globe, while 3D terrain and urban models are becoming increasingly elaborate and abundant. This makes it feasible to generate a virtual 3D flyby along a planned route. Such a flyby may be useful, either as a preview of the trip, or as an after-the-fact visual summary. However, a naively generated preview is likely to contain many boring portions, while skipping too quickly over areas worthy of attention. We have therefore proposed a general interest-driven framework that automatically computes a flyby along a planned route [9]. This flyby relies on an interest function to derive how close and how slow the camera should focus on the interesting areas, while skipping interest-less regions by using elevated smoothed camera motions. To address the problem, we devised a specific iterative solving process that incrementally approaches the optimal camera trajectory by adjusting position and speed.

7.5.2. Flashbacks in narratives

Participants: Marc Christie, Hui-Yin Wu.

The flashback is a well-known storytelling device used to invoke surprise, suspense, or fill in missing details in a story. Film literature provides a deeper and more complex grounding of flashbacks by explaining their role to stimulate the viewer's memory in order to guide and change viewer comprehension. Yet, in adapting flashback mechanisms to AI storytelling systems, existing approaches have not fully modelled the roles of a flashback event on the viewer's comprehension and memory. To expand the scope of AI generated stories, we propose a formal definition of flashbacks based on the identification of four different impacts on the viewer's beliefs. We then establish a cognitive model that can predict how viewers would perceive a flashback event. We finally design a user evaluation to demonstrate that our model correctly predicts the effects of different flashbacks. This opens great opportunities for creating compelling and temporally complex interactive narratives grounded on cognitive models [29].

7.5.3. Embedded Cinematography Patterns for film Analysis

Participants: Marc Christie, Hui-Yin Wu.

Cinematography carries messages on the plot, emotion, or more general feeling of the film. Yet cinematographic devices are often overlooked in existing approaches to film analysis. To solve this limitation, we present Embedded Constrained Patterns (ECPs), a dedicated query language to search annotated film clips for sequences that fulfill complex stylistic constraints [28]. ECPs are groups of framing and sequencing constraints defined using vocabulary in film textbooks. Using a set algorithm, all occurrences of the ECPs can be found in annotated film sequences. We use a film clip from the Lord of the Rings to demonstrate a range of ECPs that can be detected, and analyse them in relation to story and emotions in the film.

MYRIADS Project-Team

7. New Results

7.1. Scaling Clouds

7.1.1. Heterogeneous Resource Management

Participants: Baptiste Goupille-Lescar, Ancuta Iordache, Christine Morin, Manh Linh Pham, Nikos Parlavantzas, Guillaume Pierre, Arnab Sinha.

7.1.1.1. High performance in the cloud with FPGA virtualization

Participants: Ancuta Iordache, Guillaume Pierre.

Cloud platforms are becoming increasingly heterogeneous, with the availability of large numbers of virtual machine instance types as well as accelerator devices such as GPUs. In collaboration with Maxeler technologies, we have proposed a technique to virtualize FPGAs and make them available as first-class high-performance computation devices in the cloud [24]. The increasing variety of computation, storage and networking resources in the cloud is an opportunity for adjusting the provisioned resources to the individual needs of each application, but making an informed choice is extremely difficult. We therefore proposed application profiling techniques which can automatically identify the configuration which provides the best performance/cost tradeoff [49]. These two results were developed as part of the HARNESS European project, and they constitute Anca Iordache's PhD thesis [50]. FPGA virtualization is being further developed by Maxeler technologies toward commercial exploitation, and application profiling has been integrated in the open-source ConPaaS platform.

7.1.1.2. Multi-cloud application execution

Participants: Manh Linh Pham, Nikos Parlavantzas, Arnab Sinha.

Within the PaaSage European project, we improved and extended the Adapter subsystem, the part of the PaaSage platform that dynamically adapts the application deployment to changes in current runtime conditions [45]. Specifically, we added full support for causal connection between the running system and the runtime model and extended the plan validation functionality to use historical reconfiguration information. Moreover, we assisted industrial PaaSage partners with applying the PaaSage platform in diverse business scenarios.

7.1.1.3. Adaptive resource management for high-performance, multi-sensor systems

Participants: Baptiste Goupille-Lescar, Christine Morin, Nikos Parlavantzas.

In the context of our collaboration with Thales Research and Technology, we are applying cloud resource management techniques to high-performance, multi-sensor, embedded systems with real-time constraints. The objective is to increase the flexibility and efficiency of resource allocation in such systems, enabling the execution of dynamic sets of applications with strict QoS requirements. In 2016, we focused on characterising the targeted applications and platforms and developing a simulator in order to explore relevant resource management solutions. This work is performed in the context of Baptiste Goupille-Lescar's PhD work.

7.1.2. Distributed Cloud Computing

Participants: Nikos Parlavantzas, Jean-Louis Pazat, Guillaume Pierre, Genc Tato, Cédric Tedeschi, Alexandre Van Kempen.

7.1.2.1. Application self-optimization in multi-cloud environments

Participant: Nikos Parlavantzas.

Current approaches to application adaptation in multi-cloud environments are typically static, platform dependent, complex, and error prone. To address these limitations, we are combining the use of software product lines (SPLs) with models@run-time techniques. This work is performed in the context of the thesis of Carlos Ruiz Diaz, a PhD student at the University of Guadalajara, co-advised by Nikos Parlavantzas. The work focuses on the development of an SPL-based framework supporting initial cloud configuration as well as proactive, dynamic adaptation in a systematic, platform-independent way. The evaluation of this framework is currently in progress.

7.1.2.2. Edge clouds

Participants: Guillaume Pierre, Genc Tato, Cédric Tedeschi, Alexandre Van Kempen.

Mobile edge cloud computing aims to deploy cloud resources even closer to the end users, typically within mobile network access points. This is useful for hyper-interactive applications such as augmented reality which demand ultra-low network latencies (2-5 ms) between the end-user device and the cloud instances serving it. In contrast, current mobile networks exhibit network latencies in the order of 50-150 ms between the device and any cloud. We extended the ConPaaS open-source cloud platform to support the deployment of cloud applications in a distributed set of Raspberry Pi machines: instead of reaching the cloud through a wide-area network, in this setup each cloud node is also equipped with a wifi hotspot which allows local users to access it directly [53]. This work is ongoing, and a paper on this topic is currently being reviewed.

Getting closer to the edge user can be done through provisioning computing resources in Points of Presence (PoPs) within the telco's backbone network. The Discovery project [52] aims at revisiting the OpenStack Cloud stack to allow to disperse several smaller cloud facilities and connect them together to make them appear as a single Cloud entity. Genc Tato's PhD aims at proposing the building blocks on top of such an infrastructure to abstract out the network, route queries, store and retrieve objects (VMs and data). We have devised an overlay network to support such functionalities keeping in mind to maximise the laziness of the maintenance protocol to avoid any useless cost. A paper is being written on the subject.

7.1.2.3. Community Clouds

Participant: Jean-Louis Pazat.

Hosting services on an edge infrastructure based on devices owned and operated by end-users may be interesting for serving a community of users. However, these devices (such as internet boxes, disks or small computers have heterogeneous capabilities and no guaranteed availability. It is therefore challenging to ensure to the guest application a minimal hosting service level, like availability or Quality of Service. The management of the hosting service should adapt to the characteristics of the infrastructure. We are designing an architecture for a middleware capable of adapting the deployment of services on edge devices to ensure a given Quality of Service to access the service. While the middleware requires a minimal knowledge of the underlying infrastructure, its adaptation decisions are based on the feedbacks of users of the deployed service, like measured network latency. The environment relies on the use of micro-services which are composed to build the end-user services. This allows many adaptation strategies to adapt the system during run-time.

7.1.3. Scaling workflows with GinFlow

Participants: Matthieu Simonin, Cédric Tedeschi.

In 2016, we deployed GinFlow over 800 cores of the Grid'5000 platform, running Montage workflows comprising 118 tasks, and artificial workflows made of more than 3000 tasks. The ability of GinFlow to support adaptation and versioning of workflow with seamless transitions between workflow alternatives at runtime has been validated experimentally and presented on the Inria booth at SuperComputing in November 2016. These results have been presented at the IPDPS conference [32], and have been submitted to a journal special issue on workflows.

7.2. Greening Clouds

7.2.1. Energy Models

Participants: Yvon Jégou, Anne-Cécile Orgerie, Edouard Outin, Jean-Louis Pazat, Martin Quinson.

Simulating the impact of DVFS within SimGrid Simulation is a a popular approach for studying the performance of HPC applications in a variety of scenarios. However, simulators do not typically provide insights on the energy consumption of the simulated platforms. The goal of this ongoing work is to enable energy-aware experimentation within the SimGrid simulation toolkit, by introducing a model of energy consumption for computing applications making use of Dynamic Voltage and Frequency Scaling (DVFS) techniques.

Simulating Energy Consumption of Wired Networks In this work, we aim at simulating the energy consumption of wired networks which receive little attention in the Cloud computing community even though they represent key elements of these distributed architectures. To this end, we are contributing to the well-known open-source simulator ns3 by developing an energy consumption module named ECOFEN. This simulator embeds green levers: low power idle (IEEE 802.3az) and adaptive link rate. An article is currently under review on this topic.

Multicriteria scheduling for large-scale HPC environments Energy consumption is one of the main limiting factor for the design and deployment of large scale numerical infrastructures. The road towards "Sustainable Exascale" is a challenge with a target of 50 Gflops per watt. As platforms become more and more heterogeneous (co-processors, GPUs, low power processors...), an efficient scheduling of applications and services at large scale remains a challenge. In this context, we explore a multicriteria scheduling model and framework for large scale HPC systems. This work is done in collaboration with ROMA and Avalon teams from LIP in Lyon [29], [37].

Dynamic resource management for energy-efficiency The B-Com project, a joint private/public focusing on transfer, targets the design and the implementation of Watcher, a software module used to optimize an OpenStack cloud (in terms of performance, storage optimization or energy savings). This Software module is in the "Big Tent" software development process of OpenStack. In cooperation with Olivier Barais (Diverse Inria Team), we focus on dynamic management of cloud resources for energy-efficiency. Our approach relies on machine learning techniques, models@run-time and dynamic adaptation, and is intended to be included in Watcher. At regular intervals of time, we optimize the use of cloud resources by checking if a better placement of Virtual Machines on physical resources can be achieved, taking into account the migration cost. To achieve this, we have an energy model of the resources which is regularly updated using machine learning techniques that helps optimization algorithms to check if a better configuration can be reached energy-wise. This year we worked on the evaluation of the energy model [28].

7.2.2. Involving users in Energy Saving

Participants: Deborah Agarwal, Ismael Cuadrado Cordero, David Guyon, Christine Morin, Anne-Cécile Orgerie.

Energy-efficient cloud elasticity for data-driven applications Data centers hosting cloud systems consume enormous amounts of energy. Reducing this consumption becomes an urgent challenge with the rapid growth of cloud utilization. An existing solution to lower this consumption is to turn off as many servers as possible, but these solutions do not involve the user as a main lever to save energy. We introduce a system that proposes to the user to run her application with degraded performance in order to promote a better consolidation and thus to turn off more servers. Experimentation results using the Montage workflow show promising outcomes [47], [48]. We also performed a simulation-based evaluation on how much an energy-aware cloud system could save in energy consumed depending on the proportion of users selecting a green execution mode. These results based on the simulation of two typical daily uses of a data center running 3 real scientific applications will be published in Euromicro PDP 2017.

Energy-efficient and network-aware resource allocation in Cloud infrastructures The ever-growing appetite of new applications for network resources leads to an unprecedented electricity bill, and for these bandwidth-hungry applications, networks can become a significant bottleneck. Towards this end, we proposed microclouds, a fully autonomous energy-efficient subnetwork of clients of the same service, designed to keep the greenest path between its node. This semi-decentralized PaaS architecture for real-time multiple-users applications geographically distributes the computation among the clients of the cloud, moving the

computation away from the datacenter to save energy - by shutting down or downgrading non utilized resources such as routers and switches, servers, etc. - and provides lower latencies for users. In this work, we have also analyzed the use of incentives for Mobile Clouds, and proposed a new auction system adapted to the high dynamism and heterogeneity of these systems [20], [19] [46].

7.2.3. Exploiting Renewable Energy in Datacenters

Participants: Sabbir Hasan Rochi, Yunbo Li, Anne-Cécile Orgerie, Jean-Louis Pazat.

Resource allocation in a Cloud partially powered by renewable energy sources We propose here to design a disruptive approach to Cloud resource management which takes advantage of renewable energy availability to perform opportunistic tasks. This Cloud receives a fixed amount of power from the regular electric Grid. This power allows it to run usual tasks. In addition, this Cloud is also connected to renewable energy sources (such as windmills or solar cells) and when these sources produce electricity, the Cloud can use it to run more tasks. The proposed resource management system integrates a prediction model to be able to forecast these extra-power periods of time in order to schedule more work during these periods. This work is done in collaboration with Ascola team from LINA in Nantes [44], [51][9].

Creating green-energy adaptivity awareness in SaaS application In addition to "green" resource allocation at the IaaS level in Datacenters, we think that users should be involved in "greening" their energy use (SaaS level). We propose that applications should have multiple "modes" of execution, each mode using a different level of energy and providing a different service level. For example, a B2C application may provide more or less recommandations . If this applcation can be dynamically swiched between these modes depending on the availability of green energy, the IaaS can optimize resource allocation better. To enforce his, we have designed green energy aware controllers.

This work is done in collaboration with Ascola team [23], [9].

7.3. Securing Clouds

7.3.1. Security monitoring in clouds

Participants: Jean Leon Cusinato, Anna Giannakou, Fergal Martin-Tricot, Christine Morin, Jean-Louis Pazat, Louis Rilling, Amir Teshome Wonjiga.

In the INDIC project we aim at making security monitoring a dependable service for IaaS cloud customers. To this end, we study three topics:

- defining relevant SLA terms for security monitoring,
- enforcing and verifying SLA terms,
- making the SLA terms enforcement mechanisms self-adaptable to cope with the dynamic nature of clouds.

The considered enforcement and verification mechanisms should have a minimal impact on performance.

In 2016 we improved the SAIDS approach, that we proposed in 2015, and that makes a network intrusion detection system (NIDS) deployed in a cloud operator infrastructure self-adaptable. In particular, we validated that the approach is generic enough to handle signature-based NIDSs (support for Snort and Suricata was implemented) as well as event-based NIDSs (support for Bro was implemented). An experimental evaluation of SAIDS has also been started in order to submit a full paper for publication in 2017. Jean-Léon Cusinato contributed to this work during his master internship.

We also improved the AL-SAFE approach, that we proposed in 2015, and that secures an application-level firewall by isolating it from the customer virtual machine and makes it self-adaptable [36], [35]. In particular, we validated that the self-adaptation architecture introduced for SAIDS could be reused to address firewalls, and the prototype was improved to implement stateful filtering. Fergal Martin-Tricot contributed to this work during his master internship. We also evaluated AL-SAFE experimentally on the prototype as well as analytically regarding the security correctness. The design and the evaluation of AL-SAFE were published in the CloudCom 2016 conference [21].

Regarding SLA definition and enforcement, in 2016 we have studied a verification method to enable a Cloud customer to verify that an NIDS located in the operator infrastructure is configured correctly according to the Service-Level Objectives (SLO) figuring in the SLA. A simple example of SLO is being used for this study, and further work should address more complete SLO regarding NIDSs. A prototype of the proposed verification method was implemented on OpenStack and Open vSwitch, and the NIDS software used is Snort. An evaluation of the verification method has been started and will include both experiments on the Grid'5000 platform and a correctness analysis. The design and evaluation of the verification method will be submitted in a full paper for publication in 2017.

7.3.2. Risk assessment in clouds

Participant: Christine Morin.

Attack graphs are leveraged in networks to exhibit the various scenarios available to compromise the system. They allow to uncover vulnerabilities chains exploitable by attackers based on network connectivity and vulnerabilities pre-requisites. In physical infrastructures, the acquisition of the topology has been vastly addressed in existing works with either passive or active discovery methods. Considering the Cloud context, in which virtualization attacks and virtual infrastructure dynamism are introduced, new methods need to be developed. We have designed a topology builder able to keep the topology and connectivity up to date in cloud environments. Based on the use of an IaaS cloud management system and a SDN (Software-Defined Networking) controller, our approach encompasses two steps: (i) when plugged into a running system, the topology builder retrieves the current topology and builds the associated connectivity: this represents the static topology and connectivity retrieval, in which we assume the network configuration to be fixed ; (ii) the topology builder listens to change events generated inside the infrastructure and within the SDN controller in order to update the topology and connectivity previously built: this represents the dynamic topology and connectivity retrieval. A prototype has been developed based on OpenStack cloud management system and ONOS SDN open source technologies. This work is carried out in the context of Pernelle Mensah's PhD thesis and in collaboration with Nokia and CIDRE Inria project-team.

7.4. Experimenting with Clouds

7.4.1. Simulation

Participants: Simon Bihel, Martin Quinson.

Providing better interfaces to the users for Cloud Studies. Aware that the current user interface is a impediment to the adoption of our framework by the scientific community, we tried to propose a new, simplified API through the internship of Simon Bihel this summer. We identified several use cases and usage scenario that relevant to our context, and started implementing the new interface that we will provide. This work is still under progress.

Production-ready simulator of large-scale distributed systems. We are currently involved in a complete reorganization of the SimGrid implementation. The goal is two-fold: first we want reduce the tool's learning curve to help beginners. At the same time, we want to normalize the tool's internals so that power users can modify it and/or script the kernel behavior easily. Eventually, we are targeting usages in production and teaching contexts. This long term overhaul is still underway.

7.4.2. Experimentation Testbed

Participants: Anirvan Basu, Julien Lefeuvre, David Margery, Pascal Morillon.

Providing ready to use scripts to deploy popular and complex stacks. The study of complex software stacks on Grid'5000 has always been possible due to the reconfigurability properties of the testbed. Nevertheless, for newcomers with little background in system administration, automating the deployment of these stacks on Grid'5000 has always proved difficult. In 2016, we have provided scripts, that users can fork on github to customise to their needs, to deploy OpenStack, Ceph, Hadoop over Ceph or Sparkle. These have been presented to users during the 2016 winter school.

7.4.3. Use cases

Participants: Deborah Agarwal, Yvon Jégou, Nikos Parlavantzas, Manh Linh Pham, Christine Morin, Kartik Sathyanarayanan, Arnab Sinha.

7.4.3.1. Experimental Evaluation of Data Stream Processing Frameworks

We worked on evaluating data stream processing environments deployed in clouds. We compared the throughput, latency and energy consumption of Spark Streaming, Storm and Heron real-time data processing environments executed on top of Linux clusters and on top of virtual clusters deployed on top of the OpenStack IaaS cloud. The preliminary evaluation was conducted using the word count application on the twitter data stream. All experiments were conducted on Grid'5000 experimentation platform. The experimental results are described in a technical report to be published in 2017. This work was carried out by Kartik Sathyanarayanan, a student intern in Myriads team in the framework of DALHIS associate team.

7.4.3.2. Simulation framework for studying between-herd pathogen spread in a region

In our collaboration with Inra in the context of the Mihmes project, we worked on the design of decision tools to evaluate the epidemio-economic effectiveness of disease prevention and control strategies at the scales of the herd, the region and the supply chain. We developed a generic service-based framework to efficiently execute models of infection dynamics in a metapopulation of cattle herds on large-scale computing infrastructures. Our framework has been designed to execute complex regional models combining within-herds epidemiological models. The framework automatically distributes the simulation runs on multiple servers in a cluster and exploits the parallelism of the multicore servers. It relies on OpenMP for parallelizing simulation loops and deals with server heterogeneity and failures. We leveraged PaaSage software stack to deploy the framework on several IaaS clouds.

7.4.3.3. Mobile application for reliable collection of field data for Fluxnet

Critical to the interpretation of Fluxnet carbon flux data is the ancillary information and measurements taken at the tower sites. The submission and update of this data using excel sheets is difficult and error prone. In partnership with ICOS in the framework of DALHIS associate team, we are innovating the data submission and organization method through a responsive web User Interface able to run on desktop, mobile etc.; thus easing the data lookup and entry process from anywhere including the field sites. Continuing with our initial usability feedback experiences gathered last year on the application interface designs, we decided on the mobile application workflow for implementation. We developed a first prototype based on the PhoneGap [0] platform which provided the advantage of the same development code generating mobile application for IOS, Android and Windows platform simultaneously. The main functionality realized in the application prototype is that the user can download all the site data required by logging in through the application; and then view/edit them at the tower site (even in offline mode). The next logical step would be developing the synchronization and validation of data held locally in the application with the servers.

[0] http://phonegap.com/

PACAP Project-Team

7. New Results

7.1. Compiler, vectorization, interpretation

Participants: Erven Rohou, Emmanuel Riou, Arjun Suresh, André Seznec, Nabil Hallou, Sylvain Collange, Rabab Bouziane, Arif Ali Ana-Pparakkal, Stefano Cherubin.

7.1.1. Improving sequential performance through memoization

Participants: Erven Rohou, Emmanuel Riou, André Seznec, Arjun Suresh.

Many applications perform repetitive computations, even when properly programmed and optimized. Performance can be improved by caching results of pure functions, and retrieving them instead of recomputing a result (a technique called memoization).

We propose [20] a simple technique for enabling software memoization of any dynamically linked pure function and we illustrate our framework using a set of computationally expensive pure functions – the transcendental functions.

Our technique does not need the availability of source code and thus can be applied even to commercial applications as well as applications with legacy codes. As far as users are concerned, enabling memoization is as simple as setting an environment variable.

Our framework does not make any specific assumptions about the underlying architecture or compiler tool-chains, and can work with a variety of current architectures.

We present experimental results for x86-64 platform using both gcc and icc compiler tool-chains, and for ARM cortex-A9 platform using gcc. Our experiments include a mix of real world programs and standard benchmark suites: SPEC and Splash2x. On standard benchmark applications that extensively call the transcendental functions we report memoization benefits of upto 16 %, while much higher gains were realized for programs that call the expensive Bessel functions. Memoization was also able to regain a performance loss of 76 % in *bwaves* due to a known performance bug in the gcc libm implementation of *pow* function.

Initial work has been published in ACM TACO 2015 [20] and accepted for presentation at the International Conference HiPEAC 2016 in Prague.

Further developments have been accepted for publication at the Compiler Construction Conference 2017 [49].

This research is described in the PhD thesis of Arjun Suresh [24].

7.1.2. Optimization in the Presence of NVRAM

Participants: Erven Rohou, Rabab Bouziane.

Energy-efficiency is one of the most challenging design issues in both embedded and high-performance computing domains. The aim is to reduce as much as possible the energy consumption of considered systems while providing them with the best computing performance. Finding an adequate solution to this problem certainly requires a cross-disciplinary approach capable of addressing the energy/performance trade-off at different system design levels.

We proposed [42] an empirical impact analysis of the integration of Spin Transfer Torque Magnetic Random Access Memory (STT-MRAM) technologies in multicore architectures when applying some existing compiler optimizations. For that purpose, we use three well-established architecture and NVM evaluation tools: NVSim, gem5 and McPAT. Our results show that the integration of STT-MRAM at cache memory levels enables a significant reduction of the energy consumption (up to 24.2 % and 31 % on the considered multicore and monocore platforms respectively) while preserving the performance improvement provided by typical code optimizations. We also identified how the choice of the clock frequency impacts the relative efficiency of the considered memory technologies.

This research is done in collaboration with Abdoulaye Gamatié at LIRMM (Montpellier) within the context the the ANR project CONTINUUM.

7.1.3. Hardware/Software JIT Compiler

Participant: Erven Rohou.

Dynamic Binary Translation (DBT) is often used in hardware/software co-design to take advantage of an architecture model while using binaries from another one. The co-development of the DBT engine and of the execution architecture leads to architecture with special support to these mechanisms. We proposed a hardware accelerated dynamic binary translation where the first steps of the DBT process are fully accelerated in hardware. Results shows that using our hardware accelerators leads to a speed-up of 8× and a cost in energy 18× lower, compared with an equivalent software approach.

An initial version of this work has been presented at Compas'16 [51]. The latest results have been accepted for publication at DATE 2017 [44].

This research is done in collaboration with Steven Derrien and Simon Rokicki from the CAIRN team.

7.1.4. Dynamic Parallelization of Binary Programs

Participants: Erven Rohou, Emmanuel Riou, Nabil Hallou.

We address runtime automatic parallelization of binary executables, assuming no previous knowledge on the executable code. The Padrone platform is used to identify candidate functions and loops. Then we disassemble the loops and convert them to the intermediate representation of the LLVM compiler. This allows us to leverage the power of the polyhedral model for auto-parallelizing loops. Once optimized, new native code is generated just-in-time in the address space of the target process.

Our approach enables user transparent auto-parallelization of legacy and/or commercial applications with auto-parallelization.

This work has been accepted for publication in the Springer journal IJPP: "Runtime Vectorization Transformations of Binary Code".

This work is done in collaboration with Philippe Clauss (Inria CAMUS).

7.1.5. Dynamic Function Specialization

Participants: Erven Rohou, Arif Ali Ana-Pparakkal.

Compilers can do better optimization with the knowledge of run-time behaviour of the program. *Function Specialization* is an optimization technique in which different versions of a function are created according to the value of its arguments. It can be difficult to predict the exact value/behaviour of arguments during static compilation and so it is difficult for a static compiler to do efficient function specialization. In our *dynamic function specialization* technique, we capture the actual value of arguments during execution of the program and, when profitable, create specialized versions and include them at runtime.

This research is done within the context of the Nano 2017 PSAIC collaborative project.

7.1.6. Application Autotuning for Performance and Energy

Participants: Erven Rohou, Stefano Cherubin, Imane Lasri.

Due to the increasing complexity of both applications behaviors and underlying hardware, achieving reasonable (not to mention best) performance can hardly be done at compile time. Autotuning is an approach where a runtime manager is able to adapt the software to the runtime conditions. We have developed a framework and shown through a domain specific application initial exploration scenarios [32], [47].

We started characterizing applications – in particular the Parasuite benchmarks – and we will rely on split-compilation [2] embed hints and heuristics inside a binary program for dynamic adaptation and optimization.

This research is done within the context of the H2020 FET HPC collaborative project ANTAREX.

7.1.7. Customized Precision Computing

Participants: Erven Rohou, Stefano Cherubin, Imane Lasri.

Customized precision originates from the fact that many applications can tolerate some loss of quality during computation, as in the case of media processing (audio, video and image), data mining, machine learning, etc. Error-tolerating applications are increasingly common in the emerging field of real-time HPC. Thus, recent works have investigated this line of research in the HPC domain as a way to provide a breakthrough in power and performance for the Exascale era.

We aim at leveraging existing, HPC-oriented hardware architectures, while including in the precision tuning an adaptive selection of floating and fixed-point arithmetic. It is part of a wider effort to provide the programmers with an easy way to manage extra-functional properties of programs, including precision, power, and performance.

We explore tradeoffs between precision and time-to-solution, as well as precision and energy-to-solution.

This is done within the context of the ANTAREX project in collaboration with Stefano Cherubin, Cristina Silvano and Giovanni Agosta from Politecnico di Milano, and Olivier Sentieys from the CAIRN team.

7.1.8. SPMD Function Call Re-Vectorization

Participant: Sylvain Collange.

SPMD programming languages for SIMD hardware such as C for CUDA, OpenCL or ISPC have contributed to increase the programmability of SIMD accelerators and graphics processing units. However, SPMD languages still lack the flexibility offered by low-level SIMD programming on explicit vectors. To close this expressiveness gap while preserving the SPMD abstraction, we introduce the notion of Function Call Re-Vectorization (CREV) [38]. CREV allows changing the dimension of vectorization during the execution of an SPMD kernel, and exposes it as a nested parallel kernel call. CREV affords a programmability close to dynamic parallelism, a feature that allows the invocation of kernels from inside kernels, but at much lower cost. In this paper, we present a formal semantics of CREV, and an implementation of it on the ISPC compiler. To validate our idea, we have used CREV to implement some classic algorithms, including string matching, depth first search and Bellman-Ford, with minimum effort. These algorithms, once compiled by ISPC to Intel-based vector instructions, are as fast as state-of-the-art implementations, yet much simpler. As an example, our straightforward implementation of string matching beats the Knuth-Morris-Pratt algorithm by 12 %.

This work was done during the internship of Rubens Emilio in Rennes in collaboration with Sylvain Collange and Fernando Pereira (UFMG) as part of the Inria PROSPIEL Associate Team.

7.1.9. SPMD Function Call Fusion

Participant: Sylvain Collange.

The increasing popularity of Graphics Processing Units (GPUs) has brought renewed attention to old problems related to the Single Instruction, Multiple Data execution model. One of these problems is the reconvergence of divergent threads. A divergence happens at a conditional branch when different threads disagree on the path to follow upon reaching this split point. Divergences may impose a heavy burden on the performance of parallel programs.

We have proposed a compiler-level optimization to mitigate the performance loss due to branch divergence on GPUs [21]. This optimization consists in merging function call sites located at different paths that sprout from the same branch. We show that our optimization adds negligible overhead on the compiler. When not applicable, it does not slow down programs and it accelerates substantially those in which it is applicable. As an example, we have been able to speed up the well known SPLASH Fast Fourier Transform benchmark by 11 %.

This work is done in collaboration with Douglas do Couto Teixeira and Fernando Pereira from UFMG as part of the Inria PROSPIEL Associate Team.

7.1.10. SIMD programming in SPMD: application to multi-precision computations

Participant: Sylvain Collange.

GPUs are an important hardware development platform for problems where massive parallel computations are needed. Many of these problems require a higher precision than the standard double floating-point (FP) available. One common way of extending the precision is the multiple-component approach, in which real numbers are represented as the unevaluated sum of several standard machine precision FP numbers. This representation is called a FP expansion and it offers the simplicity of using directly available and highly optimized FP operations. We propose new data-parallel algorithms for adding and multiplying FP expansions specially designed for extended precision computations on GPUs [34]. These are generalized algorithms that can manipulate FP expansions of different sizes (from double-double up to a few tens of doubles) and ensure a certain worst case error bound on the results.

This work is done in collaboration with Mioara Joldes (CNRS/LAAS), Jean-Michel Muller (CNRS/LIP) and Valentina Popescu (ENS Lyon/LIP).

7.2. Processor Architecture

Participants: Pierre Michaud, Sylvain Collange, Erven Rohou, André Seznec, Arthur Perais, Sajith Kalathingal, Andrea Mondelli, Aswinkumar Sridharan, Biswabandan Panda, Fernando Endo, Kleovoulos Kalaitzidis.

Processor, cache, locality, memory hierarchy, branch prediction, multicore, power, temperature

7.2.1. Microarchitecture

7.2.1.1. Branch prediction

Participant: André Seznec.

IMLI-based predictors

The wormhole (WH) branch predictor was recently introduced to exploit branch outcome correlation in multidimensional loops. For some branches encapsulated in a multidimensional loop, their outcomes are correlated with those of the same branch in neighbor iterations, but in the previous outer loop iteration. In [18], we introduced practical predictor components to exploit this branch outcome correlation in multidimensional loops: the IMLI-based predictor components. The iteration index of the inner most loop in an application can be efficiently monitored at instruction fetch time using the Inner Most Loop Iteration (IMLI) counter. The outcomes of some branches are strongly correlated with the value of this IMLI counter. Our experiments show that augmenting a state-of-the-art global history predictor such as TAGE-SC-L [45] with IMLI-based components outperforms previous state-of-the-art academic predictors leveraging local and global history at much lower hardware complexity (i.e., smaller storage budget , smaller number of tables and simpler management of speculative states).

This study was accepted in the special issue Top Picks of the best papers in 2015 computer architecture conferences in IEEE Micro [30].

This research was done in collaboration with Joshua San Miguel and Jorge Albericio from University of Toronto

Championship Branch Prediction

The 5th Championship Branch Prediction was organized in Seoul in June 2016. The predictors submitted by the PACAP-team, respectively TAGE-SC-L and MTAGE-SC, for limited storage budgets and infinite storage budgets won the three tracks of the competition [46], [45]. These predictors are derived from our reference work [17].

7.2.1.2. Revisiting Value Prediction

Participants: Arthur Perais, André Seznec.

Value prediction was proposed in the mid 90's to enhance the performance of high-end microprocessors. From 2013 to 2016, we have progressively revived the interest in value prediction. At a first step, we showed that all predictors are amenable to very high accuracy at the cost of some loss on prediction coverage [12]. Furthermore, we proposed EOLE [13]. EOLE leverages Value Prediction to *Early Execute* simple instructions whose operands are ready in parallel with Rename and to *Late Execute* to simple predicted instructions just before Commit. EOLE allows to reduce the out-of-order issue-width by 33% without impeding performance.

An extension of the initial EOLE paper [13] was published in ACM TOCS [27].

7.2.1.3. Physical register sharing

Participants: Arthur Perais, André Seznec.

Sharing a physical register between several instructions is needed to implement several microarchitectural optimizations. However, register sharing requires modifications to the register reclaiming process: Committing a single instruction does not guarantee that the physical register allocated to the previous mapping of its architectural destination register is free-able anymore. Consequently, a form of register reference counting must be implemented. While such mechanisms (e.g., dependency matrix, per register counters) have been described in the literature, we argue that they either require too much storage, or that they lengthen branch misprediction recovery by requiring sequential rollback. As an alternative, we present the Inflight Shared Register Buffer (ISRB), a new structure for register reference counting [41]. The ISRB has low storage overhead and lends itself to checkpoint-based recovery schemes, therefore allowing fast recovery on pipeline flushes. We illustrate our scheme with Move Elimination (short-circuiting moves) and an implementation of Speculative Memory Bypassing (short-circuiting store-load pairs) that makes use of a TAGE-like predictor to identify memory dependencies. We show that the whole potential of these two mechanisms can be achieved with a small register tracking structure.

