RESEARCH CENTER
Grenoble - Rhône-Alpes

FIELD

Activity Report 2016

Section New Results

Edition: 2017-08-25

ARIC Project-Team

6. New Results

6.1. Floating-point arithmetic

6.1.1. Parallel floating-point expansions for extended-precision GPU computations

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 an FP expansion and it offers the simplicity of using directly available and highly optimized FP operations. In [30] we present new data-parallel algorithms for adding and multiplying FP expansions specially designed for extended precision computations on GPUs. 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.

6.1.2. Error analysis of the Cornea-Harrison-Tang method

Assuming floating-point arithmetic with a fused multiply-add operation and rounding to nearest, the Cornea-Harrison-Tang method aims to evaluate expressions of the form $ab + cd$ with high relative accuracy. In [12] we provide a rounding error analysis of this method, which unlike previous studies is not restricted to binary floating-point arithmetic but holds for any radix β. We show first that an asymptotically optimal bound on the relative error of this method is $\frac{2\beta u+2u^2}{\beta-2u^2} = 2u + \frac{2}{\beta}u^2 + O(u^3)$, where $u = \frac{1}{2}\beta^{1-p}$ is the unit roundoff in radix β and precision p. Then we show that the possibility of removing the $O(u^2)$ term from this bound is governed by the radix parity and the tie-breaking strategy used for rounding: if β is odd or rounding is *to nearest even*, then the simpler bound $2u$ is obtained, while if β is even and rounding is *to nearest away*, then there exist floating-point inputs a, b, c, d that lead to a relative error larger than $2u + \frac{2}{\beta}u^2 - 4u^3$. All these results hold provided underflows and overflows do not occur and under some mild assumptions on β and p satisfied by IEEE 754-2008 formats.

6.1.3. Sharp error bounds for complex floating-point inversion

In [14] we study the accuracy of the classic algorithm for inverting a complex number given by its real and imaginary parts as floating-point numbers. Our analyses are done in binary floating-point arithmetic, with an unbounded exponent range and in precision p; we also assume that the basic arithmetic operations ($+$, $-$, $\times$, $/$) are rounded to nearest, so that the unit roundoff is $u = 2^{-p}$. We bound the largest relative error in the computed inverse either in the componentwise or in the normwise sense. We prove the componentwise relative error bound $3u$ for the complex inversion algorithm (assuming $p \geq 4$), and we show that this bound is asymptotically optimal (as $p \to \infty$) when p is even, and sharp when using one of the basic IEEE 754 binary formats with an odd precision ($p = 53, 113$). This componentwise bound obviously leads to the same bound $3u$ for the normwise relative error. However, we prove that the smaller bound $2.707131u$ holds (assuming $p \geq 24$) for the normwise relative error, and we illustrate the sharpness of this bound for the basic IEEE 754 binary formats ($p = 24, 53, 113$) using numerical examples.

6.1.4. On relative errors of floating-point operations: optimal bounds and applications

Rounding error analyses of numerical algorithms are most often carried out via repeated applications of the so-called standard models of floating-point arithmetic. Given a round-to-nearest function fl and barring underflow and overflow, such models bound the relative errors $E_1(t) = |t - \mathrm{fl}(t)|/|t|$ and $E_(t) = |t - \mathrm{fl}(t)|/|\mathrm{fl}(t)|$ by the unit roundoff u. With S. M. Rump (Hamburg University of Technology), we investigate in [15] the possibility and the usefulness of refining these bounds, both in the case of an arbitrary real t and in the case where t is

the exact result of an arithmetic operation on some floating-point numbers. We show that $E_1(t)$ and $E_2(t)$ are optimally bounded by $u/(1+u)$ and u, respectively, when t is real or, under mild assumptions on the base and the precision, when $t = x \pm y$ or $t = xy$ with x, y two floating-point numbers. We prove that while this remains true for division in base $\beta > 2$, smaller, attainable bounds can be derived for both division in base $\beta = 2$ and square root. This set of optimal bounds is then applied to the rounding error analysis of various numerical algorithms: in all cases, we obtain significantly shorter proofs of the best-known error bounds for such algorithms, and/or improvements on these bounds themselves.

6.1.5. Computing floating-point logarithms with fixed-point operations

Elementary functions from the mathematical library input and output floating-point numbers. However, it is possible to implement them purely using integer/fixed-point arithmetic. This option was not attractive between 1985 and 2005, because mainstream processor hardware supported 64-bit floating-point, but only 32-bit integers. Besides, conversions between floating-point and integer were costly. This has changed in recent years, in particular with the generalization of native 64-bit integer support. The purpose of this article is therefore to reevaluate the relevance of computing floating-point functions in fixed-point. For this, several variants of the double-precision logarithm function are implemented and evaluated. Formulating the problem as a fixed-point one is easy after the range has been (classically) reduced. Then, 64-bit integers provide slightly more accuracy than 53-bit mantissa, which helps speed up the evaluation. Finally, multi-word arithmetic, critical for accurate implementations, is much faster in fixed-point, and natively supported by recent compilers. Novel techniques of argument reduction and rounding test are introduced in this context. Thanks to all this, a purely integer implementation of the correctly rounded double-precision logarithm outperforms the previous state of the art, with the worst-case execution time reduced by a factor 5. This work also introduces variants of the logarithm that input a floating-point number and output the result in fixed-point. These are shown to be both more accurate and more efficient than the traditional floating-point functions for some applications [35].

6.1.6. A library for symbolic floating-point arithmetic

To analyze a priori the accuracy of an algorithm in floating-point arithmetic, one usually derives a uniform error bound on the output, valid for most inputs and parametrized by the precision p. To show further that this bound is sharp, a common way is to build an input example for which the error committed by the algorithm comes close to that bound, or even attains it. Such inputs may be given as floating-point numbers in one of the IEEE standard formats (say, for $p = 53$) or, more generally, as expressions parametrized by p, that can be viewed as symbolic floating-point numbers. With such inputs, a sharpness result can thus be established for virtually all reasonable formats instead of just one of them. This, however, requires the ability to run the algorithm on those inputs and, in particular, to compute the correctly-rounded sum, product, or ratio of two symbolic floating-point numbers. We show in [61] how these basic arithmetic operations can be performed automatically. We introduce a way to model symbolic floating-point data, and present algorithms for round-to-nearest addition, multiplication, fused multiply-add, and division. An implementation as a Maple library is also described, and experiments using examples from the literature are provided to illustrate its interest in practice.

6.1.7. On the robustness of the 2Sum and Fast2Sum algorithms

The 2Sum and Fast2Sum algorithms are important building blocks in numerical computing. They are used (implicitly or explicitly) in many *compensated* algorithms (such as compensated summation or compensated polynomial evaluation). They are also used for manipulating floating-point *expansions*. We show in [56] that these algorithms are much more robust than it is usually believed: the returned result makes sense even when the rounding function is not round-to-nearest, and they are almost immune to overflow.

6.1.8. Tight and rigourous error bounds for basic building blocks of double-word arithmetic

In [63] we analyze several classical basic building blocks of double-word arithmetic (frequently called "double-double arithmetic" in the literature): the addition of a double-word number and a floating-point number, the addition of two double-word numbers, the multiplication of a double-word number by a floating-point number, the multiplication of two double-word numbers, the division of a double-word number by a

floating-point number, and the division of two double-word numbers. For multiplication and division we get better relative error bounds than the ones previously published. For addition of two double-word numbers, we show that the previously published bound was wrong, and we provide a relative error bound. We introduce new algorithms for division. We also give examples that illustrate the tightness of our bounds.

6.1.9. A new multiplication algorithm for extended precision using floating-point expansions

Some important computational problems must use a floating-point (FP) precision several times higher than the hardware-implemented available one. These computations critically rely on software libraries for high-precision FP arithmetic. The representation of a high-precision data type crucially influences the corresponding arithmetic algorithms. Recent work showed that algorithms for FP expansions, that is, a representation based on unevaluated sum of standard FP types, benefit from various high-performance support for native FP, such as low latency, high throughput, vectorization, threading, etc. Bailey's QD library and its corresponding Graphics Processing Unit (GPU) version, GQD, are such examples. Despite using native FP arithmetic as the key operations, QD and GQD algorithms are focused on double-double or quad-double representations and do not generalize efficiently or naturally to a flexible number of components in the FP expansion. In [45] we introduce a new multiplication algorithm for FP expansion with flexible precision, up to the order of tens of FP elements in mind. The main feature consists in the partial products being accumulated in a special designed data structure that has the regularity of a fixed-point representation while allowing the computation to be naturally carried out using native FP types. This allows us to easily avoid unnecessary computation and to present rigorous accuracy analysis transparently. The algorithm, its correctness and accuracy proofs and some performance comparisons with existing libraries are all contributions of this paper.

6.1.10. CAMPARY: Cuda Multiple Precision Arithmetic Library and Applications

Many scientific computing applications demand massive numerical computations on parallel architectures such as Graphics Processing Units (GPUs). Usually, either floating-point single or double precision arithmetic is used. Higher precision is generally not available in hardware, and software extended precision libraries are much slower and rarely supported on GPUs. We develop CAMPARY: a multiple-precision arithmetic library, using the CUDA programming language for the NVidia GPU platform. In our approach, the precision is extended by representing real numbers as the unevaluated sum of several standard machine precision floating-point numbers. We make use of error-free transforms algorithms, which are based only on native precision operations, but keep track of all rounding errors that occur when performing a sequence of additions and multiplications. This offers the simplicity of using hardware highly optimized floating-point operations, while also allowing for rigorously proven rounding error bounds. This also allows for easy implementation of an interval arithmetic. Currently, all basic multiple-precision arithmetic operations are supported. Our target applications are in chaotic dynamical systems or automatic control [34].

6.1.11. Arithmetic algorithms for extended precision using floating-point expansions

Many numerical problems require a higher computing precision than the one offered by standard floating-point (FP) formats. One common way of extending the precision is to represent numbers in a *multiple component* format. By using the so-called *floating-point expansions*, real numbers are represented as the unevaluated sum of standard machine precision FP numbers. This representation offers the simplicity of using directly available, hardware implemented and highly optimized, FP operations. It is used by multiple-precision libraries such as Bailey's QD or the analogue Graphics Processing Units (GPU) tuned version, GQD. In this article we briefly revisit algorithms for adding and multiplying FP expansions, then we introduce and prove new algorithms for normalizing, dividing and square rooting of FP expansions. The new method used for computing the reciprocal a^{-1} and the square root $\sqrt{a}$ of an FP expansion a is based on an adapted Newton-Raphson iteration where the intermediate calculations are done using "truncated" operations (additions, multiplications) involving FP expansions. We give here a thorough error analysis showing that it allows very accurate computations. More precisely, after q iterations, the computed FP expansion $x = x_0 + \cdots + x_{2^q-1}$ satisfies, for the reciprocal algorithm, the relative error bound: $|(x - a^{-1})/a^{-1}| \leq 2^{-2^q(p-3)-1}$ and, respectively, for the square root one: $|x - 1/\sqrt{a}| \leq 2^{-2^q(p-3)-1}/\sqrt{a}$, where $p > 2$ is the precision of the FP representation used ($p = 24$ for single precision and $p = 53$ for double precision) [16].

6.1.12. Comparison between binary and decimal floating-point numbers

We introduce an algorithm to compare a binary floating-point (FP) number and a decimal FP number, assuming the "binary encoding" of the decimal formats is used, and with a special emphasis on the basic interchange formats specified by the IEEE 754-2008 standard for FP arithmetic. It is a two-step algorithm: a first pass, based on the exponents only, quickly eliminates most cases, then, when the first pass does not suffice, a more accurate second pass is performed. We provide an implementation of several variants of our algorithm, and compare them [8].

6.1.13. Automatic source-to-source error compensation of floating-point programs: code synthesis to optimize accuracy and time

Numerical programs with IEEE 754 floating-point computations may suffer from inaccuracies, since finite precision arithmetic is an approximation of real arithmetic. Solutions that reduce the loss of accuracy are available, such as compensated algorithms or double-double precision floating-point arithmetic. With Ph. Langlois and M. Martel (LIRMM and Université de Perpignan), we show in [21] how to automatically improve the numerical quality of a numerical program with the smallest impact on its performance. We define and implement source code transformations in order to derive automatically compensated programs. We present several experimental results to compare the transformed programs and existing solutions. The transformed programs are as accurate and efficient as the implementations of compensated algorithms when the latter exist. Furthermore, we propose some transformation strategies allowing us to improve partially the accuracy of programs and to tune the impact on execution time. Trade-offs between accuracy and performance are assured by code synthesis. Experimental results show that user-defined trade-offs are achievable in a reasonable amount of time, with the help of the tools we present here.

6.1.14. Correctly rounded arbitrary-precision floating-point summation

We have designed a fast, low-level algorithm to compute the correctly rounded summation of several floating-point numbers in arbitrary precision in radix 2, each number (each input and the output) having its own precision. We have implemented it in GNU MPFR; it will be part of the next MPFR major release (GNU MPFR 4.0). In addition to a pen-and-paper proof, various kinds of tests are provided. Timings show that this new algorithm/implementation is globally much faster and takes less memory than the previous one (from MPFR 3.1.5): the worst-case time and memory complexity was exponential and it is now polynomial. Timings on pseudo-random inputs with various sets of parameters also show that this new implementation is even much faster than the (inaccurate) basic sum implementation in some cases. [36], [65]

6.2. Lattices: algorithms and cryptology

6.2.1. Zero-Knowledge Arguments for Lattice-Based Accumulators: Logarithmic-Size Ring Signatures and Group Signatures Without Trapdoors

An accumulator is a function that hashes a set of inputs into a short, constant-size string while preserving the ability to efficiently prove the inclusion of a specific input element in the hashed set. It has proved useful in the design of numerous privacy-enhancing protocols, in order to handle revocation or simply prove set membership. In the lattice setting, currently known instantiations of the primitive are based on Merkle trees, which do not interact well with zero-knowledge proofs. In order to efficiently prove the membership of some element in a zero-knowledge manner, the prover has to demonstrate knowledge of a hash chain without revealing it, which is not known to be efficiently possible under well-studied hardness assumptions. In [39], we provide an efficient method of proving such statements using involved extensions of Stern's protocol. Under the Small Integer Solution assumption, we provide zero-knowledge arguments showing possession of a hash chain. As an application, [39] describes new lattice-based group and ring signatures in the random oracle model. In particular, the paper obtains: (i) The first lattice-based ring signatures with logarithmic size in the cardinality of the ring; (ii) The first lattice-based group signature that does not require any GPV trapdoor and thus allows for a much more efficient choice of parameters.

6.2.2. A Lattice-Based Group Signature Scheme with Message-Dependent Opening

Group signatures are an important anonymity primitive allowing users to sign messages while hiding in a crowd. At the same time, signers remain accountable since an authority is capable of de-anonymizing signatures via a process called opening. In many situations, this authority is granted too much power as it can identify the author of any signature. Sakai et al. proposed a flavor of the primitive, called Group Signature with Message-Dependent Opening (GS-MDO), where opening operations are only possible when a separate authority (called "admitter") has revealed a trapdoor for the corresponding message. So far, all existing GS-MDO constructions rely on bilinear maps, partially because the message-dependent opening functionality inherently implies identity-based encryption. In [40], the team proposes the first GS-MDO candidate based on lattice assumptions. The construction combines the group signature of Ling, Nguyen and Wang (PKC'15) with two layers of identity-based encryption. These components are tied together using suitable zero-knowledge argument systems.

6.2.3. Practical "Signatures with Efficient Protocols" from Simple Assumptions

Digital signatures are perhaps the most important base for authentication and trust relationships in large scale systems. More specifically, various applications of signatures provide privacy and anonymity preserving mechanisms and protocols, and these, in turn, are becoming critical (due to the recently recognized need to protect individuals according to national rules and regulations). A specific type of signatures called "signatures with efficient protocols", as introduced by Camenisch and Lysyanskaya (CL), efficiently accommodates various basic protocols and extensions like zero-knowledge proofs, signing committed messages, or re-randomizability. These are, in fact, typical operations associated with signatures used in typical anonymity and privacy-preserving scenarios. To date there are no "signatures with efficient protocols" which are based on simple assumptions and truly practical. These two properties assure us a robust primitive: First, simple assumptions are needed for ensuring that this basic primitive is mathematically robust and does not require special ad hoc assumptions that are more risky, imply less efficiency, are more tuned to the protocol itself, and are perhaps less trusted. In the other dimension, efficiency is a must given the anonymity applications of the protocol, since without proper level of efficiency the future adoption of the primitives is always questionable (in spite of their need). In [41], the team presents a new CL-type signature scheme that is re-randomizable under a simple, well-studied, and by now standard, assumption (SXDH). The signature is efficient (built on the recent QA-NIZK constructions), and is, by design, suitable to work in extended contexts that typify privacy settings (like anonymous credentials, group signature, and offline e-cash). The paper demonstrates its power by presenting practical protocols based on it.

6.2.4. Functional Commitment Schemes: From Polynomial Commitments to Pairing-Based Accumulators from Simple Assumptions

In [42], the team formalizes a cryptographic primitive called functional commitment (FC) which can be viewed as a generalization of vector commitments (VCs), polynomial commitments and many other special kinds of commitment schemes. A non-interactive functional commitment allows committing to a message in such a way that the committer has the flexibility of only revealing a function $F(M)$ of the committed message during the opening phase. We provide constructions for the functionality of linear functions, where messages consist of a vectors of n elements over some domain D (e.g., $m = (m_1, ..., m_n) \in D_n$) and commitments can later be opened to a specific linear function of the vector coordinates. An opening for a function $F : D_n \rightarrow R$ thus generates a witness for the fact that $F(m)$ indeed evaluates to $y \in R$. One security requirement is called function binding and requires that no adversary be able to open a commitment to two different evaluations y, y' for the same function F. The paper [42] proposes a construction of functional commitment for linear functions based on constant-size assumptions in composite order groups endowed with a bilinear map. The construction has commitments and openings of constant size (i.e., independent of n or function description) and is perfectly hiding – the underlying message is information theoretically hidden. Our security proofs builds on the Déjà Q framework of Chase and Meiklejohn (Eurocrypt 2014) and its extension by Wee (TCC 2016) to encryption primitives, thus relying on constant-size subgroup decisional assumptions. The paper shows that the FC for linear functions are sufficiently powerful to solve four open problems. They, first, imply

polynomial commitments, and, then, give cryptographic accumulators (i.e., an algebraic hash function which makes it possible to efficiently prove that some input belongs to a hashed set). In particular, specializing the new FC construction leads to the first pairing-based polynomial commitments and accumulators for large universes known to achieve security under simple assumptions. We also substantially extend our pairing-based accumulator to handle subset queries which requires a non-trivial extension of the Déjà Q framework.