7.2.1.4. Register Sharing for Equality Prediction

Participants: Arthur Perais, Fernando Endo, André Seznec.

Recently, Value Prediction (VP) has been gaining renewed traction in the research community. VP speculates on the result of instructions to increase Instruction Level Parallelism (ILP). In most embodiments, VP requires large tables to track predictions for many static instructions. However, in many cases, it is possible to detect that the result of an instruction is produced by an older inflight instruction, but not to predict the result itself. Consequently it is possible to rely on predicting register equality and handle speculation through the renamer. To do so, we propose to use Distance Prediction [40], a technique that was previously used to perform Speculative Memory Bypassing (short-circuiting def-store-load-use chains). Distance Prediction attempts to determine how many instructions separate the instruction of interest and the most recent older instruction that produced the same result. With this information, the physical register identifier of the older instruction can be retrieved from the ROB and provided to the renamer. The implementation of Distance Prediction necessitates a hardware mechanism to handle the sharing of physical registers as the ISRB [41].

7.2.1.5. Storage-Free Memory Dependency Prediction

Participants: Arthur Perais, André Seznec.

Memory Dependency Prediction (MDP) is paramount to good out-of-order performance, but decidedly not trivial as a all instances of a given static load may not necessarily depend on all instances of a given static store. As a result, for a given load, MDP should predict the exact store instruction the load depends on, and not only whether it depends on an inflight store or not, i.e., ideally, prediction should not be binary. However, we first argue that given the high degree of sophistication of modern branch predictors, the fact that a given dynamic load depends on an inflight store can be captured using the binary prediction capabilities of the branch predictor, providing coarse MDP at zero storage overhead. Second, by leveraging hysteresis counters, we show that the precise producer store can in fact be identified. This embodiment of MDP yields performance levels that are on par with state-of-the-art, and requires less than 70 additional bits of storage over a baseline without MDP at all [28].

7.2.1.6. Compressed Caches

Participants: André Seznec, Biswabandan Panda.

The YACC compressed cache

Cache memories play a critical role in bridging the latency, bandwidth, and energy gaps between cores and off-chip memory. However, caches frequently consume a significant fraction of a multicore chip's area, and thus account for a significant fraction of its cost. Compression has the potential to improve the effective capacity of a cache, providing the performance and energy benefits of a larger cache while using less area. The design of a compressed cache must address two important issues: i) a low-latency, low-overhead compression algorithm that can represent a fixed-size cache block using fewer bits and ii) a cache organization that can efficiently store the resulting variable-size compressed blocks. This paper focuses on the latter issue. We propose YACC (Yet Another Compressed Cache), a new compressed cache design that targets improving effective cache capacity with a simple design [29]. YACC uses super-blocks to reduce tag overheads, while packing variable-size compressed blocks to reduce internal fragmentation. YACC achieves the benefits of two state-of-the art compressed caches, Decoupled Compressed Cache (DCC) [61] and Skewed Compressed Cache (SCC) [15], with a more practical and simpler design. YACC's cache layout is similar to conventional caches, with a largely unmodified tag array and unmodified data array.

This study was done in collaboration with Somayeh Sardashti and David Wood from University of Wisconsin.

The DISH compression scheme

The effectiveness of a compressed cache depends on three features: i) the compression scheme, ii) the compaction scheme, and iii) the cache layout of the compressed cache. Both SCC [15] and YACC [29] use compression techniques to compress individual cache blocks, and then a compaction technique to compact multiple contiguous compressed blocks into a single data entry. The primary attribute used by these techniques for compaction is the compression factor of the cache blocks, and in this process, they waste cache space. We propose dictionary sharing (DISH), a dictionary based cache compression scheme that reduces this wastage [39]. DISH compresses a cache block by keeping in mind that the block is a potential candidate for the compaction process. DISH encodes a cache block with a dictionary that stores the distinct 4-byte chunks of a cache block and the dictionary is shared among multiple neighboring cache blocks. The simple encoding scheme of DISH also provides a single cycle decompression latency and it does not change the cache layout of compressed caches. Compressed cache layouts that use DISH outperforms the compression schemes, such as BDI and CPACK+Z, in terms of compression ratio, system performance , and energy efficiency.

7.2.1.7. Clustered microarchitecture

Participants: Andrea Mondelli, Pierre Michaud, André Seznec.

In the last 10 years, the clock frequency of high-end superscalar processors did not increase significantly. Performance keeps being increased mainly by integrating more cores on the same chip and by introducing new instruction set extensions. However, this benefits only to some applications and requires rewriting and/or recompiling these applications. A more general way to increase performance is to increase the IPC, the number of instructions executed per cycle.

In [8], we argue that some of the benefits of technology scaling should be used to increase the IPC of future superscalar cores. Starting from microarchitecture parameters similar to recent commercial high-end cores, we show that an effective way to increase the IPC is to increase the issue width. But this must be done without impacting the clock cycle. We propose to combine two known techniques: clustering and register write specialization. The objective of past work on clustered microarchitecture was to allow a higher clock frequency while minimizing the IPC loss. This led researchers to consider narrow-issue clusters. Our objective, instead, is to increase the IPC without impacting the clock cycle, which means wide-issue clusters. We show that, on a wide-issue dual cluster, a very simple steering policy that sends 64 consecutive instructions to the same cluster, the next 64 instructions to the other cluster, and so on, permits tolerating an inter-cluster delay of several cycles. We also propose a method for decreasing the energy cost of sending results of one cluster to the other cluster.

This study published in ACM TACO in 2015 [8] and was presented at the HIPEAC 2016 conference.

7.2.1.8. Hardware data prefetching

Participant: Pierre Michaud.

Hardware prefetching is an important feature of modern high-performance processors. When an application's working set is too large to fit in on-chip caches, disabling hardware prefetchers may result in severe performance reduction. We propose a new hardware data prefetcher, the Best-Offset (BO) prefetcher. The BO prefetcher is an offset prefetcher using a new method for selecting the best prefetch offset taking into account prefetch timeliness. The hardware required for implementing the BO prefetcher is very simple. A version of the BO prefetcher won the 2015 Data Prefetching Championship. A comprehensive study of the BO prefetcher was presented at the HPCA 2016 conference [37].

7.2.1.9. Exploiting loops for lower energy consumption

Participants: Andrea Mondelli, Pierre Michaud, André Seznec.

Recent superscalar processors use a loop buffer to decrease the energy consumption in the front-end. The energy savings comes from the branch predictor, instruction cache and instruction decoder being idle when micro-ops are delivered to the back-end from the loop buffer. We explored the possibility to exploit loop behaviors for decreasing energy consumption further, in the back-end, without impacting performance. We proposed two independent optimizations requiring little extra hardware. The first optimization detects and removes from the execution redundant micro-ops producing the same result in every loop iteration. The second optimization focuses on loop loads and detects situations where a loop load needs accessing only the data cache, or only the store queue, not both.

7.2.2. Microarchitecture Performance Modeling

7.2.2.1. Optimal cache replacement

Participant: Pierre Michaud.

A cache replacement policy is an algorithm, implemented in hardware, selecting a block to evict to make room for an incoming block. This research topic has been revitalized recently, as level-2 and level-3 caches were integrated on chip. A cache replacement policy cannot be optimal in general unless it has the knowledge of future references. Unfortunately, practical replacement policies do not have this knowledge. Still, optimal replacement is an important benchmark for understanding replacement policies. Moreover, some new replacement policies proposed recently are directly inspired from algorithms for determining hits and misses under optimal replacement. Hence it is important to improve our understanding of optimal replacement.

The OPT policy, which evicts the block referenced furthest in the future, was proved optimal by Mattson et al. [57]. However, their proof is long and somewhat complicated. In collaboration with some researchers from Inha University, we found a shorter and more intuitive proof of optimality for OPT [6].

An intriguing aspect of optimal replacement, seldom mentioned in the literature, is the fact that Belady's MIN algorithm determines OPT hits and misses without the knowledge of future references [54]. Starting from this fact, we searched and found a new algorithm, different from MIN, for determining OPT hits and misses. This algorithm provides new insights about optimal replacement. We show that traces of OPT stack distances have a distinctive structure. In particular, we prove that OPT miss curves are always convex. We show that, like an LRU cache, an OPT cache cannot experience more misses as the reuse distance of references is decreased. Consequently, accessing data circularly is the worst access pattern for OPT, like it is for LRU. We discovered an equivalence between an OPT cache of associativity N with bypassing allowed and an OPT cache of associativity N+1 with bypassing disabled. A paper deriving these results was accepted in ACM TACO and will be presented at the HiPEAC 2017 conference [25].

7.2.2.2. Adaptive Intelligent Memory Systems

Participants: André Seznec, Aswinkumar Sridharan.

Multi-core processors employ shared Last Level Caches (LLC). This trend will continue in the future with large multi-core processors (16 cores and beyond) as well. At the same time, the associativity of this LLC tends to remain in the order of sixteen. Consequently, with large multicore processors, the number of cores that share the LLC becomes larger than the associativity of the cache itself. LLC management policies have been extensively studied for small scale multi-cores (4 to 8 cores) and associativity degree in the 16 range. However, the impact of LLC management on large multi-cores is essentially unknown, in particular when the associativity degree is smaller than the number of cores.

In [48], we introduce Adaptive Discrete and deprioritized Application PrioriTization (ADAPT), an LLC management policy addressing the large multi-cores where the LLC associativity degree is smaller than the number of cores. ADAPT builds on the use of the Footprint-number metric. Footprint-number is defined as the number of unique accesses (block addresses) that an application generates to a cache set in an interval of time. We propose a monitoring mechanism that dynamically samples cache sets to estimate the Footprint-number of applications and classifies them into discrete (distinct and more than two) priority buckets. The cache replacement policy leverages this classification and assigns priorities to cache lines of applications during cache replacement operations. Footprint-number is computed periodically to account the dynamic changes in applications behavior. We further find that de- prioritizing certain applications during cache replacement is beneficial to the overall performance. We evaluate our proposal on 16, 20 and 24-core multi-programmed workloads and discuss other aspects in detail.

[48] got the best paper award at the IPDPS 2016 conference.

7.2.2.3. Augmenting superscalar architecture for efficient many-thread parallel execution

Participants: Sylvain Collange, André Seznec, Sajith Kalathingal.

Threads of Single-Program Multiple-Data (SPMD) applications often exhibit very similar control flows, i.e. they execute the same instructions on different data. In [36] we propose the Dynamic Inter-Thread Vectorization Architecture (DITVA) to leverage this implicit data-level parallelism in SPMD applications by assembling dynamic vector instructions at runtime. DITVA extends an in-order SMT processor with SIMD units with an inter-thread vectorization execution mode. In this mode, multiple scalar threads running in lockstep share a single instruction stream and their respective instruction instances are aggregated into SIMD instructions. To balance thread-and data-level parallelism, threads are statically grouped into fixed-size independently scheduled warps. DITVA leverages existing SIMD units and maintains binary compatibility with existing CPU architectures. Our evaluation on the SPMD applications from the PARSEC and Rodinia OpenMP benchmarks shows that a 4-warp $\times$ 4-lane 4-issue DITVA architecture with a realistic bank-interleaved cache achieves 1.55$\times$ higher performance than a 4-thread 4-issue SMT architecture with AVX instructions while fetching and issuing 51 % fewer instructions, achieving an overall 24 % energy reduction.

Our paper [36] received the Best Paper Award of the SBAC-PAD conference.

7.2.2.4. Generalizing the SIMT execution model to general-purpose instruction sets

Participant: Sylvain Collange.

The *Single Instruction, Multiple Threads* (SIMT) execution model as implemented in NVIDIA Graphics Processing Units (GPUs) associates a multi-thread programming model with an SIMD execution model [59]. It combines the simplicity of scalar code from the programmer's and compiler's perspective with the efficiency of SIMD execution units at the hardware level. However, current SIMT architectures demand specific instruction sets. In particular, they need specific branch instructions to manage thread divergence and convergence. Thus, SIMT GPUs have remained incompatible with traditional general-purpose CPU instruction sets.

We designed Simty, an SIMT processor proof of concept that lifts the instruction set incompatibility between CPUs and GPUs [50]. Simty is a massively multi-threaded processor core that dynamically assembles SIMD instructions from scalar multi-thread code. It runs the RISC-V (RV32-I) instruction set. Unlike existing SIMD or SIMT processors like GPUs, Simty takes binaries compiled for general-purpose processors without any instruction set extension or compiler changes. Simty is described in synthesizable RTL. A FPGA prototype validates its scaling up to 2048 threads per core with 32-wide SIMD units.

7.3. WCET estimation and optimization

Participants: Isabelle Puaut, Damien Hardy, Viet Anh Nguyen, Benjamin Rouxel, Sébastien Martinez, Erven Rohou.

7.3.1. WCET estimation for many core processors

Participants: Viet Anh Nguyen, Damien Hardy, Sébastien Martinez, Isabelle Puaut, Benjamin Rouxel.

7.3.1.1. Optimization of WCETs by considering the effects of local caches

The overall goal of this research is to defined WCET estimation methods for parallel applications running on many-core architectures, such as the Kalray MPPA machine.

Some approaches to reach this goal have been proposed, but they assume the mapping of parallel applications on cores already done. Unfortunately, on architectures with caches, task mapping requires a priori known WCETs for tasks, which in turn requires knowing task mapping (i.e., co-located tasks, co-running tasks) to have tight WCET bounds. Therefore, scheduling parallel applications and estimating their WCET introduce a chicken and egg situation.

We address this issue by developing both optimal and heuristic techniques for solving the scheduling problem, whose objective is to minimize the WCET of a parallel application. Our proposed static partitioned non-preemptive mapping strategies address the effect of local caches to tighten the estimated WCET of the parallel application. Experimental results obtained on real and synthetic parallel applications show that co-locating tasks that reuse code and data improves the WCET.

This research is part of the PIA Capacités project.

7.3.1.2. Accounting for shared resource contentions to minimize WCETs

Accurate WCET analysis for multi-cores is known to be challenging, because of concurrent accesses to shared resources, such as communication through busses or Networks on Chips (NoC). Since it is impossible in general to guarantee the absence of resource conflicts during execution, current WCET techniques either produce pessimistic WCET estimates or constrain the execution to enforce the absence of conflicts, at the price of a significant hardware under-utilization. In addition, the large majority of existing works consider that the platform workload consists of independent tasks. As parallel programming is the most promising solution to improve performance, we envision that within only a few years from now, real-time workloads will evolve toward parallel programs. The WCET behavior of such programs is challenging to analyze because they consist of *dependent* tasks interacting through complex synchronization/communication mechanisms.

In this work, we propose techniques that account for interferences to access shared ressources, in order to minimize the WCET of parallel applications. An optimal and a heuristic method are proposed to map and schedule tasks on multi-cores. These methods take the structure of applications (synchronizations/communications) into consideration to tightly identify shared resource interferences and consequently tighten WCET estimates.

This work is performed in cooperation with Steven Derrien, Angeliki Kritikakou and Imen Fassi from the CAIRN research group and is part of the ARGO H2020 project.

7.3.2. Cache-Persistence-Aware Response-Time Analysis for Fixed-Priority Preemptive Systems

Participants: Damien Hardy, Isabelle Puaut.

A task can be preempted by several jobs of higher priority tasks during its execution. Assuming the worst-case memory demand for each of these jobs leads to pessimistic worst-case response time (WCRT) estimations. Indeed, there is a big chance that a large portion of the instructions and data associated with the preempting task τ_j are still available in the cache when τ_j releases its next jobs. Accounting for this observation allows the pessimism of WCRT analysis to be significantly reduced, which is not considered by existing work.

The four main contributions of this work are: 1) The concept of persistent cache blocks is introduced in the context of WCRT analysis, which allows re-use of cache blocks to be captured, 2) A cache-persistence-aware WCRT analysis for fixed-priority preemptive systems exploiting the PCBs to reduce the WCRT bound, 3) A multi-set extension of the analysis that further improves the WCRT bound and 4) An evaluation showing that our cache-persistence-aware WCRT analysis results in up to 10 % higher schedulability than state-of-the-art approaches.

This work [43] appeared at ECRTS 2016 and was selected as an outstanding paper in this conference.

This work was performed in cooperation with Syed Aftab Rashid, Geoffrey Nelissen, Benny Akesson and Eduardo Tovar from ISEP (Polytechnic Institute of Porto), Portugal.

7.4. Fault Tolerance

7.4.1. WCET estimation for architectures with faulty caches

Participants: Damien Hardy, Isabelle Puaut.

Fine-grained disabling and reconfiguration of hardware elements (functional units, cache blocks) will become economically necessary to recover from permanent failures, whose rate is expected to increase dramatically in the near future. This fine-grained disabling will lead to degraded performance as compared to a fault-free execution.

Until recently, all static worst-case execution time (WCET) estimation methods were assuming fault-free processors, resulting in unsafe estimates in the presence of faults. The first static WCET estimation technique dealing with the presence of permanent faults in instruction caches was proposed in [4]. This study probabilistically quantified the impact of permanent faults on WCET estimates. It demonstrated that the probabilistic WCET (pWCET) estimates of tasks increase rapidly with the probability of faults as compared to fault-free WCET estimates.

New results show that very simple reliability mechanisms allow mitigating the impact of faulty cache blocks on pWCETs. Two mechanisms, that make part of the cache resilient to faults are analyzed. Experiments show that the gain in pWCET for these two mechanisms are on average 48 % and 40 % as compared to an architecture with no reliability mechanism.

This work [35] appeared at DATE 2016 (best paper award for the embedded systems track).

This is joint work with Yannakis Sazeides from University of Cyprus.

PANAMA Project-Team

7. New Results

7.1. Recent results on Sparse Representations, Inverse Problems, and Dimension Reduction

Sparsity, low-rank, dimension-reduction, inverse problem, sparse recovery, scalability, compressive sensing

The team has had a substantial activity ranging from theoretical results to algorithmic design and software contributions in the fields of sparse representations, inverse problems, and dimension reduction, which is at the core of the ERC project PLEASE (Projections, Learning and Sparsity for Efficient Data Processing, see Section 9.2.1.1).

7.1.1. Theoretical results on Sparse Representations, Graph Signal Processing, and Dimension Reduction

Participants: Rémi Gribonval, Yann Traonmilin, Gilles Puy, Nicolas Tremblay, Pierre Vandergheynst.

Main collaboration: Mike Davies (University of Edinburgh), Pierre Borgnat (ENS Lyon), and members of the LTS2 lab of Pierre Vandergheynst at EPFL

Stable recovery of low-dimensional cones in Hilbert spaces: Many inverse problems in signal processing deal with the robust estimation of unknown data from underdetermined linear observations. Low dimensional models, when combined with appropriate regularizers, have been shown to be efficient at performing this task. Sparse models with the ℓ_1-norm or low rank models with the nuclear norm are examples of such successful combinations. Stable recovery guarantees in these settings have been established using a common tool adapted to each case: the notion of restricted isometry property (RIP). We established generic RIP-based guarantees for the stable recovery of cones (positively homogeneous model sets) with arbitrary regularizers. These guarantees were illustrated on selected examples. For block structured sparsity in the infinite dimensional setting, we used the guarantees for a family of regularizers which efficiency in terms of RIP constant can be controlled, leading to stronger and sharper guarantees than the state of the art. This has been published in a journal paper [21].

Recipes for stable linear embeddings from Hilbert spaces to $\mathbb{R}^m$: We considered the problem of constructing a linear map from a Hilbert space (possibly infinite dimensional) to $\mathbb{R}^m$ that satisfies a restricted isometry property (RIP) on an arbitrary signal model set. We obtained a generic framework that handles a large class of low-dimensional subsets but also *unstructured* and *structured* linear maps. We provided a simple recipe to prove that a random linear map satisfies a general RIP on the model set with high probability. We also described a generic technique to construct linear maps that satisfy the RIP. Finally, we detailed how to use our results in several examples, which allow us to recover and extend many known compressive sampling results. This has been presented at the conference EUSIPCO 2015 [90], and a journal paper is under revision [91].

Signal processing on graphs: from filtering to random sampling and robust PCA: Graph signal processing is an emerging field aiming at extending classical tools from signal processing (1D time series) and image processing (2D pixel grids, 3D voxel grids) to more loosely structured numerical data: collections of numerical values each associated to a vertex of a graph, where the graph encodes the underlying "topology" of proximities and distances. Since our pioneering contributions on this topic [4], the team regularly works on various aspects of graph signal processing, in collaboration with the LTS2 lab of Pierre Vandergheynst at EPFL. This year, we studied the problem of sampling k-bandlimited signals on graphs. We proposed two sampling strategies that consist in selecting a small subset of nodes at random. The first strategy is non-adaptive, i.e., independent of the graph structure, and its performance depends on a parameter called the graph coherence. On the contrary, the second strategy is adaptive but yields optimal results. Indeed, no more than $O(klog(k))$ measurements are sufficient to ensure an accurate and stable recovery of all k-bandlimited signals. This second strategy is based on a careful choice of the sampling distribution, which can be estimated quickly. Then, we proposed a

computationally efficient decoder to reconstruct k-bandlimited signals from their samples. We proved that it yields accurate reconstructions and that it is also stable to noise. Finally, we conducted several experiments to test these techniques. A journal paper has been published [17] accompanied by a toolbox for reproducible research (see Section 6.14). Other contributions from this year on the topic of graph signal processing include new subgraph-based filterbanks for graph signals [22], and new accelerated and robustified techniques for PCA on graphs [19], [20] (see also below our contributions in terms of new algorithms to obtain approximate Fast Graph Fourier Transforms [32], [53]).

Accelerated spectral clustering: We leveraged the proposed random sampling technique to propose a faster spectral clustering algorithm. Indeed, classical spectral clustering is based on the computation of the first k eigenvectors of the similarity matrix' Laplacian, whose computation cost, even for sparse matrices, becomes prohibitive for large datasets. We showed that we can estimate the spectral clustering distance matrix without computing these eigenvectors: by graph filtering random signals. Also, we took advantage of the stochasticity of these random vectors to estimate the number of clusters k. We compared our method to classical spectral clustering on synthetic data, and showed that it reaches equal performance while being faster by a factor at least two for large datasets of real data. Two conference papers have been presented, at ICASSP 2016 [39] and ICML 2016 [40] and a toolbox for reproducible research has been released (see Section 6.4).

7.1.2. An Alternative Framework for Sparse Representations: Sparse "Analysis" Models

Participants: Rémi Gribonval, Nancy Bertin, Srdan Kitic, Clément Gaultier.

In the past decade there has been a great interest in a synthesis-based model for signals, based on sparse and redundant representations. Such a model assumes that the signal of interest can be composed as a linear combination of *few* columns from a given matrix (the dictionary). An alternative *analysis-based* model can be envisioned, where an analysis operator multiplies the signal, leading to a *cosparse* outcome.

Building on our pioneering work on the cosparse model [7] [73], [87] successful applications of this approach to sound source localization, audio declipping and brain imaging have been developed in 2015 and 2016. In addition, new applications to audio denoising were also introduced this year.

Versatile cosparse regularization: Digging the groove of previous years' results (comparison of the performance of several cosparse recovery algorithms in the context of sound source localization [77], demonstration of its efficiency in situations where usual methods fail ([79], see paragraph 7.4.2), applicability to the hard declipping problem [78], application to EEG brain imaging [56]), a journal paper embedding the latest algorithms and results in sound source localization and brain source localization in a unified fashion was published this year [5]. This framework was also exploited to extend results on audio inpainting (see Section 7.3.2).

New results include experimental confirmation of robustness and versatility of the proposed scheme, and of its computational merits (convergence speed increasing with the amount of data). In a work presented in a workshop [44], we also proposed a multiscale strategy that aims at exploiting computational advantages of both sparse and cosparse regularization approaches, thanks to the simple yet effective all-zero initialization which the synthesis-based optimization can benefit from, while retaining the computational properties of the analysis-based approach for huge scale optimization problems arising in physics-driven settings.

Parametric operator learning for cosparse calibration: In many inverse problems, a key challenge is to cope with unknown physical parameters of the problem such as the speed of sound or the boundary impedance. In the sound source localization problem, we previously showed that the unknown speed of sound can be learned jointly in the process of cosparse recovery, under mild conditions [58], [81]. This year, we extended the formulation to the case of unknown boundary impedance, and showing that a similar biconvex formulation and optimization could solve this new problem efficiently (conference paper published in ICASSP 2016 [29], see also Section 7.3.3).

7.1.3. Algorithmic and Theoretical results on Computational Representation Learning

Participants: Rémi Gribonval, Luc Le Magoarou, Nicolas Bellot, Adrien Leman, Cassio Fraga Dantas, Igal Rozenberg.

An important practical problem in sparse modeling is to choose the adequate dictionary to model a class of signals or images of interest. While diverse heuristic techniques have been proposed in the literature to learn a dictionary from a collection of training samples, classical dictionary learning is limited to small-scale problems. Inspired by usual fast transforms, we proposed a general dictionary structure that allows cheaper manipulation, and an algorithm to learn such dictionaries together with their fast implementation. The principle and its application to image denoising appeared at ICASSP 2015 [84] and an application to speedup linear inverse problems was published at EUSIPCO 2015 [83]. A Matlab library has been released (see Section 6.6) to reproduce the experiments from the comprehensive journal paper published this year [16], which additionally includes theoretical results on the improved sample complexity of learning such dictionaries. Pioneering identifiability results have been obtained in the Ph.D. thesis of Luc Le Magoarou on this topic [85].

We further explored the application of this technique to obtain fast approximations of Graph Fourier Transforms. A conference paper on this latter topic appeared in ICASSP 2016 [32], and a journal paper has been submitted [53] where we empirically show that $\mathcal{O}(n \log n)$ approximate implementations of Graph Fourier Transforms are possible for certain families of graphs. This opens the way to substantial accelerations for Fourier Transforms on large graphs.

A C++ software library has been developed (see Section 6.6) to release the resulting algorithms.

7.2. Activities on Waveform Design for Telecommunications

Peak to Average Power Ratio (PAPR), Orthogonal Frequency Division Multiplexing (OFDM), Generalized Waveforms for Multi Carrier (GWMC), Adaptive Wavelet Packet Modulation (AWPM)

7.2.1. Characterizing and designing multi-carrier waveform systems with optimum PAPR

Participant: Rémi Gribonval.

Main collaboration: Marwa Chafii, Jacques Palicot, Carlos Bader (Equipe SCEE, Supelec, Rennes)

In the context of the TEPN (Towards Energy Proportional Networks) Comin Labs project (see Section 9.1.1.2), in collaboration with the SCEE team at Supelec (thesis of Marwa Chafii [64], defended in October this year and co-supervised by R. Gribonval), we investigated a problem related to dictionary design: the characterization of waveforms with low Peak to Average Power Ratio (PAPR) for wireless communications. This is motivated by the importance of a low PAPR for energy-efficient transmission systems. A first stage of the work consisted in characterizing the statistical distribution of the PAPR for a general family of multi-carrier systems, leading to a journal paper [67] and several conference communications [65], [66]. Our characterization of waveforms with optimum PAPR [68] has been published in a journal this year [14]. The work this year has concentrated on designing new adaptive multi-carrier waveform systems able to cope with frequency-selective channels while minimizing PAPR. This has given rise to a patent [49] and a journal paper is in preparation.

7.3. Emerging activities on Compressive Learning and Nonlinear Inverse Problems

Compressive sensing, compressive learning, audio inpainting, phase estimation

7.3.1. Phase Estimation in Multichannel Mixtures

Participants: Antoine Deleforge, Yann Traonmilin.

The problem of estimating source signals given an observed multichannel mixture is fundamentally ill-posed when the mixing matrix is unknown or when the number of sources is larger that the number of microphones. Hence, prior information on the desired source signals must be incorporated in order to tackle it. An important line of research in audio source separation over the past decade consists in using a model of the source signals' magnitudes in the short-time Fourier domain [8]. Such models can be inferred through, *e.g.*, non-negative matrix factorization [89] or deep neural networks [88]. Magnitudes estimates are often interpreted as instantaneous variances of Gaussian-process source signals, and are combined with Wiener filtering for source separation. In [50], we introduced a shift of this paradigm by considering the *Phase Unmixing* problem: how can one recover the instantaneous phases of complex mixed source signals when their magnitudes and mixing matrix are known? This problem was showed to be NP-hard, and three approaches were proposed to tackle it: a heuristic method, an alternate minimization method, and a convex relaxation into a semi-definite program. The last two approaches were showed to outperform the oracle multichannel Wiener filter in under-determined informed source separation tasks. The latter yielded best results, including the potential for exact source separation in under-determined settings.

7.3.2. Audio Inpainting and Denoising

Participants: Rémi Gribonval, Nancy Bertin, Srdan Kitic.

Inpainting is a particular kind of inverse problems that has been extensively addressed in the recent years in the field of image processing. Building upon our previous pioneering contributions (definition of the audio inpainting problem as a general framework for many audio processing tasks, application to the audio declipping or desaturation problem, formulation as a sparse recovery problem [55]), we proposed over the last two years a series of algorithms leveraging the competitive cosparse approach, which offers a very appealing trade-off between reconstruction performance and computational time [78], [80], [81]. The work on cosparse audio declipping which was awarded the Conexant best paper award at the LVA/ICA 2015 conference [80], together with the associated toolbox for reproducible research (see Section 6.8) draw the attention of a world leading company in professional audio signal processing, with which some transfer has been negotiated. In 2016, real-time implementation of the A-SPADE algorithm was obtained and demonstrated at various events (HCERES evaluation, TechnofÃ©rence # 18 « Nouvelles expériences son et vidéo », ...).

Current and future works deal with developing advanced (co)sparse decomposition for audio inpainting, including several forms of structured sparsity (*e.g.* temporal and multichannel joint-sparsity), dictionary learning for inpainting, and several applicative scenarios (declipping, denoising, time-frequency inpainting, joint source separation and declipping). In particular, we investigated the incorporation of the so-called "social" structure constraint [82] into problems regularized by a cosparse prior, including declipping and denoising. Publication of this work is currently under preparation.

7.3.3. Blind Calibration of Impedance and Geometry

Participants: Rémi Gribonval, Nancy Bertin, Srdan Kitic.

Main collaborations: Laurent Daudet, Thibault Nowakowski, Julien de Rosny (Institut Langevin)

Last year, we also investigated extended inverse problem scenarios where a "lack of calibration" may occur, i.e., when some physical parameters are needed for reconstruction but apriori unknown: speed of sound, impedance at the boundaries of the domain where the studied phenomenon propagates, or even the shape of these boundaries. In a first approach, based on our physics-driven cosparse regularization of the sound source localization problem [5] (see section 7.1.2), we managed to preserve the sound source localization performance when the speed of sound is unknown, or, equally, when the impedance is unknown, provided the shape is and under some smoothness assumptions. Unlike the previous case (gain calibration), the arising problems are not convex but biconvex, and can be solved with proper biconvex formulation of ADMM algorithm. In a second approach based on eigenmode decomposition (limited to a 2D membrane), we showed that impedance learning with known shape, or shape learning with known impedance can be expressed as two facets of the same problem, and solved by the same approach, from a small number of measurements. Two papers presenting these two sets of results appeared at ICASSP 2016 [29], [37].

7.3.4. Sketching for Large-Scale Mixture Estimation

Participants: Rémi Gribonval, Nicolas Keriven.

Main collaborations: Patrick Perez (Technicolor R&I France) Anthony Bourrier (formerly Technicolor R&I France, then GIPSA-Lab)

When fitting a probability model to voluminous data, memory and computational time can become prohibitive. We proposed during the Ph.D. thesis of Anthony Bourrier [60] a framework aimed at fitting a mixture of isotropic Gaussians to data vectors by computing a low-dimensional sketch of the data. The sketch represents empirical moments of the underlying probability distribution. Deriving a reconstruction algorithm by analogy with compressive sensing, we experimentally showed that it is possible to precisely estimate the mixture parameters provided that the sketch is large enough. The proposed algorithm provided good reconstruction and scaled to higher dimensions than previous probability mixture estimation algorithms, while consuming less memory in the case of voluminous datasets. It also provided a potentially privacy-preserving data analysis tool, since the sketch does not explicitly disclose information about individual datum it is based on [63], [61], [62]. Last year, we consolidated our extensions to non-isotropic Gaussians, with new algorithms [76] and conducted large-scale experiments demonstrating its potential for speaker verification. A conference paper appeared at ICASSP 2016 [31] and a journal version has been submitted [52], accompanied by a toolbox for reproducible research (see Section 6.12).

This year the work concentrated on extending the approach beyond the case of Gaussian Mixture Estimation. First, we showed empirically that the algorithm can be adapted to sketch a training collection while still allowing to compute clusters. The approach, called "Compressive K-means", is described in a paper accepted at ICASSP 2017 [27]. Then, we expressed a theoretical framework for sketched learning, encompassing statistical learning guarantees as well as dimension reduction guarantees. The framework already covers compressive K-means as well as compressive Principal Component Analysis (PCA), and a conference paper has been submitted. A comprehensive journal paper is under preparation, and future work will include expliciting the impact of the proposed framework on a wider set of concrete learning problems.

7.4. Source Separation and Localization

Source separation, sparse representations, probabilistic model, source localization

Source separation is the task of retrieving the source signals underlying a multichannel mixture signal.

About a decade ago, state-of-the-art approaches consisted of representing the signals in the time-frequency domain and estimating the source coefficients by sparse decomposition in that basis. These approaches rely only on spatial cues, which are often not sufficient to discriminate the sources unambiguously. Over the last years, we proposed a general probabilistic framework for the joint exploitation of spatial and spectral cues [8], which generalizes a number of existing techniques including our former study on spectral GMMs [57]. We showed how it could be used to quickly design new models adapted to the data at hand and estimate its parameters via the EM algorithm, and it became the basis of a large number of works in the field, including our own. In the last years, improvements were obtained through the use of prior knowledge about the source spatial covariance matrices [71], [75], [74], knowledge on the source positions and room characteristics [72], or a better initialization of parameters thanks to specific source localization techniques [59].