6.2.5. *Fully Secure Functional Encryption for Inner Products, from Standard Assumptions*

Functional encryption is a modern public-key paradigm where a master secret key can be used to derive sub-keys SKF associated with certain functions F in such a way that the decryption operation reveals $F(M)$, if M is the encrypted message, and nothing else. Recently, Abdalla *et al.* gave simple and efficient realizations of the primitive for the computation of linear functions on encrypted data: given an encryption of a vector y over some specified base ring, a secret key SK_x for the vector x allows computing $\langle x, y\rangle$. Their technique surprisingly allows for instantiations under standard assumptions, like the hardness of the Decision Diffie-Hellman (DDH) and Learning-with-Errors (LWE) problems. Their constructions, however, are only proved secure against selective adversaries, which have to declare the challenge messages M_0 and M_1 at the outset of the game. In [22], we provide constructions that provably achieve security against more realistic adaptive attacks (where the messages M_0 and M_1 may be chosen in the challenge phase, based on the previously collected information) for the same inner product functionality. The constructions of [22] are obtained from hash proof systems endowed with homomorphic properties over the key space. They are (almost) as efficient as those of Abdalla *et al.* and rely on the same hardness assumptions. In addition, the paper [22] obtains a solution based on Paillier's composite residuosity assumption, which was an open problem even in the case of selective adversaries. We also propose LWE-based schemes that allow evaluation of inner products modulo a prime p, as opposed to the schemes of Abdalla et al. that are restricted to evaluations of integer inner products of short integer vectors. The paper [22] finally proposes a solution based on Paillier's composite residuosity assumption that enables evaluation of inner products modulo an RSA integer $N = pq$. The paper [22] demonstrates that the functionality of inner products over a prime field is powerful and can be used to construct bounded collusion FE for all circuits.

6.2.6. *Signature Schemes with Efficient Protocols and Dynamic Group Signatures from Lattice Assumptions*

A recent line of works – initiated by Gordon, Katz and Vaikuntanathan (Asiacrypt 2010) – gave lattice-based realizations of privacy-preserving protocols allowing users to authenticate while remaining hidden in a crowd. Despite five years of efforts, known constructions remain limited to static populations of users, which cannot be dynamically updated. For example, none of the existing lattice-based group signatures seems easily extendable to the more realistic setting of dynamic groups. In [37], the team provides new tools enabling the design of anonymous authen-tication systems whereby new users can register and obtain credentials at any time. The first contribution of [37] is a signature scheme with efficient protocols, which allows users to obtain a signature on a committed value and subsequently prove knowledge of a signature on a committed message. This construction, which builds on the lattice-based signature of Böhl *et al.* (Eurocrypt'13), is well-suited to the design of anonymous credentials and dynamic group signatures. As a second technical contribution, [37] provides a simple, round-optimal joining mechanism for introducing new members in a group. This mechanism consists of zero-knowledge arguments allowing registered group members to prove knowledge of a secret short vector of which the corresponding public syndrome was certified by the group manager. This method provides similar advantages to those of structure-preserving signatures in the realm of bilinear groups. Namely, it allows group members to generate their public key on their own without having to prove knowledge of the underlying secret key. This results in a two-round join protocol supporting concurrent enrollments, which can be used in other settings such as group encryption.

6.2.7. *Zero-Knowledge Arguments for Matrix-Vector Relations and Lattice-Based Group Encryption*

Group encryption (GE) is the natural encryption analogue of group signatures in that it allows verifiably encrypting messages for some anonymous member of a group while providing evidence that the receiver is

a properly certified group member. Should the need arise, an opening authority is capable of identifying the receiver of any ciphertext. As introduced by Kiayias, Tsiounis and Yung (Asiacrypt'07), GE is motivated by applications in the context of oblivious retriever storage systems, anonymous third parties and hierarchical group signatures. In [38], we provide the first realization of group encryption under lattice assumptions. The construction of [38] is proved secure in the standard model (assuming interaction in the proving phase) under the Learning-With-Errors (LWE) and Short-Integer-Solution (SIS) assumptions. As a crucial component of our system, [38] describes a new zero-knowledge argument system allowing to demonstrate that a given ciphertext is a valid encryption under some hidden but certified public key, which incurs to prove quadratic statements about LWE relations. Specifically, the protocol of [38] allows arguing knowledge of witnesses consisting of $X \in \mathbb{Z}_q^{m\times}$, $s \in \mathbb{Z}_q^n$ and a small-norm $e \in \mathbb{Z}^m$ which underlie a public vector $b = X \cdot s + e \in \mathbb{Z}_q^m$ while simultaneously proving that the matrix $X \in \mathbb{Z}_q^{m\times n}$ has been correctly certified.

6.2.8. *Efficient Cryptosystems From 2^k-th Power Residue Symbols*

Goldwasser and Micali (1984) highlighted the importance of randomizing the plaintext for public-key encryption and introduced the notion of semantic security. They also realized a cryptosystem meeting this security notion under the standard complexity assumption of deciding quadratic residuosity modulo a composite number. The Goldwasser-Micali cryptosystem is simple and elegant but is quite wasteful in bandwidth when encrypting large messages. A number of works followed to address this issue and proposed various modifications. In [4], we revisit the original Goldwasser-Micali cryptosystem using 2^k-th power residue symbols. The so-obtained cryptosystems appear as a very natural generalization for $k \geq 2$ (the case $k = 1$ corresponds exactly to the Goldwasser-Micali cryptosystem). Advantageously, they are efficient in both bandwidth and speed; in particular, they allow for fast decryption. Further, the cryptosystems described in this paper inherit the useful features of the original cryptosystem (like its homomorphic property) and are shown to be secure under a similar complexity assumption. As a prominent application, the paper [4] describes an efficient lossy trapdoor function based thereon.

6.2.9. *Born and raised distributively: Fully distributed non-interactive adaptively-secure threshold signatures with short shares*

Threshold cryptography is a fundamental distributed computational paradigm for enhancing the availability and the security of cryptographic public-key schemes. It does it by dividing private keys into n shares handed out to distinct servers. In threshold signature schemes, a set of at least $t + 1 \leq n$ servers is needed to produce a valid digital signature. Availability is assured by the fact that any subset of $t + 1$ servers can produce a signature when authorized. At the same time, the scheme should remain robust (in the fault tolerance sense) and unforgeable (cryptographically) against up to t corrupted servers; i.e., it adds quorum control to traditional cryptographic services and introduces redundancy. Originally, most practical threshold signatures have a number of demerits: They have been analyzed in a static corruption model (where the set of corrupted servers is fixed at the very beginning of the attack); they require interaction; they assume a trusted dealer in the key generation phase (so that the system is not fully distributed); or they suffer from certain overheads in terms of storage (large share sizes). In [17], we construct practical fully distributed (the private key is born distributed), non-interactive schemes – where the servers can compute their partial signatures without communication with other servers – with adaptive security (i.e., the adversary corrupts servers dynamically based on its full view of the history of the system). The schemes of [17] are very efficient in terms of computation, communication, and scalable storage (with private key shares of size $O(1)$, where certain solutions incur $O(n)$ storage costs at each server). Unlike other adaptively secure schemes, the new schemes [17] are erasure-free (reliable erasure is hard to assure and hard to administer properly in actual systems). To the best of our knowledge, such a fully distributed highly constrained scheme has been an open problem in the area. In particular, and of special interest, is the fact that Pedersen's traditional distributed key generation (DKG) protocol can be safely employed in the initial key generation phase when the system is born although it is well-known not to ensure uniformly distributed public keys. An advantage of this is that this protocol only takes one round optimistically (in the absence of faulty player).

6.2.10. Non-Zero Inner Product Encryption with Short Ciphertexts and Private Keys

In [28], the team describes two constructions of non-zero inner product encryption (NIPE) systems in the public index setting, both having ciphertexts and secret keys of constant size. Both schemes are obtained by tweaking the Boneh-Gentry-Waters broadcast encryption system (Crypto 2005) and are proved selectively secure without random oracles under previously considered assumptions in groups with a bilinear map. Our first realization builds on prime-order bilinear groups and is proved secure under the Decisional Bilinear Diffie-Hellman Exponent assumption, which is parameterized by the length n of vectors over which the inner product is defined. By moving to composite order bilinear groups, the paper [28] obtains security under static subgroup decision assumptions following the Déjà Q framework of Chase and Meiklejohn (Eurocrypt 2014) and its extension by Wee (TCC 2016). The schemes of [28] are the first NIPE systems to achieve such parameters, even in the selective security setting. Moreover, they are the first proposals to feature optimally short private keys, which only consist of one group element. The prime-order-group realization of [28] is also the first one with a deterministic key generation mechanism.

6.2.11. More Efficient Constructions for Inner-Product Encryptions

In [48], the team describes new constructions for inner product encryption (called IPE1 and IPE2), which are both secure under the eXternal Diffie-Hellman assumption (SXDH) in asymmetric pairing groups. The IPE1 scheme of [48] has constant-size ciphertexts whereas the second one is weakly attribute hiding. The second scheme is derived from the identity-based encryption scheme of Jutla and Roy (Asiacrypt 2013), that was extended from tag-based quasi-adaptive non-interactive zero-knowledge (QA-NIZK) proofs for linear subspaces of vector spaces over bilinear groups. The verifier common reference string (CRS) in these tag-based systems are split into two parts, that are combined during verification. The paper [48] considers an alternate form of the tag-based QA-NIZK proof with a single verifier CRS that already includes a tag, different from the one defining the language. The verification succeeds as long as the two tags are unequal. Essentially, we embed a two-equation revocation mechanism in the verification. The new QA-NIZK proof system leads to IPE1, a constant-sized ciphertext IPE scheme with very short ciphertexts. Both the IPE schemes are obtained by applying the n-equation revocation technique of Attrapadung and Libert (PKC 2010) to the corresponding identity based encryption schemes and proved secure under SXDH assumption. As an application, the paper [48] shows how the new schemes can be specialized to obtain the first fully secure identity-based broadcast encryption based on SXDH with a trade-off among the public parameters, ciphertext and key sizes, all of them being sub-linear in the maximum number of recipients of a broadcast.

6.2.12. Verifiable Message-Locked Encryption

One of today's main challenge related to cloud storage is to maintain the functionalities and the efficiency of customers' and service providers' usual environments, while protecting the confidentiality of sensitive data. Deduplication is one of those functionalities: it enables cloud storage providers to save a lot of memory by storing only once a file uploaded several times. But classical encryption blocks deduplication. One needs to use a "message-locked encryption" (MLE), which allows the detection of duplicates and the storage of only one encrypted file on the server, which can be decrypted by any owner of the file. However, in most existing scheme, a user can bypass this deduplication protocol. In [27], we provide servers verifiability for MLE schemes: the servers can verify that the ciphertexts are well-formed. This property that we formally define forces a customer to prove that she complied to the deduplication protocol, thus preventing her to deviate from *the prescribed functionality* of MLE. We call it *deduplication consistency*. To achieve this deduplication consistency, we provide (i) a generic transformation that applies to any MLE scheme and (ii) an ElGamal-based deduplication-consistent MLE, which is secure in the random oracle model.

6.2.13. Privately Outsourcing Exponentiation to a Single Server: Cryptanalysis and Optimal Constructions

In [29], we address the problem of speeding up group computations in cryptography using a single untrusted computational resource. We analyze the security of an efficient protocol for securely outsourcing multi-exponentiations proposed at ESORICS 2014. We show that this scheme does not achieve the claimed security

guarantees and we present several practical polynomial-time attacks on the delegation protocol which allows the untrusted helper to recover part (or the whole) of the device secret inputs. We then provide simple constructions for outsourcing group exponentiations in different settings (e.g. public/secret, fixed/variable bases and public/secret exponents). Finally, we prove that our attacks on the ESORICS 2014 protocol are unavoidable if one wants to use a single untrusted computational resource and to limit the computational cost of the limited device to a constant number of (generic) group operations. In particular, we show that our constructions are actually optimal.

6.3. Algebraic computing and high-performance kernels

6.3.1. Algebraic Diagonals and Walks: Algorithms, Bounds, Complexity

The diagonal of a multivariate power series F is the univariate power series $\mathrm{Diag}(F)$ generated by the diagonal terms of F. Diagonals form an important class of power series; they occur frequently in number theory, theoretical physics and enumerative combinatorics. We study algorithmic questions related to diagonals in the case where F is the Taylor expansion of a bivariate rational function. It is classical that in this case $\mathrm{Diag}(F)$ is an algebraic function. We propose an algorithm that computes an annihilating polynomial for $\mathrm{Diag}(F)$. We give a precise bound on the size of this polynomial and show that generically, this polynomial is the minimal polynomial and that its size reaches the bound. The algorithm runs in time quasi-linear in this bound, which grows exponentially with the degree of the input rational function. We then address the related problem of enumerating directed lattice walks. The insight given by our study leads to a new method for expanding the generating power series of bridges, excursions and meanders. We show that their first N terms can be computed in quasi-linear complexity in N, without first computing a very large polynomial equation [6].

6.3.2. Multiple Binomial Sums

Multiple binomial sums form a large class of multi-indexed sequences, closed under partial summation, which contains most of the sequences obtained by multiple summation of products of binomial coefficients and also all the sequences with algebraic generating function. We study the representation of the generating functions of binomial sums by integrals of rational functions. The outcome is twofold. Firstly, we show that a univariate sequence is a multiple binomial sum if and only if its generating function is the diagonal of a rational function. Secondly, we propose algorithms that decide the equality of multiple binomial sums and that compute recurrence relations for them. In conjunction with geometric simplifications of the integral representations, this approach behaves well in practice. The process avoids the computation of certificates and the problem of the appearance of spurious singularities that afflicts discrete creative telescoping, both in theory and in practice [7].

6.3.3. Fast and Accurate Computation of Orbital Collision Probability for Short-Term Encounters

We provide a new method for computing the probability of collision between two spherical space objects involved in a short-term encounter under Gaussian-distributed uncertainty. In this model of conjunction, classical assumptions reduce the probability of collision to the integral of a two-dimensional Gaussian probability density function over a disk. The computational method is based on an analytic expression for the integral, derived by use of Laplace transform and D-finite functions properties. The formula has the form of a product between an exponential term and a convergent power series with positive coefficients. Analytic bounds on the truncation error are also derived and are used to obtain a very accurate algorithm. Another contribution is the derivation of analytic bounds on the probability of collision itself, allowing for a very fast and — in most cases — very precise evaluation of the risk. The only other analytical method of the literature — based on an approximation — is shown to be a special case of the new formula. A numerical study illustrates the efficiency of the proposed algorithms on a broad variety of examples and favorably compares the approach to the other methods of the literature [20].

6.3.4. Efficient Algorihtms for Mixed Creative Telescoping

Creative telescoping is a powerful computer algebra paradigm — initiated by Doron Zeilberger in the 90's — for dealing with definite integrals and sums with parameters. We address the mixed continuous-discrete case, and focus on the integration of bivariate hypergeometric-hyperexponential terms. We design a new creative telescoping algorithm operating on this class of inputs, based on a Hermite-like reduction procedure. The new algorithm has two nice features: it is efficient and it delivers, for a suitable representation of the input, a minimal-order telescoper. Its analysis reveals tight bounds on the sizes of the telescoper it produces [26].

6.3.5. Symbolic-Numeric Tools for Analytic Combinatorics in Several Variables

Analytic combinatorics studies the asymptotic behaviour of sequences through the analytic properties of their generating functions. This article provides effective algorithms required for the study of analytic combinatorics in several variables, together with their complexity analyses. Given a multivariate rational function we show how to compute its smooth isolated critical points, with respect to a polynomial map encoding asymptotic behaviour, in complexity singly exponential in the degree of its denominator. We introduce a numerical Kronecker representation for solutions of polynomial systems with rational coefficients and show that it can be used to decide several properties (0 coordinate, equal coordinates, sign conditions for real solutions, and vanishing of a polynomial) in good bit complexity. Among the critical points, those that are minimal—a property governed by inequalities on the moduli of the coordinates—typically determine the dominant asymptotics of the diagonal coefficient sequence. When the Taylor expansion at the origin has all non-negative coefficients (known as the 'combinatorial case') and under regularity conditions, we utilize this Kronecker representation to determine probabilistically the minimal critical points in complexity singly exponential in the degree of the denominator, with good control over the exponent in the bit complexity estimate. Generically in the combinatorial case, this allows one to automatically and rigorously determine asymptotics for the diagonal coefficient sequence. Examples obtained with a preliminary implementation show the wide applicability of this approach [43].

6.3.6. Tableau sequences, open diagrams, and Baxter families

Walks on Young's lattice of integer partitions encode many objects of algebraic and combinatorial interest. Chen *et al.* established connections between such walks and arc diagrams. We show that walks that start at $\varnothing$, end at a row shape, and only visit partitions of bounded height are in bijection with a new type of arc diagram — open diagrams. Remarkably, two subclasses of open diagrams are equinumerous with well known objects: standard Young tableaux of bounded height, and Baxter permutations. We give an explicit combinatorial bijection in the former case, and a generating function proof and new conjecture in the second case [9].

6.3.7. On 3-dimensional lattice walks confined to the positive octant

Many recent papers deal with the enumeration of 2-dimensional walks with prescribed steps confined to the positive quadrant. The classification is now complete for walks with steps in $\{0,\pm 1\}^2$: the generating function is differentially finite if and only if a certain group associated with the step set is finite. We explore in this paper the analogous problem for 3-dimensional walks confined to the positive octant. The first difficulty is their number: we have to examine no less than 11074225 step sets in $\{0,\pm 1\}^3$ (instead of 79 in the quadrant case). We focus on the 35548 that have at most six steps. We apply to them a combined approach, first experimental and then rigorous. On the experimental side, we try to guess differential equations. We also try to determine if the associated group is finite. The largest finite groups that we find have order 48 — the larger ones have order at least 200 and we believe them to be infinite. No differential equation has been detected in those cases. On the rigorous side, we apply three main techniques to prove D-finiteness. The algebraic kernel method, applied earlier to quadrant walks, works in many cases. Certain, more challenging, cases turn out to have a special Hadamard structure, which allows us to solve them via a reduction to problems of smaller dimension. Finally, for two special cases, we had to resort to computer algebra proofs. We prove with these techniques all the guessed differential equations. This leaves us with exactly 19 very intriguing step sets for which the group is finite, but the nature of the generating function still unclear [5].

6.3.8. Asymptotic Lattice Path Enumeration Using Diagonals

We consider d-dimensional lattice path models restricted to the first orthant whose defining step sets exhibit reflective symmetry across every axis. Given such a model, we provide explicit asymptotic enumerative formulas for the number of walks of a fixed length: the exponential growth is given by the number of distinct steps a model can take, while the sub-exponential growth depends only on the dimension of the underlying lattice and the number of steps moving forward in each coordinate. The generating function of each model is first expressed as the diagonal of a multivariate rational function, then asymptotic expressions are derived by analyzing the singular variety of this rational function. Additionally, we show how to compute subdominant growth, reflect on the difference between rational diagonals and differential equations as data structures for D-finite functions, and show how to determine first order asymptotics for the subset of walks that start and end at the origin [18].