This accumulated progress lead, in 2015, to two main achievements: a new version of the Flexible Audio Source Separation Toolbox, fully reimplemented, was released [92] and we published an overview paper on recent and going research along the path of *guided* separation in a special issue of IEEE Signal Processing Magazine devoted to source separation and its applications [10]. This two achievements formed the basis of our work in 2016, exploring intensively the concrete use of these tools and principles in real-world scenarios, in particular within the voiceHome project (see Section 6.13).

7.4.1. Towards Real-world Separation and Remixing Applications

Participants: Nancy Bertin, Frédéric Bimbot, Ewen Camberlein, Romain Lebarbenchon.

In 2015, we began a new industrial collaboration, in the context of the VoiceHome project, aiming at another challenging real-world application: natural language dialog in home applications, such as control of domotic and multimedia devices. As a very noisy and reverberant environment, home is a particularly challenging target for source separation, used here as a pre-processing for speech recognition (and possibly with stronger interactions with voice activity detection or speaker identification tasks as well). In 2016, we publicly released a realistic corpus of room impulse responses and utterances recorded in real homes, and presented it during the Interspeech conference [28]. We also continued benchmarking and adapting existing localization and separation tools to the particular context of this application, worked on a better interface between source localization and source separations steps, and investigated new means to reduce the latency and computational burden of the currently available tools (low-resolution source separation preserving speech recognition improvement, automatic selection of the best microphones, joint localization and multichannel speech / non speech classification prior to any separation).

In november 2016, we started investigating a new application of source separation to sound respatialization from Higher Order Ambisonics (HOA) signals, in the context of free navigation in 3D audiovisual contents. This work is conducted in a collaboration with the IRT b<>Com, through the Ph.D. of Mohammed Hafsati (co-supervised by Nancy Bertin, RÃ©mi Gribonval).

7.4.2. Implicit Localization through Audio-based Control for Robotics

Participant: Nancy Bertin.

Main collaborations (audio-based control for robotics): Aly Magassouba and François Chaumette (Inria, EPI LAGADIC, France)

Acoustic source localization is, in general, the problem of determining the spatial coordinates of one or several sound sources based on microphone recordings. This problem arises in many different fields (speech and sound enhancement, speech recognition, acoustic tomography, robotics, aeroacoustics...) and its resolution, beyond an interest in itself, can also be the key preamble to efficient source separation. Common techniques, including beamforming, only provides the *direction of arrival* of the sound, estimated from the *Time Difference of Arrival* (*TDOA*) [59]. This year, we have particularly investigated alternative approaches, either where the explicit localization is not needed (audio-based control of a robot) or, on the contrary, where the exact location of the source is needed and/or TDOA is irrelevant (cosparse modeling of the acoustic field, see Section 7.1.2).

In robotics, the use of aural perception has received recently a growing interest but still remains marginal in comparison to vision. Yet audio sensing is a valid alternative or complement to vision in robotics, for instance in homing tasks. Most existing works are based on the relative localization of a defined system with respect to a sound source, and the control scheme is generally designed separately from the localization system.

In contrast, the approach that we investigate over the last three years focuses on a sensor-based control approach. We proposed a new line of work, by considering the hearing sense as a direct and real-time input of a closed loop control scheme for a robotic task. Thus, and unlike most previous works, this approach does not necessitate any explicit source localization: instead of solving the localization problem, we focus on developing an innovative modeling based on sound features. To address this objective, we placed ourselves in the sensor-based control framework, especially visual servoing (VS) that has been widely studied in the past [69].

Last year, we established an analytical model linking the Interaural Time Difference (ITD) sound features and control input of the robot, defined and analyzed robotic homing tasks involving multiple sound sources, and validated the proposed approach by simulations and experiments with an actual robot [86]. This year, we consolidated these results and extended the range of applicative tasks [36] and obtained similar results (including theoretical and experimental) for the Interaural Level Difference (ILD), in combination with the absolute energy level [34]. Another set of experiments, presented during the IROS workshop [35] was successfully carried with a humanoid robot, notably without any measurement nor modeling of the robot's Head Relative Transfer Functions (HRTF). This work was mainly lead by Aly Magassouba, who defended his Ph.D. (co-supervised by Nancy Bertin and François Chaumette) in December 2016.

7.4.3. Emerging activities on Virtually-Supervised Sound Localization

Participants: Antoine Deleforge, Clément Gaultier, Saurabh Kataria.

Audio source localization consists in estimating the position of one or several sound sources given the signals received by a microphone array. It can be decomposed into two sub-tasks : (i) computing spatial auditory features from raw audio input and (ii) mapping these features to the desired spatial information.

Extracting spatial features from raw audio input: The most commonly used features in binaural (two microphones) sound source localization are frequency-dependent phase and level differences between the two microphones. To handle the presence of noise, several sources, or reverberation, most existing methods rely on some kind of aggregation of these features in the time-frequency plane, often in a heuristic way. In [25], we introduced the rectified binaural ratio as a new spatial feature. We showed that for Gaussian point-source signals corrupted by stationary Gaussian noise, this ratio follows a complex t-distribution with explicit parameters. This new formulation provides a principled, statistically sound and efficient method to aggregate these features in the presence of noise. Experiments notably showed the higher robustness of these features compared to traditional ones, in the task of localizing heavily corrupted speech signals.

Mapping features to spatial information: Existing methods to map auditory features to spatial properties divide into two categories. *Physics-driven* methods attempt to estimate an explicit mapping based on an approximate physical model of sound propagation in the considered system. *Data-driven* methods bypass the use of a physical model by learning the mapping from a training set, obtained by manually annotating features extracted from real data. We proposed a new paradigm that aims at making the best of physics-driven and data-driven approaches, referred to as *virtually-supervised acoustic space mapping* [26], [51]. The idea is to use a physics-based room-acoustic simulator to generate arbitrary large datasets of room-impulse responses corresponding to various acoustic environments, adapted to the physical audio system at hand. We demonstrated that mappings learned from these data could potentially be used to not only estimate the 3D position of a source but also some acoustical properties of the room [51]. We also showed that a virtually-learned mapping could robustly localize sound sources from real-world binaural input, which is the first result of this kind in audio source localization [26].

7.5. Music Content Processing and Information Retrieval

Music structure, music language modeling, System & Contrast model

Current work developed in our research group in the domain of music content processing and information retrieval explore various information-theoretic frameworks for music structure analysis and description [24], in particular the System & Contrast model [1].

7.5.1. Tensor-based Representation of Sectional Units in Music

Participants: Corentin Guichaoua, Frédéric Bimbot.

Following Kolmogorov'ÂÂs complexity paradigm, modeling the structure of a musical segment can be addressed by searching for the compression program that describes as economically as possible the musical content of that segment, within a given family of compression schemes.

In this general framework, packing the musical data in a tensor-derived representation enables to decompose the structure into two components : (i) the shape of the tensor which characterizes the way in which the musical elements are arranged in an n-dimensional space and (ii) the values within the tensor which reflect the content of the musical segment and minimize the complexity of the relations between its elements.

This approach is currently developed and tested for the grouping of chord sequences into sectional units for pop music songs, with very encouraging segmentation results on pop songs.

7.5.2. Minimal Transport Graphs for the Modeling of Chord Progressions

Participants: Corentin Louboutin, Frédéric Bimbot.

In this work, we model relations between chords by minimal transport and we investigate different types of dependencies within chord sequences [33]. For this purpose we use the ÂÂSystem & Contrast (S&C) model [1], designed for the description of music sectional units, to infer non-sequential structures called chord progression graphs (CPG).

Minimal transport is defined as the shortest displacement of notes, in semitones, between a pair of chords. The paper [33] present three algorithms to find CPGs for chords sequences: one is sequential, and two others are based on the S& C model. The three methods are compared using the perplexity as an efficiency measure.

The experiments on a corpus of 45 segments taken from songs of multiple genres indicate that optimization processes based on the S&C model outperform the sequential model with a decrease in perplexity over 1.0.

7.5.3. Regularity Constraints for the Fusion of Music Structure Segmentation System

Participant: Frédéric Bimbot.

Main collaborations Gabriel Sargent (EPI LinkMedia, Rennes, France)

Music structure estimation has recently emerged as a central topic within the field of Music Information Retrieval. Indeed, as music is a highly structured information stream, knowledge of how a music piece is organized represents a key challenge to enhance the management and exploitation of large music collections.

Former work carried out in our group [9] has illustrated the benefits that can be expected from a regularity constraint on the structural segmentation of popular music pieces : a constraint which favors structural segments of comparable size provides a better conditioning of the boundary estimation process.

As a further investigation, we have explored the benefits of the regularity constraint as an efficient way for combining the outputs of a selection of systems presented at MIREX between 2010 and 2015. These experiments have yielded a level of performance which is competitive to that of the state-of-the-art on the ”MIREX10" dataset (100 J-Pop songs from the RWC database) [18].

SERPICO Project-Team

7. New Results

7.1. Statistical aggregation methods for image denoising and estimation

Participants: Charles Kervrann, Frédéric Lavancier.

In the line of the Non-Local means [43] and ND-SAFIR [9], [10], [5] denoising algorithms, we have proposed a novel adaptive estimator based on the weighted average of observations taken in a neighborhood with weights depending on image data. The idea is to compute adaptive weights that best minimize an upper bound of the pointwise L_2 risk. In the framework of adaptive estimation, we show that the "oracle" weights depend on the unknown image and are optimal if we consider triangular kernels instead of the commonly-used Gaussian kernel. Furthermore, we propose a way to automatically choose the spatially varying smoothing parameter for adaptive denoising. Under conventional minimal regularity conditions, the obtained estimator converges at the usual optimal rate. The implementation of the proposed algorithm is also straightforward. The simulations show that our algorithm improves significantly the classical NL-means, and is competitive when compared to the more sophisticated NL-means filters both in terms of PSNR values and visual quality.

Previously, we investigated statistical aggregation methods which optimally combine several estimators to produce a boosted solution [11]. In this range of work, we also introduced in [24] a general method to combine estimators in order to produce a better estimate. From a theoretical point of view, we proved that this method is optimal in some sense. It is illustrated on standard statistical problems in parametric and semi-parametric models where the averaging estimator outperforms the initial estimators in most cases. This method has been subsequently adapted in [39] to models in spatial statistics. As part of an on-going work, we are applying this method to improve patch-based image denoising algorithms.

References: [24] [39]

Collaborators: Qiyu Jin (School of Mathematical Science, Inner Mongolia University, China),
Ion Grama and Quansheng Liu (University of Bretagne-Sud, Vannes),
Paul Rochet (Laboratoire de Mathématiques Jean Leray (LMJL), University of Nantes).

7.2. Algorithms for dejittering and deconvolving large fluorescence and Tissue MicroArray (TMA) images

Participants: Hoai Nam Nguyen, Giovanni Petrazzuoli, Aminata Diouf, Charles Kervrann.

In fluorescence microscopy, the image quality is limited by out-of-focus blur and high noise. Traditionally, image deconvolution is needed to estimate a good quality version of the observed image. The result of deconvolution depends heavily on the choice of the regularization term and the noise dependent fidelity term. The regularization functional should be designed to remove noise while preserving image discontinuities. Accordingly, we investigated new regularization terms to preserve fine details of underlying structures and we studied appropriate proximal algorithms. The deconvolution method has been especially dedicated to large 2D 20000×60000 images acquired with ISO scan imager (see Fig.3). The images are preliminary pre-processed to compensate non constant pixel displacement during acquisition/scanning (dejittering effect). The method has also been evaluated on 2D Vimentin filament images (UTSW, CytoDI Associated Team) to facilitate filament segmentation. The method is able to process a 512×512 image in 250 ms with a non optimized implementation.

Collaborators: Vincent Paveau and Cyril Cauchois (Innopys company),
Philippe Roudot (UTSW, Dallas, USA).

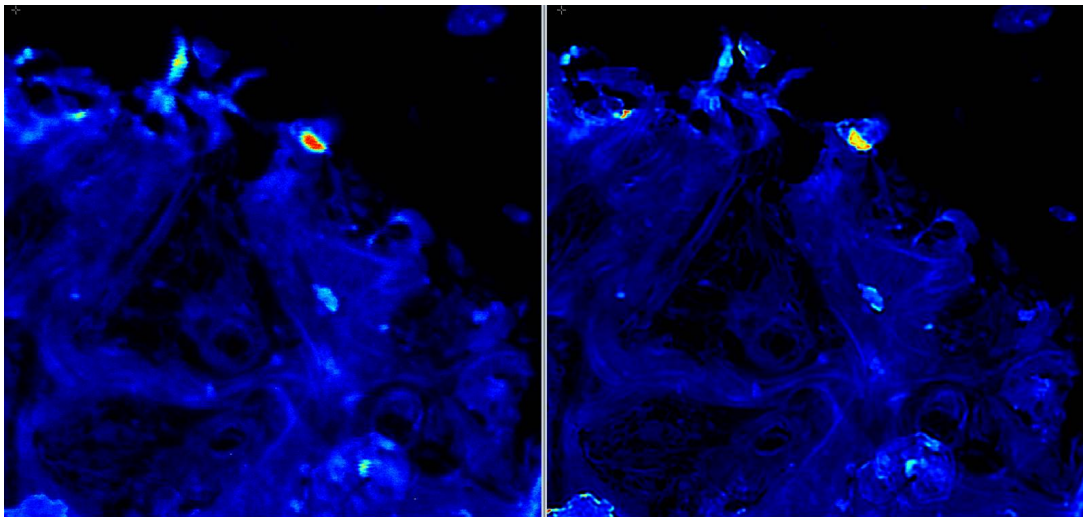


Figure 3. Illustration of image dejittering and deconvolution algorithms applied to a selected fluorescence region extracted from a large TMA image (by courtesy of Innopsys).

7.3. Quantifying the spatial distribution of intracellular events

Participants: Thierry Pécot, Charles Kervrann.

Automated processing of fluorescence microscopy data allows quantifying cell phenotypes in an objective and reproducible way. However, most computational methods are based on the complex combination of heterogeneous features such as statistical, geometrical, morphological and frequency properties, which makes difficult to draw definitive biological conclusions. Additionally, most experimental designs, especially at single cell level, pool together data coming from replicated experiments of a given condition, neglecting the biological variability between individual cells. To address these issues, we developed a generic and non-parametric framework (QuantEv) to study the spatio-temporal distribution of moving Rab6 membranes and the effect of actin disruption on Rab11 trafficking in coordination with cell shape. The main advantage of QuantEv is to process robustly and accurately homogeneous and heterogeneous populations. As demonstration, we compared the results obtained by QuantEv with those from kernel density maps, for Rab6 positive membranes on crossbow- and disk-shaped cells.

Collaborators: Jean Salamero, Jérôme Boulanger and Liu Zengzhen (UMR 144 CNRS-Institut Curie).

7.4. Correlation-based method for membrane diffusion estimation during exocytosis in TIRFM

Participants: Ancageorgiana Caranfil, Charles Kervrann.

The dynamics of the plasma membrane of the cell is not fully understood yet; one of the crucial aspects to clarify is the diffusion process during exocytosis. Several image acquisition modalities exist, including TIRFM (Total Internal Reflection Fluorescence Microscopy), that have successfully been used to determine the successive steps of exocytosis. However, computing characteristic values for plasma membrane dynamics is problematic, as the experimental conditions have a strong influence on the obtained data, and a general model of molecular interaction dynamics cannot be determined.

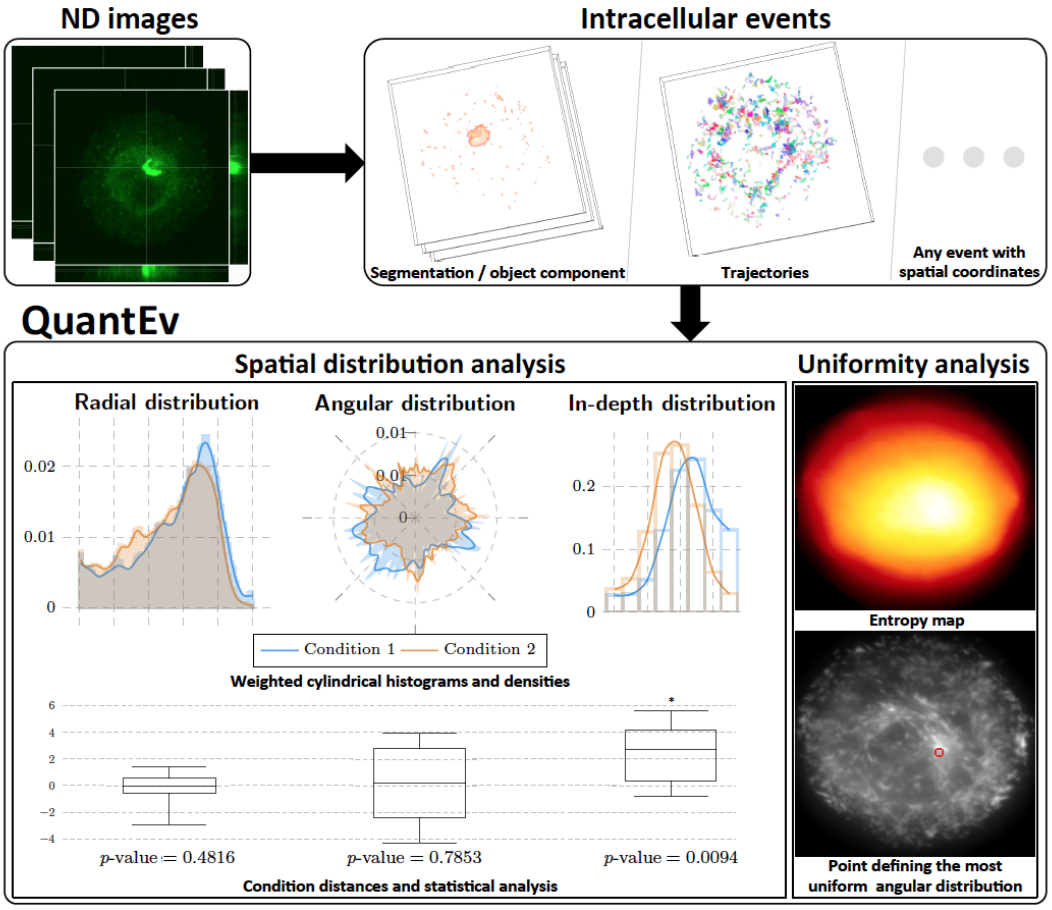


Figure 4. Overview of QuantEv approach.

This year, we have continued our study of correlation-based methods for local diffusion estimation in TIRFM images. Our original method was tested on both synthetic and real images showing an isolated diffusion event, and a robust algorithm was developed to cope with noisy data. Our first model was linear and had only two parameters to estimate. Diffusion coefficient estimation was accurate on synthetic images even with moderate to low signal-to-noise ratio, and within reasonable margins of error on real images with little noise. We have then extended our mathematical model by using a global approach subject to initial local diffusion conditions. Isolated diffusion events are well described, but this new model can also handle the case of noisy images with non-uniform background, and the case of two or more diffusion events in the region of interest. The extended model is non-linear but has few parameters to estimate. An iterative minimization method is used to fit the model parameters to the data points (see Fig. 5). Despite non-linearity, results are accurate on images with pure diffusion events and show robustness to background. The quality of parameters estimation is barely influenced by the length and size of the input TIRFM sequence, which is not the case with standard correlation methods. We have thus developed a correlation-based method that is able to estimate diffusion in a variety of cases in TIRFM images (Fig. 5).

Collaborators: Francois Waharte (UMR 144 CNRS-Institut Curie, PICT-IBiSA),
Perrine Paul-Gilloteaux (UMS 3556, IRS-UN, Nantes).

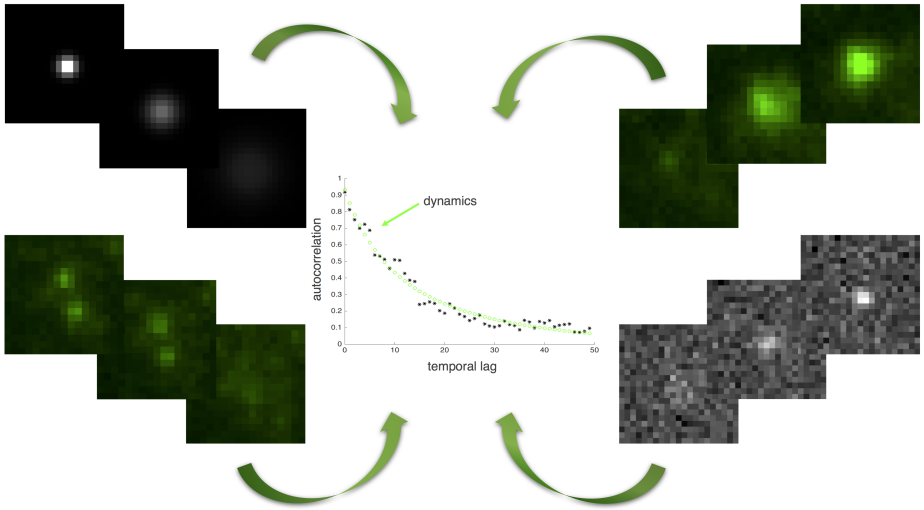


Figure 5. Local diffusion estimation in both synthetic (black and white spots) and real TIRF images (green spots, UMR 144 CNRS-Institut Curie, PICT-IBiSA). Only three frames of each sequence are shown. The computed autocorrelation values (in black) for different temporal lag values, and the fitting function for these values (in green) are given (plot in the middle) for the upper-right sequence.

7.5. Colocalization for live cell and super-resolution imaging

Participants: Frédéric Lavancier, Thierry Pécot, Charles Kervrann.

In the context of bioimaging, colocalization refers to the detection of emissions from two or more fluorescent molecules within the same pixel of the image. This approach enables to quantify the protein-protein interactions inside the cell, just at the resolution limit of the microscope. In statistics, this amounts to characterizing the joint spatial repartition and the spatial overlap between different fluorescent labels. Two distinct categories of colocalization approaches are considered to address this issue: intensity-based methods and object-based methods. The popular (intensity-based) Pearson's correlation method, which returns values between -1 and +1, is known to be sensitive to high intensity backgrounds and provides errors if the signal-to-noise ratio (SNR) is typically low. The object-based method, recommended in single molecule imaging, analyses the spatial distribution of the two sets of detected spots by using point process statistics.

Accordingly, we developed an original, fast, robust-to-noise and versatile approach that reconciles intensity-based and object-based methods for both conventional diffraction-limited microscopy and sub-resolved microscopy. The procedure is only controlled by a p-value and tests whether the Pearson correlation between two binary images is significantly positive. This amount to quantifying the interaction strength by the area/volume of the intersection between the two binary images viewed as random distributions of geometrical objects. Under mild assumptions, it turns out that the appropriately normalized Pearson correlation follows a standard normal distribution under the null hypothesis if the number of image pixels is large. Unlike previous methods, the method handles 2D and 3D images, variable SNRs and any kind of cell shapes. It is able to colocalize large regions with small dots, as it is the case in TIRF-PALM experiments and to detect negative colocalization. The typical processing time is two milliseconds per image pair in 2D and a few seconds in 3D, with no dependence on the number of objects per image. Finally, the method provides maps to geocolocalize molecule interactions in specific image regions.

Collaborators: Jean Salamero and Liu Zengzhen (UMR 144 CNRS-Institut Curie).

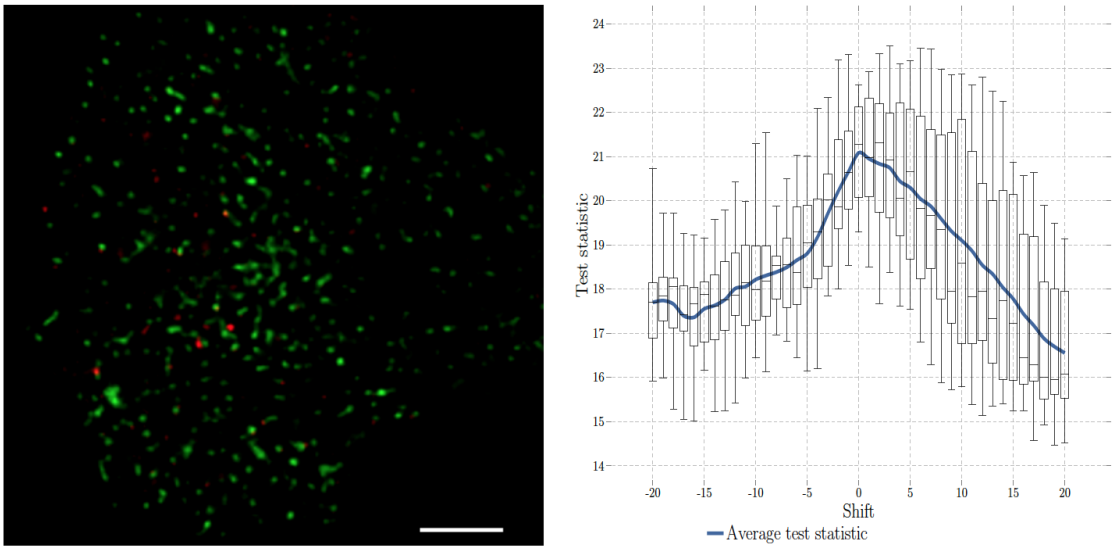


Figure 6. Illustration of statistical spatio-temporal colocalization. a) Average intensity projection of a 3D TIRF acquisition showing m-Cherry Langerin (red channel) and GFP Rab11A proteins (green channel) (courtesy of UMR 144 CNRS-Institut Curie); b) test statistic value with respect to the shift between frames.

7.6. Classification of diffusion dynamics from particle trajectories

Participants: Vincent Briane, Charles Kervrann.

In this study, we are currently interested in describing the dynamics of particles inside live cell. We assume that the motions of particles follow a certain class of random process: the diffusion processes. In 2015, we developed a statistical test to classify the intracellular motions into three groups : free diffusion (*i-e* Brownian motion), subdiffusion and superdiffusion. This method is an alternative to the commonly used Mean Square Displacement (MSD) analysis. This year, we have studied theoretical properties of our procedure. We have shown that it behaves well asymptotically, that is when we observe the particle trajectory for a very long time, for certain parametric models. The models on which we assess our procedure are representative of the three classes aforementioned and extensively used in the literature. Among them we can cite Brownian motion with drift, Ornstein-Uhlenbeck process and fractional Brownian motion. An illustration of the testing procedure is shown in Fig. 7 .

We also extend our method to address two different questions. First, we are interested in testing a large number of trajectories. The first version of our test is a single test procedure. It is known that applying multiple times a test without care leads to a high number of false positives. Then, we modify our initial method to overcome this problem. Secondly, in the case in which we observe very long trajectories, it is likely that the particle motion changes over time. Therefore, we are currently adapting our initial procedure to detect change-point along a single trajectory.

Collaborators: Myriam Vimond (ENSAI Rennes),
Jean Salamero (UMR 144 CNRS-Institut Curie).

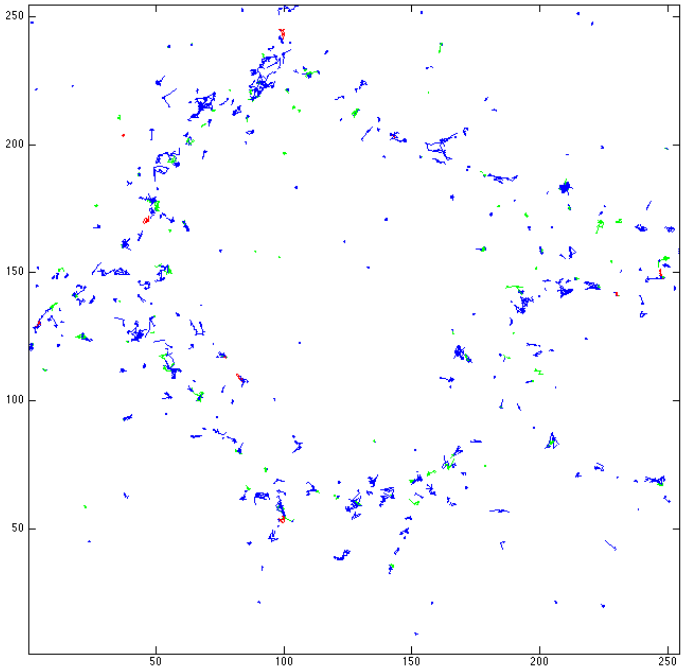


Figure 7. Labelling of the dynamics of trajectories in Single Particle Tracking PALM (Courtesy of Institut Interdisciplinaire de Neurosciences CNRS UMR 5297). The color code is green for subdiffusion, blue for Brownian motion and red for superdiffusion.

7.7. Inference for spatial Gibbs point processes

Participant: Frédéric Lavancier.

Gibbs point processes are popular and widely used models in spatial statistics to describe the repartition of points or geometrical structures in space. They initially arose from statistical physics where they are models for interacting particles. They are now used in as different domains as astronomy, biology, computer science, ecology, forestry, image analysis and materials science.

Assuming a parametric form of the Gibbs interaction, the natural method to estimate the parameters is likelihood inference. Since its first use in the 80's, this method is conjectured to be consistent and efficient. However the theoretical properties of maximum likelihood for Gibbs point processes remain largely unknown. In [17], we partly solved this 30 years old conjecture by proving the consistency of the likelihood procedure for a large class of Gibbs models. As important examples, we deduce the consistency of the maximum likelihood estimator for all parameters of the Strauss model, the hardcore Strauss model, the Lennard-Jones model and the area-interaction model, which are commonly used models in practice.

A practical issue of likelihood estimation yet is that this method depends on an intractable normalizing constant that has to be approximated by simulation. To avoid this problem, other methods of estimation have been introduced, including pseudo-likelihood estimation. The theoretical properties of the pseudo-likelihood method are fairly well known in the case of finite-range Gibbs interactions. However this setting rules out some major Gibbs model as the Lennard-Jones model. In [16], we extend the pseudo-likelihood procedure to infinite range Gibbs interactions and we prove its consistency and its asymptotic normality.

References: [16] [17]

Collaborators: David Dereudre (Laboratoire Paul Painlevé (UMR 8524), University of Lille 1),
Jean-François Coeurjolly (Laboratoire Jean Kutzmann, University of Grenoble).

7.8. Statistical aspects of Determinantal Point Processes

Participant: Frédéric Lavancier.

Determinantal point processes (DPPs) have been introduced in their general form by Macchi (1975) and have been extensively studied from a probabilistic point of view in the 2000's (one of the main reason being their central role in random matrix theory). In a previous work, we have demonstrated that DPPs provide useful models for the description of spatial point pattern datasets where nearby points repel each other.

In [15], we have addressed the question of how repulsive a stationary DPP can be, in order to assess the range of practical situations this promising class of models may model. We determine the most repulsive DPP (in some sense) and we introduce new parametric families of stationary DPPs that can cover a large range of DPPs, from the stationary Poisson process (the case of no interaction) to the most repulsive DPP. Some theoretical aspects of inference for stationary DPPs are tackled in [13] and [14]. In the former study we establish the Brillinger mixing property of stationary DPPs, a first important step toward asymptotic inference. In the latter contribution, we exploit this result to deduce the consistency and asymptotic properties of contrast estimators for stationary DPPs.

References: [13] [15] [14]

Collaborators: Christophe Ange Napoléon Biscio (LMJL, University of Nantes),
Jesper Möller (Department of Mathematical Sciences, Aalborg University, Denmark),
Ege Rubak (Department of Mathematical Sciences, Aalborg University, Denmark).

7.9. Modeling aggregation and regularity in spatial point pattern datasets

Participant: Frédéric Lavancier.

In the spatial point process literature, analysis of spatial point pattern datasets are often classified into three main cases: (i) Regularity (or inhibition or repulsiveness), modelled by Gibbs point processes, hard core processes like Matern hard core models, and determinantal point processes; (ii) Complete spatial randomness, modelled by Poisson point processes; (iii) Aggregation (or clustering), modelled by Poisson cluster processes and Cox processes. For applications the classification (i)-(iii) can be too simplistic, and there is a lack of useful spatial point process models with, loosely speaking, aggregation on the large scale and regularity on the small scale. For instance, we may be interested in such a model for the repartition of the centres of vesicles in a cell, that exhibit some spatial clustering at large scales while having a minimal distance between them.

In [23], we have considered a dependent thinning of a regular point process with the aim of obtaining aggregation on the large scale and regularity on the small scale in the resulting target point process of retained points. Various parametric models for the underlying processes are suggested and the properties of the target point process are studied. Simulation and inference procedures are discussed when a realization of the target point process is observed, depending on whether the thinned points are also observed or not.

Reference: [23]
Collaborator: Jesper Möller (Department of Mathematical Sciences, Aalborg University, Denmark).

7.10. Multi-scale spot segmentation with automatic selection of image scales

Participants: Bertha Mayela Toledo Acosta, Patrick Bouthemy.

Detecting spot-like objects of different sizes in images is required for many applications. A spot detection framework can be divided in three sub-steps : first, image preprocessing to smooth out noise; second, signal enhancement to highlight spots; third, spot detection by thresholding; the two first ones being often merged in a single operator. However, elements of interest do not all correspond to the same image scale, if the collection includes subgroups of different sizes or if perspective effects occur. Then, the need is not merely the selection of the optimal image scale, but of all the meaningful scales. We dealt with the problem of multi-scale spot detection while automatically selecting the meaningful scales. Our primary interest is to detect particles in microscopy images, but our method can be applied to other types of images as well. We defined an original criterion based on the a contrario approach and the LoG scale-space framework to automatically select the meaningful scales. We designed a coarse-to-fine multi-scale spot segmentation scheme involving a locally adaptive thresholding across scales, to come up with the final map of segmented spots. We carried out experimental results on simulated and real images of different types, and we demonstrated that our method outperforms other existing methods.

Reference: paper accepted, ICASSP'2017.
Collaborator: Antoine Basset (CNES, Toulouse).

7.11. Multi-modal registration for correlative light-electron microscopy

Participants: Bertha Mayela Toledo Acosta, Patrick Bouthemy, Charles Kervrann.

We pursue our work on correlative light-electron microscopy (CLEM), which combines the strengths of two different imaging modalities, light microscopy (LM) and electron microscopy (EM), to jointly study intracellular dynamics and ultrastructure of a biological sample. CLEM registration is an important and difficult problem given the significant differences between LM and EM images regarding resolution, field of view, image size and appearance. We designed an automated approach for retracing and registering CLEM images, by implementing a patch-based search using a common Laplacian of Gaussian (LoG) image representation of the LM and EM images. We have redefined the geometry of the patch, opting for a disk-shaped patch. The search (or retracing) step uses histogram-based methods as they are invariant to rotation, and it provides a pre-registration by producing the estimate of the translation component. Usually, there is a large disparity on the orientation of EM and LM images. To handle this problem, we have implemented a mutual information-based method to compute the rotation between the EM and LM patches and to refine the registration. We have also tackled the registration issue in both directions (LM to EM, and EM to LM), and compared our approach to a correlation-based method.

We have tested our approach on a larger set of real CLEM images (provided by Institut Curie) presenting a large diversity in content, image size, and appearance, further validating our method (see Fig. 8). We are currently exploring how our automated CLEM registration method could be exploited to guide EM acquisition within a coarse-to-fine framework.

Reference: [35]
Collaborators: Xavier Heiligenstein (UMR 144 CNRS-Institut Curie),
Grégoire Malandain (Inria, Morpheme EPC, Sophia-Antipolis).

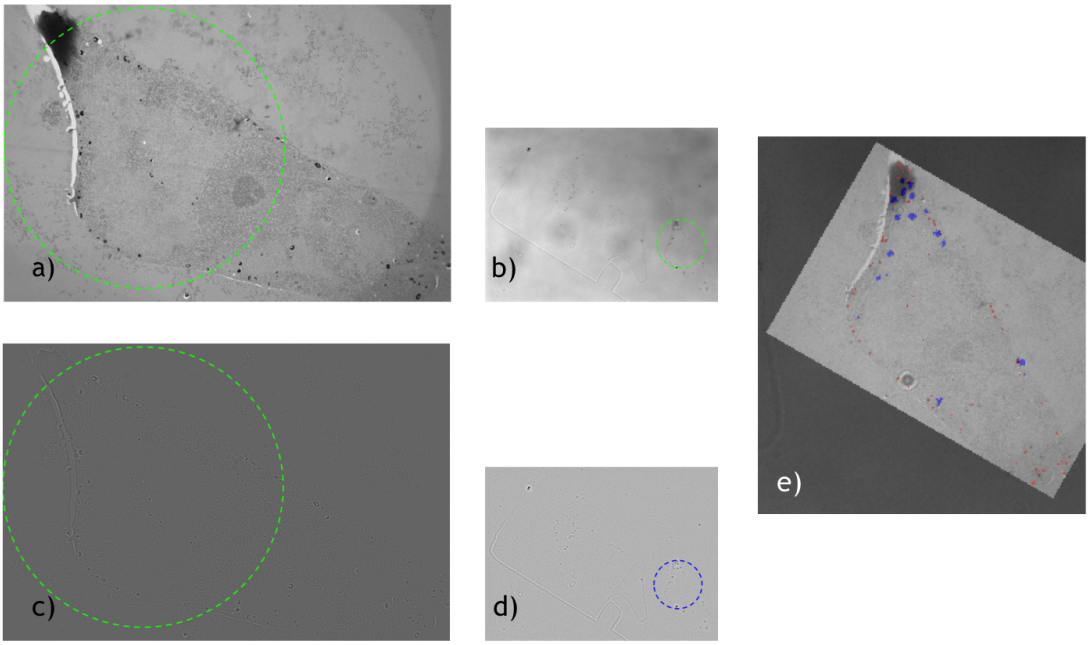


Figure 8. Figure 1. CLEM experiment (images from UMR 144 CNRS-Institut Curie, Xavier Heiligenstein): a) EM image with Region of Interest (ROI) framed in green; b) ground-truth location of the corresponding LM patch framed in green; c) ROI delineated in the Log-EM image, framed in green; d) selected patch (SP) in the LoG-LM image, framed in blue to be comaprd with the green disk in b); f) overlay (after registration).

7.12. Denoising and compensation of the missing wedge in cryo electron tomography

Participants: Emmanuel Moebel, Charles Kervrann.

In this study, we address two important issues in cryo electron tomography (CET) images: reduction of noise and restoration of information in the missing wedge (MW). The MW is responsible for several type of imaging artifacts, and arises because of limited angle tomography: it is observable in the Fourier domain and is depicted by a region where Fourier coefficient values are unknown (see Fig. 9). The proposed stochastic method tackles the restoration problem by filling up the MW by iterating following steps : adding noise into the MW (step 1) and applying a denoising algorithm (step 2). The role of the first step is to propose candidates for the missing Fourier coefficients and the second step acts as a regularizer. A constraint is added in the spectral domain by imposing the known Fourier coefficients to be unchanged through iterations.

Several aspects of the method have been studied in order to gain a deeper understanding of this strategy: different kinds of noise as well as several denoising algorithms (BM3D, NL-Bayes, NL-means, Total Variation...) have been evaluated. Furthermore, different kinds of transforms have been tested in order to apply the constraint (Fourier transform, Cosine transform, pseudo-polar Fourier transform). Also, a process has been set up in order to evaluate the performance of the proposed method on experimental data. Thus, convincing results on experimental data have been achieved (see Fig. 9) using the Fourier Shell Correlation (FSC) as an evaluation metric. In order to measure the quality of the recovered MW only, we also compute the FSC over the MW support ("constrained FSC").

Collaborators: Damien Larivière (Fondation Fourmentin-Guilbert),
Julio Ortiz (Max-Planck Institute, Martinsried, Germany).

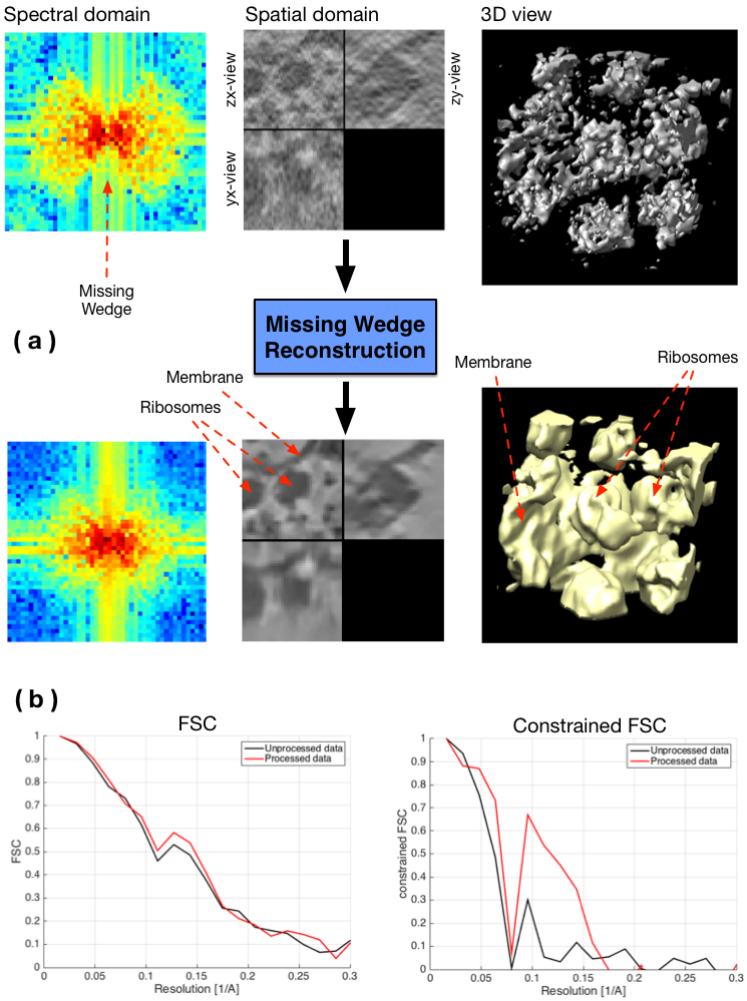


Figure 9. Experimental sub-tomogram containing ribosomes attached to a membrane. (a) Top row: original data in the spectral (left) and spatial (middle) domains and 3D view of the thresholded data (right). Bottom row: denoised data shown as above. (b) FSC and constrained FSC measures of the method input (in black) and output (in red). All measures are wrt the same reference.

7.13. Spatially-variant kernel for optical flow under low signal-to-noise ratios

Participant: Charles Kervrann.

Local and global approaches can be identified as the two main classes of optical flow estimation methods. This year, we have proposed a framework to combine the advantages of these two principles, namely robustness to noise of the local approach and discontinuity preservation of the global approach. The idea is to adapt spatially the local support of the local parametric constraint in the combined local-global model [42]. To this end, we jointly estimate the motion field and the parameters of the spatial support. We apply our approach to the case of Gaussian filtering, and we derive efficient minimization schemes for usual data terms. The estimation of a spatially varying standard deviation map prevents from the smoothing of motion discontinuities, while ensuring robustness to noise. We validated our method for a standard model and demonstrated how a baseline approach with pixel-wise data term can be improved when integrated in our framework. The method has been evaluated on the Middlebury benchmark with ground truth and on real fluorescence microscopy data for which noise is the main limitation for usual optical flow methods.

Collaborator: Denis Fortun (EPFL-BIG, Lausanne, Switzerland)
Noémie Debroux (Laboratory of Mathematics, INSA Rouen, Normandie University)

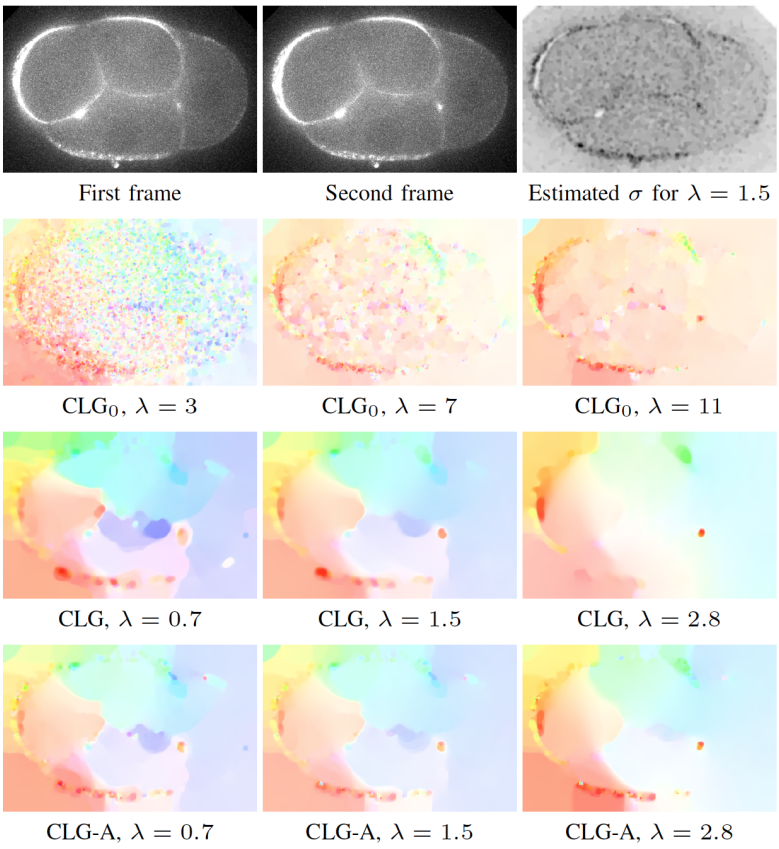


Figure 10. Visual results obtained with variants of CLG ("Combined Local-Global") methods for several values of regularization parameters λ: CLG-A (our adaptive CLG), CLG_0 (pointwise method) and CLG ([42]) on a image pair from a CElegans sequence.

7.14. Frame-based hierarchical motion decomposition and segmentation

Participants: Juan Manuel Perez Rua, Patrick Bouthemy.

A number of applications in video analysis rely on a per-frame motion segmentation of the scene as key preprocessing step. Moreover, different settings in video production require extracting segmentation masks of multiple moving objects and object parts in a hierarchical fashion. In order to tackle this problem, we propose to analyze and exploit the compositional structure of scene motion to provide a segmentation which is not purely driven by local image information. Specifically, we leveraged a hierarchical motion-based partition of the scene to capture a mid-level understanding of the dynamic video content. To recover the decomposition tree, we formulated the problem as a per-pixel label selection interleaved with motion models estimation. The labels represent the set of nodes from the initial proposal tree which are selected to explain globally the input correspondence field. We carried out experimental results showing the strengths of this approach in comparison to current video segmentation approaches. Indeed, they demonstrated the superior ability of our method to capture the main moving objects of the scene in the first layer of the tree, and to segment them in moving parts in deeper layers. As such, we believe our segmentation method is closer to the complex needs of video editing than current hierarchical segmentation approaches.

Reference: [34]
Collaborators: Tomas Crivelli and Patrick Pérez (Technicolor).

7.15. Trajectory-based discovery of motion hierarchies in video sequences

Participants: Juan Manuel Perez Rua, Patrick Bouthemy.

The dynamic content of physical scenes is largely compositional, that is, the movements of the objects and of their parts are hierarchically organized and relate through composition along this hierarchy. This structure also prevails in the apparent 2D motion that a video captures. Visual motion in the scene is roughly organized along a tree, with the dominant motion (typically induced by camera motion) at the root, and motion components adding up along the branches. Accessing this visual motion hierarchy is important to get a better understanding of dynamic scenes and is useful for video manipulation. We proposed to capture it through learned, tree-structured sparse coding of point trajectories. We found that dictionary learning and sparse coding provide appealing tools to disentangle this latent hierarchical structure. More precisely, we introduced a new tree-structured dictionary learning method that allows describing each track with a few basis functions, all but one being inherited from its parent in the structure. The sparse codes thus associated to the tracks capture the desired structure and lend themselves naturally to hierarchical clustering of the collection. We showed through experiments on motion capture data that our model is able to extract moving segments along with their organization. We also obtained competitive results on the task of segmenting objects in real video sequences from trajectories.

Reference: [33]
Collaborators: Tomas Crivelli and Patrick Pérez (Technicolor).

SIROCCO Project-Team

7. New Results

7.1. Analysis and modeling for compact representation

3D modelling, multi-view plus depth videos, light-fields, 3D meshes, epitomes, image-based rendering, inpainting, view synthesis

7.1.1. Visual attention

Participant: Olivier Le Meur.

Visual attention is the mechanism allowing to focus our visual processing resources on behaviorally relevant visual information. Two kinds of visual attention exist: one involves eye movements (overt orienting) whereas the other occurs without eye movements (covert orienting). Our research activities deal with the understanding and modeling of overt attention.

Saccadic model: Previous research showed the existence of systematic tendencies in viewing behavior during scene exploration. For instance, saccades are known to follow a positively skewed, long-tailed distribution, and to be more frequently initiated in the horizontal or vertical directions. In 2016, we investigated the fact that these viewing biases are not universal, but are modulated by the semantic visual category of the stimulus. We showed that the joint distribution of saccade amplitudes and orientations significantly varies from one visual category to another. These joint distributions turn out to be, in addition, spatially variant within the scene frame. We demonstrated that a saliency model based on this better understanding of viewing behavioral biases and blind to any visual information outperforms well-established saliency models. We also proposed an extension of the saccadic model developed in 2015. The improvement consists in accounting for spatially-variant and context-dependent viewing biases. This model outperforms state-of-the-art saliency models, and provides scanpaths in close agreement with human behavior.

Inference of age from eye movements: We have presented evidence that information derived from eye gaze can be used to infer observers' age. From simple features extracted from the sequence of fixations and saccades, we predict the age of an observer. To reach this objective, we used the eye data from 101 observers split in 4 age groups (adults, 6-10 year-old, 4-6 year-old. and 2 year-old) to train a computational model. Participant's eye movements were monitored while participants were instructed to explore color pictures taken from children books for 10 seconds. The analysis of eye gaze provided evidence of age-related differences in viewing patterns. Fixation durations decreased with age while saccades turned out to be shorter when comparing children with adults. We combine several features, such as fixation durations, saccade amplitudes, and learn a direct mapping from those features to age using Gentle AdaBoost classifiers. Experimental results show that the proposed method succeeds in predicting reasonably well the observer's age.

7.1.2. Graph structure in the rays space for fast light fields segmentation

Participants: Christine Guillemot, Matthieu Hog.

In collaboration with Technicolor (Neus Sabater), we have introduced a novel graph representation for interactive light field segmentation using Markov Random Field (MRF). The greatest barrier to the adoption of MRF for light field processing is the large volume of input data. The proposed graph structure exploits the redundancy in the ray space in order to reduce the graph size, decreasing the running time of MRF-based optimisation tasks. The concepts of free rays and ray bundles with corresponding neighbourhood relationships are defined to construct the simplified graph-based light field representation. We have then developed a light field interactive segmentation algorithm using graph-cuts based on such ray space graph structure, that guarantees the segmentation consistency across all views. Our experiments with several datasets show results that are very close to the ground truth, competing with state of the art light field segmentation methods in terms of accuracy and with a significantly lower complexity. They also show that our method performs well on both densely and sparsely sampled light fields [18] (see Figure 1).

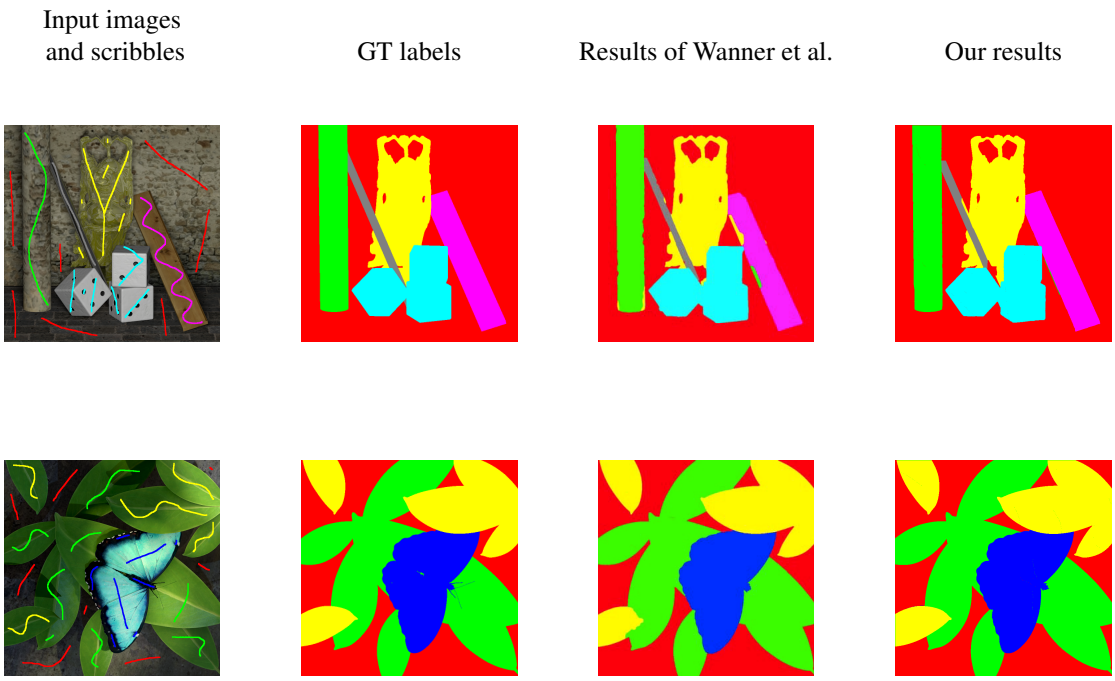


Figure 1. Light-field segmentation results obtained with synthetic light-fields. From left to right, we show, the input central view with scribbles, the ground truth labelling, the results of Wanner et al. and our results.

7.2. Rendering, inpainting and super-resolution

image-based rendering, inpainting, view synthesis, super-resolution

7.2.1. Joint color and gradient transfer through Multivariate Generalized Gaussian Distribution

Participants: Hristina Hristova, Olivier Le Meur.

Multivariate generalized Gaussian distributions (MGGDs) have aroused a great interest in the image processing community thanks to their ability to describe accurately various image features, such as image gradient fields, wavelet coefficients, etc. However, so far their applicability has been limited by the lack of a transformation between two of these parametric distributions. In collaboration with FRVSense (Rémi Cozot and Kadi Bouatouch), we have proposed a novel transformation between MGGDs, consisting of an optimal transportation of the second-order statistics and a stochastic-based shape parameter transformation. We employ the proposed transformation in both color and gradient transfers between images. We have also proposed a new simultaneous transfer of color and gradient.

7.2.2. High-Dynamic-Range Image Recovery from Flash and Non-Flash Image Pairs

Participants: Hristina Hristova, Olivier Le Meur.

In 2016, in collaboration with FRVSense (Rémi Cozot and Kadi Bouatouch), we have proposed a novel method for creating High Dynamic Range (HDR) images from only two images - flash and non-flash images. The proposed method consists of two main steps, namely brightness gamma correction and bi-local chromatic adaptation transform (CAT). First, the brightness gamma correction performs series of increases and decreases of the brightness of the non-flash image and that way yields multiple images with various exposure values. Second, a proposed CAT method, called bi-local CAT enhances the quality of the computed images, by recovering details in the under-/over-exposed regions, using detail information from the flash image. The

final multiple exposure images are then merged together to compute an HDR image. Evaluation shows that our HDR images, obtained by using only two LDR images, are close to HDR images, obtained by combining five manually taken multi-exposure images. The proposed method does not require the usage of a tripod and it is suitable for images of non-still objects, such as people, candle flames, etc. Figure 2 illustrates some results of the proposed method. The HDR-VDP-2 color-coded map (right-most image) shows the main luminance differences (the red areas) between our HDR result and the real HDR image. Snippets (a) and (b) show that the proposed method sharpens fine details, e.g. the net on the lamp. The net on the lamp of the real HDR image is blurry, due to a movement in the real multi-exposure images.

Figure 2. HDR image recovery from two input images, i.e. flash and non-flash images. Our HDR result and the real HDR image are tone-mapped for visualization on an LDR display.

7.2.3. Depth inpainting

Participant: Olivier Le Meur.

To tackle the disocclusion inpainting of RGB-D images appearing when synthesizing new views of a scene by changing its viewpoint, in collaboration with Pierre Buyssens from the Greyc laboratory from the Caen University, we have developed a new examplar-based inpainting method of depth map. The proposed method is based on two main components. First, a novel algorithm to perform the depth-map disocclusion inpainting has been proposed. In particular, this intuitive approach is able to recover the lost structures of the objects and to inpaint the depth-map in a geometrically plausible manner. Then, a depth-guided patch-based inpainting method has been defined in order to fill-in the color image. Depth information coming from the reconstructed depth-map is added to each key step of the classical patch-based algorithm from Criminisi et al. in an intuitive manner. Relevant comparisons to state-of-the-art inpainting methods for the disocclusion inpainting of both depth and color images have illustrated the effectiveness of the proposed algorithms.

7.2.4. Super-resolution and inpainting for face recognition

Participants: Reuben Farrugia, Christine Guillemot.

Most face super-resolution methods assume that low- and high-resolution manifolds have similar local geometrical structure, hence learn local models on the low-resolution manifold (e.g. sparse or locally linear embedding models), which are then applied on the high-resolution manifold. However, the low- resolution manifold is distorted by the one-to-many relationship between low- and high- resolution patches.

We have developed a method which learns linear models based on the local geometrical structure on the high-resolution manifold rather than on the low-resolution manifold. For this, in a first step, the low-resolution patch is used to derive a globally optimal estimate of the high-resolution patch. The approximated solution is shown to be close in Euclidean space to the ground-truth but is generally smooth and lacks the texture details needed by state-of-the-art face recognizers. Unlike existing methods, the sparse support that best estimates the first approximated solution is found on the high-resolution manifold. The derived support is then used to extract the atoms from the coupled dictionaries that are most suitable to learn an upscaling function between the low- and high- resolution patches.

The proposed solution has also been extended to compute face super-resolution of non-frontal images. Experimental results show that the proposed method out- performs six face super-resolution and a state-of-the-art cross- resolution face recognition method. These results also reveal that the recognition and quality are significantly affected by the method used for stitching all super-resolved patches together, where quilting was found to better preserve the texture details which helps to achieve higher recognition rates. The proposed method was shown to be able to super-resolve facial images from the IARPA Janus Benchmark A (IJB-A) dataset which considers a wide range of poses and orientations.

A method has also been developed to inpaint occluded facial regions with unconstrained pose and orientation. This approach first warps the facial region onto a reference model to synthesize a frontal view [15]. A modified Robust Principal Component Analysis (RPCA) approach is then used to suppress warping errors. It then uses a novel local patch-based face inpainting algorithm which hallucinates missing pixels using a dictionary of face images which are pre-aligned to the same reference model. The hallucinated region is then warped back onto the original image to restore missing pixels. Experimental results on synthetic occlusions demonstrate that the proposed face inpainting method has the best performance achieving PSNR gains of up to 0.74dB over the second-best method. Moreover, experiments on the COFW dataset and a number of real-world images show that the proposed method successfully restores occluded facial regions in the wild even for Closed-Circuit Television (CCTV) quality images.

7.2.5. Light-field inpainting

Participants: Christine Guillemot, Xiaoran Jiang, Mikael Le Pendu.

Building up on the advances in low rank matrix completion, we have developed a novel method for propagating the inpainting of the central view of a light field to all the other views. After generating a set of warped versions of the inpainted central view with random homographies, both the original light field views and the warped ones are vectorized and concatenated into a matrix. Because of the redundancy between the views, the matrix satisfies a low rank assumption enabling us to fill the region to inpaint with low rank matrix completion. To this end, a new matrix completion algorithm, better suited to the inpainting application than existing methods, has also been developed. Unlike most of the existing light field inpainting algorithms, our method does not require any depth prior. Another interesting feature of the low rank approach is its ability to cope with color and illumination variation between the input views of the light field (see Fig.3 . As it can be seen in Figure 3 , the proposed method yields inpainting consistency across views.

Figure 3. Illustration of our inpainting propagation method : (a) Original central view. (b) Inpainted central view. (c) Another view of the light field inpainted with a state-of-the-art 2D image inpainting method. (d) Propagated inpainting from central view to a different view with the developed low rank method.

7.3. Representation and compression of large volumes of visual data

Sparse representations, data dimensionality reduction, compression, scalability, perceptual coding, rate-distortion theory

7.3.1. Graph-based multi-view video representation

Participants: Christine Guillemot, Thomas Maugey, Mira Rizkallah, Xin Su.

One of the main open questions in multiview data processing is the design of representation methods for multiview data, where the challenge is to describe the scene content in a compact form that is robust to lossy data compression. Many approaches have been studied in the literature, such as the multiview and multiview plus depth formats, point clouds or mesh-based techniques. All these representations contain two types of data: i) the color or luminance information, which is classically described by 2D images; ii) the geometry information that describes the scene 3D characteristics, represented by 3D coordinates, depth maps or disparity vectors. Effective representation, coding and processing of multiview data partly rely on a proper representation of the geometry information. The multiview plus depth (MVD) format has become very popular in recent years for 3D data representation. However, this format induces very large volumes of data, hence the need for efficient compression schemes. On the other hand, lossy compression of depth information in general leads to annoying rendering artefacts especially along the contours of objects in the scene. Instead of lossy compression of depth maps, we consider the lossless transmission of a geometry representation that captures only the information needed for the required view reconstructions.

The goal is thus to develop a Graph-Based Representation (GBR) for geometry information, where the geometry of the scene is represented as connections between corresponding pixels in different views. In this representation, two connected pixels are neighboring points in the 3D scene. The graph connections are derived from dense disparity maps and provide just enough geometry information to predict pixels in all the views that have to be synthesized. GBR drastically simplifies the geometry information to the bare minimum required for view prediction. This "task-aware" geometry simplification allows us to control the view prediction accuracy before coding compared to baseline depth compression methods. In 2015, we have first considered multi-view configurations, in which cameras are parallel.

In 2016, we have developed the extension of GBR to complex camera configurations. In [21], Xin Su has implemented a generalized Graph-Based Representation handling two views with complex translations and rotations between them (Fig. 4). The proposed approach uses the epipolar segments to have a row-wise description of the geometry that is as simple as for rectified views. This generalized GBR has been further extended to handle multiple views and scalable description of the geometry, *i.e.*, a geometry data that is coded as a function of the user navigation among the views.

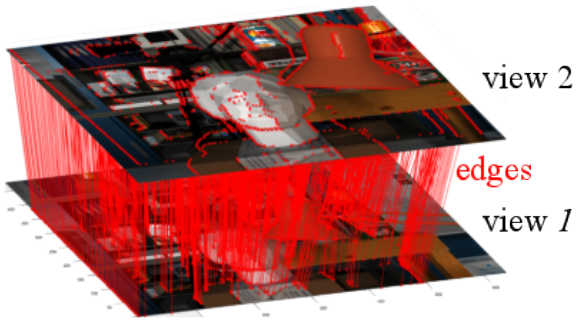


Figure 4. The proposed GBR (i) provides edges describing the geometry information and (ii) link pixels that are neighbors in the 3D scene.

The graph described above links neighboring pixels in the 3D scene as 3D meshes do. This meaningful structure might be used to code the color pixels lying on it. This can be done thanks to the new processing tools developed for signals lying on graphs. These tools rely however on covariance models that are assumed to be suited for the processed data. The PhD work of Mira Rizkallah is currently focussing on the effect of errors in the correlation models on the efficiency of the graph-based transforms.

7.3.2. Sparse and low rank approximation of light fields

Participants: Christine Guillemot, Xiaoran Jiang, Mikael Le Pendu.

We have studied the problem of low rank approximation of light fields for compression. A homography-based approximation method has been proposed which jointly searches for homographies to align the different views of the light field together with the low rank approximation matrices. We have first considered a global homography per view and shown that depending on the variance of the disparity across views, the global homography is not sufficient to well-align the entire images. In a second step, we have thus considered multiple homographies, one per region, the region being extracted using depth information. We have first shown the benefit of the joint optimization of the homographies together with the low-rank approximation. The resulting compact representation compressed using HEVC yields compression performance significantly superior to those obtained by directly applying HEVC on the light field views re-structured as a video sequence.

7.3.3. Deep learning, autoencoders and neural networks for sparse representation and compression

Participants: Thierry Dumas, Christine Guillemot, Aline Roumy.

Deep learning is a novel research area that attempts to extract high level abstractions from data by using a graph with multiple layers. One could therefore expect that deep learning might allow efficient image compression based on these high level features. However, deep learning, as classical machine learning, consists in two phases: (i) build a graph that can make a good representation of the data (i.e. find an architecture usually made with neural nets), and (ii) learn the parameters of this architecture from large-scale data. As a consequence, neural nets are well suited for a specific task (text or image recognition) and require one training per task. The difficulty to apply machine learning approach to image compression is that it is important to deal with a large variety of patches, and with also various compression rates. To test the ability of neural networks to compress images, we studied shallow sparse autoencoders (AE) for image compression in [14]. A performance analysis in terms of rate-distortion trade-off and complexity is conducted, comparing sparse AEs with LARS-Lasso, Coordinate Descent (CoD) and Orthogonal Matching Pursuit (OMP). A Winner Take All Auto-encoder (WTA AE) is proposed where image patches compete with one another when computing their sparse representation. This allows to spread the sparsity constraint on the whole image. Since the learning is made for this WTA AE, the neural network also learns to deal with various patches, which helps building a general-purpose AE. Finally, we showed that, WTA AE achieves the best rate-distortion trade-off, is robust to quantization noise and it is less complex than LARS-Lasso, CoD and OMP.

7.3.4. Data geometry aware local basis selection

Participants: Julio Cesar Ferreira, Christine Guillemot.

Local learning of sparse image models has proven to be very effective to solve a variety of inverse problems in many computer vision applications. To learn such models, the data samples are often clustered using the K-means algorithm with the Euclidean distance as a dissimilarity metric. However, the Euclidean distance may not always be a good dissimilarity measure for comparing data samples lying on a manifold.

In 2015, we have developed, in collaboration with Elif Vural (now Prof. at METU in Ankara, former postdoc in the team), two algorithms for determining a local subset of training samples from which a good local model can be computed for reconstructing a given input test sample, where we take into account the underlying geometry of the data. The first algorithm, called Adaptive Geometry-driven Nearest Neighbor search (AGNN), is an adaptive scheme which can be seen as an out-of-sample extension of the replicator graph clustering method for local model learning. The second method, called Geometry-driven Overlapping Clusters (GOC), is

a less complex nonadaptive alternative for training subset selection. The AGNN and GOC methods have been evaluated in image super-resolution, deblurring and denoising applications and shown to outperform spectral clustering, soft clustering, and geodesic distance based subset selection in most settings. The selected patches are used for learning good local bases using the traditional PCA method. PCA is considered an efficient tool to recover the tangent space of the patch manifold when the manifold is sufficiently regular.

However, when the patch manifold has high curvature, which is observed to be the case for images with high frequencies, PCA may not be suitable. It can be seen in Figure 5 that the PCA basis with respect to a manifold fails to approximate the tangent space as the manifold bends over itself. In other words, PCA basis is not adapted when the curvature is too high. On the other hand, it can be seen in Figure 5 that a union of subspaces with respect to a manifold might generate a local model that yields a more efficient local representation of data.

In 2016, we have proposed a strategy to choose between these two kinds of bases locally depending on the local data geometry. This function is defined as the variability of the tangent space in each cluster.