6.3.9. Asymptotics of lattice walks via analytic combinatorics in several variables

We consider the enumeration of walks on the two-dimensional non-negative integer lattice with steps defined by a finite set $S \subset \{0, \pm 1\}^2$. Up to isomorphism there are 79 unique two-dimensional models to consider, and previous work in this area has used the kernel method, along with a rigorous computer algebra approach, to show that 23 of the 79 models admit D-finite generating functions. In 2009, Bostan and Kauers used Padé-Hermite approximants to guess differential equations which these 23 generating functions satisfy, in the process guessing asymptotics of their coefficient sequences. In this article we provide, for the first time, a complete rigorous verification of these guesses. Our technique is to use the kernel method to express 19 of the 23 generating functions as diagonals of tri-variate rational functions and apply the methods of analytic combinatorics in several variables (the remaining 4 models have algebraic generating functions and can thus be handled by univariate techniques). This approach also shows the link between combinatorial properties of the models and features of its asymptotics such as asymptotic and polynomial growth factors. In addition, we give expressions for the number of walks returning to the x-axis, the y-axis, and the origin, proving recently conjectured asymptotics of Bostan, Chyzak, van Hoeij, Kauers, and Pech [44].

6.3.10. Linear Time Interactive Certificates

With J.G. Dumas (LJK, Grenoble), E. Kaltofen (NCSU, USA), and E. Thomé (Inria Nancy) we work on interactive certificates. Computational problem certificates are additional data structures for each output, which can be used by a (possibly randomized) verification algorithm that proves the correctness of each output. In [32] we give a new certificate for the minimal polynomial of sparse or structured matrices whose Monte Carlo verification complexity requires a single matrix-vector multiplication and a linear number of extra field operations (sufficiently large cardinality field). We also propose a novel preconditioner that ensures irreducibility of the characteristic polynomial of the generically preconditioned matrix. This preconditioner takes linear time to be applied and uses only two random entries. We combine these two techniques to give algorithms that compute certificates for the determinant, and thus for the characteristic polynomial, whose Monte Carlo verification complexity is therefore also linear.

6.3.11. Computing minimal interpolation bases

With É. Schost (U. Waterloo, Canada), we consider the problem of computing minimal bases of solutions for a general interpolation problem, which encompasses Hermite-Padé approximation and constrained multivariate interpolation, and has applications in coding theory and security. The problem is classically solved using iterative algorithms based on recurrence relations. First, we discuss in [62] a fast, divide-and-conquer version of this recurrence, taking advantage of fast matrix computations over the scalars and over the polynomials. This new algorithm is deterministic, and for computing shifted minimal bases of relations between m vectors of size σ it uses $\widetilde{O}(m^{\omega-1}(\sigma + |s|))$ field operations, where the notation $\widetilde{O}(\cdot)$ indicates that logarithmic terms are omitted, $\omega \in [2, 2.38]$ is the exponent of matrix multiplication, and $|s|$ is the sum of the entries of the input shift s, with $\min(s) = 0$. This complexity bound improves in particular on earlier algorithms in the case of bivariate interpolation for soft decoding, while matching fastest existing algorithms for simultaneous Hermite-Padé approximation. Then we propose in [33] an algorithm for the computation of an interpolation

basis in shifted-Popov normal form with a cost of $\widetilde{O}(m^{\omega-1}\sigma)$ field operations. Previous works, in the case of Hermite-Padé approximation and in the general interpolation case, compute non-normalized bases. Since for arbitrary shifts such bases may have size $\Theta(m^2\sigma)$, the cost bound $\widetilde{O}(m^{\omega-1}\sigma)$ was feasible only with restrictive assumptions on the shift that ensure small output sizes. The question of handling arbitrary shifts with the same complexity bound was left open. To obtain the target cost for any shift, we strengthen the properties of the output bases, and of those obtained during the course of the algorithm: all the bases are computed in shifted Popov form, whose size is always $O(m\sigma)$. Then, we design a divide-and-conquer scheme. We recursively reduce the initial interpolation problem to sub-problems with more convenient shifts by first computing information on the degrees of the intermediate bases.

6.3.12. Fast computation of shifted Popov forms of polynomial matrices via systems of modular polynomial equations

In [46] we give a Las Vegas algorithm which computes the shifted Popov form of an $m \times m$ nonsingular polynomial matrix of degree d in expected $\widetilde{O}(m^\omega d)$ field operations, where ω is the exponent of matrix multiplication and $\widetilde{O}(\cdot)$ indicates that logarithmic factors are omitted. This is the first algorithm in $\widetilde{O}(m^\omega d)$ for shifted row reduction with arbitrary shifts. Using partial linearization, we reduce the problem to the case $d \leq \lceil \sigma/m \rceil$ where σ is the generic determinant bound, with σ/m bounded from above by both the average row degree and the average column degree of the matrix. The cost above becomes $\widetilde{O}(m^\omega \lceil \sigma/m \rceil)$, improving upon the cost of the fastest previously known algorithm for row reduction, which is deterministic. Our algorithm first builds a system of modular equations whose solution set is the row space of the input matrix, and then finds the basis in shifted Popov form of this set. We give a deterministic algorithm for this second step supporting arbitrary moduli in $\widetilde{O}(m^{\omega-1}\sigma)$ field operations, where m is the number of unknowns and σ is the sum of the degrees of the moduli. This extends previous results with the same cost bound in the specific cases of order basis computation and M-Padé approximation, in which the moduli are products of known linear factors.

6.3.13. Fast, deterministic computation of the Hermite normal form and determinant of a polynomial matrix

With G. Labahn and W. Zhou (U. Waterloo, Canada) we give in [64] fast and deterministic algorithms to compute the determinant and Hermite normal form of a nonsingular $n \times n$ matrix of univariate polynomials over a field $\mathbb{K}$. Our algorithms use $\widetilde{O}(n^\omega \lceil s \rceil)$ operations in $\mathbb{K}$, where s is bounded from above by both the average of the degrees of the rows and that of the columns of the matrix and ω is the exponent of matrix multiplication. The soft-O notation indicates that logarithmic factors in the big-O are omitted while the ceiling function indicates that the cost is $\widetilde{O}(n^\omega)$ when $s = o(1)$. Our algorithms are based on a fast and deterministic triangularization method for computing the diagonal entries of the Hermite form of a nonsingular matrix.

6.3.14. Fast Computation of the Rank Profile Matrix and the Generalized Bruhat Decomposition

The row (resp. column) rank profile of a matrix describes the stair-case shape of its row (resp. column) echelon form. With J. G. Dumas and Z. Sultan (LJK, Grenoble), we propose in [11] a new matrix invariant, the rank profile matrix, summarizing all information on the row and column rank profiles of all the leading sub-matrices. We show that this normal form exists and is unique over any ring, provided that the notion of McCoy's rank is used, in the presence of zero divisors. We then explore the conditions for a Gaussian elimination algorithm to compute all or part of this invariant, through the corresponding PLUQ decomposition. This enlarges the set of known Elimination variants that compute row or column rank profiles. As a consequence a new Crout base case variant significantly improves the practical efficiency of previously known implementations over a finite field. With matrices of very small rank, we also generalize the techniques of Storjohann and Yang to the computation of the rank profile matrix, achieving an $(r^\omega + mn)^{1+o(1)}$ time complexity for an $m \times n$ matrix of rank r, where ω is the exponent of matrix multiplication. Finally, by give connections to the Bruhat decomposition, and several of its variants and generalizations. Thus, our algorithmic improvements for the PLUQ factorization, and their implementations, directly apply to these decompositions. In particular, we show how a PLUQ decomposition revealing the rank profile matrix also reveals both a row and

a column echelon form of the input matrix or of any of its leading sub-matrices, by a simple post-processing made of row and column permutations.

6.3.15. Computing with quasiseparable matrices

The class of quasiseparable matrices is defined by a pair of bounds, called the quasiseparable orders, on the ranks of the sub-matrices entirely located in their strictly lower and upper triangular parts. These arise naturally in applications, as e.g. the inverse of band matrices, and are widely used for they admit structured representations allowing to compute with them in time linear in the dimension. In [47] we show the connection between the notion of quasiseparability and the rank profile matrix invariant of Dumas et al. This allows us to propose an algorithm computing the quasiseparable orders (r_L, r_U) in time $O(n^2 s^{\omega-2})$, where $s = \max(r_L, r_U)$ and ω is the exponent of matrix multiplication. We then present two new structured representations, a binary tree of PLUQ decompositions, and the Bruhat generator, using respectively $O(ns\log(n/s))$ and $O(ns)$ field elements instead of $O(ns^2)$ for the classical generator and $O(nslogn)$ for the hierarchically semiseparable representations. We present algorithms computing these representations in time $O(n^2 s^{\omega-2})$. These representations allow a matrix-vector product in time linear in the size of their representation. Lastly we show how to multiply two such structured matrices in time $O(n^2 s^{\omega-2})$.

6.3.16. A Real QZ Algorithm for Structured Companion Pencils

With Y. Eidelman (U. Tel Aviv) and L. Gemignani (U. Pisa), we design in [54] a fast implicit real QZ algorithm for eigenvalue computation of structured companion pencils arising from linearizations of polynomial rootfind-ing problems. The modified QZ algorithm computes the generalized eigenvalues of an $N \times N$ structured matrix pencil using $O(N^2)$ flops and $O(N)$ memory storage. Numerical experiments and comparisons confirm the effectiveness and the stability of the proposed method.

6.3.17. Efficient Solution of Parameter Dependent Quasiseparable Systems and Computation of Meromorphic Matrix Functions

In [55], with Y. Eidelman (U. Tel Aviv) and L. Gemignani (U. Pisa), we focus on the solution of shifted quasiseparable systems and of more general parameter dependent matrix equations with quasiseparable representations. We propose an efficient algorithm exploiting the invariance of the quasiseparable structure under diagonal shifting and inversion. This algorithm is applied to compute various functions of matrices. Numerical experiments show the effectiveness of the approach.

COMPSYS Team

7. New Results

7.1. Handling Polynomials for Program Analysis and Transformation

Participant: Paul Feautrier.

As is well known in natural language processing, the first step in translating a text from one language to another is to understand it. The situation is the same for formal languages. A language processor has to "understand" a program before translating or optimizing or verifying it. Such understanding takes the form of a *model*, usually a mathematical representation whose natural operations mimic the behavior of its program. The polyhedral model is such a representation. However, the set of programs it can represent is too restricted, and the hunt for more powerful models has been under way since the millennium.

An obvious ideas is to replace affine formulas by polynomials, and hence polyhedra by semi-algebraic sets. Polynomials are ubiquitous in HPC and embedded system programming. For instance, the so-called "linearizations" (replacing a multi-dimensional object by a one-dimensional one) generate polynomial access functions. These polynomials then reappear in dependence testing, scheduling, and invariant construction. It may also happen that polynomials are absent from the source program, but are created either by an enabling analysis, as for OpenStream (see Section 7.2), or are imposed by complexity consideration. Lastly, polynomials may be native to the underlying algorithm, as when distances are computed by the usual Euclidean formula. What is needed here is a replacement for the familiar emptiness tests and for Farkas lemma (deciding whether an affine form is positive inside a polyhedron). Recent mathematical results by Handelman and Schweighofer on the *Positivstellensatz* allow one to devise algorithms that are able to solve these problems. The difference is that one gets only sufficient conditions, and that complexity is much higher than in the affine cases.

A paper presenting applications of these ideas to three use cases – dependence testing, scheduling, and transitive closure approximation – was presented at (IMPACT'15) [14]. A tool to manipulate polynomials, polynomial constraints and objective functions, needed for the derivation of polynomial schedules, complements this work (see Section 6.2). It implements Farkas lemma and its generalization with Handelman & Schweighofer formulations, and is in constant development, including improvements on the objective functions, in particular to make schedule selection more stable, independently on the degree of the polynomial schedule.

7.2. Static Analysis of OpenStream Programs

Participants: Albert Cohen [Inria Parkas team], Alain Darte, Paul Feautrier.

In the context of the ManycoreLabs project, we started to study the applicability of polyhedral techniques to the parallel language OpenStream [19]. When applicable, polyhedral techniques are indeed invaluable for compile-time debugging and for generating efficient code well suited to a target architecture. OpenStream is a two-level language in which a control program directs the initialization of parallel task instances that communicate through *streams*, with possibly multiple writers and readers. It has a fairly complex semantics in its most general setting, but we restricted ourselves to the case where the control program is sequential, which is representative of the majority of the OpenStream applications.

In contrast to the language X10, which we studied in previous years, this restriction offers deterministic concurrency by construction, but deadlocks are still possible. We showed that, if the control program is polyhedral, one may statically compute, for each task instance, the read and write indices to each of its streams, and thus reason statically about the dependences among task instances (the only scheduling constraints in this polyhedral subset). If the control program has nested loops, communications use one-dimensional channels in a form of linearization, and these indices may be polynomials of arbitrary degree, thus requiring to extend to polynomials the standard polyhedral techniques for dependence analysis, scheduling, and deadlock detection. Modern SMT allow to solve polynomial problems, albeit with no guarantee of success; the approach previously developed by P. Feautrier [14], and recalled in Section 7.1 , offers an alternative solution.

The usual way of disproving deadlocks is by exhibiting a schedule for the program operations, a well-known problem for polyhedral programs where dependences can be described by affine constraints. In the case of OpenStream, we established two important results related to deadlocks: 1) a characterization of deadlocks in terms of dependence paths, which implies that streams can be safely bounded as soon as a schedule exists with such sizes, 2) the proof that deadlock detection is undecidable, even for polyhedral OpenStream. Details of this work have been published at the international workshop IMPACT'16 [1].

Some further developments are in progress for scheduling OpenStream programs using polynomial techniques (with a corresponding prototype scheduling tool, specific to OpenStream, see Section 6.3). In particular, we made some progress for parsing a simplified version of OpenStream, exhibiting the relevant structure, and on the properties and construction of schedules with bounded streams and bounded delays, and on the analysis of the "foot bath", i.e., the pool of tasks that are created (already requiring some resources) but not activated yet (because they need to wait for the termination of other tasks due to dataflow semantics). This work should have interesting connections with the way runtime systems of tasks are managed.

7.3. Liveness Analysis in Explicitly-Parallel Programs

Participants: Alain Darte, Alexandre Isoard, Tomofumi Yuki.

In the light of the parallel specifications encountered in our other work – kernel offloading with pipelined communications [10], automatic parallelization, analysis of X10 [22], [23] and of OpenStream (see Section 7.2), intra-array reuse (see Section 7.4) – we revisited scalar and array element-wise liveness analysis for programs with parallel specifications. In earlier work on memory allocation/contraction (register allocation or intra- and inter-array reuse in the polyhedral model), a notion of "time" or a total order among the iteration points was used to compute the liveness of values. In general, the execution of parallel programs is not a total order, and hence the notion of time is not applicable.

We first revised how conflicts are computed by using ideas from liveness analysis for register allocation, studying the structure of the corresponding conflict/interference graphs. Instead of considering the conflict between two pairs of live ranges, we only consider the conflict between a live range and a write. This simplifies the formulation from having four instances involved in the test down to three, and also improves the precision of the analysis in the general case. Then we extended the liveness analysis to work with partial orders so that it can be applied to many different parallel languages/specifications with different forms of parallelism. An important result is that the complement of the conflict graph with partial orders is directly connected to memory reuse, even in presence of races. However, programs with conditionals do not even have a partial order, and our next step will be to handle such cases with more accuracy. Details of this work have been published at the international workshop IMPACT'16 [3].

Some new developments are in progress to explore even further the properties of such liveness analysis and the construction of conflict sets, in the general case (with connections with the concept of trace monoid) or for some common situations such as series-parallel graphs, appearing in languages such as X10 or OpenMP.

7.4. Extended Lattice-Based Memory Allocation

Participants: Alain Darte, Alexandre Isoard, Tomofumi Yuki.

We extended lattice-based memory allocation [11], an earlier work on memory (array) reuse analysis. The main motivation is to handle in a better way the more general forms of specifications we see today, e.g., with loop tiling, pipelining, and other forms of parallelism available in explicitly parallel languages. Our extension has two complementary aspects. We showed how to handle more general specifications where conflicting constraints (those that describe the array indices that cannot share the same location) are specified as a (non-convex) union of polyhedra. Unlike convex specifications, this also requires to be able to choose suitable directions (or basis) of array reuse. For that, we extended two dual approaches, previously proposed for a fixed basis, into optimization schemes to select suitable basis. Our final approach relies on a combination of the two, also revealing their links with, on one hand, the construction of multi-dimensional schedules for parallelism and tiling (but with a fundamental difference that we identify) and, on the other hand, the construction of

universal reuse vectors (UOV), which was only used so far in a specific context, for schedule-independent mapping.

This algorithmic work, connected to our previous work on parametric tiling [10] and the liveness analysis results of Section 7.3 , is complemented by a set of prototype scripting tools, see Section 6.1 . Details of this work have been published at the 2016 International Conference on Compiler Construction [2].

7.5. Stencil Accelerators

Participants: Steven Derrien [University of Rennes 1, Inria/CAIRN], Sanjay Rajopadhye [Colorado State University], Tomofumi Yuki.

Stencil computations have been known to be an important class of programs for scientific calculations. Recently, various architectures (mostly targeting FPGAs) for stencils are being proposed as hardware accelerators with high throughput and/or high energy efficiency. There are many different challenges for such design: How to maximize compute-I/O ratio? How to partition the problem so that the data fits on the on-chip memory? How to efficiently pipeline? How to control the area usage? We seek to address these challenges by combining techniques from compilers and high-level synthesis tools.

One project in collaboration with the CAIRN team and Colorado State University targets stencils with regular dependence patterns. Although many architectures have been proposed for this type of stencils, most of them use a large number of small processing elements (PE) to achieve high throughput. We are exploring an alternative design that aims for a single, large, deeply-pipelined PE. The hypothesis is that the pipelined parallelism is more area-efficient compared to replicating small PEs. We have published a work-in-progress paper on this topic at IMPACT'16 [4].

7.6. Efficient Mapping of Irregular Memory Accesses on FPGA

Participants: Xinyu Niu [Imperial College London], Tomofumi Yuki.

In a collaboration with Imperial College, we looked at efficiently implementing dynamic dependences on FPGAs. The collaboration is in the context of the EURECA project [0] where the dynamic reconfigurability of modern FPGAs is used to efficiently handle dynamic access patterns. We worked on analyzing data dependent array accesses to identify regularities within irregular memory accesses to reduce the cost of a dynamic memory reconfiguration module.

One part of this work has been published at the 2016 International Conference on Field Programmable Logic and Applications [5].

7.7. PolyApps

Participant: Tomofumi Yuki.

Loop transformation frameworks using the polyhedral model have gained increased attention since the rise of the multi-core era. We now have several research tools that have demonstrated their power on important kernels found in scientific computations. However, there remains a large gap between the typical kernels used to evaluate these tools and the actual applications used by the scientists.

PolyApps is an effort to collect applications from other domains of science to better establish the link between the compiler tools and "real" applications. The applications are modified to bypass some of the front-end issues of research tools, while keeping the ability to produce the original output. The goal is to assess how the state-of-the-art automatic parallelizers perform on full applications, and to identify new opportunities that only arise in larger pieces of code.

We showed that, with a few enhancements, the current tools will be able to reach and/or exceed the performance of existing parallelizations of the applications. One of the most critical element missing in current tools is the ability to modify the memory mappings.

[0] http://www.doc.ic.ac.uk/~nx210/2015/09/01/eureca.html

CONVECS Project-Team

6. New Results

6.1. New Formal Languages and their Implementations

The ability to compile and verify formal specifications with complex, user-defined operations and data structures is a key feature of the CADP toolbox since its very origins. A long-run effort has been recently undertaken to ensure a uniform treatment of types, values, and functions across all the various CADP tools.

6.1.1. Translation from LNT to LOTOS

Participants: Hubert Garavel, Frédéric Lang, Wendelin Serwe.

LNT is a next generation formal description language for asynchronous concurrent systems, which attempts to combine the best features of imperative programming languages and value-passing process algebras. LNT is increasingly used by CONVECS for industrial case studies and applications (see § 6.5) and serves also in university courses on concurrency, in particular at ENSIMAG (Grenoble) and at Saarland University.

In 2016, the long-term effort to enhance the LNT language and its tools has been pursued. LNT has been enriched with a new statement "use X" that suppresses compiler warnings when a variable X is assigned but not used. The syntax of LNT expressions has been modified so that field selections ("E.X"), field updates ("E1.X = E2"), and array accesses ("E1 [E2]") can now be freely combined without extra parentheses. LNT programs can now import predefined libraries, and two such libraries (BIT.lnt and OCTET.lnt) have been introduced.

A move towards "safer" LNT exceptions has started. The syntax for exceptions in function declarations has been modified and the semantics of LNT has shifted from "unchecked" to "checked" exceptions: exception parameters, if any, must be explicitly mentioned when a function is called. Such exception parameters can now be passed using either the named style or the positional style.

A few static-semantics constraints have been relaxed; for instance, it is no longer required that actual gate parameters be different when calling a process. Various bugs have been fixed. Several error/warning messages have been made more precise, and the format of LNT error/warning messages has been aligned on that of GCC. Finally, the LNT2LOTOS Reference Manual has been updated and enhanced.

6.1.2. Translation from LOTOS NT to C

Participants: Hubert Garavel, Sai Srikar Kasi, Wendelin Serwe.

The TRAIAN compiler is used to build many compilers and translators of the CADP toolbox. This compiler itself is built using the FNC-2 compiler generator that, unfortunately, is no longer available. For this reason, TRAIAN only exists in 32-bit version, and sometimes hits the 3-4 GByte RAM limit when dealing with complex compilers such as LNT2LOTOS.

In 2016 we addressed this issue, in several steps. As a first step, we released a stable version 2.8 of TRAIAN. Then, we gathered all programs written in LOTOS NT, the input language of TRAIAN, and organized them in non-regression test bases. We entirely scrutinized the source code of TRAIAN, which consists in a large collection of attribute grammars, deleting all parts of code corresponding to those features of the LOTOS NT language that were either not fully implemented or seldom used in practice. This reduced the source code of TRAIAN by 40% and divided by two the size of TRAIAN executables. A few other bugs have been fixed and the reference manual of TRAIAN was entirely revised and updated.

In parallel, we undertook a complete rewrite of TRAIAN to get rid of the FNC-2 deprecated attribute grammar tool. We developed lexical and syntactic descriptions of the input language using the SYNTAX compiler-generation system developed at Inria Paris. The syntax tree of LOTOS NT and the library of predefined LOTOS NT types and functions are now themselves defined in LOTOS NT, as we plan to follow a bootstraping approach, using the current version of TRAIAN to build the next one. To this aim, a large fraction of the TRAIAN attribute grammars has been rewritten in LOTOS NT.

6.1.3. Translation from LOTOS to Petri nets and C

Participants: Hubert Garavel, Wendelin Serwe.

The LOTOS compilers CAESAR and CAESAR.ADT, which were once the flagship of CADP, now play a more discrete role since LNT (rather than LOTOS) has become the recommended specification language of CADP. Thus, CAESAR and CAESAR.ADT are mostly used as back-end translators for LOTOS programs automatically generated from LNT or other formalisms such as Fiacre, and are only modified when this appears to be strictly necessary.

In 2016, following the writing of the new CADP manual page for LOTOS, the common front-end of CAESAR and CAESAR.ADT was carefully inspected, which led to various bug fixes regarding type signatures, error messages for overloaded functions, renaming/actualization of sorts and operations, equations for renamed operations, C-language reserved keywords, implementation of numeral sorts, and iterators over these sorts. Another bug was fixed for the "`-external`" option of CAESAR and a new "`-numeral`" option was introduced. Also, the C identifiers automatically generated by CAESAR.ADT for sorts, operations, tester and selector macros have been simplified; as the new conventions are not backward compatible, migration tools were developed to ease transitioning the existing LOTOS and C files.

6.1.4. NUPN

Participant: Hubert Garavel.

Nested-Unit Petri Nets (NUPNs) is an upward-compatible extension of P/T nets, which are enriched with structural information on their concurrent structure. Such additional information can easily be produced when NUPNs are generated from higher-level specifications (e.g., process calculi); quite often, such information allows logarithmic reductions in the number of bits required to represent states, thus enabling verification tools to perform better. The principles of NUPNs are exposed in [33] and its PNML representation is described here [0].

In 2016, the NUPN principles have been presented in an invited talk at D-CON, the German national conference on concurrency theory. The collection of NUPN models used for experimentation has been enlarged and reorganized; it now contains more than $10,000$ models. A new beta-version of the VLPN (*Very Large Petri Nets*) benchmark suite, which contains 350 large models has been produced. Also, new prototype tools have been developed that try to convert P/T nets into NUPNs, which requires to automatically infer the concurrent structure of flat, unstructured nets.

The CAESAR.BDD tool that analyzes NUPN models and serves to prepare the yearly Model Checking Contest [0] has been enhanced with two new options "`-initial-places`" and "`-initial-tokens`". It now properly handles the case where the initial marking contains more than 2^{31} tokens. The output of the "`-mcc`" option has been made more precise when the NUPN under study is conservative or sub-conservative.

6.1.5. Translation from BPMN to LNT

Participants: Gwen Salaün, Ajay Muroor-Nadumane.

Evolution has become a central concern in software development and in particular in business processes, which support the modeling and the implementation of software as workflows of local and inter-process activities. We advocate that business process evolution can be formally analyzed in order to compare different versions of processes, identify precisely the differences between them, and ensure the desired consistency.

In collaboration with Pascal Poizat (LIP6, Paris), we worked on checking the evolution of BPMN processes. To promote its adoption by business process designers, we have implemented it in a tool, VBPMN, that can be used through a Web application. We have defined different kinds of atomic evolutions that can be combined and formally verified. We have defined a BPMN to LNT model transformation, which, using the LTS operational semantics of LNT, enables us to automate our approach using existing LTS model checking and equivalence checking tools, such as those provided by CADP. We have applied our approach to many examples for evaluation purposes. These results have been published in an international conference [23].

[0] http://mcc.lip6.fr/nupn.php
[0] http://mcc.lip6.fr/

6.1.6. Translation from GRL to LNT

Participants: Hubert Garavel, Fatma Jebali, Jingyan Jourdan-Lu, Frédéric Lang, Eric Léo, Radu Mateescu, Wendelin Serwe.

In the context of the Bluesky project (see § 8.2.2.1), we study the formal modeling of GALS (*Globally Asynchronous, Locally Synchronous*) systems, which are composed of several synchronous subsystems evolving cyclically, each at its own pace, and communicating with each other asynchronously. Designing GALS systems is challenging due to both the high level of (synchronous and asynchronous) concurrency and the heterogeneity of computations (deterministic and nondeterministic). To bring our formal verification techniques and tools closer to the GALS paradigm, we designed a new formal language named GRL (*GALS Representation Language*), as an intermediate format between GALS models and purely asynchronous concurrent models. GRL combines the main features of synchronous dataflow programming and asynchronous process calculi into one unified language, while keeping the syntax homogeneous for better acceptance by industrial GALS designers. GRL allows a modular composition of synchronous systems (blocks), environmental constraints (environments), and asynchronous communication mechanisms (mediums), to be described at a level of abstraction that is appropriate to verification. GRL also supports external C and LNT code. A translator named GRL2LNT has been developed, allowing an LNT program to be generated from a GRL specification automatically. Additionally, an OPEN/CAESAR-compliant compiler named GRL.OPEN (based on GRL2LNT and LNT.OPEN) enables the on-the-fly exploration of the LTS underlying a GRL specification using CADP.

In 2016, a new version 3.3 of the GRL2LNT translator has been released, with an improved LNT code generation exploiting the "`use`" construct newly added to LNT. Also, a non-regression test base containing hundreds of GRL specifications has been set up. This also contributes to the non-regression testing of the compilation chain for LNT by providing new LNT descriptions generated automatically by GRL2LNT.

An overview paper about GRL and its translation to LNT was published in an international journal [14]. The complete definition of GRL and its applications to GALS systems are available in F. Jebali's PhD thesis [44].

6.1.7. Translation of Term Rewrite Systems

Participants: Hubert Garavel, Lina Marsso, Mohammad-Ali Tabikh.

In 2016, we pursued the development undertaken in 2015 of a software platform for systematically comparing the performance of rewrite engines and pattern-matching implementations in algebraic specification and functional programming languages. Our platform reuses the benchmarks of the three Rewrite Engine Competitions (2006, 2009, and 2010). Such benchmarks are term-rewrite systems expressed in a simple formalism named REC, for which we developed automated translators that convert REC benchmarks into various languages.

In 2016, we corrected a number of benchmarks and added many new ones, to reach a total of 85 benchmarks in December 2016. Among these new benchmarks, one can mention a formalization of arithmetic operations on signed integers, a collection of (8-bit, 16-bit, and 32-bit) binary adders and multipliers, and a complete model of the MAA ("Message Authenticator Algorithm"), a Message Authentication Code used for financial transactions (ISO 8731-2) between 1987 and 2002.

The existing translators (for Haskell, LOTOS, Maude, mCRL, OCAML, Opal, Rascal, Scala, SML-NJ, and Tom) have been enhanced and new translators (for AProVE, Clean, LNT, MLTON, Stratego/XT) have been developed. Tools for automatically extracting and synthesizing performance statistics have also been developed.

6.2. Parallel and Distributed Verification

6.2.1. Distributed State Space Manipulation

Participants: Hubert Garavel, Hugues Evrard, Wendelin Serwe.

For distributed verification, CADP provides the PBG format, which implements the theoretical concept of *Partitioned LTS* [38] and provides a unified access to an LTS distributed over a set of remote machines.

In 2016, the code of the CAESAR_NETWORK_1 library, which is a building block for the distributed verification tools of CADP, has been carefully scrutinized and split into logically-independent modules. Nine problems have been detected and solved, among which a flaw in the distributed termination algorithm: the entire network could freeze if a worker process crashed too early, before the opening of TCP sockets. From now on, a better distributed termination algorithm is used, which supports the coexistence of several networks, ensures that all connections are closed before terminating, and produces more informative traces indicating which worker has triggered termination. Also, the improved CAESAR_NETWORK_1 library checks that all workers operate in directories that are pairwise distinct, mutually disjoint, and different from the working directory of the coordinator process.

6.2.2. Distributed Code Generation for LNT

Participants: Hugues Evrard, Frédéric Lang, Wendelin Serwe.

Rigorous development and prototyping of a distributed algorithm using LNT involves the automatic generation of a distributed implementation. For the latter, a protocol realizing process synchronization is required. As far as possible, this protocol must itself be distributed, so as to avoid the bottleneck that would inevitably arise if a unique process would have to manage all synchronizations in the system. Using a synchronization protocol that we verified formally in 2013, we developed a prototype distributed code generator, named DLC (*Distributed LNT Compiler*), which takes as input the model of a distributed system described as a parallel composition of LNT processes.

In 2016, we improved the user documentation of the DLC distribution, and added support for structured data types, enabling experiments of DLC on the LNT model of the CAESAR_SOLVE_2 library (see § 6.2.3). An overview paper about DLC has been accepted in an international journal [12].

6.2.3. Distributed Resolution of Boolean Equation Systems

Participant: Wendelin Serwe.

The BES_SOLVE tool of CADP enables to solve BESs (*Boolean Equation Systems*) using the various resolution algorithms provided by the CAESAR_SOLVE library (see 5.1), including a distributed on-the-fly resolution algorithm described in pseudo-code in [45].

In 2016, we modeled the pseudo-code of the distributed resolution algorithm in LNT (about $1,000$ lines). For a set of BES examples (encoded as LNT data types and functions), we experimented the generation of the LTS corresponding to the distributed resolution algorithm applied to each BES. We also experimented with the DLC tool [21] to generate a prototype distributed implementation of the resolution algorithm from its LNT specification. These experiments uncovered some errors in the original pseudo-code.

We also simplified the C implementation included in the BES_SOLVE tool to closer match the corrected LNT model, mainly by removing additional synchronization messages. We started to evaluate the simplified implementation using our non-regression test base (more than $15,000$ BESs), with promising results.

6.2.4. Stability of Communicating Systems

Participant: Gwen Salaün.

Analyzing systems communicating asynchronously via reliable FIFO buffers is an undecidable problem. A typical approach is to check whether the system is bounded, and if not, the corresponding state space can be made finite by limiting the presence of communication cycles in behavioral models or by imposing an upper bound for the size of communication buffers.

In 2016, our focus was on systems that are likely to be unbounded and therefore result in infinite systems. We did not want to restrict the system by imposing any arbitrary bound. We introduced a notion of stability and proved that once the system is stable for a specific buffer bound, it remains stable whatever larger bounds are chosen for buffers. This enables one to check certain properties on the system for that bound and to ensure that the system will preserve them for arbitrarily larger buffer bounds. We also proved that computing this bound is undecidable but we showed how we succeed in computing these bounds for many examples using heuristics and equivalence checking. These results have been published in an international conference [18].

In collaboration with Carlos Canal (University of Málaga, Spain), we have also shown how the stability approach can be used for composition and adaptation of component-based software. This led to a publication in an international conference [20].

6.2.5. Debugging of Concurrent Systems

Participants: Gianluca Barbon, Gwen Salaün.

Model checking is an established technique for automatically verifying that a model satisfies a given temporal property. When the model violates the property, the model checker returns a counterexample, which is a sequence of actions leading to a state where the property is not satisfied. Understanding this counterexample for debugging the specification is a complicated task for several reasons: (i) the counterexample can contain hundreds of actions, (ii) the debugging task is mostly achieved manually, and (iii) the counterexample does not give any clue on the state of the system (e.g., parallelism or data expressions) when the error occurs.

In 2016, we proposed a new approach that improves the usability of model checking by simplifying the comprehension of counterexamples. Our solution aims at keeping only actions in counterexamples that are relevant for debugging purposes. To do so, we first extract in the model all the counterexamples. Second, we define an analysis algorithm that identifies actions that makes the behaviour skip from incorrect to correct behaviours, making these actions relevant from a debugging perspective. Our approach is fully automated by a tool that we implemented and applied on real-world case studies from various application areas for evaluation purposes. A paper presenting these results has been accepted at an international conference.

6.3. Timed, Probabilistic, and Stochastic Extensions

6.3.1. On-the-fly Model Checking for Extended Regular Probabilistic Operators

Participant: Radu Mateescu.

In the context of the SENSATION project (see § 8.3.1.1), we study the specification and verification of quantitative properties of concurrent systems, which requires expressive and user-friendly property languages combining temporal, data-handling, and quantitative aspects.

In 2016, in collaboration with José Ignacio Requeno (Univ. Zaragoza, Spain), we aimed at facilitating the quantitative analysis of systems modeled as PTSs (Probabilistic Transition Systems) labeled by actions containing data values and probabilities. We proposed a new regular probabilistic operator that computes the probability measure of a path specified by a generalized regular formula involving arbitrary computations on data values. This operator, which subsumes the Until operators of PCTL (Probabilistic Computation Tree Logic) [41] and their action-based counterparts, can provide useful quantitative information about paths having certain (e.g., peak) cost values. We integrated the regular probabilistic operator into MCL and we devised an associated on-the-fly model checking method, based on a combined local resolution of linear and Boolean equation systems. We implemented the method in a prototype extension of the EVALUATOR model checker and experimented it on realistic PTSs modeling concurrent systems. This work led to a publication [22].

6.4. Component-Based Architectures for On-the-Fly Verification

6.4.1. Compositional Verification

Participant: Frédéric Lang.

The CADP toolbox contains various tools dedicated to compositional verification, among which EXP.OPEN, BCG_MIN, BCG_CMP, and SVL play a central role. EXP.OPEN explores on the fly the graph corresponding to a network of communicating automata (represented as a set of BCG files). BCG_MIN and BCG_CMP respectively minimize and compare behavior graphs modulo strong or branching bisimulation and their stochastic extensions. SVL (*Script Verification Language*) is both a high-level language for expressing complex verification scenarios and a compiler dedicated to this language.

In 2016, the *n among m* parallel composition operator "`par`" of the EXP language has been extended. Before the extension, the set of m processes among which any subset of size n could be synchronized on a given action was necessarily the set of all parallel processes composed by the "`par`" operator. From now on, by a slight extension of the syntax, this set of m processes can be defined as a subset of the parallel processes. Also, while n had to be strictly greater than 1, it can now also have value 0 (meaning that none of the m processes can perform the corresponding action) or 1 (meaning that each process can perform the corresponding action on its own, without synchronization). A bug in EXP.OPEN has been fixed and better messages are now emitted to warn the user about dubious usage of the "`par`" operator.

The SVL language has been extended to include the extended "`par`" operator. Two bugs have also been corrected.

6.4.2. Other Component Developments

Participants: Hubert Garavel, Frédéric Lang, Radu Mateescu, Wendelin Serwe.