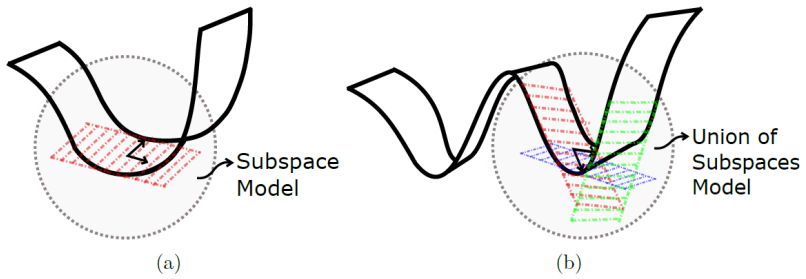


Figure 5. Subspaces computed with data sampled from a neighborhood on a manifold; (a): PCA basis which fails to approximate the subspace as the manifold curvature is too high; (b): union of subspaces generating a local model more coherent with the manifold geometry.

7.3.5. Rate-distortion optimized tone curves for HDR video compression

Participants: David Gommelet, Christine Guillemot, Aline Roumy.

High Dynamic Range (HDR) images contain more intensity levels than traditional image formats. Instead of 8 or 10 bit integers, floating point values requiring much higher precision are used to represent the pixel data. These data thus need specific compression algorithms. In collaboration with Ericsson [17], we have developed a novel compression algorithm that allows compatibility with the existing Low Dynamic Range (LDR) broadcast architecture in terms of display, compression algorithm and datarate, while delivering full HDR data to the users equipped with HDR display. The developed algorithm is thus a scalable video compression offering a base layer that corresponds to the LDR data and an enhancement layer, which together with the base layer corresponds to the HDR data. The novelty of the approach relies on the optimization of a mapping called Tone Mapping Operator (TMO) that maps efficiently the HDR data to the LDR data. The optimization has been carried out in a rate-distortion sense: the distortion of the HDR data is minimized under the constraint of minimum sum datarate (for the base and enhancement layer), while offering LDR data that are closed to some "aesthetic" a priori. Taking into account the aesthetic of the scene in video compression is indeed novel, since video compression is traditionally optimized to deliver the smallest distortion with the input data at the minimum datarate.

7.3.6. Cloud-based image compression

Participants: Jean Begaint, Christine Guillemot.

The emergence of cloud applications and web services has led to an increasing use of online resources. Image processing applications can benefit from this vast storage and distribution capacity. In collaboration with Technicolor, we investigate the use of this mass of redundant data to enhance image compression schemes. A region-based registration algorithm has been developped to capture complex deformations between two images. The registration method is then used to exploit both global and local correspondences between pairs of images of the same scene. The region-based registration yields a better prediction (hence reduced prediction errors, see Fig.6) which in turn yields a significant rate-distortion performance gain compared to current image coding solutions.

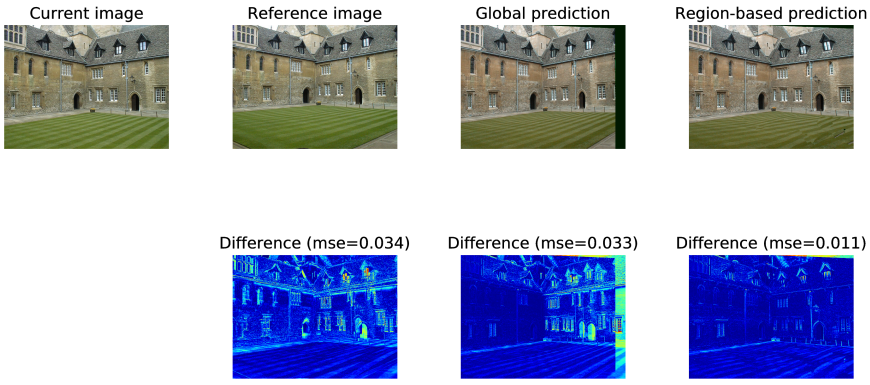


Figure 6. Image registration with global and region-based homographies and corresponding prediction error.

7.4. Distributed processing and robust communication

Information theory, stochastic modelling, robust detection, maximum likelihood estimation, generalized likelihood ratio test, error and erasure resilient coding and decoding, multiple description coding, Slepian-Wolf coding, Wyner-Ziv coding, information theory, MAC channels

7.4.1. Interactive Coding for Navigation in 3D scenes (ICON 3D)

Participants: Thomas Maugey, Aline Roumy.

In order to have performing FTV systems, the data transmission has to take into account the interactivity of the user, *i.e.,* the viewpoint that is requested. In other words, a FTV system transmits to the visualisation support only what needs to be updated when a user changes its viewpoint angle (*i.e.*, the new information appearing in its vision field). The Sirocco has recently proposed some promising work using channel coding for interactive data coding. This coding scheme focusses on multi-view plus depth format only. In order to extend this approach to other formats, we have started a collaboration with the I3S laboratory in Nice, expert in 3D mesh compression.

The project ICON 3D funded by the GdR-Isis will be divided into two parts. First, we will study and develop new geometry prediction algorithms for surface meshes. Given a part of a mesh, the prediction algorithm should be able to estimate a neighboring mesh subset corresponding to the one newly visible after user viewpoint angle change (Fig. 7). The prediction error will be characterized. Then, we will study the channel coding method that should be developed to correct this error.

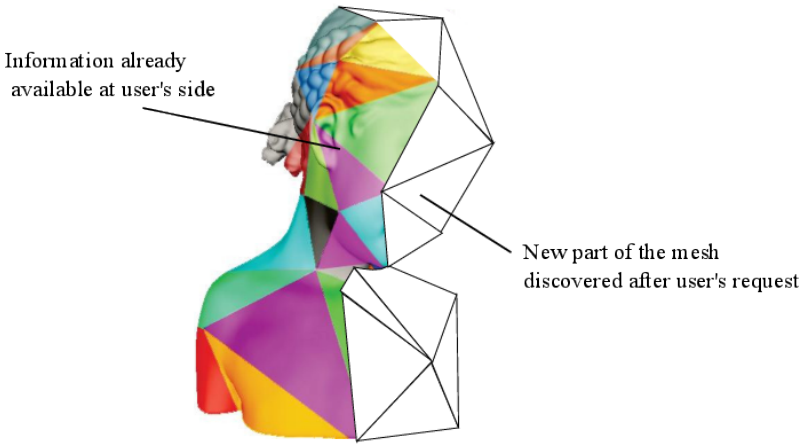


Figure 7. When a user changes his viewpoint angle, he discovers new part of the mesh that has to be transmitted.

SUMO Project-Team

7. New Results

7.1. Analysis and verification of quantitative systems

7.1.1. Quantitative verification of distributions of stochastic models

Participant: Blaise Genest.

In [24], we obtained conditions under which quantitative verification of distributions of stochastic systems is decidable. This is a challenging question as for general Markov Chains, verification of distribution is Skolem-complete, a problem on linear recurrence sequences whose decidability is a long-standing problem open for 40 years. In this paper, we approach this problem by studying the languages generated by Markov Chains, whose regularity would entail the decidability of quantitative verification. Given an initial distribution, we represent the trajectory of Markov Chain over time as an infinite word over a finite alphabet, where the n^{th} letter represents a probability range after n steps. We extend this to a language of trajectories (a set of words), one trajectory for each initial distribution from a (possibly infinite) set. We show that if the eigenvalues of the transition matrix associated with the Markov Chain are all distinct positive real numbers, then the language is *effectively regular*. Further, we show that this result is at the boundary of regularity, as non-regular languages can be generated when the restrictions are even slightly relaxed. The regular representation of the language allows us to reason about more general properties, e.g., robustness of a regular property in a neighbourhood around a given distribution.

7.1.2. Diagnosability of repairable faults

Participants: Éric Fabre, Loïc Hélouët, Hervé Marchand, Engel Lefaucheux.

For (partially observable) discrete event systems, diagnosability characterizes the ability to detect the occurrence of a permanent fault in bounded time after it occurs, given the observations available on that system. Diagnosability can be decided in polynomial time, relying on the so-called twin-machine construction. We have examined the case of repairable faults, and a notion of diagnosability that requires the detection of the fault before it is repaired. It was proved in [35]that diagnosability is a PSpace complete problem.

7.1.3. Diagnosability of stochastic systems

Participants: Éric Fabre, Blaise Genest, Hugo Bazille, Ocan Sankur.

Diagnosis of partially observable stochastic systems prone to faults was introduced in the late nineties. Diagnosability, i.e. the existence of a diagnoser, may be specified in different ways: (1) exact diagnosability (called A-diagnosability) requires that almost surely a fault is detected and that no fault is erroneously claimed while (2) approximate diagnosability (called ε-diagnosability) allows a small probability of error when claiming a fault and (3) accurate approximate diagnosability (called AA-diagnosability) requires that this error threshold may be chosen arbitrarily small. In a recent work [27], we focused on approximate diagnoses. We first refined the almost sure requirement about finite delay introducing a uniform version and showing that while it does not discriminate between the two versions of exact diagnosability this is no more the case in approximate diagnosis. We then gave a complete picture of relations between the different diagnosability specifications for probabilistic systems and establish characterisations for most of them in the finite-state case. Based on these characterisations, we developped decision procedures, studied their complexity and proved their optimality. We also designed synthesis algorithms to construct diagnosers and we analysed their memory requirements. Finally we established undecidability of the diagnosability problems for which we provided no characterisation. Notably, we proved the AA-diagnosability problem to be undecidable, answering a longstanding open question.

In another work [28], we investigated semantical and computational issues for exact notions of diagnosability in the context of infinite-state probabilistic systems. We first showed established a characterisation of the so-called FF-diagnosability using a $G\delta$ set (instead of an open set for finite-state systems) and also for two other notions, IF- and IA-diagnosability, when models are finitely branching. We also proved that surprisingly the last notion, FA-diagnosability, cannot be characterised in this way even in the finitely branching case. Then we applied our characterisations for a partially observable probabilistic extension of visibly pushdown automata, yielding EXPSPACE procedures for solving diagnosability problems. In addition, we establish some computational lower bounds and show that slight extensions of these probabilistic visibly pushdown automata lead to undecidability.

7.1.4. Analysing decisive stochastic processes

Participant: Nathalie Bertrand.

In 2007, Abdulla et al. introduced the elegant concept of decisive Markov chain. Intuitively, decisiveness allows one to lift the good properties of finite Markov chains to infinite Markov chains. For instance, the approximate quantitative reachability problem can be solved for decisive Markov chains (enjoying reasonable effectiveness assumptions) including probabilistic lossy channel systems and probabilistic vector addition systems with states. In a recent work [26], we extended the concept of decisiveness to more general stochastic processes. This extension is non trivial as we consider stochastic processes with a potentially continuous set of states and uncountable branching (common features of real-time stochastic processes). This allowed us to obtain decidability results for both qualitative and quantitative verification problems on some classes of real-time stochastic processes, including generalized semi-Markov processes and stochastic timed automata.

7.1.5. Concurrent timed systems

Participants: Loïc Hélouët, Blaise Genest.

Adding real time information to Petri net models often leads to undecidability of classical verification problems such as reachability and boundedness. For instance, models such as Timed-Transition Petri nets (TPNs) [47] are intractable except in a bounded setting. On the other hand, the model of Timed-Arc Petri nets [50] enjoys decidability results for boundedness and control-state reachability problems at the cost of disallowing urgency (the ability to enforce actions within a time delay).

We have addressed semantics variants of time and timed Petri nets to obtain concurrent models with interesting expressive power, but yet allowing decidability of verification and robustness questions. Robustness of timed systems aims at studying whether infinitesimal perturbations in clock values can result in new discrete behaviors. A model is robust if the set of discrete behaviors is preserved under arbitrarily small (but positive) perturbations.

In [25] we have considered time in Petri nets under a strong semantics with multiple enabling of transitions. We focus on a structural subclass of unbounded TPNs, where the underlying untimed net is free-choice, and show that it enjoys nice properties under a multi-server semantics. In particular, we showed that the questions of fireability (whether a chosen transition can fire), and termination (whether the net has a non-terminating run) are decidable for this class. We then consider the problem of robustness under guard enlargement [48], i.e., whether a given property is preserved even if the system is implemented on an architecture with imprecise time measurement. Unlike in [15], where decidability of several problems is obtained for bounded classes of nets, we showed that robustness of fireability is decidable for unbounded free choice TPNs with a multi-server semantics.

The robustness of time Petri nets was addressed in [15] by considering the model of parametric guard enlargement which allows time-intervals constraining the firing of transitions in TPNs to be enlarged by a (positive) parameter. We show that TPNs are not robust in general and checking if they are robust with respect to standard properties (such as boundedness, safety) is undecidable. We then extend the marking class timed automaton construction for TPNs to a parametric setting, and prove that it is compatible with guard enlargements. We apply this result to the (undecidable) class of TPNs which are robustly bounded (i.e., whose finite set of reachable markings remains finite under infinitesimal perturbations): we provide two decidable

robustly bounded subclasses, and show that one can effectively build a timed automaton which is timed bisimilar even in presence of perturbations. This allows us to apply existing results for timed automata to these TPNs and show further robustness properties.

The goal of [23] is to investigate decidable classes of Petri nets with time that capture some urgency and still allow unbounded behaviors, which go beyond finite state systems. We have shown, up to our knowledge, the first decidability results on reachability and boundedness for Petri net variants that combine unbounded places, time, and urgency. For this, we have introduced the class of Timed-Arc Petri nets with restricted Urgency, where urgency can be used only on transitions consuming tokens from bounded places. We showed that control-state reachability and boundedness are decidable for this new class, by extending results from Timed-Arc Petri nets (without urgency) [43]. Our main result concerns (marking) reachability, which is undecidable for both TPNs (because of unrestricted urgency) [46] and Timed-Arc Petri Nets (because of infinite number of "clocks") [49]. We obtained decidability of reachability for unbounded TPNs with restricted urgency under a new, yet natural, timed-arc semantics presenting them as Timed-Arc Petri Nets with restricted urgency. Decidability of reachability under the intermediate marking semantics is also obtained for a restricted subclass.

7.1.6. Petri nets realizability

Participants: Loïc Hélouët, Abd El Karim Kecir.

We considered in [30] the realizability of urban train schedules by stochastic concurrent timed systems. Schedules are high level views of desired timetables that a metro system should implement. They are represented as partial orders decorated with timing constraints. Train systems are represented as elementary stochastic time Petri nets. We have first considered logical realizability: a schedule is realizable by a net $\mathcal{N}$ if it embeds in a time process of $\mathcal{N}$ that satisfies all its constraints. However, with continuous time domains, the probability of a time process that realizes a schedule is null. We have extended the former notion of realizability to consider probabilistic realizability of schedules up to some imprecision α. This probabilistic realizability holds if the probability that $\mathcal{N}$ logically realizes S with constraints enlarged by α time units is strictly positive. We have shown that upon a sensible restriction guaranteeing time progress (systems can not perform an arbitrary number of actions within a single time unit), logical and probabilistic realizability of a schedule can be checked on the finite set of symbolic prefixes extracted from a bounded unfolding of the net. We have provided a construction technique for these prefixes and shown that they represent all time processes of a net occurring up to a given maximal date. We have then shown how to verify existence of an embedding and compute the probability of its realization.

7.2. Control of quantitative systems

7.2.1. Smart regulation for urban trains

Participants: Éric Fabre, Loïc Hélouët, Hervé Marchand, Abd El Karim Kecir.

The regulation of subway lines consists in accomodating small random perturbations in transit times as well as more impacting incidents, by playing on continuous commands (transit times and dwell times) and by making more complex decisions (insertions or extractions of trains, changes of missions, overpassing, shorter returns, etc.). The objectives are multiple : ensuring the regularity and punctuality of trains, adapting to transportation demand, minimizing energy consumption, etc. We have developed an event-based control strategy that aims at equalizing headways on a line. This distributed control strategy is remarquably robust to perturbations and reactive enough to accomodate train insertions/extractions. We have also developed another approach based on event graphs in order to optimally interleave trains at a junction.

7.2.2. Games and reactive synthesis

Participant: Ocan Sankur.

In game theory, a strategy is *dominated* by another one if the latter systematically yields a payoff as good as the former, while also yielding a better payoff in some cases. A strategy is *admissible* if it is not dominated. This notion is well studied in game theory and is useful to describe the set of strategies that are "reasonable" whose choice can be justified. Recent works studied this notion in graph games with omega-regular objectives and investigated its applications in controller synthesis. For multi-agent controller synthesis, admissibility can be used as a hypothesis on the behaviors of each agent, thus enabling a compositional reasoning framework for controller synthesis. In [29], we investigate this framework for quantitative graph games. We characterize admissible strategies, study their existence, and give an effective characterization of the set of paths that are compatible with admissible payoffs. This is then used to derive algorithms for model checking under admissibility, but also assume-admissible synthesis.

In [21], we present the reactive synthesis competition (SYNTCOMP), a long-term effort intended to stimulate and guide advances in the design and application of synthesis procedures for reactive systems. The first iteration of SYNTCOMP is based on the controller synthesis problem for finite-state systems and safety specifications. We provide an overview of this problem and existing approaches to solve it, and report on the design and results of the first SYNTCOMP. This includes the definition of the benchmark format, the collection of benchmarks, the rules of the competition, and the five synthesis tools that participated. We present and analyze the results of the competition and draw conclusions on the state of the art. Finally, we give an outlook on future directions of SYNTCOMP.

In the invited [22], we summarize new solution concepts useful for the synthesis of reactive systems that we have introduced in several recent publications. These solution concepts are developed in the context of non-zero sum games played on graphs. They include the assume-admissible synthesis on Boolean games, synthesis under multiple environments for Markov decision processes, and multi-objective synthesis with probability thresholds for Markov decision processes with multi-dimensional weights. They are part of the contributions obtained in the inVEST project funded by the European Research Council.

7.2.3. Runtime enforcement

Participants: Hervé Marchand, Thierry Jéron.

In the [20] we generalize our line of work on runtime enforcement for timed properties. Runtime enforcement is a verification/validation technique aiming at correcting possibly incorrect executions of a system of interest. In this work we consider enforcement monitoring for systems where the physical time elapsing between actions matters. Executions are thus modelled as timed words (i.e., sequences of actions with dates). We consider runtime enforcement for timed specifications modelled as timed automata. Our enforcement mechanisms have the power of both delaying events to match timing constraints, and suppressing events when no delaying is appropriate, thus possibly allowing for longer executions. To ease their design and their correctness-proof, enforcement mechanisms are described at several levels: enforcement functions that specify the input-output behaviour in terms of transformations of timed words, constraints that should be satisfied by such functions, enforcement monitors that describe the operational behaviour of enforcement functions, and enforcement algorithms that describe the implementation of enforcement monitors.

This year we went one step ahead [33] and consider predictive runtime enforcement, where the system is not entirely black-box, but we know something about its behavior. This *a priori* knowledge about the system allows to output some events immediately, instead of delaying them until more events are observed, or even blocking them permanently. This in turn results in better enforcement policies. We also show that if we have no knowledge about the system, then the proposed enforcement mechanism reduces to a classical non-predictive runtime enforcement framework. All our results are formalized and proved in the Isabelle theorem prover.

7.2.4. Decentralized control

Participant: Hervé Marchand.

In collaboration with Laurie Ricker, we have been interested in decentralized control of discrete event systems. In decentralized discrete-event system (DES) architectures, agents fuse their local decisions to arrive at the global decision. The contribution of each agent to the final decision is never assessed; however, it may be the case that only a subset of agents, i.e., a (static) coalition, perpetually contribute towards the correct final decisions. In casting the decentralized DES control (with and without communication) problem as a cooperative game, it is possible to quantify the average contribution that each agent makes towards synthesizing the overall correct control strategy. Specifically, we explore allocations that assess contributions of non-communicating and communicating controllers for this class of problems. This allows a quantification of the contribution that each agent makes to the coalition with respect to decisions made solely based on its partial observations and decisions made based on messages sent to another agent(s) to facilitate a correct control decision [34].

7.3. Management of large distributed systems

7.3.1. Non-interference in partial order models

Participant: Loïc Hélouët.

We obtained new results on security issues such as non-interference [41]. Noninterference (NI) is a property of systems stating that confidential actions should not cause effects observable by unauthorized users. Several variants of NI have been studied for many types of models but rarely for true concurrency or unbounded models. In [45], we had already demonstrated the discriminating power of partial orders, and investigated NI for High-level Message Sequence Charts (HMSCs), a partial order language for the description of distributed systems. We had proposed a general definition of security properties in terms of equivalence among observations of behaviors, and showed that equivalence, inclusion, and NI properties were undecidable for HMSCs. We defined a new formalism called *partial order automata*, that captures natural observations of distributed systems, and in particular observations of HMSCs. It generalizes HMSCs and permits assembling partial orders. We have then considered subclasses of partial order automata and HMSCs for which Non-Interference is decidable. This allowed us to exhibit more classes of HMSCs for which NI is decidable. Finally, we have defined weaker local Non-interfernce properties, describing situations where a system is attacked by a single agent, and shown that local NI is decidable. We have then refined local NI to a finer notion of causal NI that emphasizes causal dependencies between confidential actions and observations and extended it to causal NI with (selective) declassification of confidential events, which allows to consider that confidential actions need can be kept secret for a limited duration and can then be declassified. Checking whether a system satisfies local and causal NI and their declassified variants are PSPACE-complete problems.

7.3.2. Simulations for stochastic abstractions of large systems

Participants: Éric Fabre, Blaise Genest, Matthieu Pichené.

In [32], we developped a new simulation strategy to accurately simulate DBNs (Dynamic Bayesian Networks) obtained as stochastic abstractions of large systems. The DBN abstractions are given under the form of probability tables, describing the probability for a variable to take a given value given the values of some variables at the previous time point. To be able to handle large systems with many variables, there is a table for each variable (coupling between variable is not explicitely represented). This creates discrepencies when simulating variables independently. Our new algorithm simulates tuples of variables together by looking ahead for such discrepencies in order to avoid them. Such simulations are still efficient, and match more faithfully the original systems.

7.4. Data driven systems

7.4.1. Structured data nets

Participants: Éric Badouel, Loïc Hélouët, Christophe Morvan.

In [16] we proposed a Petri net extension, called Structured Data Nets (SDN), that describes transactional systems with data. In these nets, tokens are semi-structured documents. Each transition is attached to a query, guarded by patterns, (logical assertions on the contents of its preset) and transforms tokens.

We define SDNs and their semantics and consider their formal properties: coverability of a marking, termination and soundness of transactions.

Unrestricted SDNs are Turing complete, so these properties are undecidable. We thus used an order on documents, and showed that under reasonable restrictions on documents and on the expressiveness of patterns and queries, SDNs are well-structured transition systems, for which coverability, termination and soundness are decidable.

7.4.2. An active workspace model for disease surveillance

Participant: Éric Badouel.

Flexibility and change at both design- and run-time are fast becoming the Rule rather than the Exception in Business Process Models. This is attributed to the continuous advances in domain knowledge, the increase in expert knowledge, and the diverse and heterogeneous nature of contextual variables. Such processes are characterized by collaborative work and decision making between users with heterogeneous profiles on a processes designed on-the-fly. A model for such processes should thus natively support human interactions. We showed in [31] how the Active Workspaces model proposed [44] for distributed collaborative systems supports these interactions.

TACOMA Team

6. New Results

6.1. RFID for pervasive computing environments

Participants: Nebil Ben Mabrouk, Frédéric Weis, Paul Couderc [contact].

Here the principle is to implement distributed data structure over a set of RFID tags, enabling a complex object (made of various parts) or a set of objects belonging to a given logical group to "self-describe" itself and the relation between the various physical elements. Some applications examples includes waste management, assembling and repair assistance, prevention of hazards in situations where various products / materials are combined etc. The key property of self-describing objects is, like for coupled objects, that the vital data are self-hosted by the physical element themselves (typically in RFID chips), not an external infrastructure like most RFID systems. This property provides the same advantages as in coupled objects, namely high scalability, easy deployment (no interoperability dependence/interference), and limited risk for privacy. However, given the extreme storage limitation of RFID chips, designing such systems is difficult:

- Data structures must be very frugal in terms of space requirements, both for the structure and for the coding.
- Data structures must be robust and able to survive missing or corrupted elements if we want to ensure the self-describing property for a damaged or incorrect object.

In the context of RFID system, the resiliency property of such data structures enables new information architecture and autonomous (offline) operation, which is very important for some RFID applications. We previously applied the self-describing objects approach to the waste management domain, which has shown to be a specially challenging situation for RFID. This challenge is found more generally in pervasive computing scenarios involving RFID reading in uncontrolled environments (see section 4.4).

We achieved the following results:

- We showed the importance of diversity in the context of challenging RFID reading. A reconfigurable antenna was designed to support dynamic reading protocols.
- A software approach based on error correcting code was developed to support robust data storage in groups of RFID.
- An innovative RFID testbed for experimenting a large range of RFID situations/applications was operational (minus some features to be completed), supported by a simulation environment and a control environment.
- A patent was filed and some contacts made with RFID companies.

However, the supports for implementing dynamic reading protocols were lacking, both on the software and the radio side. The following further progress were made:

- The four elements diversity antenna designed in first phase was implemented.
- The control software has been greatly improved. A new environment was designed, offering powerful and flexible programming capabilities for easy prototyping of RFID reading scenarios and collecting experiments results. A simulator of the testbed was also developed, allowing off-site developments. This work is supported by the RFID-Lab ADT.
- Motion-induced improvements of RFID reliability were experimented, as shown below in Figure 4 .
- A significant dissemination efforts toward the industry was made, and we have good hope that some of the contacts will lead to perspectives.

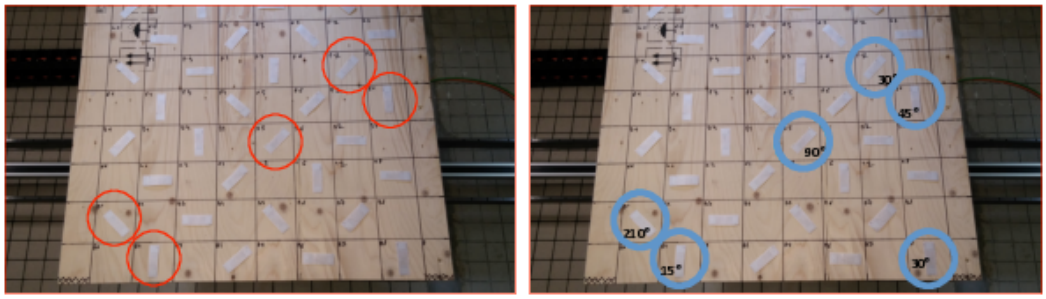


Figure 4. (a) initial read, 20% of the tags are missing (b) After 210 deg of rotation, all the tags are recovered

An example of motion-assisted RFID readings implemented is shown in figure 4 : a matrix of 32 RFID tags are arranged in reduced power conditions, so that the tags are near their sensivity limit. In such conditions, 20% of the tags failed to be read by the reader. By coupling the reading with a rotation of 210 deg, we show that all the missing tags are progressively recovered.

6.2. Building an extensible information sharing mechanism

Participants: Adrien Capaine, Yoann Maurel, Frédéric Weis [contact].

Context aware applications adapt their behavior based on information they can collect on their surrounding environment. Most of these data are provided by third-party software, sensors or computed by the application itself. A striking challenge facing the building of comprehensive pervasive system is the lack of integration between the different services provided by third parties. In this project, we intend to study and to provide mechanisms to enhance information sharing between applications and more specifically to augment information on the surrounding environment. The idea is to endow applications with the capability to increase or augment information on the physical world they are interacting with and to retrieve and share these data seamlessly depending on their location. Such mechanism aims at providing a complementary source of information in order to improve the process of choosing the best service/information provider and to help them keeping additional information on physical resources such as environment specific configuration (e.g., calibration data).

The idea of augmenting information on the physical world is not new. It has been done for centuries in the real world. For instance, the Little Thumb sowed pebbles to find his way just as hikers use cairns so as not to get lost. In daily life, people use sticky notes on pieces of hardware or objects to keep relevant information on their use or capabilities. Applied to IT, such ideas have been pushed by the augmented reality domain where users can access a personalized view of the real world that helps them to carry out their activities. Although this idea has already been implemented in some ad hoc solutions (to exchange ratings for instance), we aim to provide a more generic solution. Our solution must be applicable to nowadays devices and applications with little adjustment to the underlying architectures. It should then be flexible enough to deal with the lack of standards in the domain without imposing architectural choices. Such lack of standard is very common in IT and mainly due to well known factors : (1) for technical reasons, developers tends to think that their âstandardâ are better suited for their current use-case, or/and (2) for commercial reasons companies want to keep a closed siloed system to capture their users, or/and (3) because the domain is still new and evolving and no standard as emerged yet, or/and finally (4) because the problem is too complex to be standardized and most proposed standards tend to be bloated and hard to use.

We are currently implementing these ideas by designing and experimenting two architectures/prototypes:

- **Matriona** is a global distributed framework developed on top of OSGi. This project has been described in more details in the previous activity report. It is meant to be a comprehensive framework for exposing devices as REST-like resources. Resources functionalities can be extended through the mean of decorators. The system also provides access control mechanisms. The main interest of matriona concerning the information enrichment is its ability to support dynamic extension of resource meta-information by application and to provide means to share these meta-information with others. It implements the concept of group of interest with access control on meta-information. The concepts described in Matriona are in the process to be published.
- **Little Thumb Registry (LTReg)** is an independent resource registry that provides the same enrichment capabilities than Matriona but impose less constraints on the architecture of application. Although the prototype is operational, Matriona does not comply with the principle advocated herebefore: it supposes the use of a pivot technology (REST) and assumes that application developers will develop their application on top of on OSGi based platform. The idea behind LTReg is to decouple the registry from the framework and to provide a registry in the manner of Consul [0] with meta-information enrichment and sharing mechanisms. By focussing only on the discovery mechanism and information sharing, LTReg imposes fewer constraints on application and comply more with the goal of being ready to use in actual application. This is still a work in progress.

6.3. Modeling activities to promote self-consumption of locally produced energy

Participants: Jean-Marie Bonnin, Alexandre Rio, Yoann Maurel [contact].

Traditional electricity distribution schemes decouple the production sources from the consumers so that it is necessary to transport energy over long distances. This type of organization is illustrated by the consumption of region such as Brittany, where 91% of the energy consumed is imported. It induces inherent inefficiencies due to the line losses and the transformation steps and therefore induces a high infrastructure and distribution cost. To face these problems and in order to reduce the environmental impacts associated with the use of energy, recent years have seen the development of initiatives to produce energy locally.

The sources of renewable energies are good candidates for this because they are varied and adapt easily to the different geographical situations. The infrastructures necessary for their implementation also impose fewer constraints in terms of installation and safety. One of the main obstacles to the unique use of these technologies comes from their strong dependence on physical and meteorological characteristics, which makes it more difficult to foresee production capacities. These characteristics vary from one facility to another and from one region to another. The combined use of these technologies therefore appears to be necessary to ensure that there will always be available energy at the lowest possible cost. In this context, OKWind proposes to deploy self-production units directly where the consumption is done and has developed expertise in multi-source energy production (see section 8.1).

In 2016, we started to study a solution favoring maximum autonomy of the instrumented sites from the traditional channels energy production by modeling business processes and using learning algorithms to shift demanding activities according to local production capacities. For example, the system should be able to anticipate a potential consumption of hot water (and thus of the energy needed for its production) in order to produce it at the best time when the renewable energy is available. It should also choose the best storage solution for this energy: hot water could be directly stored by the heat pump for instance. The system must implement policies that will intelligently shift demanding activities according to the predictions of energy production. It thus requires:

- **capabilities to predict the production of energy**. A lot of theoretical work has been done in the literature to predict the production of renewable sources of energy. In addition, in order to

[0] https://www.consul.io/

evaluate the production of energy and its consumption over time, OKWind has developed data retrieval mechanisms on each deployed sites. They produce accurate statistics on production and consumption. Both approaches should be used as inputs of our decision processes and model. One of our goals is to evaluate the precision of the theoretical prediction models against these real-world data to determine which are the most relevant for the implementation of our approach.

- **capabilities to model the consumption on energy**. Numerous works of the literature are interested in similar problems but focuses mainly on building electricity consumption model of machine tools [10]. We propose to focus instead on activity and business processes. In a related domain, modeling work has been conducted on water consumption of farms [7]. The objective was to predict the water consumption of an operating farm by modeling business processes. Our goal is to propose a similar model for electricity targeting a broader scope of economic sectors.
- **capabilities to schedule activities in order to match production and consumption so as to promote self-consumption.** This requires developing control loop that will proactively analyze and predict consumption and take measure to shift demand. This can, in a first approach, be done by assisting the consumers and providing them guidance on when to perform certain tasks. Assisted demand shifting have already been developed for the residential domain [6] but this project focused on uses mainly and little on the modeling of business processes. Ultimately, we would like to develop automated process transparently when possible. The learning algorithms will be developed in collaboration with Ubiant [0], a company specialized in artificial intelligence to smart-buildings.

To validate the approach and to understand business processes, we have started a field study targeting two types of activities (e.g. farm or hotel). We also want to develop tools to simulate a site so that we can quickly evaluate our policies over simulated long periods of time.

6.4. Definition of a Smart Energy Aware architecture

Participant: Jean-Marie Bonnin [contact].

In the past years, energy demand has increased and shifted especially towards electricity as the form of consuming energy. As the number of electric devices constantly grows and energy production must increasingly rely on renewable sources, this leads into noteworthy disparity between electricity production and consumption. Within the ITEA2 12004 Smart Energy Aware Systems (SEAS) project (see section 1), we proposed the "SEAS Reference Architecture Model" (S-RAM). This architecture relies on four distributed services that enable to interconnect any energy actors and give them the opportunity to provide new energy services. The benefits of S-RAM have been studied on a specific use case, which aims to provide a service for estimating local photovoltaic production. It particularly helps energy management systems better plan electric consumption. The main principles of this architecture have been published and we developed several profs of concept that have been demonstrated in the project consortium. Our partner continue to develop the components of the architecture that will be demonstrated in the final review of the project.