Sustained effort was made to improve the documentation of the CADP toolbox. Various corrections have been brought to the CADP manual pages. A 27-page manual page explaining how the LOTOS language is implemented has been written, and the manual pages of the CAESAR and CAESAR.ADT compilers have been also updated. To make documentation more readable, the EVALUATOR3, and EVALUATOR4 manual pages have been splitted each in two parts, so as to better distinguish between the languages (namely, MCL3 and MCL) and their model checkers. The CADP distribution has been made leaner by keeping only the essential papers, and the "publications" and "tutorial" pages of the CADP Web site have been enriched and reorganized.

The conventions for string notations to represent "raw" values (i.e., values whose type is not a predefined one, but a custom type defined by the user) have been improved, together with the associated conversion algorithms for reading/writing raw values from/to strings. The BCG_WRITE manual page has been updated to more accurately describe how label fields of type "raw" are parsed. The behaviour of the functions `bcg_character_scan()`, `bcg_string_scan()`, `bcg_real_scan()`, and `bcg_raw_scan()` has been carefully revised, and all the BCG libraries and tools (especially BCG_IO) have been modified to follow the new conventions and emit better diagnostics when label fields contain incorrect notations of raw values. Also, BCG_IO has been enhanced so that very long execution sequences can be converted into SEQ or LOTOS files without causing stack overflow.

Finally, enhancements and bug fixes have been brought to other CADP components, including CADP_MEMORY, EUCALYPTUS, INSTALLATOR, OCIS, RFL, TST, and XTL. The style files for the various editors supported by CADP have been updated to take into account the recent features added to LNT. The predefined MCL libraries of the EVALUATOR model checker have been modified to generate more explanatory diagnostics for the inevitability operators.

Although CADP is mostly used on Linux, specific effort has been made to target other execution platforms. Concerning macOS: CADP now supports the recent versions 10.10 ("Yosemite"), 10.11 ("El Capitan"), and 10.12 ("Sierra"). Concerning Windows: changes have been brought to support Windows 10 and the 64-bit version of Cygwin (previously, only the 32-bit version was supported). Other adaptations were required to handle the recent versions of Cygwin packages, MinGW C compiler, and Mintty shell, as well as the case where Cygwin is not installed in "`C:\`", but in either "`C:\Cygwin`" or "`C:\Cygwin64`".

6.5. Real-Life Applications and Case Studies

6.5.1. Reconfiguration and Resilience of Distributed Cloud Applications

Participants: Umar Ozeer, Gwen Salaün.

In the context of a collaboration with Orange Labs, an Ensimag student (Bakr Derrazi) supervised by Xavier Etchevers and Gwen Salaün, has made his internship from February 2016 to July 2016 at Orange Labs. As a result, we have proposed a first solution and prototype for detecting and repairing failures of data-centric applications distributed in the cloud. A PhD thesis (Umar Ozeer) has started on this subject in November 2016.

6.5.2. Activity Detection in a Smart Home

Participants: Frédéric Lang, Radu Mateescu.

In collaboration with Paula-Andrea Lago-Martinez and Claudia Roncancio (SIGMA team, LIG) and with Nicolas Bonnefond (PERVASIVE INTERACTION team, Inria and LIG), we study how formal methods can help to analyze logs of events obtained from the many sensors and actuators installed in the Amiqual4Home smart home.

In 2016, we considered using the MCL temporal logic to detect the start and end of activities in a log, such as cooking or taking a shower. We applied our tools on a log containing about $140,000$ events that had been generated over 10 days of living in the smart home. This preliminary study has shown that the MCL temporal logic is sufficiently rich to enable an easy specification of the searched activities, notably thanks to its multiple extensions such as macro definitions, parameterized fixed point operators, and data handling mechanisms. The particularly long length of the analyzed logs also enabled us to improve some of the CADP tools, so that they better scale up. This work led to an article submitted to an international conference.

6.5.3. Other Case Studies

Participant: Hubert Garavel.

The demo examples of CADP, which have been progressively accumulated since the origins of the toolbox, are a showcase for the multiple capabilities of CADP, as well as a test bed to assess the new features of the toolbox. In 2016, the effort to maintain and enhance these demos has been pursued. The demo 12 (Message Authentication Algorithm) and demo 31 (SCSI-2 bus arbitration protocol) have been manually translated from LOTOS to LNT. Additionally, demo 12 has been deeply revised by simplifying its LOTOS, LNT, and C code, by taking advantage of the imperative-programming features of LNT, and by enriching the LNT specification with the test cases contained in the original MAA description. This allowed to detect and correct a mistake in the C code implementing function `HIGH_MUL()`. Other CADP demos (namely demos 05, 16, and 36) have also been simplified and/or enhanced in various ways.

CORSE Project-Team

6. New Results

6.1. Simplification and Run-time Resolution of Data Dependence Constraints for Loop Transformations

Participants: Diogo Nunes Sampaio, Alain Ketterlin [Inria CAMUS], Louis-Noël Pouchet [CSU, USA], Fabrice Rastello.

Loop optimizations such as tiling, thread-level parallelization or vectorization are essential transformations to improve performance. It is needed to compute dependence information at compile-time to assess their validity, but in many real situations, static dependence analysis fails to provide precise enough information. Part of the reason for this failure comes from the need to handle polynomial constraints in the dependence computation problem: such polynomial constraints can arise from linearized array accesses, typical in compilers IR such as LLVM-IR. In this scenario, the compiler will often be unable to apply aggressive transformations due to lack of conclusive static dependence analysis. This work tackles the problem of eliminating quantifiers in systems of inequalities using polynomial constraints. In particular, we design a quantifier elimination scheme on integer multivariate-polynomials, which can aid application of off-the-shelf polyhedral transformations on a larger class of programs, that holds polynomial memory access and affine loop bounds. We make a significant leap in accuracy compared to prior approaches, enabling to implement a hybrid optimizing compilation scheme. In this scheme, a test is evaluated at run-time to determine the legality of the program transformation chosen by the compiler, falling back to executing the original code if the test fails. This test integrates all may-dependences, involving polynomial inequalities, and is simplified by quantifier elimination at compile-time using our techniques. The preciseness of the presented scheme and the low run-time overhead of the test are key to make this approach realistic. We experimentally validate our technique on 25 benchmarks using complex loop transformations, achieving negligible overhead. Preciseness is assessed by the observed success of generated test in practical cases. We compare our variable elimination technique to other existing tools and demonstrate we achieve better precision when dealing with polynomial memory accesses.

This work is the fruit of the collaboration 8.4 with OSU.

6.2. A bounded memory allocator for software-defined global address spaces

Participants: François Gindraud, Fabrice Rastello, Albert Cohen [ENS Ulm], Francois Broquedis.

This work is about the design of a memory allocator targeting manycore architectures with distributed memory. Among the family of Multi Processor System on Chip (MPSoC), these devices are composed of multiple nodes linked by an on-chip network; most nodes have multiple processors sharing a small local memory. While MPSoC typically excel on their performance-per-Watt ratio, they remain hard to program due to multilevel parallelism, explicit resource and memory management, and hardware constraints (limited memory, network topology).

Typical programming frameworks for MPSoC leave much target-specific work to the programmer: combining threads or node-local OpenMP, software caching, explicit message passing (and sometimes, routing), with non-standard interfaces. More abstract, automatic frameworks exist, but they target large-scale clusters and do not model the hardware constraints of MPSoC.

This memory allocator is one component of a larger runtime system, called Givy 5.3 , to support dynamic task graphs with automatic software caching and data-driven execution on MPSoC. To simplify the programmer's view of memory, both runtime and program data objects live in a Global Address Space (GAS). To avoid address collisions when objects are dynamically allocated, and to manage virtual memory mappings across nodes, a GAS-aware memory allocator is required. This work proposes such an allocator with the following properties: (1) it is free of inter-node synchronizations; (2) its node-local performance match that of state-of-the-art shared-memory allocators; (3) it provides node-local mechanisms to implement inter-node software caching within a GAS; (4) it is well suited for small memory systems (a few MB per node).

This work has been presented at the international conference ISMM 2016 [16].

6.3. On Fusing Recursive Traversals of K-d Trees

Participants: Samyam Rajbhandari [OSU, USA], Jinsung Kim [OSU, USA], Sriram Krishnamoorthy [PNNL, USA], Louis-Noel Pouchet [CSU, USA], Fabrice Rastello, Robert J. Harrison [Stony Brook, USA], P. Sadayappan [OSU, USA].

Loop fusion is a key program transformation for data locality optimization that is implemented in production compilers. But optimizing compilers for imperative languages currently cannot exploit fusion opportunities across a set of recursive tree traversal computations with producer-consumer relationships. In this work, we develop a compile-time approach to dependence characterization and program transformation to enable fusion across recursively specified traversals over k-d trees. We present the FuseT source-to-source code transformation framework to automatically generate fused composite recursive operators from an input program containing a sequence of primitive recursive operators. We use our framework to implement fused operators for MADNESS, Multiresolution Adaptive Numerical Environment for Scientific Simulation. We show that locality optimization through fusion can offer significant performance improvement.

This work is the fruit of the collaboration 8.4 with OSU. The specific work on FuseT has been presented to the international conference CC 2016 [32] and the more general work on the improvement of MADNESS at the ACM/IEEE international conference SC 2016 [20].

6.4. Effective Padding of Multidimensional Arrays to Avoid Cache Conflict Misses

Participants: Changwan Hong [OSU, USA], Wenlei Bao [OSU, USA], Albert Cohen [Inria PARKAS], Sriram Krishnamoorthy [PNNL, USA], Louis-Noel Pouchet [CSU, USA], Fabrice Rastello, J. Ramanujam [LSU, USA], P. Sadayappan [OSU, USA].

Caches are used to significantly improve performance. Even with high degrees of set associativity, the number of accessed data elements mapping to the same set in a cache can easily exceed the degree of associativity. This can cause conflict misses and lower performance, even if the working set is much smaller than cache capacity. Array padding (increasing the size of array dimensions) is a well-known optimization technique that can reduce conflict misses. In this work, we develop the first algorithms for optimal padding of arrays aimed at a set-associative cache for arbitrary tile sizes. In addition, we develop the first solution to padding for nested tiles and multi-level caches. Experimental results with multiple benchmarks demonstrate a significant performance improvement from padding.

This work is the fruit of the collaboration 8.4 with OSU. It has been presented at the ACM international conference PLDI 2016 [29].

6.5. PolyCheck: Dynamic Verification of Iteration Space Transformations on Affine Programs

Participants: Sriram Krishnamoorthy [PNNL], Bao Wenlei [OSU], Louis-Noël Pouchet [UCLA], P. Sadayappan [OSU], Fabrice Rastello.

High-level compiler transformations, especially loop transformations, are widely recognized as critical optimizations to restructure programs to improve data locality and expose parallelism. Guaranteeing the correctness of program transformations is essential, and to date three main approaches have been developed: proof of equivalence of affine programs, matching the execution traces of programs, and checking bit-by-bit equivalence of program outputs. Each technique suffers from limitations in the kind of transformations supported, space complexity, or the sensitivity to the testing dataset. In this work, we take a novel approach that addresses all three limitations to provide an automatic bug checker to verify any iteration reordering transformations on affine programs, including non-affine transformations, with space consumption proportional to the original

program data and robust to arbitrary datasets of a given size. We achieve this by exploiting the structure of affine program control- and data-flow to generate at compile-time lightweight checker code to be executed within the transformed program. Experimental results assess the correctness and effectiveness of our method and its increased coverage over previous approaches.

This work is the fruit of the collaboration 8.4 with OSU and was presented at ACM POPL'16 [14].

6.6. Modularizing Crosscutting Concerns in Component-Based Systems

Participants: Antoine El-Hokayem, Yliès Falcone, Mohamad Jaber [American University of Beirut, Lebanon].

We define a method to modularize crosscutting concerns in the Behavior Interaction Priority (BIP) component-based framework. Our method is inspired from the Aspect Oriented Programming (AOP) paradigm which was initially conceived to support the separation of concerns during the development of monolithic systems. BIP has a formal operational semantics and makes a clear separation between architecture and behavior to allow for compositional and incremental design and analysis of systems. We thus distinguish local from global aspects. Local aspects model concerns at the component level and are used to refine the behavior of components. Global aspects model concerns at the architecture level, and hence refine communications (synchronization and data transfer) between components. We formalize global aspects as well as their integration into a BIP system through rigorous transformation primitives and overview local aspects. We present AOP-BIP, a tool for Aspect-Oriented Programming of BIP systems, and demonstrate its use to modularize logging, security, and fault-tolerance in a network protocol.

This work results of the collaboration with American University of Beirut (Lebanon) and was presented at SEFM 2016 [15].

6.7. Predictive runtime enforcement

Participants: Srinivas Pinisetty [Aalto University, Finland], Viorel Preoteasa [Aalto University, Finland], Stavros Tripakis [Aalto University, Finland], Thierry Jéron [Inria Rennes, France], Yliès Falcone, Hervé Marchand [Inria Rennes, France].

Runtime enforcement (RE) is a technique to ensure that the (untrustworthy) output of a black-box system satisfies some desired properties. In RE, the output of the running system, modeled as a stream of events, is fed into an enforcement monitor. The monitor ensures that the stream complies with a certain property, by delaying or modifying events if necessary. This work deals with 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 RE framework. All our results are formalized and proved in the Isabelle theorem prover.

This work was presented at SAC-SVT 2016 [19].

6.8. Third International Competition on Runtime Verification

Participants: Giles Reger [University of Manchester, UK], Sylvain Hallé [The University of Québec at Chicoutimi, Canada], Yliès Falcone.

We report on the Third International Competition on Runtime Verification (CRV-2016). The competition was held as a satellite event of the 16th International Conference on Runtime Verification (RV'16). The competition consisted of two tracks: offline monitoring of traces and online monitoring of Java programs. The intention was to also include a track on online monitoring of C programs but there were too few participants to proceed with this track. This report describes the format of the competition, the participating teams, the submitted benchmarks and the results. We also describe our experiences with transforming trace formats from other tools into the standard format required by the competition and report on feedback gathered from current and past participants and use this to make suggestions for the future of the competition.

This work was presented at RV 2016 [13].

6.9. Monitoring Multi-threaded Component-Based Systems

Participants: Hosein Nazarpour [Verimag, France], Yliès Falcone, Saddek Bensalem [Verimag, France], Marius Bozga [Verimag, France], Jacques Combaz [Verimag, France].

This work addresses the monitoring of logic-independent linear-time user-provided properties on multi-threaded component-based systems. We consider intrinsically independent components that can be executed concurrently with a centralized coordination for multiparty interactions. In this context, the problem that arises is that a global state of the system is not available to the monitor. A naive solution to this problem would be to plug a monitor which would force the system to synchronize in order to obtain the sequence of global states at runtime. Such solution would defeat the whole purpose of having concurrent components. Instead, we reconstruct on-the-fly the global states by accumulating the partial states traversed by the system at runtime. We define formal transformations of components that preserve the semantics and the concurrency and, at the same time, allow to monitor global-state properties. Moreover, we present RVMT-BIP, a prototype tool implementing the transformations for monitoring multi-threaded systems described in the BIP (Behavior, Interaction, Priority) framework, an expressive framework for the formal construction of heterogeneous systems. Our experiments on several multi-threaded BIP systems show that RVMT-BIP induces a cheap runtime overhead.

This work was presented at iFM 2016 [18].

6.10. Decentralized Enforcement of Artifact Lifecycles

Participants: Sylvain Hallé [The University of Québec at Chicoutimi, Canada], Raphaël Khoury [The University of Québec at Chicoutimi, Canada], Antoine El-Hokayem, Yliès Falcone.

Artifact-centric workflows describe possible executions of a business process through constraints expressed from the point of view of the documents exchanged between principals. A sequence of manipulations is deemed valid as long as every document in the workflow follows its prescribed lifecycle at all steps of the process. So far, establishing that a given workflow complies with artifact lifecycles has mostly been done through static verification, or by assuming a centralized access to all artifacts where these constraints can be monitored and enforced. We propose an alternate method of enforcing document lifecycles that requires neither static verification nor single-point access. Rather, the document itself is designed to carry fragments of its history, protected from tampering using hashing and public-key encryption. Any principal involved in the process can verify at any time that a document's history complies with a given lifecycle. Moreover, the proposed system also enforces access permissions: not all actions are visible to all principals, and one can only modify and verify what one is allowed to observe.

This work was presented at EDOC 2016 [17].

6.11. Runtime enforcement of regular timed properties by suppressing and delaying events

Participants: Yliès Falcone, Thierry Jéron [Inria Rennes, France], Hervé Marchand [Inria Rennes, France], Srinivas Pinisetty [Aalto University, Finland].

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. The feasibility of enforcement monitoring for timed properties is validated by prototyping the synthesis of enforcement monitors from timed automata.

This work was published in the journal Science of Computer Programming [8].

6.12. Organising LTL monitors over distributed systems with a global clock

Participants: Christian Colombo [University of Malta, Malta], Yliès Falcone.

Users wanting to monitor distributed systems often prefer to abstract away the architecture of the system by directly specifying correctness properties on the global system behaviour. To support this abstraction, a compilation of the properties would not only involve the typical choice of monitoring algorithm, but also the organisation of submonitors across the component network. Existing approaches, considered in the context of LTL properties over distributed systems with a global clock, include the so-called orchestration and migration approaches. In the orchestration approach, a central monitor receives the events from all subsystems. In the migration approach, LTL formulae transfer themselves across subsystems to gather local information. We propose a third way of organising submonitors: choreography, where monitors are organised as a tree across the distributed system, and each child feeds intermediate results to its parent. We formalise choreography-based decentralised monitoring by showing how to synthesise a network from an LTL formula, and give a decentralised monitoring algorithm working on top of an LTL network. We prove the algorithm correct and implement it in a benchmark tool. We also report on an empirical investigation comparing these three approaches on several concerns of decentralised monitoring: the delay in reaching a verdict due to communication latency, the number and size of the messages exchanged, and the number of execution steps required to reach the verdict.

This work was published in the journal Formal Methods in System Design [6].

6.13. Decentralised LTL monitoring

Participants: Andreas Bauer [TU Munich, Software and Systems Engineering Munich, Germany], Yliès Falcone.

Users wanting to monitor distributed or component-based systems often perceive them as monolithic systems which, seen from the outside, exhibit a uniform behaviour as opposed to many components displaying many local behaviours that together constitute the system's global behaviour. This level of abstraction is often reasonable, hiding implementation details from users who may want to specify the system's global behaviour in terms of a linear-time temporal logic (LTL) formula. However, the problem that arises then is how such a specification can actually be monitored in a distributed system that has no central data collection point, where all the components' local behaviours are observable. In this case, the LTL specification needs to be decomposed into sub-formulae which, in turn, need to be distributed amongst the components' locally attached monitors, each of which sees only a distinct part of the global behaviour. The main contribution of this work is an algorithm for distributing and monitoring LTL formulae, such that satisfaction or violation of specifications can be detected by local monitors alone. We present an implementation and show that our algorithm introduces only a negligible delay in detecting satisfaction/violation of a specification. Moreover, our practical results show that the communication overhead introduced by the local monitors is generally lower than the number of messages that would need to be sent to a central data collection point. Furthermore, our experiments strengthen the argument that the algorithm performs well in a wide range of different application contexts, given by different system/communication topologies and/or system event distributions over time.

This work was published in the journal Formal Methods in System Design [4].

6.14. Using data dependencies to improve task-based scheduling strategies on NUMA architectures

Participants: Philippe Virouleau, François Broquedis, Thierry Gautier [Inria, AVALON], Fabrice Rastello.

The recent addition of data dependencies to the OpenMP 4.0 standard provides the application programmer with a more flexible way of synchronizing tasks. Using such an approach allows both the compiler and the runtime system to know exactly which data are read or written by a given task, and how these data will be used through the program lifetime. Data placement and task scheduling strategies have a significant impact on performances when considering NUMA architectures. While numerous studies focus on these topics, none of them has made extensive use of the information available through dependencies. One can use this information to modify the behavior of the application at several levels : during initialization to control data placement and during the application execution to dynamically control both the task placement and the tasks stealing strategy, depending on the topology. This work introduces several heuristics for these strategies, their implementations in the xkaapi OpenMP runtime system and the performances on linear algebra applications executed on a 192-core NUMA machine. Such approaches report noticeable performance improvement when considering both the architecture topology and the tasks data dependencies.

This work has been presented at the international conference EuroPar'2016 [22].

6.15. Description, Implementation and Evaluation of an Affinity Clause for Task Directives

Participants: Philippe Virouleau, Adrien Roussel [IFPEN], François Broquedis, Thierry Gautier [Inria, AVALON], Fabrice Rastello, Jean-Marc Gratien [IFPEN].

This work extends the affinity-based scheduling we proposed at the Europar 2016 conference to fit the philosophy of OpenMP programming. On this topic, OpenMP does not provide a lot of flexibility to the programmer yet, which lets the runtime system decide where a task should be executed. In this work, we propose our own interpretation of the new affinity clause for the task directive, which is being discussed by the OpenMP Architecture Review Board. This clause enables the programmer to give hints to the runtime about tasks placement during the program execution, which can be used to control the data mapping on the architecture. In our proposal, the programmer can express affinity between a task and the following resources: a thread, a NUMA node, and a data. We provide an implementation of this proposal in the Clang-3.8 compiler, and an implementation of the corresponding extensions in the xkaapi OpenMP runtime system.

This work has been presented at the international workshop on OpenMP IWOMP'2016 [23].

6.16. Design methodology for workload-aware loop scheduling strategies based on genetic algorithm and simulation

Participants: Pedro H. Penna [PUC Minas], Márcio Castro [UFSC], Henrique C. Freitas [PUC Minas], François Broquedis, Jean-François Méhaut.

In high-performance computing, the application's workload must be evenly balanced among threads to deliver cutting-edge performance and scalability. In OpenMP, the load balancing problem arises when scheduling loop iterations to threads. In this context, several scheduling strategies have been proposed, but they do not take into account the input workload of the application and thus turn out to be suboptimal. In this work, we introduce a design methodology to propose, study, and assess the performance of workload-aware loop scheduling strategies. In this methodology, a genetic algorithm is employed to explore the state space solution of the problem itself and to guide the design of new loop scheduling strategies, and a simulator is used to evaluate their performance. As a proof of concept, we show how the proposed methodology was used to propose and study a new workload-aware loop scheduling strategy named smart round-robin (SRR). We implemented this strategy into GNU Compiler Collection's OpenMP runtime. We carry out several experiments to validate the simulator and to evaluate the performance of SRR. Our experimental results show that SRR may deliver up to 37.89% and 14.10% better performance than OpenMP's dynamic loop scheduling strategy in the simulated environment and in a real-world application kernel, respectively.

This work is presented in the CCPE journal [9].

6.17. The Mont-Blanc prototype: An Alternative Approach for HPC Systems

Participants: Brice Videau, Kevin Pouget, Jean-François Méhaut.

The evolution of High-Performance Computing (HPC) systems is driven by the need of reducing time-to-solution and increasing the resolution of models and problems being solved by a particular program. Important milestones from the HPC system performance perspective were achieved using commodity technology. Examples are the ASCI Red and the Roadrunner supercomputers, which broke the 1 TFLOPS and 1 PFLOPS barriers, respectively. These systems showed how commodity technology could be used to take the next step in HPC system architecture.

Driven by a much larger market, commodity components evolve faster than their special-purpose counterparts, eventually achieving the same performance and eventually surpassing or replacing them. For this reason, RISC processors displaced vector processors, and x86 displaced RISC.

Nowadays commodity is in the embedded / mobile processor segment. Mobile processors develop fast, and are still not at a point of diminishing performance improvements from new designs. Furthermore, they progressively incorporate the capabilities required for HPC.

The embedded market size and endless customer requirements allow for constant investments into innovative designs, and rapid testing and adoption of new technologies. For example, LPDDR memory technology was first introduced in the mobile domain and has recently been proposed as a memory solution for energy proportional servers.

The Mont-Blanc project aims at providing an alternative HPC system solution based on the current commodity technology: mobile chips. As a demonstrator of such an approach, the project designed, built, and set-up a 1080-node HPC cluster made of Samsung Exynos 5250 SoCs. The Mont-Blanc project established the following goals: to design and deploy a sufficiently large HPC prototype system based on the current mobile commodity technology; to port and optimize the software stack, and enable its use for HPC; to port and optimize a set of HPC applications to be run at this HPC system.

Comparing the Mont-Blanc prototype to a contemporary supercomputer, MareNostrum III, reveals that a single-socket Mont-Blanc node is 9x slower than a dual-socket MareNostrum III node, while saving up to 40% of energy. MPI parallel applications show a 3.5x slowdown when running with the same number of MPI ranks on both machines, while consuming 9% less energy on the Mont-Blanc prototype on average. When targeting the same execution time, the Mont-Blanc prototype offers 12.5% space savings.

This work was funded by the European Commission with the Mont-Blanc projects 8.3.1.1 . This scientific result was presented at the SuperComputing Conference SC'2016 in Salt Lake City [31]. The paper was selected as a *best paper finalist*.

6.18. Control of Autonomid Parallelism on Software Transactional Memory

Participants: Naweiluo Zhou, Gwenaël Delaval [Univ. Grenoble Alpes, Associate Professor, Ctrl-A Inria team], Bogdan Robu [Univ. Grenoble Alpes, Associate Professor, Gipsa Laboratory], Eric Rutten [Inria, Rsearcher, Ctrl-A Inria team], Jean-François Méhaut.

Parallel programs need to manage the trade-off between the time spent in synchronization and computation. A high parallelism may decrease computing time while increase synchronization cost among threads. A way to improve program performance is to adjust parallelism to balance conflicts among threads. However, there is no universal rule to decide the best parallelism for a program from an offline view. Furthermore, an offline tuning is error-prone. Hence, it becomes necessary to adopt a dynamic tuning-configuration strategy to better manage a STM system. Software Transactional Memory (STM) has emerged as a promising technique, which bypasses locks, to address syn- chronization issues through transactions. Autonomic computing offers designers a framework of methods and techniques to build automated systems with well-mastered behaviours. Its key idea is to implement feedback control loops to design safe, efficient and predictable controllers, which enable monitoring and adjusting controlled systems dynamically while keeping overhead low. We propose to design feedback control loops to automate the choice of parallelism level at runtime to diminish program execution time.

This work is funded by the Persyval laboratory (LabEx) and the HPES team 8.1.2 . This scientific result is part of the Naweiluo Zhou's thesis. The thesis was defended in October 2016 [2]. This work was presented in the HPCS conference [25]. The paper was selected as *best paper finalist*. The Naweiluo Zhou's work is also presented at the ICAC conference.

6.19. Evaluating the SEE sensitivity of a 45nm SOI Multi-core Processor due to 14 MeV Neutrons

Participants: Pablo Ramos [Univ. Grenoble Alpes and ESPE Ecuador, PhD student TIMA Laboratory], Vanessa Vargas [Univ. Grenoble Alpes and ESPE Ecuador, PhD student TIMA Laboratory], Maud Baylac [CNRS, IN2P3, LSPSC Laboratory], Francesca Villa [CNRS, IN2P3, LSPSC Laboratory], Nacer-Eddine Zergainoh [Univ. Grenoble Alpes, Associate Professor, TIMA Laboratory], Jean-François Méhaut, Raoul Velazco [CNRS, Senior Scientist, TIMA Laboratory].

The aim of this work is to evaluate the SEE sensitivity of a multi-core processor having implemented ECC and parity in their cache memories. Two different application scenarios are studied. The first one configures the multi-core in Asymmetric Multi-Processing mode running a memory-bound application, whereas the second one uses the Symmetric Multi-Processsing mode running a CPU-bound application. The experiments were validated through radiation ground testing performed with 14 MeV neutrons on the Freescale P2041 multi-core manufactured in 45nm SOI technology. A deep analysis of the observed errors in cache memories was carried-out in order to reveal vulnerabilities in the cache protection mechanisms. Critical zones like tag addresses were affected during the experiments. In addition, the results show that the sensitivity strongly depends on the application and the multi-processsing mode used.

This work is part of the STIC Amsud EnergySFE project 8.4.3 . These results are published in the IEEE Transactions on Nuclear Science [10].

DICE Team

6. New Results

6.1. Intermediation platforms

Our study of the geopolitics of intermediation aims at grasping the balance of power between platforms, as well as between states - in their relation to platforms - and between platforms and states. We have extended our studies with insights from law in [1] and economics in [2]. We have tuned the metrics we already had in order to better grasp the economic weight of intermediation economy. As we did so, we improved our understanding of the social weight of intermediation platforms and the legal issues which they raise.

Our focus has turned to the analysis of public and private policies and their relation to the development of intermediation platforms. In [3], we study a set of cases ranging from the Safe Harbour to the right to be forgotten. Using the "coalition framework" as our analysis framework, we identify the actors influencing policy-making and potential reasons for the success of failure of policies. Such failures include forbidding innovation or preventing public bodies from stepping up their digital capabilities.

Our work has been intrinsically interdisciplinary, the main result of our work is a global modal of intermediation platforms and their economy, presented in [6]. This model helps to understand the current issues raised by the ubderization for instance.

6.2. Development of platforms

Dice team designs software architectures for intermediation platforms. C3PO and BitBallot targets spontaneous and ephemeral social networks whereas Jumplyn focuses on pure central based system. All these architectures share a common JavaScript layout both at the client and the server sides. In the research context we validate state-of-the art technologies promoted by web leaders such as Google AngularJS, Facebook ReactJS and many others such as Netflix, Wallmart, or the Linux foundation for node.js. The Web environment raises many big issues since all equipments are basically connected to the Internet and the balance between end-user equipment cost and processing power is still a work in progress. Our main research track in such context is to find proper software toolkits hiding Web complexity. We mainly focus on time jitter, cornerstone of Web development, since it implies both end-user and network TCP indecisions. Due to this jitter combination the Web programming model has mutated toward the promises paradigm. It is a complex event based development model provided without external API help. It handles future execution whether successful or not, in a time jittered context. AngularJS, ReactJS, CoffeeScript, NodeJS, MongoDB, ElasticSearch are all time jitter compliant technologies designed for the Web constrains and revolutionising the way we build intermediation platforms.

In C3PO, we tested application in real conditions during the marathon of Vannes and the semi-marathon of Beaune. A few hundred users have downloaded and use the application. The returns on this one are rather positive. [4].

Our joint work with Worldline explores the promises paradigm model to enable automation extraction of independent micro-service. These micro-services called fluxion [9], from the contraction of flow and functions, may be dynamically and transparently moved over a cluster of servers. Our novelty resides in the fact that the original code is not redesigned for the cluster architecture. Fluxion are extracted from the initial code, and an equivalence is maintained between the initially promissified code and the fluxionized one. Code has two facets, a promise one, used to express software services and a fluxion one, used to express software bottlenecks [5].

Eventually our work with Jumplyn explores complex centralised social network. We want to design a software system to later support our technical research hot topics. The target theme is a software platform that helps students handle their projects. University depends more and more on external resources to teach students. Theses resources are both known by students and their teachers, and the pace and range of explored technologies leads to difficulties in teaching state-of-the-art subjects. The more dedicated a professor needs to be in his research activity, the more broad knowledge he has to teach. For instance 20 years ago one could cope software development teaching with one or two programming languages. Nowadays, a single code involves more then four programming languages to be fully understood. This technology spreading issue stands still in many teaching domains, since past technologies are still actives and future one are promising. We build Jumplyn to cope with this unbalanced game. To help student improving their project and avoid working with obsolete technologies, and to help teacher face the universal and inexpensive availability of knowledge. Jumplyn is a complex JavaScript development stack that collects resources for improving student work and providing services to help them from day to day activities. The current stack integrates the following technologies : MaterialDesign, AngularJS, CoffeeScript, NodeJs, MongoDb, ElasticSearch. Managing and developing software service above this stack is a complex research issue for a small sized development team. We do not have any publication on Jumplyn since our first goal is to build a support intermediation platform to study classical issues such as recommendation or web crawling, scraping and indexation with our own sources of raw data.

PRIVATICS Project-Team

6. New Results

6.1. MobileAppScrutinator: A Simple yet Efficient Dynamic Analysis Approach for Detecting Privacy Leaks across Mobile OSs

Participants: Jagdish Achara, Vincent Roca, Claude Castelluccia.

Smartphones, the devices we carry everywhere with us, are being heavily tracked and have undoubtedly become a major threat to our privacy. As " Tracking the trackers " has become a necessity, various static and dynamic analysis tools have been developed in the past. However, today, we still lack suitable tools to detect, measure and compare the ongoing tracking across mobile OSs. To this end, we propose MobileAppScrutinator [24], based on a simple yet efficient dynamic analysis approach, that works on both Android and iOS (the two most popular OSs today). To demonstrate the current trend in tracking, we select 140 most representative Apps available on both Android and iOS AppStores and test them with MobileAppScrutinator. In fact, choosing the same set of apps on both Android and iOS also enables us to compare the ongoing tracking on these two OSs. Finally, we also discuss the effectiveness of privacy safeguards available on Android and iOS. We show that neither Android nor iOS privacy safeguards in their present state are completely satisfying.

6.2. MyTrackingChoices: Pacifying the Ad-Block War by Enforcing User Privacy Preferences

Participants: Jagdish Achara, Claude Castelluccia.

Free content and services on the Web are often supported by ads. However, with the proliferation of intrusive and privacy-invasive ads, a significant proportion of users have started to use ad blockers. As existing ad blockers are radical (they block all ads) and are not designed taking into account their economic impact, ad-based economic model of the Web is in danger today. In this paper, we target privacy-sensitive users and provide them with fine-grained control over tracking. Our working assumption is that some categories of web pages (for example, related to health, religion, etc.) are more privacy-sensitive to users than others (education, science, etc.). Therefore, our proposed approach consists in providing users with an option to specify the categories of web pages that are privacy-sensitive to them and block trackers present on such web pages only. As tracking is prevented by blocking network connections of third-party domains, we avoid not only tracking but also third-party ads. Since users will continue receiving ads on web pages belonging to non-sensitive categories, our approach essentially provides a trade-off between privacy and economy. To test the viability of our solution, we implemented it as a Google Chrome extension, named MyTrackingChoices (available on Chrome Web Store). Our real-world experiments with MyTrackingChoices [23] show that the economic impact of ad blocking exerted by privacy-sensitive users can be significantly reduced.

6.3. Security or privacy?

Participants: Amrit Kumar, Cédric Lauradoux.

Security softwares such as anti-viruses, IDS or filters help Internet users to protect their privacy from hackers. The cost of this protection is that the users privacy is stripped away by the companies providing these security solutions. Currently, Internet users can choose between no security with the risk of being hacked or use security software and lose all personal data to security companies. As a example of this dilemma, we analyze the solution proposed by Google and Yandex for Safe Browsing [8] and shows that their privacy policies do not match the reality: Google can perform tracking.

6.4. Near-Optimal Fingerprinting with Constraints

Participants: Gabor Gulyas, Gergely Acs, Claude Castelluccia.

Several recent studies have demonstrated that people show large behavioural uniqueness. This has serious privacy implications as most individuals become increasingly re-identifiable in large datasets or can be tracked while they are browsing the web using only a couple of their attributes, called as their fingerprints. Often, the success of these attacks depend on explicit constraints on the number of attributes learnable about individuals, i.e., the size of their fingerprints. These constraints can be budget as well as technical constraints imposed by the data holder. For instance, Apple restricts the number of applications that can be called by another application on iOS in order to mitigate the potential privacy threats of leaking the list of installed applications on a device. In [15], we address the problem of identifying the attributes (e.g., smartphone applications) that can serve as a fingerprint of users given constraints on the size of the fingerprint. We give the best fingerprinting algorithms in general, and evaluate their effectiveness on several real-world datasets. Our results show that current privacy guards limiting the number of attributes that can be queried about individuals is insufficient to mitigate their potential privacy risks in many practical cases.

6.5. Data anonymization Evaluation

Participants: Claude Castelluccia, Gergely Acs, Daniel Le Metayer.

Anonymization is a critical issue because data protection regulations such as the European Directive 95/46/EC and the European General Data Protection Regulation (GDPR) explicitly exclude from their scope anonymous information" and personal data rendered anonymous"1. However, turning this general statement into effective criteria is not an easy task. In order to facilitate the implementation of this provision, the Working Party 29 (WP29) has published in April 2014 an Opinion on Anonymization Techniques. This Opinion puts forward three criteria corresponding to three risks called respectively "singling out", "linkability" and "inference". In this work, we first evaluated these criteria and showed that they are neither necessary nor effective to decide upon the robustness of an anonymization algorithm. Then we proposed an alternative approach relying on the notions of acceptable versus unacceptable inferences in [4] and we introduced differential testing, a practical way to implement this approach using machine learning techniques.

6.6. Wi-Fi and privacy

Participants: Mathieu Cunche, Celestin Matte.