6.5. Context modeling for Smart Spaces

Participants: Yoann Maurel, Frédéric Weis [contact].

To provide services for Smart Building, automation based on pre-set scenarios is ineffective: human behavior is hardly predictable and application should be able to adapt their behavior at runtime depending on the context. We focused on recognizing user's activities to adapt applications behaviors. Our aim is to compute small pieces of context we called *context attributes*. Those context attributes are diverse, for example a presence in a room, the number of people in a room etc. Building efficient and accurate context information using inexpensive and non-invasive sensors was and is still a great challenge. We proved, through the use of dedicated algorithms and a layered architecture that it is achievable when the targeted space (controlled environment) is known - due to the specific and non automated calibration process we used. Among all the available theories, we used the Belief Function Theory (BFT) [8] [9] as it allows to express **uncertainty** and **imprecision**.

[0] https://www.ubiant.com/en/about/

Context is computed by a chain if three a tasks:

- The transition between a raw sensor value and a belief function is made through the use of a belief model which maps a sensor value to a belief function. A belief function represents the degree of belief associated to each possible value of the context attribute.
- Then a set of belief functions (corresponding to a set of sensors) can be combined (fused).
- Finally the system can decide what is the "best" value for the context attribute.

Currently the BFT theories requires a huge calibration process. In 2016, we obtained new results on the semi-automated building of belief functions, that have to be provided by each sensor, using our BFT Java implementation (see section 5.1).

6.6. Towards Metamorphic Housing: the on-demand room

Participants: Frédéric Weis, Michele Dominici [contact].

6.6.1. A concrete example of Metamorphic Housing: the on-demand room

The research activities related to the research program on Metamorphic Housing mainly focused on defining the detailed architecture and functionalities of the selected case study, the on-demand room. We conducted an iterative co-design process, involving the partners of the chair "Habitat Intelligent et Innovation". Valuable input was also obtained by collaborating with Delta Dore, LOUSTIC, Université de Bretagne Occidentale, etc. The result was the identification of the needs of end users, building owners and managers with respect to the on-demand room. To satisfy these requirements, we proposed a system architecture, combining computer and mobile applications with domotic equipments and novel interaction means for end users.

These are inspired by the Pervasive Computing and Interactive Architecture principles, where a continuous and implicit interaction between occupants and the physical world is made possible by augmented architectural structures, which sense the natural actions of people and respond accordingly. In this way, the occupants of the dwellings equipped with on-demand room experience a new form of housing, stimulating social interactions between neighbors and satisfying periodic needs of additional housing surface, as we illustrated in [4]. We submitted our system architecture, novel interaction means and augmented structure designs to the industrial property services of Inria and University of Rennes 1, which are currently evaluating the possibility of establishing patent protection on these inventions.

6.6.2. Experimentation of Metamorphic Housing on social housing

We helped Néotoa, a social landlord, preparing and initiating an experimentation of the on-demand room on one of their residential buildings. For this, we built and coordinated a consortium of partners working on the project: Veolia, CCI Rennes, Cardinal Edifice, Rennes Métropole, Néotoa, Delta Dore, LOUSTIC, Université de Bretagne Occidentale, MobBI platform (University of Rennes 1), Inria, Institut de Gestion de Rennes. We took a user-centered approach to the problem, studying it from several points of view and mobilizing several disciplines: psychology and ergonomics (LOUSTIC), sociology (Université de Bretagne Occidentale), marketing (Institut de Gestion de Rennes). We conducted user interviews, initially leveraging the demonstrator of the on-demand room that we previously built via the Immersia virtual reality platform. Then, we ran on-line inquiries to reach a larger audience. We took into account the lessons that we learned in the design and development of a computing and domotic system, leveraging the expertise of valuable partners (Delta Dore, MobBI platform, Inria), as detailed in section 5.3 .

TAMIS Team

7. New Results

7.1. Results for Axis 1: Vulnerability analysis

Statistical model checking employs Monte Carlo methods to avoid the state explosion problem of probabilistic (numerical) model checking. To estimate probabilities or rewards, SMC typically uses a number of statistically independent stochastic simulation traces of a discrete event model. Being independent, the traces may be generated on different machines, so SMC can efficiently exploit parallel computation. Reachable states are generated on the fly and SMC tends to scale polynomially with respect to system description. Properties may be specified in bounded versions of the same temporal logics used in probabilistic model checking. Since SMC is applied to finite traces, it is also possible to use logics and functions that would be intractable or undecidable for numerical techniques.

Several model checking tools have added SMC as a complement to exhaustive model checking. This includes the model checker UPPAAL, for timed automata, the probabilistic model checker PRISM, and the model checker Ymer, for continuous time Markov chains. Plasma Lab [29] is the first platform entirely dedicated to SMC. Contrary to other tools, that target a specific domain and offer several analysis techniques, including basic SMC algorithms, Plasma Lab is designed as a generic platform that facilitates multiple SMC algorithms, multiple modelling and query languages and has multiple modes of use. This allows us to apply statistical model checking techniques to a wide variety of problems, reusing existing simulators. With this process we avoid the task of rewriting a model of a system in a language not ideally design to do it. This complex task often leads either to an approximation of the original system or to a more complex model harder to analyze. The task needed to support a new simulator is to implement an interface plugin between our platform Plasma Lab and the existing tool, using the public API of our platform. This task has to be performed only once to analyze all the systems supported by the existing simulator.

Plasma Lab can already be used with the PRISM language for continuous and discrete time Markov chains and biological models. During the last years we have developed several new plugins to support SystemC language [50], Simulink models [70], dynamic software architectures language [41], [14], and train interlocking systems [64]. They have been presented in recent publications.

[50] Transaction-level modeling with SystemC has been very successful in describing the behavior of embedded systems by providing high-level executable models, in which many of them have an inherent probabilistic behavior, i.e., random data, unreliable components. It is crucial to evaluate the quantitative and qualitative analysis of the probability of the system properties. Such analysis can be conducted by constructing a formal model of the system and using probabilistic model checking. However, this method is unfeasible for large and complex systems due to the state space explosion. In this paper, we demonstrate the successful use of statistical model checking to carry out such analysis directly from large SystemC models and allows designers to express a wide range of useful properties.

[70] We present an extension of the statistical model checker Plasma Lab capable of analyzing Simulink models.

[41], Dynamic software architectures emerge when addressing important features of contemporary systems, which often operate in dynamic environments subjected to change. Such systems are designed to be reconfigured over time while maintaining important properties, e.g., availability, correctness, etc. Verifying that reconfiguration operations make the system to meet the desired properties remains a major challenge. First, the verification process itself becomes often difficult when using exhaustive formal methods (such as model checking) due to the potentially infinite state space. Second, it is necessary to express the properties to be verified using some notation able to cope with the dynamic nature of these systems. Aiming at tackling these issues, we introduce

DynBLTL, a new logic tailored to express both structural and behavioral properties in dynamic software architectures. Furthermore, we propose using statistical model checking (SMC) to support an efficient analysis of these properties by evaluating the probability of meeting them through a number of simulations. In this paper, we describe the main features of DynBLTL and how it was implemented as a plug-in for PLASMA, a statistical model checker.

[14] The critical nature of many complex software-intensive systems calls for formal, rigorous architecture descriptions as means of supporting automated verification and enforcement of architectural properties and constraints. Model checking has been one of the most used techniques to automatically analyze software architectures with respect to the satisfaction of architectural properties. However, such a technique leads to an exhaustive exploration of all possible states of the system under verification, a problem that becomes more severe when verifying dynamic software systems due to their typical non-deterministic runtime behavior and unpredictable operation conditions. To tackle these issues, we propose using statistical model checking (SMC) to support the analysis of dynamic software architectures while aiming at reducing effort, computational resources, and time required for this task. In this paper, we introduce a novel notation to formally express architectural properties as well as an SMC-based toolchain for verifying dynamic software architectures described in π-ADL, a formal architecture description language. We use a flood monitoring system to show how to express relevant properties to be verified, as well as we report the results of some computational experiments performed to assess the efficiency of our approach.

[64] , accepted at HASE 2017 In the railway domain, an interlocking is the system ensuring safe train traffic inside a station by controlling its active elements such as the signals or points. Modern interlockings are configured using particular data, called application data, reflecting the track layout and defining the actions that the interlocking can take. The safety of the train traffic relies thereby on application data correctness, errors inside them can cause safety issues such as derailments or collisions. Given the high level of safety required by such a system, its verification is a critical concern. In addition to the safety, an interlocking must also ensure that availability properties, stating that no train would be stopped forever in a station, are satisfied. Most of the research dealing with this verification relies on model checking. However, due to the state space explosion problem, this approach does not scale for large stations. More recently, a discrete event simulation approach limiting the verification to a set of likely scenarios, was proposed. The simulation enables the verification of larger stations, but with no proof that all the interesting scenarios are covered by the simulation. In this paper, we apply an intermediate statistical model checking approach, offering both the advantages of model checking and simulation. Even if exhaustiveness is not obtained, statistical model checking evaluates with a parameterizable confidence the reliability and the availability of the entire system.

7.1.1. Verification of Dynamic Software Architectures

Participants: Axel Legay, Jean Quilbeuf, Louis-Marie Traonouez.

Dynamic software architectures emerge when addressing important features of contemporary systems, which often operate in dynamic environments subjected to change. Such systems are designed to be reconfigured over time while maintaining important properties, e.g., availability, correctness, etc. π-ADL is a formal, well-founded theoretically language intended to describe software architectures under both structural and behavioral viewpoints. In order to cope with dynamicity concerns, π-ADL is endowed with architectural level primitives for specifying programmed reconfiguration operations, i.e., foreseen, pre-planned changes described at design time and triggered at runtime by the system itself under a given condition or event. Additionally, code source in the Go programming language is automatically generated from π-ADL architecture descriptions, thereby allowing for their execution.

We have developed with Plasma Lab a toolchain [14] for verifying dynamic software architectures described in π-ADL. The architecture description in π-ADL is translated towards generating source code in Go. As π-ADL architectural models do not have a stochastic execution, they are linked to a stochastic scheduler parameterized by a probability distribution for drawing the next action. Furthermore, we use existing probability distribution

Go libraries to model inputs of system models as user functions. The program resulting from the compilation of the generated Go source code emits messages referring to transitions from addition, attachment, detachment, and value exchanges of architectural elements. Additionally we have introduced DynBLTL [41] a new logic tailored to express both structural and behavioral properties in dynamic software architectures.

We have developed two plugins atop the PLASMA platform, namely (i) a simulator plug-in that interprets execution traces produced by the generated Go program and (ii) a checker plugin that implements DynBLTL. With this toolchain, a software architect is able to evaluate the probability of a π-ADL architectural model to satisfy a given property specified in DynBLTL.

7.1.2. Statistical Model-Checking of Scheduling Systems

Participants: Axel Legay, Louis-Marie Traonouez.

Cyber-Physical Systems (CPS) are software implemented control systems that control physical objects in the real world. These systems are being increasingly used in many critical systems, such as avionics and automotive systems. They are now integrated into high performance platforms, with shared resources. This motivates the development of efficient design and verification methodologies to assess the correctness of CPS.

Schedulability analysis is a major problem in the design of CPS. Software computations that implements the commands sent to the CPS are split into a set of hard real-time tasks, often periodic. These tasks are associated to strict deadlines that must be satisfied. A scheduler is responsible for dispatching a shared resource (usually CPU computation time) among the different tasks according to a chosen scheduling policy. The schedulability analysis consists in verifying that the tasks always meet their deadlines.

Over the years, the schedulability of CPS have mainly been performed by analytical methods. Those techniques are known to be effective but limited to a few classes of scheduling policies. In a series of recent work, it has been shown that schedulability analysis of CPS could be performed with a model-based approach and extensions of verification tools such as UPPAAL. It shows that such models are flexible enough to embed various types of scheduling policies that go beyond those in the scope of analytical tools.

We have extended these works to include more complex features in the design of these systems and we have experimented the use of statistical model checking as a lightweight verification technique for these systems.

We also extended the approach to statistical model checking of products lines. Our first contribution has been to propose models to design software product lines (SPL) of preemptive real-time systems [25]. Software Product Line Engineering (SPLE) allows reusing software assets by managing the commonality and variability of products. Recently, SPLE has gained a lot of attention as an approach for developing a wide range of software products from non-critical to critical software products, and from application software to platform software products.

Real-time software products (such as real-time operating systems) are a class of systems for which SPLE techniques have not drawn much attention from researchers, despite the need to efficiently reuse and customize real-time artifacts. We have proposed a formal SPLE framework for real-time systems. It focuses on the formal analysis of real-time properties of an SPL in terms of resource sharing with time dependent functionalities. Our framework provides a structural description of the variability and the properties of a real time system, and behavioral models to verify the properties using formal techniques implemented in the tools UPPAAL symbolic model checker and UPPAAL statistical model checker. For the specification of an SPL, we propose an extension of a feature model, called Property Feature Model (PFM). A PFM explicitly distinguishes features and properties associated with features, so that properties are analyzed in the context of the relevant features. We also define a non-deterministic decision process that automatically configures the products of an SPL that satisfy the constraints of a given PFM and the product conditions of customers. Finally we analyze the products against the associated properties. For analyzing real-time properties, we provide feature behavioral models of the components of a scheduling unit, i.e. tasks, resources and schedulers. Using these feature behavioral models, a family of scheduling units of an SPL is formally analyzed against the designated properties with model checking techniques.

[25] This paper presents a formal analysis framework to analyze a family of platform products w.r.t. real-time properties. First, we propose an extension of the widely-used feature model, called Property Feature Model (PFM), that distinguishes features and properties explicitly. Second, we present formal behavioral models of components of a real-time scheduling unit such that all real-time scheduling units implied by a PFM are automatically composed to be analyzed against the properties given by the PFM. We apply our approach to the verification of the schedulability of a family of scheduling units using the symbolic and statistical model checkers of UPPAAL.

7.1.3. Model-based Framework for Hierarchical Scheduling Systems

Participants: Axel Legay, Louis-Marie Traonouez, Mounir Chadli.

In order to reduce costs in the design of modern CPS, manufacturers devote strong efforts to maximize the number of components that can be integrated on a given platform. This can be achieved by minimizing the resource requirements of individual components. A hierarchical scheduling systems (HSS) integrates a number of components into a single system running on one execution platform. Hierarchical scheduling systems have been gaining more attention by automotive and aircraft manufacturers because they are practical in minimizing the cost and energy of operating applications.

Several papers have proposed model-based compositional framework for HSS. In [4] we proposed a methodology for optimizing the resource requirement of a component of an HSS using model checking techniques. Our methodology consists of using a lightweight statistical model checking method and a costly but absolute certain symbolic model checking method that operates on identical models.

In another work [15] we have proposed stochastic extension of HSS that allows us to capture tasks whose real-time attributes, such as deadline, execution time or period, are also characterized by probability distributions. This is particularly useful to describe mixed-critical systems and make assumptions on the hardware domain. These systems combine hard real-time periodic tasks, with soft real-time sporadic tasks. Classical scheduling techniques can only reason about worst case analysis of these systems, and therefore always return pessimistic results. Using tasks with stochastic period we can better quantify the occurrence of these tasks. Similarly, using stochastic deadlines we can relax timing requirements, and stochastic execution times are used to model the variation of the computation time needed by the tasks. These distributions can be sampled from executions or simulations of the system, or set as requirements from the specifications. For instance in avionics, display components have lower criticality. They can include sporadic tasks generated by users requests. Average user demand is efficiently modeled with a probability distribution.

We have also developed a graphical high-level language to represent scheduling units and complex hierarchical scheduling systems. In order to bridge the gap between the formalisms, we exploit Cinco, a generator for domain specific modeling tools to generate an interface between this language and the one of UPPAAL. Cinco allows to specify the features of a graphical interface in a compact meta-model language. This is a flexible approach that could be extended to any formal model of scheduling problem.

[4] Compositional reasoning on hierarchical scheduling systems is a well-founded formal method that can construct schedulable and optimal system configurations in a compositional way. However, a compositional framework formulates the resource requirement of a component, called an interface, by assuming that a resource is always supplied by the parent components in the most pessimistic way. For this reason, the component interface demands more resources than the amount of resources that are really sufficient to satisfy sub-components. We provide two new supply bound functions which provides tighter bounds on the resource requirements of individual components. The tighter bounds are calculated by using more information about the scheduling system. We evaluate our new tighter bounds by using a model-based schedulability framework for hierarchical scheduling systems realized as UPPAAL models. The timed models are checked using model checking tools UPPAAL and UPPAAL SMC, and we compare our results with the state of the art tool CARTS.

[15] Over the years, schedulability of Cyber-Physical Systems (CPS) have mainly been performed by analytical methods. Those techniques are known to be effective but limited to a few classes of scheduling policies. In a series of recent work, we have shown that schedulability analysis of

CPS could be performed with a model-based approach and extensions of verification tools such as UPPAAL. One of our main contribution has been to show that such models are flexible enough to embed various types of scheduling policies that go beyond those in the scope of analytical tools. In this paper, we go one step further and show how our formalism can be extended to account for stochastic information, such as sporadic tasks whose attributes depend on the hardware domain. Our second contribution is to make our tools accessible to average users that are not experts in formal methods. For doing so, we propose a graphical and user-friendly language that allows us to describe scheduling problems. This language is automatically translated to formal models by exploiting a meta-model approach. The principle is illustrated on a case study.

7.1.4. Verification of Interlocking Systems

Participants: Axel Legay, Louis-Marie Traonouez, Jean Quilbeuf.

An interlocking is a system that controls the train traffic by acting as an interface between the trains and the railway track components. The track components are for example, the signals that allow the train to proceed, or the points that guide the trains from one track to another. The paths followed by the trains are called routes. Modern interlockings are computerized systems that are composed of generic software and application data.

We have proposed in collaboration with Université Catholique de Louvain and Alstom a method to automatically verify an interlocking using simulation and statistical model checking [64]. We use a simulator developed by Université Catholique de Louvain that is able to generate traces of the interlocking systems from a track layout and application data. This simulator is plug with Plasma Lab using a small interface developed with Plasma Lab's API. Then, the traces generated by the simulator have been used by Plasma Lab SMC algorithms to measure the correctness of the system. We have used Monte-Carlo and importance splitting algorithms to verify this system.

7.1.5. Advanced Statistical Model Checking

Participants: Axel Legay, Sean Sedwards, Louis-Marie Traonouez.

Statistical model checking (SMC) addresses the state explosion problem of numerical model checking by estimating quantitative properties using simulation. Rare events, such as software bugs, are often critical to the performance of systems but are infrequently observed in simulations. They are therefore difficult to quantify using SMC. Nondeterministic systems deliberately leave parts of system behaviour undefined, hence it is not immediately possible to simulate them. Our ongoing work thus pushes the boundaries of the cutting edge of SMC technology by focusing on rare event verification and the optimisation of nondeterminism.

7.1.5.1. Optimizing Nondeterministic Systems

[17] Probabilistic timed automata (PTA) generalize Markov decision processes (MDPs) and timed automata (TA), both of which include nondeterminism. MDPs have discrete nondeterministic choices, while TA have continuous nondeterministic time. In this work we consider finding *schedulers* that resolve all nondeterministic choices in order to maximize or minimize the probability of a time-bounded LTL property. Exhaustive numerical approaches often fail due to state explosion, hence we present a new lightweight on-the-fly algorithm to find near-optimal schedulers. To discretize the continuous choices we make use of the classical region and zone abstractions from timed automata model checking. We then apply our recently developed "smart sampling" technique for statistical verification of Markov decision processes. On standard case studies our algorithm provides good estimates for both maximum and minimum probabilities. We compare our new approach with alternative techniques, first using tractable examples from the literature, then motivate its scalability using case studies that are intractable to numerical model checking and challenging for existing statistical techniques.

7.1.5.2. Rare Event Verification

[3] Importance sampling is a standard technique to significantly reduce the computational cost of quantifying rare properties of probabilistic systems. It works by weighting the original distribution of the system to make the rare property appear more frequently in simulations, then compensating the resulting estimate by the weights. This can be done on the fly with minimal storage, but the challenge is to find *near optimal* importance sampling distributions efficiently, where optimal means that paths that do not satisfy the property are never seen, while paths that satisfy the property appear in the same proportion as in the original distribution.

Our approach uses a tractable cross-entropy minimization algorithm to find an optimal parameterized importance sampling distribution. In contrast to previous work, our algorithm uses a naturally defined low dimensional vector to specify the distribution, thus avoiding an explicit representation of a transition matrix. Our parametrisation leads to a unique optimum and is shown to produce many orders of magnitude improvement in efficiency on various models. In this work we specifically link the existence of optimal importance sampling distributions to time-bounded logical properties and show how our parametrisation affects this link. We also motivate and present simple algorithms to create the initial distribution necessary for cross-entropy minimization. Finally, we discuss the open challenge of defining error bounds with importance sampling and describe how our optimal parameterized distributions may be used to infer qualitative confidence.

[10] In this work we consider rare events in systems of Stochastic Timed Automata (STA) with time-bounded reachability properties. This model may include rarity arising from explicit discrete transitions, as well as more challenging rarity that results from the intersection of timing constraints and continuous distributions of time. Rare events have been considered with simple combinations of continuous distributions before, e.g., in the context of queueing networks, but we present an automated framework able to work with arbitrarily composed STA. By means of symbolic exploration we first construct a zone graph that excludes unfeasible times. We then simulate the system within the zone graph, avoiding "dead ends" on the fly and proportionally redistributing their probability to feasible transitions. In contrast to many other importance sampling approaches, our "proportional dead end avoidance" technique is guaranteed by construction to reduce the variance of the estimator with respect to simulating the original system. Our results demonstrate that in practice we can achieve substantial overall computational gains, despite the symbolic analysis.

[49] In this invited paper we outline some of our achievements in quantifying rare properties in the context of SMC. In addition to the importance sampling techniques described above, we also describe our work on importance *splitting*. Importance splitting works by decomposing the probability of a rare property into a product of probabilities of sub-properties that are easier to estimate. The sub-properties are defined by *levels* of a *score function* that maps states of the system$\times$property product automaton to values. We have provided the first general purpose implementation of this approach, using user-accessible "observers" that are compiled automatically from the property. These observers may be used by both fixed and adaptive level importance splitting algorithms and are specifically designed to make distribution efficient.

7.1.6. Side-channel Analysis of Cryptographic Substitution Boxes

Participants: Axel Legay, Annelie Heuser.

With the advent of the Internet of Things, we are surrounded with smart objects (aka things) that have the ability to communicate with each other and with centralized resources. The two most common and widely noticed artefacts are RFID and Wireless Sensor Networks which are used in supply-chain management, logistics, home automation, surveillance, traffic control, medical monitoring, and many more. Most of these applications have the need for cryptographic secure components which inspired research on cryptographic algorithms for constrained devices. Accordingly, lightweight cryptography has been an active research area over the last 10 years. A number of innovative ciphers have been proposed in order to optimize various performance criteria and have been subject to many comparisons. Lately, the resistance against side-channel attacks has been considered as an additional decision factor.

Side-channel attacks analyze physical leakage that is unintentionally emitted during cryptographic operations in a device (e.g., power consumption, electromagnetic emanation). This side-channel leakage is statistically dependent on intermediate processed values involving the secret key, which makes it possible to retrieve the secret from the measured data.

Side-channel analysis (SCA) for lightweight ciphers is of particular interest not only because of the apparent lack of research so far, but also because of the interesting properties of substitution boxes (S-boxes). Since the nonlinearity property for S-boxes usually used in lightweight ciphers (i.e., 4×4) can be maximally equal to 4, the difference between the input and the output of an S-box is much smaller than for instance for AES. Therefore, one could conclude that from that aspect, SCA for lightweight ciphers must be more difficult. However, the number of possible classes (e.g., Hamming weight (HW) or key classes) is significantly lower, which may indicate that SCA must be easier than for standard ciphers. Besides the difference in the number of classes and consequently probabilities of correct classification, there is also a huge time and space complexity advantage (for the attacker) when dealing with lightweight ciphers.

In [23], [67] we give a detailed study of lightweight ciphers in terms of side-channel resistance, in particular for software implementations. As a point of exploitation we concentrate on the non-linear operation (S-box) during the first round. Our comparison includes SPN ciphers with 4-bit S-boxes such as KLEIN, PRESENT, PRIDE, RECTANGLE, Mysterion as well as ciphers with 8-bit S-boxes: AES, Zorro, Robin. Furthermore, using simulated data for various signal-to-noise ratios (SNR) we present empirical results for Correlation Power Analysis (CPA) and discuss the difference between attacking 4-bit and 8-bit S-boxes.

Following this direction current studies evaluate and connect cryptographic properties with side-channel resistance. More precisely, in an ideal setting a cipher should be resilient against cryptanalyses as well as side-channel attacks and yet easy and cheap to be implemented. However, since that does not seem to be possible at the current level of knowledge, we are required to make a number of trade-offs. Therefore, we investigate several S-box parameters and connect them with well known cryptographic properties of S-boxes. Moreover, when possible we give clear theoretical bounds on those parameters as well as expressions connecting them with properties like nonlinearity and δ-uniformity. We emphasize that we select the parameters to explore on the basis of their possible connections with the side-channel resilience of S-boxes.

To this end, we divide the primary goal into several sub-problems. First, we discuss what is the maximal number of fixed points one can have in an optimal S-box. The question of the maximal number of fixed points for an optimal S-box is of interest on its own, but also in the side-channel context since intuitively an S-box with many fixed points should consume less power and therefore have less leakage. Moreover, the preservation of Hamming weight and a small Hamming distance between x and $F(x)$ are two more properties each of which could strengthen the resistance to SCA. Our findings show that notably in the case when exactly preserving the Hamming weight, the confusion coefficient reaches good value and consequently the S-box has good SCA resilience. We show that the S-boxes with no differences in the Hamming weight of their input and output pairs (and even, S-boxes F such that $F(x)$ have Hamming weight near the Hamming weight of x, on average) or S-boxes such that $F(x)$ lies at a small Hamming distance from x cannot have high nonlinearity (although the obtainable values are not too bad for $n = 4, 8$) and therefore are not attractive in practical applications. Note that an S-box with many fixed points is also a particular case of an S-box that preserves the Hamming weight/distance between the inputs and outputs. Furthermore, our study includes involutive functions since they have a particular advantage over general pseudo-permutations. In particular, not only from an implementation viewpoint but also their side-channel resilience is the same regardless if an attacker considers the encryption or decryption phase as well as attacking the first or the last round. Next, we find a theoretical expression connecting the confusion coefficient with that of preserving the Hamming weight of inputs and outputs.

In the practical part, we first confirm our theoretical findings about the connection between preserving Hamming weight and the confusion coefficient. Besides that, we give a number of S-box examples of size 4×4 intended to provide more insight into possible trade-offs between cryptographic properties and side-channel resilience. However, our study shows that mostly preserving Hamming weight might not automatically result in a small minimum confusion coefficient and thus in higher side-channel resistance. We therefore in

detail examine the influence of F on the confusion coefficient in general by concentrating on the input (in which key hypothesis are made) and the minimum value of the confusion coefficient. Following, we evaluate a number of S-boxes used in today's ciphers and show that their SCA resilience can significantly differ. Finally, we point out that non-involutive S-boxes might bring a significant disadvantage in case an attacker combines the information about F and F^{-1} by either targeting both first and last round of an algorithm or encryption and decryption.

[67] Side-channel Analysis of Lightweight Ciphers: Current Status and Future Directions

[23] Side-channel Analysis of Lightweight Ciphers: Does Lightweight Equal Easy?

7.1.7. Binary Code Analysis: Formal Methods for Fault Injection Vulnerability Detection

Participants: Axel Legay, Thomas Given-Wilson, Nisrine Jafri, Jean-Louis Lanet.

Formal methods such as model checking provide a powerful tool for checking the behaviour of a system. By checking the properties that define correct system behaviour, a system can be determined to be correct (or not).

Increasingly fault injection is being used as both a method to attack a system by a malicious attacker, and to evaluate the dependability of the system. By finding fault injection vulnerabilities in a system, the resistance to attacks or faults can be detected and subsequently addressed.

A process is presented that allows for the automated simulation of fault injections. This process proceeds by taking the executable binary for the system to be tested, and validating the properties that represent correct system behaviour using model checking. A fault is then injected into the executable binary to produce a mutant binary, and the mutant binary is model checked also. A different result to the validation of the executable binary in the checking of the mutant binary indicates a fault injection vulnerability.

This process has been automated with existing tools, allowing for easy checking of many different fault injection attacks and detection of fault injection vulnerabilities. This allows for the detection of fault injection vulnerabilities to be fully automated, and broad coverage of the system to be formally shown.

7.1.8. Security at the hardware and software boundaries

Participants: Axel Legay, Nisrine Jafri, Jean-Louis Lanet, Ronan Lashermes, Hélène Le Bouder.

7.1.8.1. IoT security

IoT security has to face all the challenges of the mainstream computer security but also new threats. When an IoT device is deployed, most of the time it operates in a hostile environment, i.e. the attacker can perform any attack on the device. If secure devices use tamper resistant chip and are programmed in a secure manner, IoT use low cost micro-controllers and are not programmed in a secure way. We developed new attacks but also evaluate how the code polymorphism can be used against these attacks. In [45] [27] we developed a template attack to retrieve the value of a PIN code from a cellphone. We demonstrated that the maximum trials to retrieve the four bytes of secret PIN is 8 and in average 3 attempts are sufficient. A supervised learning algorithm is used.

Often smart phones allow up to 10 attempts before locking definitely the memory. We used an embedded code generator [16], [45] dedicated to a given security function using a DSL to increase the security level of a non tamper resistant chip. We brought to the fore that a design of the software for protecting against fault attacks decreases the security against SCA. Fault attack is a mean to execute a code that is slightly different from the one that has been loaded into the device. Thus, to be sure that a genuine code cannot be dynamically transformed, one needs to analyze any possibility of a code to be transformed.

The work presented in [34] made possible to design an extremely effective architecture to achieve Montgomery modular multiplication. The proposed solution combines a limited resource consumption with the lowest latency compared with the literature. This allows to envisage new applications of asymmetric cryptography in systems with few resources. In order to find a cryptographic key using hidden channels, most attacks use the a priori knowledge of texts sent or received by the target. The proposed analysis presented in [28] does not use these assumptions. A belief propagation technique is used to cross the information obtained from leaked information with the equations governing the targeted algorithm.

7.1.8.2. Safe update mechanism for IoT

One of the challenges for IoT is the possibility to update the code through a network. This is done by stopping the system, loading the new version, verifying the signature of the firmware and installing it into the memory. Then, the memory must be cleaned to eliminate the code and the data of the previous version. Some IoT (sensor acquisition and physical system control) requires to never stop while executing the code. We have developed a complete architecture that performs such an update with real time capabilities. If one wants to use this characteristic in a real world it should pass certification. In particular he has to demonstrate that the system performs as expected. We used formal methods (mainly Coq) to demonstrate that the semantics of the code is preserved during the update. In [30], we paid attention to the detection of the Safe Update Point (SUP) because our implementation had some time an unstable behavior. We demonstrated that in a specific case, while several threads using code to be updated, the system enters into a deadlock. After discovering the bug, we patched our system.

7.1.8.3. Reverse engineering of firmware

Reverse engineering has two aspects; code reverse for which the literature is abundant and data reverse i.e. understanding the meaning of a structure and its usage has been less studied. The first step in reverse engineering consists in getting access of the code. In the case of romized code in a SoC, the access to the code is protected by a MMU mechanism and thus is an efficient mitigation mechanism against reverse engineering. In [8], [2] and [33] we set up several attacks to get access to the code even in presence of a MMU. The attack in [8] uses a vulnerability in the API where an object can be used instead of an array. This allows to read and write the code area leading to the possibility to execute arbitrary code in memory. In [33], we use the attack tree paradigm to explore all the possibilities to mount an attack on a given product. In [2], we used a ROP (Return Oriented Programming) attack to inject a shell code in the context of another application. Due to the fact that the shell code is executed in the context of the caller, the firewall is unable to detect the access to the secure container of the targeted application. This allows us to retrieve the content of the secure containers.

Once the dump has been obtained, one can try to retrieve code and data. Retrieving code is not obvious but several tools exist to help the analyst. These tools require that all the ISA (Instruction Set Architecture) is known. Sometime, the ISA is not known and in particular when one wants to obfuscate the code, he can use a virtual machine to execute dedicated byte code. In [32], we developed a methodology to infer the missing byte code, then we execute a data type inference to understand the memory management algorithm.

7.2. Results for Axis 2: Malware analysis

The detection of malicious programs is a fundamental step to be able to guarantee system security. Programs that exhibit malicious behavior, or *malware*, are commonly used in all sort of cyberattacks. They can be used to gain remote access on a system, spy on its users, exfiltrate and modify data, execute denial of services attacks, etc.

Significant efforts are being undertaken by software and data companies and researchers to protect systems, locate infections, and reverse damage inflicted by malware. Malware analysis can be divided in the following three main problems:

7.2.1. Malware Detection

Participants: Axel Legay, Fabrizio Biondi, Olivier Decourbe, Mike Enescu, Thomas Given-Wilson, Annelie Heuser, Nisrine Jafri, Jean-Louis Lanet, Jean Quilbeuf.

Given a file or data stream, the malware detection problem consists of understanding if the file or data stream contain traces of malicious behavior. For binary executable files in particular, this requires reverse engineering the file's behavior to understand if it is malicious. The main reverse engineering techniques are categorized as:

Static Analysis This refers to techniques that analyze the file without executing it. It includes disassembling the file's executable code and analyzing other static features of the binary, like its import/export table, hash, etc. The file's control flow and system flow graphs can be retrieved statically (unless they are obfuscated; see below) and used to guide the exploration of the file's semantics in the search of

malicious behavior. Information flow can be tracked since hostile applications often try to transmit private information to distant servers (this form of malware are now widely spread in the mobile world). The challenge consists in detecting into a file that a private information does not leak to the external world. The verification can be done statically, dealing with storage channel (implicit or explicit), but not with side channel.