- **Geolocation spoofing attack** We present several novel techniques to track (unassociated) mobile devices by abusing features of the Wi-Fi standard. This shows that using random MAC addresses, on its own, does not guarantee privacy. First, we show that information elements in probe requests can be used to fingerprint devices. We then combine these fingerprints with incremental sequence numbers, to create a tracking algorithm that does not rely on unique identifiers such as MAC addresses. Based on real-world datasets, we demonstrate that our algorithm can correctly track as much as 50% of devices for at least 20 minutes. We also show that commodity Wi-Fi devices use predictable scrambler seeds. These can be used to improve the performance of our tracking algorithm. Finally, we present two attacks that reveal the real MAC address of a device, even if MAC address randomization is used. In the first one, we create fake hotspots to induce clients to connect using their real MAC address. The second technique relies on the new 802.11u standard, commonly referred to as Hotspot 2.0, where we show that Linux and Windows send Access Network Query Protocol (ANQP) requests using their real MAC address.
- **Extraction of semantical information from Wi-Fi network identifiers** MAC address randomization in Wi-Fi-enabled devices has recently been adopted to prevent passive tracking of mobile devices. However, Wi-Fi frames still contain fields that can be used to fingerprint devices and potentially allow tracking. Panoptiphone is a tool inspired by the web browser fingerprinting tool Panopticlick, which aims to show the identifying information that can be found in the frames broadcast by

a Wi-Fi-enabled device. Information is passively collected from devices that have their Wi-Fi interface enabled, even if they are not connected to an access point. Panoptiphone uses this information to create a fingerprint of the device and empirically evaluate its uniqueness among a database of fingerprints. The user is then shown how much identifying information its device is leaking through Wi-Fi and how unique it is.

6.7. Formal and legal issues of privacy

Participant: Daniel Le Metayer.

- **Privacy by design** Based on our previous work on the use of formal methods to reason about privacy properties of system architectures, we have proposed a logic to reason about properties of architectures including group authentication functionalities. By group authentication, we mean that a user can authenticate on behalf of a group of users, thereby keeping a form of anonymity within this set. Then we show that this extended framework can be used to reason about privacy properties of a biometric system in which users are authenticated through the use of group signatures.
- **Privacy Risk Analysis** Privacy Impact Assessments (PIA) are recognized as a key step to enhance privacy protection in new IT products and services. They will be required for certain types of products in Europe when the future General Data Protection Regulation becomes effective. From a technical perspective, the core of a PIA is a privacy risk analysis (PRA), which has so far received relatively less attention than organizational and legal aspects of PIAs. We have proposed a rigorous and systematic methodology for conducting a PRA and illustrated it with a quantified-self use-case.

  The smart grid initiative promises better home energy management. However, there is a growing concern that utility providers collect, through smart meters, highly granular energy consumption data that can reveal a lot about the consumer's personal life. This exposes consumers to a large number of privacy harms, of various degrees of severity and likelihood: surveillance by the government and law-enforcement bodies, various forms of discrimination etc. A privacy impact assessment is vital for early identification of potential privacy breaches caused by an IT product or service and for choosing the most appropriate protection measures. So, a data protection impact assessment (DPIA) template for smart grids has been developed by the Expert Group 2 (EG2) of the European Commission's Smart Grid Task Force (SGTF). To carry out a true privacy risk analysis and go beyond a traditional security analysis, it is essential to distinguish the notions of feared events and their impacts, called "privacy harms" here, and to establish a link between them. The Working Party 29 highlights the importance of this link in its feedback on EG2's DPIA. We have provided in [11] a clear relationship among harms, feared events, privacy weaknesses and risk sources and described their use in the analysis of smart grid systems.

  Although both privacy by design and privacy risk analysis have received the attention of researchers and privacy practitioners during the last decade, to the best of our knowledge, no method has been documented yet to establish a clear connection between these two closely related notions. We have proposed a methodology to help designers select suitable architectures based on an incremental privacy risk analysis. The analysis proceeds in three broad phases: 1) a generic privacy risk analysis phase depending only on the specifications of the system and yielding generic harm trees; 2) an architecture-based privacy risk analysis that takes into account the definitions of the possible architectures of the system and yields architecture-specific harm trees by refining the generic harm trees and 3) a context-based privacy risk analysis that takes into account the context of deployment of the system (e.g., a casino, an office cafeteria, a school) and further refines the architecture-specific harm trees to yield context-specific harm trees which can be used to take decisions about the most suitable architectures. To illustrate our approach, we have considered the design of a biometric access control system. Such systems are now used commonly in many contexts such as border security controls, work premises, casinos, airports, chemical plants, hospitals, schools, etc. However, the collection, storage and processing of biometric data raise complex privacy issues. To deal with these privacy problems in biometric access control, a wide array of dedicated techniques (such as

secure sketches or fuzzy vaults) as well as adaptations of general privacy preserving techniques (such as encryption, homomorphic encryption, secure multi-party computation) have been proposed. However, each technique solves specific privacy problems and is suitable in specific contexts. Therefore, it is useful to provide guidance to system designers and help them select a solution and justify it with respect to privacy risks. We have used as an illustration of context a deployment in casinos. The verification of the identities of casino customers is required by certain laws (to prevent access by minors or individuals on blacklists) which can justify the implementation of a biometric access control system to speed up the verification process.

6.8. Building blocks

Participants: Marine Minier, Vincent Roca.

- **Symmetric cryptography**

  In [7], we introduce Constraint Programming (CP) models to solve a cryptanalytic problem: the chosen key differential attack against the standard block cipher AES. The problem is solved in two steps: In Step 1, bytes are abstracted by binary values; In Step 2, byte values are searched. We introduce two CP models for Step 1: Model 1 is derived from AES rules in a straightforward way; Model 2 contains new constraints that remove invalid solutions filtered out in Step 2. We also introduce a CP model for Step 2. We evaluate scale-up properties of two classical CP solvers (Gecode and Choco) and a hybrid SAT/CP solver (Chuffed). We show that Model 2 is much more efficient than Model 1, and that Chuffed is faster than Choco which is faster than Gecode on the hardest instances of this problem. Furthermore, we prove that a solution claimed to be optimal in two recent cryptanalysis papers is not optimal by providing a better solution.

  Using dedicated hardware is common practice in order to accelerate cryptographic operations: complex operations are managed by a dedicated co-processor and RAM/crypto-engine data transfers are fully managed by DMA operations. The CPU is therefore free for other tasks, which is vital in embedded environments with limited CPU power. In this work we discuss and benchmark XTS-AES, using either software or mixed approaches, using Linux and dm-crypt, and a low-power At-mel(tm) board. This board featurs an AES crypto-engine that supports ECB-AES but not the XTS-AES mode. We show that the `dm-crypt` module used in Linux for full disk encryption has limitations that can be relaxed when considering larger block sizes. In particular we demonstrate in [14] that performance gains almost by a factor two are possible, which opens new opportunities for future use-cases.

6.9. Other results

Participants: Mathieu Cunche, Vincent Roca.

- **Error-correcting codes**

  Recent work have shown that Reed-Muller (RM) codes achieve the erasure channel capacity. However, this performance is obtained with maximum-likelihood decoding which can be costly for practical applications. In [12], we propose an encoding/decoding scheme for Reed-Muller codes on the packet erasure channel based on Plotkin construction. We present several improvements over the generic decoding. They allow, for a light cost, to compete with maximum-likelihood decoding performance, especially on high-rate codes, while significantly outperforming it in terms of speed.

  In [3], we provide fundamentals in the design and analysis of Generalized Low Density Parity Check (GLDPC)-Staircase codes over the erasure channel. These codes are constructed by extending an LDPC-Staircase code (base code) using Reed Solomon (RS) codes (outer codes) in order to benefit from more powerful decoders. The GLDPC-Staircase coding scheme adds, in addition to the LDPC-Staircase repair symbols, extra-repair symbols that can be produced on demand and in large quantities, which provides small rate capabilities. Therefore, these codes are extremely flexible as they can be tuned to behave either like predefined rate LDPC-Staircase codes at one extreme, or like a single RS code at another extreme, or like small rate codes. Concerning the code design,

we show that RS codes with " quasi " Hankel matrix-based construction fulfill the desired structure properties, and that a hybrid (IT/RS/ML) decoding is feasible that achieves Maximum Likelihood (ML) correction capabilities at a lower complexity. Concerning performance analysis, we detail an asymptotic analysis method based on Density evolution (DE), EXtrinsic Information Transfer (EXIT) and the area theorem. Based on several asymptotic and finite length results, after selecting the optimal internal parameters, we demonstrate that GLDPC-Staircase codes feature excellent erasure recovery capabilities, close to that of ideal codes, both with large and very small objects. From this point of view they outperform LDPC-Staircase and Raptor codes, and achieve correction capabilities close to those of RaptorQ codes. Therefore all these results make GLDPC-Staircase codes a universal Application-Layer FEC (AL-FEC) solution for many situations that require erasure protection such as media streaming or file multicast transmission.

SPADES Project-Team

6. New Results

6.1. Components and contracts

Participants: Alain Girault, Christophe Prévot, Sophie Quinton, Jean-Bernard Stefani.

6.1.1. Contracts for the negotiation of embedded software updates

We address the issue of change after deployment in safety-critical embedded system applications in collaboration with Thales and also in the context of the CCC project (http://ccc-project.org/).

The goal of CCC is to substitute lab-based verification with in-field formal analysis to determine whether an update may be safely applied. This is challenging because it requires an automated process able to handle multiple viewpoints such as functional correctness, timing, etc. For this purpose, we propose an original methodology for contract-based negotiation of software updates. The use of contracts allows us to cleanly split the verification effort between the lab and the field. In addition, we show how to rely on existing viewpoint-specific methods for update negotiation. We have validated our approach on a concrete example inspired by the automotive domain in collaboration with our German partners from TU Braunschweig [19].

In collaboration with Thales we mostly focus on timing aspects with the objective to anticipate at design time future software evolutions and identify potential schedulability bottlenecks. This year we have presented an approach to quantify the flexibility of a system with respect to timing. In particular we have shown that it is possible under certain conditions to identify the task that will directly induce the limitations on a possible software update. If performed at design time, such a result can be used to adjust the system design by giving more slack to the limiting task [21].

6.1.2. Location graphs

The design of configurable systems can be streamlined and made more systematic by adopting a component-based structure, as demonstrated with the FRACTAL component model [2]. However, the formal foundations for configurable component-based systems, featuring higher-order capabilities where components can be dynamically instantiated and passivated, and non-hierarchical structures where components can be contained in different composites at the same time, are still an open topic. We have recently introduced the location graph model [79], where components are understood as graphs of locations hosting higher-order processes, and where component structures can be arbitrary graphs.

We have continued the development of location graphs, revisiting the underlying structural model (hypergraphs instead of graphs), and simplifying its operational semantics while preserving the model expressivity. Towards the development of a behavioral theory of location graphs, we have defined different notions of bisimilarity for location graphs and shown them to be congruences, although a fully fledged co-inductive characterization of contextual equivalence for location graphs is still in the works. This work has not yet been published.

6.2. Real-Time multicore programming

Participants: Pascal Fradet, Alain Girault, Gregor Goessler, Xavier Nicollin, Sophie Quinton.

6.2.1. Time predictable programming languages

Time predictability (PRET) is a topic that emerged in 2007 as a solution to the ever increasing unpredictability of today's embedded processors, which results from features such as multi-level caches or deep pipelines [52]. For many real-time systems, it is mandatory to compute a strict bound on the program's execution time. Yet, in general, computing a tight bound is extremely difficult [82]. The rationale of PRET is to simplify both the programming language and the execution platform to allow more precise execution times to be easily computed [34].

Following our past results on the PRET-C programming language [32], we have proposed a time predictable synchronous programming language for multicores, called FOREC. It extends C with a small set of ESTEREL-like synchronous primitives to express concurrency, interaction with the environment, looping, and a synchronization barrier [83] (like the `pause` statement in ESTEREL). FOREC threads communicate with each other via shared variables, the values of which are *combined* at the end of each tick to maintain deterministic execution. We provide several deterministic combine policies for shared variables, in a way similar as concurrent revisions [45]. Thanks to this, it benefits from a deterministic semantics. FOREC is compiled into threads that are then statically scheduled for a target multicore chip. Our WCET analysis takes into account the access to the shared TDMA bus and the necessary administration for the shared variables. We achieve a very precise WCET (the over-approximation being less than 2%) thanks to a reachable space exploration of the threads' states [15]. We have published a research report presenting the complete semantics and the compiler [27], and submitted it to a journal.

Furthermore, we have extended the PRET-C compiler [32] in order to make it energy aware. To achieve this, we use dynamic voltage and frequency scaling (DFVS) and we insert DVFS control points in the control flow graph of the PRET-C program. The difficulty is twofold: first the control flow graph is concurrent, and second resulting optimization problem is in the 2D space (time,energy). Thanks to a novel ILP formulation and to a bicriteria heuristic, we are able to address the two objectives jointly and to compute, for each PRET-C program, the Pareto front of the non-dominated solutions in the 2D space (time, energy) [20].

This is a collaboration with Eugene Yip from Bamberg University, and with Partha Roop and Jiajie Wang from the University of Auckland.

6.2.2. Modular distribution of synchronous programs

Synchronous programming languages describe functionally centralized systems, where every value, input, output, or function is always directly available for every operation. However, most embedded systems are nowadays composed of several computing resources. The aim of this work is to provide a language-oriented solution to describe *functionally distributed reactive systems*. This research started within the Inria large scale action SYNCHRONICS and is a joint work with Marc Pouzet (ENS, PARKAS team from Rocquencourt) and Gwenaël Delaval (UGA, CTRL-A team from Grenoble).

We are working on defining a *fully-conservative* extension of a synchronous data-flow programming language (the HEPTAGON language, inspired from LUCID SYNCHRONE [46]). The extension, by means of *annotations* adds *abstract location parameters* to functions, and *communications* of values between locations. At deployment, every abstract location is assigned an actual one; this yields an executable for each actual computing resource. Compared to the PhD of Gwenaël Delaval [50], [51], the goal here is to achieve *modular* distribution even in the presence of non-static clocks, *i.e.*, clocks defined according to the value of inputs.

By *fully-conservative*, we have three aims in mind:

1. A non-annotated (*i.e.*, centralized) program will be compiled exactly as before;
2. An annotated program eventually deployed onto only one computing location will behave exactly as its centralized couterpart;
3. The input-output semantics of a distributed program is the same as its centralized counterpart.

By *modular*, we mean that we want to compile each function of the program into a single function capable of running on any computing location. At deployment, the program of each location may be optimized (by simple Boolean-constant-propagation, dead-code and unused-variable elimination), yielding different optimized code for each computing location.

We have formalized the type-system for inferring the location of each variable and computation. In the presence of local clocks, added information is computed from the existing clock-calculus and the location-calculus, to infer necessary communication of clocks between location. All pending theorical and technical issues have been answered, and the new compiler is being implemented, with new algorithms for deployment (and code optimization), achieving the three aims detailed above.

6.2.3. Parametric dataflow models

Recent data-flow programming environments support applications whose behavior is characterized by dynamic variations in resource requirements. The high expressive power of the underlying models (*e.g.*, Kahn Process Networks or the CAL actor language) makes it challenging to ensure predictable behavior. In particular, checking *liveness* (*i.e.*, no part of the system will deadlock) and *boundedness* (*i.e.*, the system can be executed in finite memory) is known to be hard or even undecidable for such models. This situation is troublesome for the design of high-quality embedded systems.

Recently, we have introduced the *Schedulable Parametric Data-Flow* (SPDF) MoC for dynamic streaming applications [55], which extends the standard dataflow model by allowing rates to be parametric, and the *Boolean Parametric Data Flow* (BPDF) MoC [38], [37] which combines integer parameters (to express dynamic rates) and boolean parameters (to express the activation and deactivation of communication channels). In the past years, several other parametric dataflow MoCs have been presented. All these models aim at providing an interesting trade-off between analyzability and expressiveness. They offer a controlled form of dynamism under the form of parameters (*e.g.*, parametric rates), along with run-time parameter configuration.

We have written a survey which provides a comprehensive description of the existing parametric dataflow MoCs (constructs, constraints, properties, static analyses) and compares them using a common example [11]. The main objectives are to help designers of streaming applications to choose the most suitable model for their needs and to pave the way for the design of new parametric MoCs.

We have also studied *symbolic* analyses of data-flow graphs [24], [16], [17], [12]. Symbolic analyses express the system performance as a function of parameters (*i.e.*, input and output rates, execution times). Such functions can be quickly evaluated for each different configuration or checked *w.r.t.* different quality-of-service requirements. These analyses are useful for parametric MoCs, partially specified graphs, and even for completely static SDF graphs. We provide symbolic analyses for computing the maximal throughput of acyclic synchronous dataflow graphs, the minimum required buffers for which as soon as possible (asap) scheduling achieves this throughput, and finally the corresponding input-output latency of the graph. We first investigate these problems for a single parametric edge. The results are then extended to general acyclic graphs using linear approximation techniques. We assess the proposed analyses experimentally on both synthetic and real benchmarks.

6.2.4. Synthesis of switching controllers using approximately bisimilar multiscale abstractions

The use of discrete abstractions for continuous dynamics has become standard in hybrid systems design (see *e.g.*, [80] and the references therein). The main advantage of this approach is that it offers the possibility to leverage controller synthesis techniques developed in the areas of supervisory control of discrete-event systems [75]. The first attempts to compute discrete abstractions for hybrid systems were based on traditional systems behavioral relationships such as simulation or bisimulation, initially proposed for discrete systems most notably in the area of formal methods. These notions require inclusion or equivalence of observed behaviors which is often too restrictive when dealing with systems observed over metric spaces. For such systems, a more natural abstraction requirement is to ask for closeness of observed behaviors. This leads to the notions of approximate simulation and bisimulation introduced in [56].

These approaches are based on sampling of time and space where the sampling parameters must satisfy some relation in order to obtain abstractions of a prescribed precision. In particular, the smaller the time sampling parameter, the finer the lattice used for approximating the state-space; this may result in abstractions with a very large number of states when the sampling period is small. However, there are a number of applications where sampling has to be fast; though this is generally necessary only on a small part of the state-space. We have been exploring two approaches to overcome this state-space explosion [5].

We are currently investigating an approach using mode sequences of given length as symbolic states for our abstractions. By using mode sequences of variable length we are able to adapt the granularity of our abstraction to the dynamics of the system, so as to automatically trade off precision against controllability of the abstract states.

6.2.5. Schedulability of weakly-hard real-time systems

We focus on the problem of computing tight deadline miss models for real-time systems, which bound the number of potential deadline misses in a given sequence of activations of a task. In practical applications, such guarantees are often sufficient because many systems are in fact not hard real-time [4].

Our major contribution this year is the extension of our method for computing deadline miss models, called Typical Worst-Case Analysis (TWCA), to systems with task dependencies. This allows us to provide bounds on deadline misses for systems which until now could not be analyzed [18].

In parallel, we have developed an extension of sensitivity analysis for budgeting in the design of weakly-hard real-time systems. During design, it often happens that some parts of a task set are fully specified while other parameters, e.g. regarding recovery or monitoring tasks, will be available only much later. In such cases, sensitivity analysis can help anticipate how these missing parameters can influence the behavior of the whole system so that a resource budget can be allocated to them. We have developed an extension of sensitivity analysis for deriving task budgets for systems with hard and weakly-hard requirements. This approach has been validated on synthetic test cases and a realistic case study given by our partner Thales. This work will be submitted soon.