Dynamic Analysis This refers to techniques that actually executed the file in a sandbox (usually a virtualized environment) and analyze its interaction with the sandbox. This technique is effective in understanding the file's actual interactions with the system, making it easy to detect malicious behavior. However, malware often implements sandbox detection techniques to detect when it is being run in a virtualized environment, when functions or system calls are hooked by the analyst, or when the sandbox does not look like a normal user's machine (e.g. because it does not contain any document). Dynamic tracking of information flow makes it possible to cope with side channel attacks. With temporal side channel, the challenge lies in the potential declassification procedure used by malware to escape the analysis. We extend the TaintDroid framework to cope with native code invocation [47]. This approach reduces the false positive warning drastically. Recently we have extended this work to cope with timing side channels [under submission]. We are developing a new malware that declassifies the labels thanks to the audio system of the smart-phone. This is a joint work with Telecom Bretagne.

Hybrid Analysis This refers to technique that combine both static and dynamic behavior, i.e. both code analysis and execution. While more complex to implement, these techniques are able to overcome many of the shortcomings of full static and full dynamic analysis. The best example of a hybrid technique is concolic (a portmanteau for CONCrete + symbOLIC) analysis.

To contribute to concolic analysis, we are working on the state-of-the-art `angr` concolic execution engine to make it fast and efficient enough to analyze large executable malware files efficiently. We are improving `angr` 's parallelism and allowing it to precompute semantic stubs of function and system calls, allowing it to focus its analysis on the main file without having to branch in the rest of the operative system. We plan to contribute our improvements to the main `angr` branch, so that the whole community can benefit from them.

7.2.2. Malware Deobfuscation

Participants: Axel Legay, Fabrizio Biondi, Olivier Decourbe, Mike Enescu, Thomas Given-Wilson, Annelie Heuser, Nisrine Jafri, Jean-Louis Lanet, Jean Quilbeuf.

Given a file (usually a portable executable binary or a document supporting script macros), deobfuscation refers to the preparation of the file for the purposes of further analysis. Obfuscation techniques are specifically developed by malware creators to hinder detection reverse engineering of malicious behavior. Some of these techniques include:

Packing Packing refers to the transformation of the malware code in a compressed version to be dynamically decompressed into memory and executed from there at runtime. Packing techniques are particularly effective against static analysis, since it is very difficult to determine statically the content of the unpacked memory to be executed, particularly if packing is used multiple times. The compressed code can also be encrypted, with the key being generated in a different part of the code and used by the unpacking procedure, or even transmitted remotely from a command and control (C&C) server.

Control Flow Flattening This technique aims to hinder the reconstruction of the control flow of the malware. The malware's operation are divided into basic blocks, and a dispatcher function is created that calls the blocks in the correct order to execute the malicious behavior. Each block after its execution returns control to the dispatcher, so the control flow is flattened to two levels: the dispatcher above and all the basic blocks below.

To prevent reverse engineering of the dispatcher, it is often implemented with a cryptographic hash function. A more advanced variant of this techniques embed a full virtual machine with a randomly generated instruction set, a virtual program counted, and a virtual stack in the code, and uses the machine's interpreter as the dispatcher.

Virtualization is a very effective technique to prevent reverse engineering. To contrast it, we are implementing state-of-the-art devirtualization algorithms in `angr` , allowing it to detect and ignore the virtual machine code and retrieving the obfuscated program logic. Again, we plan to contribute our improvements to the main `angr` branch, thus helping the whole security community fighting virtualized malware.

Opaque Constants and Conditionals Reversing packing and control flow flattening techniques requires understanding of the constants and conditionals in the program, hence many techniques are deployed to obfuscate them and make them unreadable by reverse engineering techniques. Such techniques are used e.g. to obfuscate the decryption keys of packed encrypted code and the conditionals in the control flow.

We have proven the efficiency of dynamic synthesis in retrieving opaque constant and conditionals, compared to the state-of-the-art approach of using SMT (Satisfiability Modulo Theories) solvers, when the input space of the opaque function is small enough. We are developing techniques based on fragmenting and analyzing by brute force the input space of opaque conditionals, and SMT constraints in general, to be integrated in SMT solvers to improve their effectiveness.

7.2.3. Malware Classification

Participants: Axel Legay, Fabrizio Biondi, Olivier Decourbe, Mike Enescu, Thomas Given-Wilson, Annelie Heuser, Nisrine Jafri, Jean-Louis Lanet, Jean Quilbeuf.

Once malicious behavior has been located, it is essential to be able to classify the malware in its specific family to know how to disinfect the system and reverse the damage inflicted on it.

While it is rare to find an actually previously unknown malware, morphic techniques are employed by malware creators to ensure that different generations of the same malware behave differently enough than it is hard to recognize them as belonging to the same family. In particular, techniques based on the syntax of the program fails against morphic malware, since syntax can be easily changed.

To this end, semantic signatures are used to classify malware in the appropriate family. Semantic signatures capture the malware's behavior, and are thus resistant to morphic and differentiation techniques that modify the malware's syntactic signatures. We are investigating semantic signatures based on the program's System Call Dependency Graph (SCDG), which have been proven to be effective and compact enough to be used in practice. SCDGs are often extracted using a technique based on pushdown automata that is ineffective against obfuscated code; instead, we are applying concolic analysis via the `angr` engine to improve speed and coverage of the extraction.

Once a semantic signature has been extracted, it has to be compared against large database of known signatures representing the various malware families to classify it. The most efficient way to obtain this is to use a supervised machine learning classifier. In this approach, the classifier is trained with a large sample of signatures malware annotated with the appropriate information about the malware families, so that it can learn to quickly and automatically classify signatures in the appropriate family. Our work on machine learning classification focuses on using SCDGs as signatures. Since SCDGs are graphs, we are investigating and adapting algorithms for the machine learning classification of graphs, usually based on measures of shared subgraphs between different graphs.

In malware detection and classification, it is fundamental to have a false positive rate (i.e. rate of cleanware classified as malware) approaching zero, otherwise the classification system will classify hundred or thousands of cleanware files as malware, making it useless in practice. To decrease the false positive rate, the classifier is also trained with a large and representative database of cleanware, so that it can discriminate between signatures of cleanware and malware with a minimal false positive rate. We use a large database of malware and cleanware to train our classifier, thus guaranteeing a high detection rate with a small false positive rate.

7.2.4. Papers

This section gathers papers that are results common to all sections above pertaining to Axis 2.

[57] Black-box synthesis is more efficient than SMT deobfuscation on predicates obfuscated with Mixed-Boolean Arithmetics.

[66] Recently fault injection has increasingly been used both to attack software applications, and to test system robustness. Detecting fault injection vulnerabilities has been approached with a variety of methods, yielding varied results. This paper proposes a general process using model checking to detect fault injection vulnerabilities in binaries. The process is implemented and used to detect a variety of different kinds of fault injection vulnerabilities in binaries.

[59] Fault-injection exploits hardware weaknesses to perturbate the behaviour of embedded devices. Here, we present new model-based techniques and tools to detect such attacks developed at the High-Security Laboratory at Inria.

[52] We proposed to use a bare metal approach without virtualization and a method to let the system stop the execution while the malware has been deployed in memory.

[51] We present our framework to grab sample from the net, evaluate it on victim PC and detect its presence thanks to our counter measures.

[53] In this paper, two counter measures are presented. The first one is related with the mode ECB of the AES cryptographic algorithm and the second is related with the usage of the crypto API. We developed a cryptographic provider which intercepts the key generation and store it in a safe place. Then we are able to decipher any files that the malware should have encrypted.

7.3. Results for Axis 3: Building a secure network stack

7.3.1. Private set intersection cardinality

Participants: Jeffrey Burdges, Alvaro Garcia Recuero, Christian Grothoff.

We designed new efficient protocol for privacy-preserving signed set intersection cardinality using blinded BLS signatures over bilinear maps and demonstrated its utility in machine learning for abuse detection in decentralised online social networks. The paper was presented at DPM 2016 [21].

7.3.2. Cell tower privacy

Participants: Christian Grothoff, Neal Walfield.

We analyzed real-world mobility data based on cell tower traces, and illustrated how cell tower trace data can be used to identify patterns of life. We then used these results to predict future locations over a 24h period in 15 minute intervals with 80% accuracy [43].

7.3.3. Taler protocol improvements

Participants: Jeffrey Burdges, Florian Dold, Christian Grothoff, Marcello Stanisci.

We improved the Taler payment system protocol [13] to (1) reduce storage requirements for the exchange, which was the dominant cost, and (2) reduce security assumptions by avoiding the use of AES entirely.

We adapted the payment handshake to work even if JavaScript is disabled for the Web page, and adjusted the protocol to match discussions for future Web payment protocols from W3c. The protocol was extended with accounting functions to allow merchants to trace payments for their back office requirements. The user interface of the Taler wallet was streamlined, the wallet can finally get change, and the extension was made to work with Firefox. A public demonstrator was launched at https://demo.taler.net/.

7.4. Other research results: Information-Theoretical Quantification of Security Properties

Participants: Axel Legay, Fabrizio Biondi, Mounir Chadli, Thomas Given-Wilson.

Information theory provides a powerful quantitative approach to measuring security and privacy properties of systems. By measuring the *information leakage* of a system, security properties can be quantified, validated, or falsified. When security concerns are non-binary, information theoretic measures can quantify exactly how much information is leaked. The knowledge of such information is strategic in the developments of component-based systems.

The quantitative information-theoretical approach to security models the correlation between the secret information of the system and the output that the system produces. Such output can be observed by the attacker, and the attacker tries to infer the value of the secret information by combining this information with their prior knowledge of the system.

Armed with the produced output of the system, the attacker tries to infer information about the secret information that produced the output. The quantitative analysis we consider defines and computes how much information the attacker can expect to infer (typically measured in bits). This expected leakage of bits is the information leakage of the system.

The quantitative approach generalizes the qualitative approach and thus provides superior analysis. In particular, a system respects non-interference if and only if its leakage is equal to zero. In practice very few systems respect non-interference, and for those that don't it is imperative to be able to distinguish between the systems leaking very small amounts of secret information and systems leaking a significant amount of secret information, since only the latter are considered to pose a security vulnerability to the system.

Applied to shared-key cryptosystems, this approach allows precise reasoning about the information leakage of the secret key when the attacker knows the encoder function and information about the distribution of messages. In such scenarios, this work has generalised perfect secrecy, and so provides a more useful measure for unconditional cryptosystems (results that are safe against future advances in computing capabilities and theoretical breakthroughs in unsolved problems).

This work also explored scenarios where the attacker has less information about the cryptosystem; such as not knowing the encoder function, or not knowing the message distribution. Results here formalised that the attacker can never improve their attacks by having bad prior information, thus ensuring misinformation is always useful. Also, results show that the choice of encoder function may strengthen the cryptosystem against being learned by the attacker through observation. In particular, we showed that a well designed encoder function (represented as a matrix) has an infinitude of freedom for the attacker. Thus, the attacker cannot accurately learn all the secret information merely by observation.

There are several different scenarios where the attacker is trying to learn the secret information about the system. Here this is explored by considering what the secret information is, or equivalently, what prior knowledge the attacker has about the system.

Our new results in information leakage computation include implementing a hybrid precise-statistical computation algorithm for our QUAIL tool. The new algorithm bridges the gap between statistical and formal techniques by using static program analysis to extract structural information about the program to be analyze and decide whether each part of it would be analyzed more efficiently with precise or statistical analysis. Then each part is analyzed with the most appropriate technique, and all analyses are combined into a final result. This new hybrid method outperforms precise and statistical analysis in computation time and precision, and is a clear example of the advantages of combining precise and statistical techniques.We refer to the tools section for more details.

Additionally, we have considered how the scheduling of privileged and unprivileged processes on a shared memory could allow an unprivileged process to access confidential information temporarily stored in the memory by a privileged process. This is for instance the case in cache attacks. We have developed a general model of information leakage for scheduled systems. Our model considers a finer granularity than previous attempts on the subject, allowing us to schedule processes with small leakage, and schedule sets of processes that were considered unschedulable with no leakage by the state of the art.

[1] Preserving the privacy of private communication is a fundamental concern of computing addressed by encryption. Information-theoretic reasoning models unconditional security where the strength of

the results does not depend on computational hardness or unproven results. Usually the information leaked about the message by the ciphertext is used to measure the privacy of a communication , with perfect secrecy when the leakage is 0. However this is hard to achieve in practice. An alternative measure is the equivocation, intuitively the average number of message/key pairs that could have produced a given cipher-text. We show a theoretical bound on equivocation called max-equivocation and show that this generalizes perfect secrecy when achievable, and provides an alternative measure when perfect secrecy is not achievable. We derive bounds for max-equivocation for symmetric encoder functions and show that max-equivocation is achievable when the entropy of the ciphertext is minimized. We show that max-equivocation easily accounts for key re-use scenarios, and that large keys relative to the message perform very poorly under equivocation. We study encoders under this new perspective, deriving results on their achievable maximal equivocation and showing that some popular approaches such as Latin squares are not optimal. We show how unicity attacks can be naturally modeled, and how relaxing encoder symmetry improves equivocation. We present some algorithms for generating encryption functions that are practical and achieve 90 to 95% of the theoretical best, improving with larger message spaces.

[24] Analysis of a probabilistic system often requires to learn the joint probability distribution of its random variables. The computation of the exact distribution is usually an exhaustive precise analysis on all executions of the system. To avoid the high computational cost of such an exhaustive search, statistical analysis has been studied to efficiently obtain approximate estimates by analyzing only a small but representative subset of the system's behavior. In this paper we propose a hybrid statistical estimation method that combines precise and statistical analyses to estimate mutual information and its confidence interval. We show how to combine the analyses on different components of the system with different precision to obtain an estimate for the whole system. The new method performs weighted statistical analysis with different sample sizes over different components and dynamically finds their optimal sample sizes. Moreover it can reduce sample sizes by using prior knowledge about systems and a new abstraction-then-sampling technique based on qualitative analysis. We show the new method outperforms the state of the art in quantifying information leakage.

[12] The protection of users' data conforming to best practice and legislation is one of the main challenges in computer science. Very often, large-scale data leaks remind us that the state of the art in data privacy and anonymity is severely lacking. The complexity of modern systems make it impossible for software architect to create secure software that correctly implements privacy policies without the help of automated tools. The academic community needs to invest more effort in the formal modeling of security and anonymity properties, providing a deeper understanding of the underlying concepts and challenges and allowing the creation of automated tools to help software architects and developers. This research track provides numerous contributions to the formal modeling of security and anonymity properties and the creation of tools to verify them on large-scale software projects.

[62] High-security processes typically have to load confidential information, such as encryption keys or private data, into memory as part of their operation. In systems with a single shared memory, when high-security processes are switched out due to context switching, confidential information may remain in memory and be accessible to low-security processes. This paper considers this problem from the perspective of scheduling. A formal model supporting preemption is introduced that allows: reasoning about leakage between high-and low-security processes, and producing information-leakage aware schedulers. Several information-leakage aware heuristics are presented in the form of compositional pre-and postprocessors as part of a more general scheduling approach. The effectiveness of such heuristics is evaluated experimentally, showing them to achieve significantly better schedulability than the state of the art.

TASC Project-Team

7. New Results

7.1. Discrete Convexity

We introduce a propagator for pairs of Sum constraints, where the expressions in the sums respect a form of convexity. This propagator is parametric and can be instantiated for various concrete pairs, including Deviation, Spread, and the conjunction of Linear($\leq$) and Among. We show that despite its generality, our propagator (see Figure 1) is competitive in theory and practice with state-of-the-art propagators. (see AI journal paper).

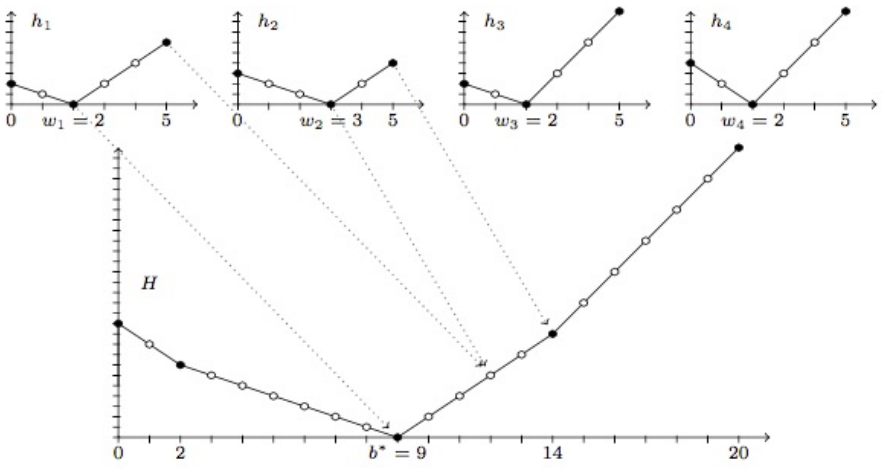


Figure 1. Illustration of the filtering wrt h function

7.2. Transducers

We describe a large family of constraints for structural time series by means of function composition. These constraints are on aggregations of features of patterns that occur in a time series, such as the number of its peaks, or the range of its steepest ascent. The patterns and features are usually linked to physical properties of the time series generator, which are important to capture in a constraint model of the system, i.e. a conjunction of constraints that produces similar time series. We formalise the patterns using finite transducers, whose output alphabet corresponds to semantic values that precisely describe the steps for identifying the occurrences of a pattern. Based on that description, we automatically synthesise automata with accumulators, as well as constraint checkers. The description scheme not only unifies the structure of the existing 30 time-series constraints in the Global Constraint Catalogue, but also leads to over 600 new constraints, with more than 100,000 lines of synthesised code. (see Constraint journal paper)

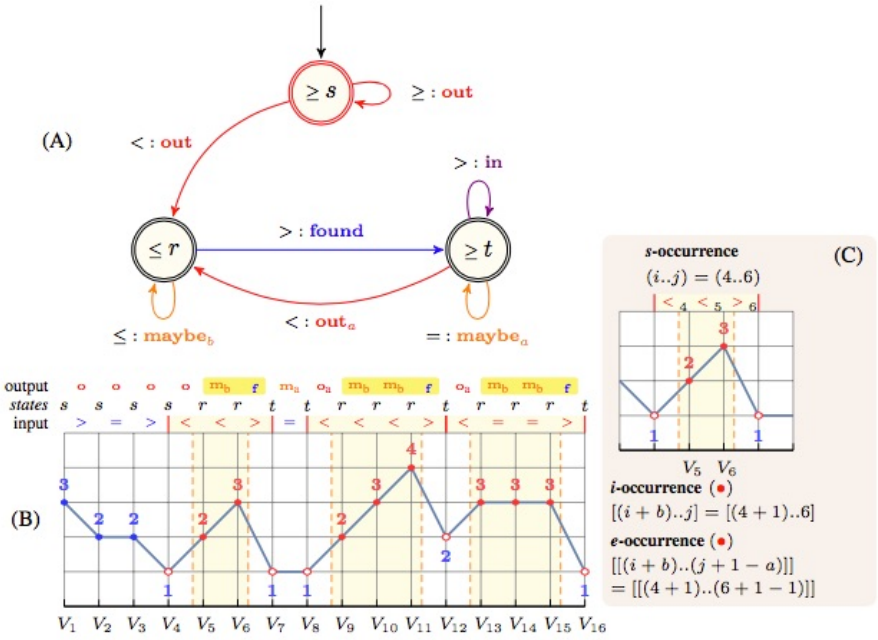


Figure 2. Transducer for the peak pattern and its execution on a sequence

7.3. Compositional Glue Matrix and Bound for Time-Series Constraints

Integer time series are often subject to constraints on the aggregation of the integer features of all occurrences of some pattern within the series. For example, the number of inflexions may be constrained, or the sum of the peak maxima, or the minimum of the peak widths. It is currently unknown how to maintain domain consistency efficiently on such constraints. We propose parametric ways of systematically deriving glue constraints (see Figures 3 and 4 for the parametric and concrete glue constraints), which are a particular kind of implied constraints, as well as aggregation bounds (see Figure 5) that can be added to the decomposition of time-series constraints. We evaluate the beneficial propagation impact of the derived implied constraints and bounds, both alone and together. (see CP conference paper)

	s	r	t
s	$\phi_g(\overrightarrow{C}, \overleftarrow{C})$	$\phi_g(\overrightarrow{C}, \overleftarrow{C})$	$\phi_g(\overrightarrow{C}, \overleftarrow{C})$
r	$\phi_g(\overrightarrow{C}, \overleftarrow{C})$	$\phi_f(\overrightarrow{D}, \overleftarrow{D}, \delta_f^i)$	$\phi_f(\overleftarrow{C}, \overrightarrow{D}, \overleftarrow{D}, \delta_f^i)$
t	$\phi_g(\overrightarrow{C}, \overleftarrow{C})$	$\phi_f(\overrightarrow{C}, \overrightarrow{D}, \overleftarrow{D}, \delta_f^i)$	$\phi_g(\overrightarrow{C}, \overleftarrow{C})$

Figure 3. Parametrised glue matrix for the peak pattern expressed in term of parametrised functions depending on the states pairs between the prefix and the suffix of a sequence

7.4. Reformulation of time-series constraint in MIP

	s	r	t
s	$\overrightarrow{C}+\overleftarrow{C}$	$\overrightarrow{C}+\overleftarrow{C}$	$\overrightarrow{C}+\overleftarrow{C}$
r	$\overrightarrow{C}+\overleftarrow{C}$	1	1
t	$\overrightarrow{C}+\overleftarrow{C}$	1	$\overrightarrow{C}+\overleftarrow{C}$

Figure 4. Concrete glue matrix for the number peak constraint expressed in term of concrete functions depending on the states pairs between the prefix and the suffix of a sequence

$\mathtt{sv} \leq 3 \vee \mathtt{rv} \leq 1 \Rightarrow \mathtt{VALUE} = 0$
$\mathtt{VALUE} \geq 0$
$\mathtt{rv} = 2 \Rightarrow \mathtt{VALUE} \leq \lfloor \mathtt{sv}/4 \rfloor$①
$\mathtt{rv} \geq 3 \Rightarrow \mathtt{VALUE} \leq \mathtt{max}(0, \lfloor (\mathtt{sv}-1)/3 \rfloor)$②
required(VARIABLES, var)
where
$\mathtt{sv} = |\mathtt{VARIABLES}|$
$\mathtt{rv} =$range(VARIABLES.var)

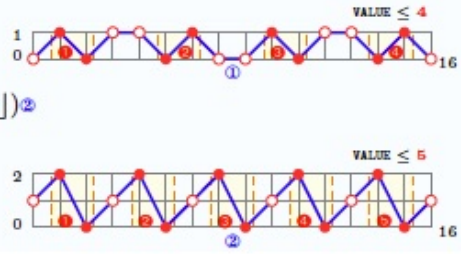


Figure 5. Upper bound on the number of zigzag depending on the domain range being equal to 2 or greater than or equal to 3

A checker for a constraint on a variable sequence can often be compactly specified by an automaton, possibly with accumulators, that consumes the sequence of values taken by the variables; such an automaton can also be used to decompose its specified constraint into a conjunction of logical constraints. The inference achieved by this decomposition in a CP solver can be boosted by automatically generated implied constraints on the accumulators, provided the latter are updated in the automaton transitions by linear expressions. Automata with non-linear accumulator updates can be automatically synthesised for a large family of time-series constraints. In this paper, we describe and evaluate extensions to those techniques. First, we improve the automaton synthesis to generate automata with fewer accumulators. Second, we decompose a constraint specified by an automaton with accumulators into a conjunction of linear inequalities, for use by a MIP solver. Third, we generalise the implied constraint generation to cover the entire family of time-series constraints. The newly synthesised automata for time-series constraints outperform the old ones, for both the CP and MIP decompositions, and the generated implied constraints boost the inference, again for both the CP and MIP decompositions. We evaluate CP and MIP solvers on a prototypical application modelled using time-series constraints. (see CPAIOR conference paper)

7.5. Scheduling Constraint for Video Summarisation

Given a sequence of tasks T subject to precedence constraints between adjacent tasks, and given a set of fixed intervals I, the TaskIntersection (T,I,o,inter) constraint restricts the overall intersection of the tasks of T with the fixed intervals of I to be greater than or equal or less than or equal to a given limit inter. We provide a bound(Z)-consistent cost filtering algorithm wrt the starts and the ends of the tasks for the TaskIntersection constraint and evaluate the constraint on the video summarisation problem. (see CPAIOR conference paper)

7.6. A Model Seeker for Learning Constraints Models from Positive Samples

We describe a system which generates finite domain constraint models from positive example solutions (e.g. see Figure 6 giving a season schedule of the Bundesliga), for highly structured problems. The system is based on the global constraint catalog, providing the library of constraints that can be used in modeling, and the Constraint Seeker tool, which finds a ranked list of matching constraints given one or more sample call patterns (e.g. see Figure 7 giving the model learned for the input data of Figure 6). We have tested the modeler with 230 examples, ranging from 4 to 6,500 variables, using between 1 and 7,000 samples. These examples come from a variety of domains, including puzzles, sports-scheduling, packing and placement, and design theory. When comparing against manually specified canonical models for the examples, we achieve a hit rate of 50 percent, processing the complete benchmark set in less than one hour on a laptop. Surprisingly, in many cases the system finds usable candidate lists even when working with a single, positive example. (see Book chapter of Data Mining and Constraint Programming)

7.7. Global Constraint Catalog Volume II: Time-Series Constraints

First this report presents a restricted set of 22 finite transducers used to synthesise structural time-series constraints described by means of a multi-layered function composition scheme. Second it provides the corresponding synthesised catalogue of structural time-series constraints where each of the 626 constraints is explicitly described in terms of automata with accumulators, see Figure 8 for the syntesised automaton of the sum surf peak constraint. (arXiv 1609.08925)

7.8. Probabilistic Model for Binary CSP

This work introduces a probabilistic-based model for binary CSP that provides a fine grained analysis of its internal structure. Assuming that a domain modification could occur in the CSP, it shows how to express, in a predictive way, the probability that a domain value becomes inconsistent, then it express the expectation of the number of arc-inconsistent values in each domain of the constraint network. Thus, it express the expectation of the number of arc-inconsistent values for the whole constraint network. Next, it provides bounds for each of these three probabilistic indicators. Finally, a polytime algorithm, which propagates the probabilistic information, is presented. (see arXiv 1606.03894 or [19])

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
8	1	14	11	4	7	2	15	12	13	6	9	10	3	18	5	16	17
3	14	17	2	13	6	5	12	9	16	11	18	1	4	15	8	7	10
...																	
18	17	2	1	4	3	6	5	10	9	16	15	14	13	12	11	8	7
13	12	11	14	17	16	15	2	9	6	1	8	7	4	5	18	3	10
...																	

Figure 6. Input data corresponding to a flat sample (a sequence of integer values) giving a one year season schedule of the Bundesliga

-	Sequence Generator	Projection	Constraint Conjunction
1	scheme(612,34,18,34,1)	id	alldifferent*18
2	scheme(612,34,18,2,2)	id	alldifferent*153
3	scheme(612,34,18,1,18)	id	alldifferent*34
4	scheme(612,34,18,1,18)	absolute_value	symmetric_alldifferent([1..18])*34
5	scheme(612,34,18,17,1)	absolute_value	alldifferent*36
6	repart(612,34,18,34,9)	id	sum_ctr(0)*306
7	repart(612,34,18,34,9)	id	twin*1
8	repart(612,34,18,34,9)	id	elements([i,-i])*1
9	first(9,[1,3,5,7,9,11,13,15,17])	id	strictly_increasing*1
10	vector(612)	id	global_cardinality([-18.. -1-17,0-0,1..18-17])*1
11	repart(612,34,18,34,9)	id	sum_powers5_ctr(0)*306
12	repart(612,34,18,34,9)	id	sum_cubes_ctr(0)*306
13	repart(612,34,18,34,3)	sign	global_cardinality([-1-3,0-0,1-3])*102
14	scheme(612,34,18,34,1)	sign	global_cardinality([-1-17,0-0,1-17])*18
15	repart(612,34,18,17,9)	sign	global_cardinality([-1-2,0-0,1-2])*153
16	repart(612,34,18,2,9)	sign	global_cardinality([-1-17,0-0,1-17])*18
17	scheme(612,34,18,1,18)	sign	global_cardinality([-1-9,0-0,1-9])*34
18	repart(612,34,18,34,9)	sign	sum_ctr(0)*306
19	repart(612,34,18,34,9)	sign	twin*1
20	repart(612,34,18,34,9)	absolute_value	twin*1
21	repart(612,34,18,34,9)	sign	elements([i,-i])*1
22	scheme(612,34,18,34,1)	sign	among_seq(3,[-1])*18
23	repart(612,34,18,34,9)	absolute_value	elements([i,i])*1
24	first(9,[1,3,5,7,9,11,13,15,17])	absolute_value	strictly_increasing*1
25	first(6,[1,4,7,10,13,16])	absolute_value	strictly_increasing*1
26	scheme(612,34,18,34,1)	absolute_value	nvalue(17)*18

Figure 7. Model, i.e. conjunction of global constraints, learned from the single flat sample

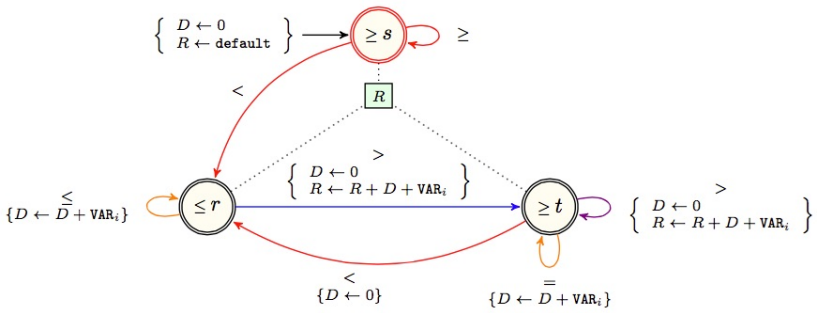


Figure 8. Synthesised automaton with accumulator of the sum surf peak constraint obtained from the transducer of the peak pattern

7.9. Estimating parallel runtimes for randomized algorithms in constraint solving

We present a detailed analysis of the scalability and parallelisation of Local Search algorithms for constraint-based and SAT (Boolean satisfiability) solvers. We propose a framework to estimate the parallel performance of a given algorithm by analyzing the runtime behavior of its sequential version. Indeed, by approximating the runtime distribution of the sequential process with statistical methods, the runtime behavior of the parallel process can be predicted by a model based on order statistics. We apply this approach to study the parallel performance of a constraint-based Local Search solver (Adaptive Search), two SAT Local Search solvers (namely Sparrow and CCASAT), and a propagation-based constraint solver (Gecode, with a random labeling heuristic). We compare the performance predicted by our model to actual parallel implementations of those methods using up to 384 processes. We show that the model is accurate and predicts performance close to the empirical data. Moreover, as we study different types of problems, we observe that the experimented solvers exhibit different behaviors and that their runtime distributions can be approximated by two types of distributions: exponential (shifted and non-shifted) and lognormal. Our results show that the proposed framework estimates the runtime of the parallel algorithm with an average discrepancy of 21 percent w.r.t. the empirical data across all the experiments with the maximum allowed number of processors for each technique. (see Journal of Heuristics)

7.10. ghost: A Combinatorial Optimization Framework for Real-Time Problems

We presents GHOST, a combinatorial optimization framework that a real-time strategy (RTS) AI developer can use to model and solve any problem encoded as a constraint satisfaction/optimization problem (CSP/COP). We show a way to model three different problems as a CSP/COP, using instances from the RTS game StarCraft as test beds. Each problem belongs to a specific level of abstraction (the target selection as reactive control problem, the wall-in as a tactics problem, and the build order planning as a strategy problem). In our experiments, GHOST shows good results computed within some tens of milliseconds. We also show that GHOST outperforms state-of-the-art constraint solvers, matching them on the resources allocation problem, a common combinatorial optimization problem. (see IEEE Transactions on Computational Intelligence and AI in games journal)

7.11. TorchCraft: a Library for Machine Learning Research on Real-Time Strategy Games

We present TorchCraft, a library that enables deep learning research on Real-Time Strategy (RTS) games such as StarCraft: Brood War, by making it easier to control these games from a machine learning framework, here Torch. This white paper argues for using RTS games as a benchmark for AI research, and describes the design and components of TorchCraft. (see arXiv 1611.00625)

7.12. POSL: A Parallel-Oriented metaheuristic-based Solver Language

For a couple of years, all processors in modern machines are multi-core. Massively parallel architectures, so far reserved for super-computers, become now available to a broad public through hardware like the Xeon Phi or GPU cards. This architecture strategy has been commonly adopted by processor manufacturers, al- lowing them to stick with Moore's law. However, this new architecture implies new ways to design and implement algorithms to exploit its full potential. This is in particular true for constraint-based solvers dealing with combinatorial optimiza- tion problems. Here we propose a Parallel-Oriented Solver Language (POSL, pronounced "puzzle"), a new framework to build interconnected meta-heuristic based solvers working in parallel. The novelty of this approach lies in looking at solver as a set of components with specific goals, written in a parallel-oriented language based on operators. A major feature in POSL is the possibility to share not only information, but also behaviors, allowing solver modifications during runtime. Our framework has been designed to easily build constraint-based solvers and reduce the developing effort in the context of parallel architecture. POSL's main advan- tage is to allow solver designers to quickly test different heuristics and parallel com- munication strategies to solve combinatorial optimization problems, usually time- consuming and very complex technically, requiring a lot of engineering.

7.13. Towards Automated Strategies in Satisfiability Modulo Theory

SMT solvers include many heuristic components in order to ease the theorem proving process for different logics and problems. Handling these heuristics is a non-trivial task requiring specific knowledge of many theories that even a SMT solver developer may be unaware of. This is the first barrier to break in order to allow end-users to control heuristics aspects of any SMT solver and to successfully build a strategy for their own purposes. We present a first attempt for generating an automatic selection of heuristics in order to improve SMT solver efficiency and to allow end-users to take better advantage of solvers when unknown problems are faced. Evidence of improvement is shown and the basis for future works with evolutionary and/or learning-based algorithms are raised (see Genetic Programming conference paper).

7.14. Using CP for the Urban Transit Crew Rescheduling Problem

Scheduling urban and trans-urban transportation is an important issue for industrial societies. The Urban Transit Crew Scheduling Problem is one of the most important optimization problem related to this issue. It mainly relies on scheduling bus drivers workday respecting both collective agreements (see Figure 9 for an example of regulation rule) and the bus schedule needs. If this problem has been intensively studied from a tactical point of view, its operational aspect has been neglected while the problem becomes more and more complex and more and more prone to disruptions. In this way, this paper presents how the constraint programming technologies are able to recover the tactical plans at the operational level in order to efficiently help in answering regulation needs after disruptions (see CP conference paper).

7.15. Traveling salesman and the tree: the importance of search in CP

The traveling salesman problem (TSP) is a challenging optimization problem for CP and OR that has many industrial applications. Its generalization to the degree constrained minimum spanning tree problem (DCMSTP) is being intensively studied by the OR community. In particular, classical solution techniques for the TSP are being progressively generalized to the DCMSTP. Recent work on cost-based relaxations has improved CP models for the TSP. However, CP search strategies have not yet been widely investigated for these problems. The contributions of this paper are twofold. We first introduce a natural generalization of the weighted cycle constraint (WCC) to the DCMSTP. We then provide an extensive empirical evaluation of various search strategies. In particular, we show that significant improvement can be achieved via our graph interpretation of the state-of-the-art Last Conflict heuristic. (see Constraints journal)

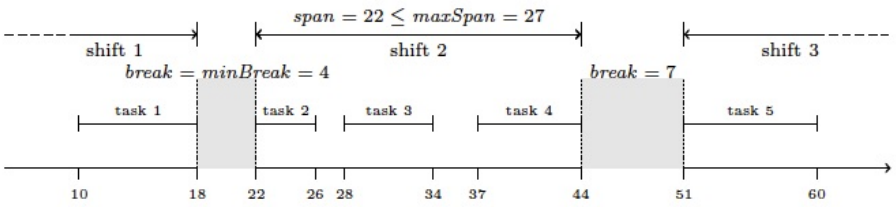


Figure 9. Illustration of typical regulation rule in the Labcom project; Shift 2 of an employee is composed of tasks 2, 3, and 4. It is between shifts 1 and 3 of the same employee. A break of duration 4 is scheduled between task 1 of shift 1 and task 2 of shift 2, because the gap between these tasks is at least a break of 4. Similarly, a break of duration 7 is scheduled after task 4 of shift 2 and task 5 of shift 3. No other breaks can be scheduled between the tasks of shift 2 because of the minimum break duration. The span of shift 2 is 22 and does not exceed 27: it is composed of tasks 2, 3, and 4, as well as of the two gaps between these tasks.

7.16. Event Selection Rules to Compute Explanations

Explanations have been introduced in the previous century. Their interest in reducing the search space is no longer questioned. Yet, their efficient implementation into CSP solver is still a challenge. In this paper, we introduce ESeR, an Event Selection Rules algorithm that filters events generated during propagation. This dynamic selection enables an efficient computation of explanations for intelligent backtracking algorithms. We show the effectiveness of our approach on the instances of the last three MiniZinc challenges. (see arXiv 1608.08015 or [20])

7.17. Towards energy-proportional Clouds partially powered by renewable energy

With the emergence of the Future Internet and the dawning of new IT models such as cloud computing, the usage of data centers (DC), and consequently their power consumption, increase dramatically. Besides the ecological impact, the energy consumption is a predominant criterion for DC providers since it determines the daily cost of their infrastructure. As a consequence, power management becomes one of the main challenges for DC infrastructures and more generally for large-scale-distributed systems. In this paper, we present the EpoCloud prototype, from hardware to middleware layers. This prototype aims at optimizing the energy consumption of mono-site Cloud DCs connected to the regular electrical grid and to renewable-energy sources (see Journal of Computing).

TEA Project-Team

7. New Results

7.1. Toward a distribution of ADFG

Participants: Alexandre Honorat, Jean-Pierre Talpin, Thierry Gautier, Loïc Besnard.

The ADFG tool is being developed in the context of the ADT "Opama" in order to serve both as scheduler synthesis tool from AADL specifications and ordinary tasksets. ADFG has been partly rewritten in order to target more users : it is now freely available online and comes with a complete documentation. These improvements imply that ADFG does not anymore provide Safety Critical Java application generation; its main purpose of scheduler synthesis is now available from an Eclipse plugin, as a command-line interface, and also in Polychrony (as part of the AADL to Signal translation process). Moreover ADFG accepts and exports several file formats with related scheduling tools: SDF3, Yartiss and soon Cheddar.

The Eclipse interface has changed significantly with a dialog window and a console to present the results (as shown in the figure 4). Also the graphical data-flow graph editor is still present but has been simplified. An other big change (not seen by the end-user) is the internal use of the free LpSolve linear programming software instead of CPLEX. Finally, it will soon be possible to use this software not only as a scheduling synthesizer but also as a scheduling checker (with timing properties given by the user).

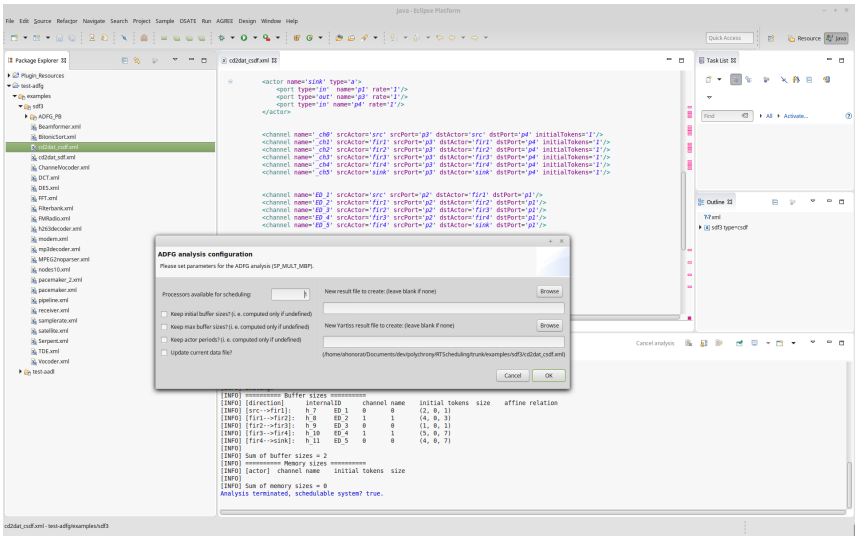


Figure 4. ADFG under Eclipse

7.2. Modular verification of cyber-physical systems using contract theory

Participants: Jean-Pierre Talpin, Benoit Boyer, David Mentre, Simon Lunel.

The primary goal of our project, in collaboration with Mitsubishi Electronics Research Centre Europe (MERCE), is to ensure correctness-by-design in realistic cyber-physical systems, i.e., systems that mix software and hardware in a physical environment, e.g., Mitsubishi factory automation lines or water-plant factory. To achieve that, we develop a verification methodology based on contract reasoning.

We have first performed a state of the art of the research and the work of A. Platzer with the Differential Dynamic Logic ($d\mathcal{L}$) retained our attention [0]. This a formalism built on the Dynamic Logic of V. Pratt augmented with the possibility of expressing Ordinary Differential Equations (ODEs). ODEs are the usual way to model physical behaviors in physics and $d\mathcal{L}$ permits to accurately model cyber-physical systems. But this logic can also express properties on real arithmetic and there is proof system associated, under the form of a sequent calculus, which let us a mean to prove specifications. To finish, it is very natural to use contract to specify systems since it was the primary goal of the work of V. Pratt. To conclude, $d\mathcal{L}$ is particularly fit to our purpose.

We have some preliminary results about a design-by-composition methodology: we have defined a syntactic composition operator in $d\mathcal{L}$, which enjoys associativity and commutativity. We have then characterized the conditions under which we can derive automatically a proof of the contract of our composition. To exemplified our ideas, we are currently studying a simplified water-tank system, which will serve as a basis for more realistic case studies. We plan to provide refinement and abstraction mechanisms to ultimately allow a mix between vertical and horizontal design.

7.3. Runtime verification and trace analysis

Participants: Vania Joloboff, Daian Yue, Frédéric Mallet.

When engineers design a new cyber physical system, there are well known requirements that can be translated as system properties that must be verified. These properties can be expressed in some formalism and when the model has been designed, the properties can be checked at the model level, using model checking techniques or other model verification techniques.

This requires that the properties are well specified at the time the virtual prototype is assembled. However it is also the case that many intrinsic properties are actually unforeseen when the virtual prototype is assembled, for example that some hardware buffer overflow should not remain unnoticed by the software. In most cases, during system design the simulation fails: the engineers then must investigate the cause of the failure.

A widely used technique for that consists in storing all of the trace data of simulation sessions into trace files, which are analyzed later with specialized trace analyzer tools. Such trace files have become huge, possibly hundred of Gigabytes as all data are stored into the trace files, and have become intractable by human manual handling.

In order to better identify the reason for such failures and capture the missing properties that the system should verify we have started to work on a new run time verification approach based on trace analysis. Approaches like PSL requires that the properties to verify are known before hand. Our approach is attempting for the engineers to experiment various property verification of failing simulations without re-building the virtual prototype. We have established a technique that makes it possible to investigate properties either statically working from a trace file or dynamically by introducing a dynamic verification component into the virtual prototype, or actually the real system.

The Trace Runtime Analysis Platform (TRAP) provides a model-based framework and implements the corresponding tool chain to support runtime analysis and verification of traces generated by virtual prototypes or cyber-physical systems. The main goal is to make it easy for engineers to define system properties that should be satisfied and verify them at system runtime (or from a recorded session). The property verification tools proposed do not require a detailed knowledge of the system implementation, do not require any modification or recompilation of the system to investigate different properties, and do not require the engineers to be familiar with temporal logic. TRAP proposes Domain Specific Languages (DSL's) integrated within the Eclipse Modeling Framework to express the properties. The DSL tool-chain uses the concept of Logical Clock defined by CCSL and takes advantage of CCSL clock algebra as the underlying formal support. The DSL's compilers eventually generate C++ code to verify the properties at run time, making usage of dynamically loaded code.

[0] *Differential Dynamic Logic for Hybrid Systems*, André Platzer, http://symbolaris.com/logic/dL.html

This year we have investigated and implemented this approach, using Eclipse EMF. The STML and TPSL compilers are implemented in Java and generate C++ code. Results of this work have been published at the FDL'16 conference referenced on IEEE Explore. [17]

7.4. Polychronous automata and formal validation of AADL models

Participants: Loïc Besnard, Thierry Gautier, Alexandre Honorat, Clément Guy, Jean-Pierre Talpin.

We have defined a model of *polychronous automata* based on clock relations [7]. A specificity of this model is that an automaton is submitted to clock constraints: these finite-state automata define transition systems to express explicit reactions together with properties, in the form of Boolean formulas over logical time, to constrain their behavior. This allows one to specify a wide range of control-related configurations, either reactive, or restrictive with respect to their control environment. A semantic model is defined for these polychronous automata, that relies on a Boolean algebra of clocks. Polychronous automata integrate smoothly with data-flow equations in the polychronous model of computation.

This polychronous MoC has been used previously as semantic model for systems described in the core AADL standard. The core AADL is extended with annexes, such as the Behavior Annex, which allows to specify more precisely architectural behaviors. The translation from AADL specifications into the polychronous model should take into account these behavior specifications, which are based on description of automata.

For that purpose, the AADL state transition systems are translated as Signal automata (a slight extension of the Signal language has been defined to support the model of polychronous automata). States are declared as Signal labels. Transitions are expressed using a call to a specific Signal process `Automaton\_Transition` which takes as parameters the labels of the source and destination states, and the condition expression corresponding to the AADL guard of the transition. The transition processes implicitly declare the equations that are required to compute the firing instants of the transitions. These processes, viewed as macros, are replaced during Signal compilation with a set of Signal equations handling current state and transition firing.

Once the AADL model of a system transformed into a Signal program, one can analyze the program using the Polychrony framework in order to check if timing, scheduling and logical requirements over the whole system are met.

We have implemented the translation and experimented it using a concrete case study, which is the AADL modeling of an Adaptive Cruise Control (ACC) system, a highly safety-critical system embedded in recent cars.

7.5. Formal Semantics of Behavior Specifications in the Architecture Analysis and Design Language Standard

Participants: Loïc Besnard, Thierry Gautier, Clément Guy, Jean-Pierre Talpin.

In system design, an architecture specification or model serves, among other purposes, as a repository to share knowledge about the system being designed. Such a repository enables automatic generation of analytical models for different aspects relevant to system design (timing, reliability, security, etc.). The Architecture Analysis and Design Language (AADL) is a standard proposed by SAE to express architecture specifications and share knowledge between the different stakeholders about the system being designed. To support unambiguous reasoning, formal verification, high-fidelity simulation of architecture specifications in a model-based AADL design work-flow, we have defined a formal semantics for the behavior specification of the AADL. Since it began being discussed in the AADL standard committee, our formal semantics evolved from a synchronous model of computation and communication to a semantic framework for time and concurrency in the standard: asynchronous, synchronous or timed, to serve as a reference for model checking, code generation or simulation tools uses with the standard [14]. These semantics are simple, relying on the structure of automata present in the standard already, yet provide tagged, trace semantics framework to establish formal relations between (synchronous, asynchronous, timed) usages or interpretations of behavior.

We define the model of computation and communication of a behavior specification by the synchronous, timed or asynchronous traces of automata with variables. These constrained automata are derived from *polychronous automata* defined within the polychronous model of computation and communication [7].

States of a behavior annex transition system can be either observable from the outside (`initial`, `final` or `complete` states), that is states in which the execution of the component is paused or stopped and its outputs are available; or non observable execution states, that is internal states. We thus define two kinds of steps in the transition system: *small steps*, that is non-observable steps from or to an internal state; and *big steps*, that is observable steps from a *complete* state to another, through a number of small steps). The semantics of the AADL considers the observable states of the automaton. The set of states S_A of automaton A (used to interpret the behavior annex) thus only contains states corresponding to these observable states and the set of transitions T_A big-step transitions from an observable state to another (by opposition with small-step transitions from or to an execution state). The action language of the behavior annex defines actions performed during transitions. Actions associated with transitions are action blocks that are built from basic actions and a minimal set of control structures (sequences, sets, conditionals and loops). Typically, a behavior action sequence is represented by concatenating the transition systems of its elements; a behavior action set is represented by composing the transition systems of its elements.

For our semantics, we considered a significant subset of the behavioral specification annex of the AADL. This annex allows one to attach a behavior specification to any components of a system modeled using the AADL, and can be then analyzed for different purposes which could be, for example, the verification of logical, timing or scheduling requirements.

7.6. Integration of Polychrony with QGen

Participants: Loïc Besnard, Thierry Gautier, Christophe Junke, Jean-Pierre Talpin.

The FUI project P gave birth to the GGen qualifiable model compiler, developed by Adacore. The tool accepts a discrete subset of Simulink expressed in a language called P and produces C or Ada code.

Our contribution was about providing a semantic bridge between Polychrony and QGen [15]. Our objective was to use Polychrony to compute fined-grained static scheduling of computations and communications for P models based on architectural properties. This work was twofold. First, we defined an alternative unambiguous static block scheduler for QGen, which can compute both partial and total orders based on user preferences. The purpose of this sequencer is to allow QGen to inter-operate with external sequencing tools while providing guarantees about the compatibility of external block execution orders with respect to both QGen's compilation scheme and user expectations. On the other hand, we developed a transformation function from the P language, more precisely, from the System Model subset of P, to the Signal meta-model, SSME. This work is based on a high-level API designed on top of SSME and can be used to transform a subset of Simulink to Signal. We validated our approach with the test suite used by QGen which is composed of over two-hundred small-sized Simulink models. We tested both block sequencing and model transformations. We ran the conversion tool and the set of models used by QGen for its regression tests and successfully converted medium to large models. The P language is capable of representing a useful subset of Simulink. That is why it is an interesting tool to help interpreting Simulink models and possibly architectural properties as executable Signal programs. The programs currently produced with our transformation tool can be compiled by Polychrony and reorganized as clusters of smaller processes.

7.7. Code generation for poly-endochronous processes

Participants: Loïc Besnard, Thierry Gautier, Jean-Pierre Talpin.

The synchronous modeling paradigm provides strong correctness guarantees for embedded system design while requiring minimal environmental assumptions. In most related frameworks, global execution correctness is achieved by ensuring the insensitivity of (logical) time in the program from (real) time in the environment. Tis property, called endochrony, can be statically checked, making it fast to ensure design correctness. Unfortunately, it is not preserved by composition, which makes it difficult to exploit with component-based design concepts in mind. It has been shown that compositionality can be achieved by weakening the objective of endochrony: a weakly endochronous system is a deterministic system that can perform independent computations and communications in any order as long as this does not alter its global state. Moreover, the non-blocking composition of weakly endochronous processes is isochronous, which means that the synchronous and asynchronous compositions of weakly endochronous processes accept the same behaviors. Unfortunately, testing weak endochrony needs state-space exploration, which is very costly in the general case. Then, a particular case of weak endochrony, called polyendochrony, was defined, which allows static checking thanks to the existing clock calculus. The clock hierarchy of a polyendochronous system may have several trees, with synchronization relations between clocks placed in different trees, but the clock expressions of the clock system must be such that there is no clock expression (especially, no root clock expression) defined by symmetric difference: root clocks cannot refer to absence. In other words, the clock system must be in disjunctive form [10].

We have now implemented code generation for poly-endochronous systems in Polychrony. This generation reuses techniques of distributed code generation, with rendez-vous management for synchronization constraints on clocks which are not placed in the same tree of clocks. The approach has been validated on several use cases running in parallel with time to time synchronization.

VISAGES Project-Team

7. New Results

7.1. Image Computing: Detection, Segmentation, Registration and Analysis

7.1.1. Quantitative analysis of T2/T2 relaxation time alteration*

Participants: Benoit Combès, Anne Kerbrat, Olivier Commowick, Christian Barillot.

T2 and T2* relaxometric data [0] becomes a standard tool for the quantitative assessment of brain tissues and of their changes along time or after the infusion of a contrast agent. Being able to detect significant changes of T2/T2* relaxation time is an important issue. Generally, such a task is performed by comparing the variability level in the regions of interest to the variability in the normal appearance white matter. However, in the case of T2 and T2* relaxometry, this solution is highly problematic. Indeed the level of noise in the normal appearance white matter is significanlty smaller than the level of noise in more intense region (e.g. MS lesions). Our aim is to provide a Bayesian analysis of T2/T2* relaxometry estimation and alteration. More specifically, we build posterior distributions for the relaxation time and the relaxation offset by elucidating the dedicated Jeffreys priors. Then the resulting posterior distributions can be evaluated using a Monte Carlo Markov Chain algorithm. Such an analysis has three main advantages over the classical estimation procedure. First it allows in a simple way to compute many estimators of the posterior including the mode, the mean, the variance and confidence intervals. Then, it allows to include prior information. Finally, becauce one can extract confidence interval from the posterior, testing properly whether the true relaxometry time is included within a certain range of value given a confidence level is simple. This work was published as a conference paper in MICCAI 2016 [22].

7.1.2. Block-Matching Distortion Correction of Echo-Planar Images with Opposite Phase Encoding Directions

Participants: Renaud Hédouin, Olivier Commowick, Élise Bannier, Christian Barillot.

By shortening the acquisition time of MRI, Echo Planar Imaging (EPI) enables the acquisition of a large number of images in a short time, compatible with clinical constraints as required for diffusion or functional MRI. However such images are subject to large, local distortions disrupting their correspondence with the underlying anatomy. The correction of those distortions is an open problem, especially in regions where large deformations occur. We have proposed a new block-matching registration method to perform EPI distortion correction based on the acquisition of two EPI with opposite phase encoding directions (PED). It relies on new transformations between blocks adapted to the EPI distortion model, and on an adapted optimization scheme to ensure an opposite symmetric transformation. We have produced qualitative and quantitative results of the block-matching correction using different metrics on a phantom dataset and on in-vivo data. We have shown the ability of the block-matching to robustly correct EPI distortion even in strongly affected areas. This work has been accepted for publication in IEEE Transactions in Medical Imaging 2017.

7.1.3. An a contrario approach for the detection of patient-specific brain perfusion abnormalities with arterial spin labelling

Participants: Pierre Maurel, Jean-Christophe Ferré, Christian Barillot.

[0] https://en.wikipedia.org/wiki/Relaxation_(NMR)

In this work, we introduce a new locally multivariate procedure to quantitatively extract voxel-wise patterns of abnormal perfusion in individual patients. This a contrario approach uses a multivariate metric from the computer vision community that is suitable to detect abnormalities even in the presence of closeby hypo- and hyper-perfusions. This method takes into account local information without applying Gaussian smoothing to the data. Furthermore, to improve on the standard a contrario approach, which assumes white noise, we introduce an updated a contrario approach that takes into account the spatial coherency of the noise in the probability estimation. Validation is undertaken on a dataset of 25 patients diagnosed with brain tumors and 61 healthy volunteers. We show how the a contrario approach outperforms the massively univariate General Linear Model usually employed for this type of analysis. This work as been published in Neuroimage [14].

7.1.4. Dictionary Learning for Pattern Classification in Medical Imaging: Why Does Size Matter?

Participants: Hrishikesh Deshpande, Pierre Maurel, Christian Barillot.

Sparse representation based dictionary learning (DL) technique has proved to be an effective tool for image classification. While standard DL methods are effective in data representation, several discriminative DL methods have been proposed for learning dictionaries better suited for classification. Majority of these methods, in pattern recognition applications, learn the dictionaries for each class and com- pare the error terms of sparse reconstruction for each dictionary. However this raises a question that is still an open problem in the sparsity community: What role does the size of each dictionary play in the classification process? In this work, we prove that this parameter is pivotal, especially in cases where there are variability differences between classes. We illustrate our assertion on standard and discriminative DL techniques in two applications: Lips detection in face images and the classification of multiple sclerosis lesions in multi-channel brain MR images.

7.2. Image processing on Diffusion Weighted Magnetic Resonance Imaging

7.2.1. Maximum Likelihood Estimators of Brain White Matter Microstructure

Participant: Olivier Commowick.

Diffusion MRI is a key in-vivo non invasive imaging capability that can probe the microstructure of the brain. However, its limited resolution requires complex voxelwise generative models of the diffusion. Diffusion Compartment (DC) models divide the voxel into smaller compartments in which diffusion is homogeneous. We developed a comprehensive framework for maximum likelihood estimation (MLE) of such models that jointly features ML estimators of (i) the baseline MR signal, (ii) the noise variance, (iii) compartment proportions, and (iv) diffusion-related parameters. ML estimators are key to providing reliable mapping of brain microstructure as they are asymptotically unbiased and of minimal variance. We compare our algorithm (which efficiently exploits analytical properties of MLE) to alternative implementations and a state-of-the-art strategy. Simulation results show that our approach offers the best reduction in computational burden while guaranteeing convergence of numerical estimators to the MLE. In-vivo results also reveal remarkably reliable microstructure mapping in areas as complex as the centrum semiovale. Our ML framework accommodates any DC model and is available freely for multi-tensor models as part of the ANIMA software. This work was published as a conference paper in MICCAI 2016 [24].

7.3. EEG and MR Imaging

7.3.1. Multi-Modal EEG and fMRI Source Localization using Sparse Constraints

Participants: Saman Noorzadeh, Pierre Maurel, Christian Barillot.

In this work a multi-modal approach is introduced to estimate the brain neuronal sources based on EEG and fMRI. These two imaging techniques can provide complementary information about the neuronal activities of the brain. Each of these data modalities are first modeled linearly based on the sources. The sources are then estimated with a high spatio-temporal resolution based on a symmetrical integrated approach of these models. For a better estimation, a sparse constraint is also applied to the method based on the physiological knowledge that we have about the brain function. The results which are validated on the real data, shows the reconstruction of neuronal sources with the high spatio-temporal resolution. This is a joint work with Remi Gribonval.

7.3.2. Unimodal versus bimodal EEG-fMRI neurofeedback of a motor imagery task

Participants: Lorraine Perronnet, Marsel Mano, Élise Bannier, Christian Barillot.

In the context of the HEMISFER project, we proposed a simultaneous EEG-fMRI experimental protocol in which 10 healthy participants performed a motor-imagery task in unimodal and bimodal neurofeedback conditions. With this protocol we were able to compare for the first time the effects of unimodal EEG-neurofeedback and fMRI-neurofeedback versus bimodal EEG-fMRI-neurofeedback by looking both at EEG and fMRI activations. We also introduced a new feedback metaphor for bimodal EEG-fMRI-neurofeedback that integrates both EEG and fMRI signal in a single bi-dimensional feedback (a ball moving in 2D). Such a feedback is intended to relieve the cognitive load of the subject by presenting the bimodal neurofeedback task as a single regulation task instead of two. Additionally, this integrated feedback metaphor gives flexibility on defining a bimodal neurofeedback target. Participants were able to regulate activity in their motor regions in all neurofeedback conditions. Moreover, motor activations as revealed by offline fMRI analysis were stronger during EEG-fMRI-neurofeedback than during EEG-neurofeedback. This result suggests that EEG-fMRI-neurofeedback could be more specific or more engaging than EEG-neurofeedback. Our results also suggest that during EEG-fMRI-neurofeedback, participants tended to regulate more the modality that was harder to control. Taken together our results shed light on the specific mechanisms of bimodal EEG-fMRI-neurofeedback and on its added-value as compared to unimodal EEG-neurofeedback and fMRI-neurofeedback.

This work was done in collaboration with the Inria Hybrid and Athena teams. Experiments were conducted at the Neurinfo MRI research facility from University of Rennes 1. This was presented during the poster session of the 2016 Organization for Human Brain Mapping (OHBM) conference.

7.3.3. Brain training with Neurofeedback

Participants: Lorraine Perronnet, Christian Barillot.

We published a book chapter called Brain training with Neurofeedback in the book “Brain Computer Interfaces 1: Methods and Perspectives” (published in French and English) [26]. The first section of the chapter defines the concept of neurofeedback and gives an overall view of the current status in this domain. The second section describes the design of a NF training program and the typical course of a NF session, as well as the learning mechanisms underlying NF. The third section retraces the history of NF, explaining the origin of its questionable reputation and providing a foothold for understanding the diversity of existing approaches. The fourth section discusses how the fields of NF and BCIs might potentially overlap in future with the development of "restorative" BCIs. Finally, the fifth and last section presents a few applications of NF and summarizes the state of research of some of its major clinical applications.

7.3.4. Design of an Experimental Platform for Hybrid EEG-fMRI Neurofeedback Studies

Participants: Marsel Mano, Élise Bannier, Lorraine Perronnet, Christian Barillot.

During a neurofeedback (NF) experiment one or more brain activity measuring technologies are used to estimate the changes of the acquired neural signals that reflect the changes of the subject’s brain activity in real-time. There exist a variety of NF research applications that use only one type of neural signals (i.e. uni-modal) like EEG or fMRI, but there are very few NF researches that use two or more neural signals (i.e. multi-modal). This is primarily because of the associated technical burdens.

We have developed, installed and successfully conducted used a hybrid EEG-fMRI platform for bi-modal NF experiments, as part of the project Hemisfer. Our system is based on the integration and the synchronization of an MR-compatible EEG and fMRI acquisition subsystems. The EEG signals are acquired with a 64 channel MR-compatible solution from Brain Products and the MR imaging is performed on a 3T Verio Siemens scanner (VB17) with a 12-ch head coil. We have developed two real-time pipelines for EEG and fMRI that handle all the necessary signal processing, the Joint NF module that calculates and fuses the NF and a visualize module that displays the NF to the subject. The control and the synchronization of both subsystems with eachother and with the experimental protocol is handled by the NF Control.

Our platform showed very good real-time performance with various pre-processing, filtering, and NF estimation and visualization methods. The entire fMRI process from acquisition to NF takes always less than 200ms, well below the TR of regular EPI sequences (2s). The same process for EEG, with NF update cycles varying 2-5Hz, is done in virtually real time (50Hz). Various NF tasks scenarios for regulating the measured brain activity were tested with subjects. In particular, the platform was used for a NF study on 10 subjects with over 50 sessions using three NF protocols based on motor imagery related brain activity: a) fMRI-NF, b) EEG-NF and c) EEG and fMRI-NF; and two online brain activity regulating protocols without NF. Our hybrid EEG-fMRI NF platform has been a very reliable environment for the NF experiments in our project. Its modular architecture is easily adaptable to different experimental environments, and offers high efficiency for optimal real-time NF applications.

7.4. Applications in Neuroradiology and Neurological Disorders

7.4.1. Imaging biomarkers in Multiple Sclerosis: from image analysis to population imaging

Participants: Christian Barillot, Gilles Edan, Olivier Commowick.

The production of imaging data in medicine increases more rapidly than the capacity of computing models to extract information from it. The grand challenges of better understanding the brain, offering better care for neurological disorders, and stimulating new drug design will not be achieved without significant advances in computational neuroscience. The road to success is to develop a new, generic, computational methodology and to confront and validate this methodology on relevant diseases with adapted computational infrastructures. This new concept sustains the need to build new research paradigms to better understand the natural history of the pathology at the early phase; to better aggregate data that will provide the most complete representation of the pathology in order to better correlate imaging with other relevant features such as clinical, biological or genetic data. In this context, one of the major challenges of neuroimaging in clinical neurosciences is to detect quantitative signs of pathological evolution as early as possible to prevent disease progression, evaluate therapeutic protocols or even better understand and model the natural history of a given neurological pathology. Many diseases encompass brain alterations often not visible on conventional MRI sequences, especially in normal appearing brain tissues (NABT). MRI has often a low specificity for differentiating between possible pathological changes which could help in discriminating between the different pathological stages or grades. The objective of medical image analysis procedures is to define new quantitative neuroimaging biomarkers to track the evolution of the pathology at different levels. We have published a position paper in Medical Image Analysis [2] that illustrates this issue in one acute neuro-inflammatory pathology: Multiple Sclerosis (MS). It exhibits the current medical image analysis approaches and explains how this field of research will evolve in the next decade to integrate larger scale of information at the temporal, cellular, structural and morphological levels.

7.4.2. Multiple Sclerosis lesion segmentation using an automated multimodal Graph Cut

Participants: Jérémy Beaumont, Olivier Commowick, Christian Barillot.

In this work, we present an algorithm for Multiple Sclerosis (MS) lesion segmentation. Our method is fully automated and includes three main steps: 1. the computation of a rough total lesion load in order to optimize the parameter set of the following step; 2. the detection of lesions by graph cut initialized with a robust Expectation-Maximization (EM) algorithm; 3. the application of rules to remove false positives and to adjust the contour of the detected lesions. This work was part of the FLI 2016 MSSEG challenge data organized at MICCAI 2016 [25].

7.4.3. Automatic Multiple Sclerosis lesion segmentation from Intensity-Normalized multi-channel MRI

Participants: Jérémy Beaumont, Olivier Commowick, Christian Barillot.

In the context of the FLI MICCAI 2016 MSSEG challenge for lesion segmentation, we present a fully automated algorithm for Multiple Sclerosis (MS) lesion segmentation. Our method is composed of three main steps. First, the MS patient images are registered and intensity normalized. Then, the lesion segmentation is done using a voxel-wise comparison of multi-channel Magnetic Resonance Images (MRI) against a set of controls. Finally, the segmentation is refined by applying several lesion appearance rules. This work was part of the FLI 2016 MSSEG challenge data organized at MICCAI 2016 [21].

7.5. Management of Information in Neuroimaging

Participants: Michael Kain, Olivier Commowick, Élise Bannier, Inès Fakhfakh, Justine Guillaumont, Florent Leray, Yao Chi, Christian Barillot.

The major topic that is addressed in this period concern the sharing of data and processing tools in neuroimaging (through the “Programme d’Investissement d’Avenir” project such as OFSEP and FLI-IAM) which led to build a suitable architecture to share images and processing tools, started from the NeuroBase project (supported by the French Ministry of Research). Our overall goal within these projects is to set up a computer infrastructure to facilitate the sharing of neuroimaging data, as well as image processing tools, in a distributed and heterogeneous environment. These consortium gathered expertise coming from several complementary domains of expertise: image processing in neuroimaging, workflows and GRID computing, ontology development and ontology-based mediation. This enables a large variety of users to diffuse, exchange or reach neuroimaging information with appropriate access means, in order to be able to retrieve information almost as easily as if the data were stored locally by means of the "cloud computing" Storage as a Service (SaaS) concept As an example, the Shanoir environment has been succesfully deployed to the Neurinfo platform were it is routinely used to manage images of the research studies. It is also currently being deployed for two large projets: OFSEP ("Observatoire Français de la Sclérose en Plaques") where up to 30000 patients will be acquired on a ten years frame, and the Image Analysis and Mangment (IAM) node of the France Life Imaging national infrastructure (FLI-IAM). Our team fulfills multiple roles in this nation-wide FLI project. Christian Barillot is the chair of the IAM node, Olivier Commowick is participating in the working group workflow and image processing and Michael Kain is the technical manager of the node. Apart from the team members, software solutions like medInria and Shanoir are part of the final software platform.