Finally, in collaboration with TU Braunschweig and Daimler we have investigated the use of TWCA in conjunction with the Logical Execution Time paradigm [68] according to which data are read and written at predefined time instants. In particular, we have extended TWCA to different deadline miss handling strategies. This work has not been published yet.

6.3. Language Based Fault-Tolerance

Participants: Pascal Fradet, Alain Girault, Yoann Geoffroy, Gregor Goessler, Jean-Bernard Stefani, Martin Vassor, Athena Abdi.

6.3.1. Fault Ascription in Concurrent Systems

The failure of one component may entail a cascade of failures in other components; several components may also fail independently. In such cases, elucidating the exact scenario that led to the failure is a complex and tedious task that requires significant expertise.

The notion of causality *(did an event e cause an event e'?)* has been studied in many disciplines, including philosophy, logic, statistics, and law. The definitions of causality studied in these disciplines usually amount to variants of the counterfactual test "e is a cause of e' if both e and e' have occurred, and in a world that is as close as possible to the actual world but where e does not occur, e' does not occur either". In computer science, almost all definitions of logical causality — including the landmark definition of [63] and its derivatives — rely on a causal model that may not be known, for instance in presence of black-box components. For such systems, we have been developing a framework for blaming that helps us establish the causal relationship between component failures and system failures, given an observed system execution trace. The analysis is based on a formalization of counterfactual reasoning [7].

In his PhD thesis, Yoann Geoffroy proposed a generalization of our fault ascription technique to systems composed of black-box and white-box components. For the latter a faithful behavioral model is given but no specification. The approach leverages results from game theory and discrete controller synthesis to define several notions of causality.

We are currently working on an instantiation of our general semantic framework for fault ascription in [60] to acyclic models of computation, in order to compare our approach with the standard definition of *actual causality* proposed by Halpern and Pearl.

6.3.2. Tradeoff exploration between energy consumption and execution time

We have continued our work on multi-criteria scheduling, in two directions. First, in the context of dynamic applications that are launched and terminated on an embedded homogeneous multi-core chip, under execution time and energy consumption constraints, we have proposed a two layer adaptive scheduling method. In the first layer, each application (represented as a DAG of tasks) is scheduled statically on subsets of cores: 2 cores, 3 cores, 4 cores, and so on. For each size of these sets (2, 3, 4, ...), there may be only one topology or several topologies. For instance, for 2 or 3 cores there is only one topology (a "line"), while for 4 cores there are three distinct topologies ("line", "square", and "T shape"). Moreover, for each topology, we generate statically several schedules, each one subject to a different total energy consumption constraint, and consequently with a different Worst-Case Reaction Time (WCRT). Coping with the energy consumption constraints is achieved thanks to Dynamic Frequency and Voltage Scaling (DVFS). In the second layer, we use these pre-generated static schedules to reconfigure dynamically the applications running on the multi-core each time a new application is launched or an existing one is stopped. The goal of the second layer is to perform a dynamic global optimization of the configuration, such that each running application meets a pre-defined quality-of-service constraint (translated into an upper bound on its WCRT) and such that the total energy consumption be minimized. For this, we *(i)* allocate a sufficient number of cores to each active application, *(ii)* allocate the unassigned cores to the applications yielding the largest gain in energy, and *(iii)* choose for each application the best topology for its subset of cores (*i.e.*, better than the by default "line" topology). This is a joint work with Ismail Assayad (U. Casablanca, Morocco) who visited the team in September 2015.

Second, in the context of a static application (again represented a DAG of tasks) running on an homogeneous multi-core chip, we have worked on the static scheduling minimizing the WCRT of the application under the multiple constraints that the reliability, the power consumption, and the temperature remain below some given thresholds. There are multiple difficulties: *(i)* the reliability is not an invariant measure w.r.t. time, which makes it impossible to use backtrack-free scheduling algorithms such as list scheduling [33]; to overcome this, we adopt instead the Global System Failure Rate (GSFR) as a measure of the system's reliability, which is invariant with time [57]; *(ii)* keeping the power consumption under a given threshold requires to lower the voltage and frequency, but this has a negative impact both on the WCRT and on the GSFR; keeping the GSFR below a given threshold requires to replicate the tasks on multiple cores, but this has a negative impact both on the WCRT, on the power consumption, and on the temperature; *(iii)* keeping the temperature below a given threshold is even more difficult because the temperature continues to increase even after the activity stops, so each scheduling decision must be assessed not based on the current state of the chip (*i.e.*, the temperature of each core) but on the state of the chip at the end of the candidate task, and cooling slacks must be inserted. We have proposed a multi-criteria scheduling heuristics to address these challenges. It produces a static schedule of the given application graph and the given architecture description, such that the GSFR, power, and temperature thresholds are satisfied, and such that the execution time is minimized. We then combine our heuristic with a variant of the ε-constraint method [62] in order to produce, for a given application graph and a given architecture description, its entire Pareto front in the 4D space (exec. time, GSFR, power, temp.). This is a joint work with Athena Abdi and Hamid Zarandi from Amirkabir U., Iran, who have visited the team in 2016.

6.3.3. Automatic transformations for fault tolerant circuits

In the past years, we have studied the implementation of specific fault tolerance techniques in real-time embedded systems using program transformation [1]. We are now investigating the use of automatic transformations to ensure fault-tolerance properties in digital circuits. To this aim, we consider program transformations for hardware description languages (HDL). We consider both single-event upsets (SEU) and single-event transients (SET) and fault models of the form *"at most 1 SEU or SET within n clock cycles"*.

We have expressed several variants of triple modular redundancy (TMR) as program transformations. We have proposed a verification-based approach to minimize the number of voters in TMR [25]. Our technique guarantees that the resulting circuit *(i)* is fault tolerant to the soft-errors defined by the fault model and *(ii)* is functionally equivalent to the initial one. Our approach operates at the logic level and takes into account the input and output interface specifications of the circuit. Its implementation makes use of graph traversal algorithms, fixed-point iterations, and BDDs. Experimental results on the ITC'99 benchmark suite indicate that

our method significantly decreases the number of inserted voters which entails a hardware reduction of up to 55% and a clock frequency increase of up to 35% compared to full TMR. We address scalability issues arising from formal verification with approximations and assess their efficiency and precision. As our experiments show, if the SEU fault-model is replaced with the stricter fault-model of SET, it has a minor impact on the number of removed voters. On the other hand, BDD-based modeling of SET effects represents a more complex task than the modeling of an SEU as a bit-flip. We propose solutions for this task and explain the nature of encountered problems. We discuss scalability issues arising from formal verification with approximations and assess their efficiency and precision.

6.3.4. Concurrent flexible reversibility

Reversible concurrent models of computation provide natively what appears to be very fine-grained checkpoint and recovery capabilities. We have made this intuition clear by formally comparing a distributed algorithm for checkpointing and recovery based on causal information, and the distributed backtracking algorithm that lies at the heart of our reversible higher-order pi-calculus. We have shown that (a variant of) the reversible higher-order calculus with explicit rollback can faithfully encode a distributed causal checkpoint and recovery algorithm. The reverse is also true but under precise conditions, which restrict the ability to rollback a computation to an identified checkpoint. This work has currently not been published.

BIPOP Project-Team

6. New Results

6.1. The contact complementarity problem, and Painlevé paradoxes

Participants: Bernard Brogliato, Florence Bertails-Descoubes, Alejandro Blumentals.

The contact linear complementarity problem is an set of equalities and complementarity conditions whose unknowns are the acceleration and the contact forces. It has been studied in a frictionless context with possibly singular mass matrix and redundant constraints, using results on well-posedness of variational inequalities obtained earlier by the authors. This is also the topic of the first part of the Ph.D. thesis of Alejandro Blumentals where the frictional case is treated as a perturbation of the frictionless case [22]. With R. Kikuuwe from Kyushu University, we have also proposed a new formulation of the Baumgarte's stabilisation method, for unilateral constraints and Coulomb's friction , which sheds new light on Painlevé paradoxes [27]. It relies on a particular limiting process of normal cones.

6.2. Discrete-time sliding mode control

Participants: Vincent Acary, Bernard Brogliato, Olivier Huber.

This topic concerns the study of time-discretized sliding-mode controllers. Inspired by the discretization of nonsmooth mechanical systems, we propose implicit discretizations of discontinuous, set-valued controllers [3]. This is shown to result in preservation of essential properties like simplicity of the parameters tuning, suppression of numerical chattering, reachability of the sliding surface after a finite number of steps, and disturbance attenuation by a factor h or h^2 [25]. This work was part of the ANR project CHASLIM. Within the framework of CHASLIM we have performed many experimental validations on the electropneumatic setup of IRCCyN (Nantes), which nicely confirm our theoretical and numerical predictions: the implicit implementation of sliding mode control, drastically improves the input and output chattering behaviours, both for the classical order-one ECB-SMC and the twisting algorithms [26], [25], [39]. In particular the high frequency bang-bang controllers which are observed with explicit discretizations, are completely suppressed. The implicit discretization has been applied to the classical equivalent-based-control SMC, and also to the twisting sliding-mode controller. Incidentally an error in a previous article is corrected in [19]. The previous results deal with disturbances which are matched and uniformly upperbounded. In [48], [49] they are extended to the case of parametric uncertainties, which are more difficult to handle because they may yield unmatched equivalent disturbances, and these disturbances are not uniformly upperbounded by a constant. Finally the results in [20] deal with the numerical analysis (and not the discrete-time control, which is a different problem) of Lagrangian systems with set-valued controllers. An implicit Euler method is used, and the convergence is shown.

6.3. Lur'e set-valued dynamical systems

Participants: Bernard Brogliato, Christophe Prieur, Alexandre Vieira.

Lur'e systems are quite popular in Automatic Control since the fifties. Set-valued Lur'e systems possess a static feedback nonlinearity that is a multivalued function. We study in [31] state observers for particular Lur'e systems which are Moreau's sweeping processes modelling Lagrange dynamics with frictionless unilateral constraints. The observers are themselves set-valued (first order sweeping process with measures), a complete analysis (existence of solutions, stability of the error system) is led. In [51], we extend previous results in the team and also more recently by Camlibel and Schumacher, to solve the problem of output regulation for evolution variational inequalities (in a convex analysis setting). In the PhD thesis of A. Vieira, we attack the problem of optimal control of linear complementarity systems. In the first part of this thesis, the case when the LCS is equivalent to an ODE with Lipschitz continuous right-hand side, is treated. Starting from first-order necessary conditions stated in a broad context by Clarke, we show that the Pontryagin's conditions are a mixed LCS, that yield so-called MPEC problems.

6.4. Numerical analysis of multibody mechanical systems with constraints

This scientific theme concerns the numerical analysis of mechanical systems with bilateral and unilateral constraints, with or without friction [2]. They form a particular class of dynamical systems whose simulation requires the development of specific simulators.

6.4.1. Numerical time–integration methods for event-detecting schemes.

Participants: Vincent Acary, Bernard Brogliato, Mounia Haddouni.

The CIFRE thesis of M. Haddouni concerns the numerical simulation of mechanical systems subject to holonomic bilateral constraints, unilateral constraints and impacts. This work is performed in collaboration with ANSYS and the main goal is to improve the numerical time–integration in the framework of event-detecting schemes. Between nonsmooth events, time integration amounts to numerically solving a differential algebraic equations (DAE) of index 3. We have compared dedicated solvers (Explicit RK schemes, Half-explicit schemes, generalizes α-schemes) that solve reduced index formulations of these systems. Since the drift of the constraints is crucial for the robustness of the simulation through the evaluation of the index sets of active contacts, we have proposed some recommendations on the use of the solvers of dedicated to index-2 DAE. A manuscript has been submitted to Multibody System Dynamics.

6.4.2. Multibody systems with clearances (dynamic backlash)

Participants: Vincent Acary, Bernard Brogliato, Narendra Akadkhar.

The PhD thesis of N. Akadkhar under contract with Schneider Electric concerns the numerical simulation of mechanical systems with unilateral constraints and friction, where the presence of clearances in imperfect joints plays a crucial role. A first work deals with four-bar planar mechanisms with clearances at the joints, which induce unilateral constraints and impacts, rendering the dynamics nonsmooth. The objective is to determine sets of parameters (clearance value, restitution coefficients, friction coefficients) such that the system's trajectories stay in a neighborhood of the ideal mechanism (*i.e.* without clearance) trajectories. The analysis is based on numerical simulations obtained with the projected Moreau-Jean time-stepping scheme. These results have been reported in [21]. It is planned to extend these simulations to frictional cases and to mechanisms of circuit breakers.

6.5. Nonlinear waves in granular chains

Participants: Guillaume James, Bernard Brogliato.

Granular chains made of aligned beads interacting by contact (e.g. Newton's cradle) are widely studied in the context of impact dynamics and acoustic metamaterials. While much effort has been devoted to the theoretical and experimental analysis of solitary waves in granular chains, there is now an increasing interest in the study of breathers (spatially localized oscillations) in granular systems. Due to their oscillatory nature and associated resonance phenomena, static or traveling breathers exhibit much more complex dynamical properties compared to solitary waves. Such properties have strong potential applications for the design of acoustic metamaterials allowing to efficiently damp or deviate shocks and vibrations. In the work [29], the existence of static breathers is analyzed in granular metamaterials consisting of hollow beads with internal masses. Using multiple scale analysis and exploiting the unilateral character of Hertzian interactions, we show that long-lived breather solutions exist but time-periodic breathers do not (breather solutions actually disperse on long time scales). In [28], we consider the effect of adding precompression to the above system and establish that the envelope of small ampliude oscillations is governed by a nonlinear Schrödinger equation. This allows us to show that, depending on the applied precompression, normal modes can become modulationally unstable and evolve towards traveling breathers. Moreover, in a collaboration with Y. Starosvetsky and D. Meimukhin (Technion), we numerically study the persistence of traveling breathers in granular chains with local potentials under the effect of contact damping. Using a viscoelastic damping model (Hertz-Kuwabara-Kono model), we show that breathers can be generated by simple impacts in granular chains made from various materials (breathers propagate over a significant number of sites before being damped). The design of an experimental setup to test these theoretical predictions is underway. Another work in progress concerns more specifically

the modeling and numerical analysis of dissipative impacts (James, Brogliato). The methodology is based on the introduction of appropriate variables and simplifications for different models of contact damping. A postdoctoral fellow will work on this topic in the team, starting January 2017.

6.6. Travelling waves in a spring-block chain sliding down a slope

Participants: Guillaume James, Jose Eduardo Morales Morales, Arnaud Tonnelier.

In this work we study the dynamics of an infinite chain of identical blocks sliding on a slope under the effet of gravity. Each block is coupled to its nearest neighbour through linear springs and is subjected to a nonlinear friction force. For a piecewise-linear spinodal friction law, a closed-form expression of front waves is derived. Pulse waves are obtained as the matching of two travelling fronts with identical wave speeds. Explicit formulas are obtained for the wavespeed and the wave form in the anti-continuum limit. The link with propagating phenomena in the Burridge-Knopoff model is briefly discussed. These results have been reported in [44].

6.7. Solitary waves in the excitable Burridge-Knopoff model

Participants: Guillaume James, Jose Eduardo Morales Morales, Arnaud Tonnelier.

The Burridge-Knopoff model is a lattice differential equation describing a chain of blocks connected by springs and pulled over a surface. This model was originally introduced to investigate nonlinear effects arising in the dynamics of earthquake faults. One of the main ingredients of the model is a nonlinear velocity-dependent friction force between the blocks and the fixed surface. For some classes of non-monotonic friction forces, the system displays a large response to perturbations above a threshold, which is characteristic of excitable dynamics. Using extensive numerical simulations, we show that this response corresponds to the propagation of a solitary wave for a broad range of friction laws (smooth or nonsmooth) and parameter values. These solitary waves develop shock-like profiles at large coupling (a phenomenon connected with the existence of weak solutions in a formal continuum limit) and propagation failure occurs at low coupling. We introduce a simplified piecewise linear friction law (reminiscent of the McKean nonlinearity for excitable cells) which allows us to obtain analytical expression of solitary waves and study some of their qualitative properties, such as wavespeed and propagation failure. We propose a possible physical realization of this system as a chain of impulsively forced mechanical oscillators. In certain parameter regimes, non-monotonic friction forces can also give rise to bistability between the ground state and limit-cycle oscillations and allow for the propagation of fronts connecting these two stable states. These results have been reported in [45]. In addition, an existence theorem for solitary waves in the Burridge-Knopoff model is proved in the weak coupling limit and for a piecewise-linear friction force.

6.8. Propagation in space-discrete excitable systems

Participant: Arnaud Tonnelier.

We introduce a simplified model of excitable systems where the response of an isolated cell to an incoming signal is described by a fixed pulse-shape function. When the total activity of the cell reaches a given threshold a signal is sent to its N nearest neighbors. We show that a chain of such excitable cells is able to propagate a set of simple traveling waves where the time interval between the firing of two successive cells remains constant. A comprehensive study is done for a transmission line with $N = 2$ and $N = 3$. It is shown that, depending on initial copnditions, the network may propagate signals with different velocities. Some necessary conditions for multistationarity are derived for an arbitrary N.

6.9. Direct and inverse modeling of thin elastic rods and shells

6.9.1. Experimental validation of the inverse statics of a thin elastic rod

Participants: Florence Bertails-Descoubes, Victor Romero.

In collaboration with Arnaud Lazarus (UPMC, Laboratoire Jean le Rond d'Alembert), we have built an experimental set-up to fabricate thin elastic rods and measure their deformation, with the aim to validate our full process for inverse static design. This work is still ongoing.

6.9.2. Strain-based modeling of inextensible and developable shells

Participants: Florence Bertails-Descoubes, Romain Casati, Alejandro Blumentals.

We have worked out the analagous of a super-helix element for modeling an inextensible and developable shell patch, using only two material curvatures. As for the super-helix model, the terms of the dynamics can be integrated formally, leading to a rich and efficient dynamical model [36]. How to connect different patches together is a topic for future work.

6.9.3. Inverse statics of plates and shells with frictional contact

Participants: Florence Bertails-Descoubes, Romain Casati, Gilles Daviet.

We study the problem of cloth inverse design, relying on a nodal shell model for modeling garments. We have shown how to formulate draping as a local constrained minimization problem, and we have generalized the adjoint method to handle constrained cases, e.g., frictional contact between the garment and the body [43].

6.10. Continnum modeling of granular materials

6.10.1. Continuum modeling of granular materials

Participants: Florence Bertails-Descoubes, Gilles Daviet.

We have proposed a new numerical framework for the continuous simulation of dilatable materials with pressure-dependent (Coulomb) yield stress, such as sand or cement. Relying upon convex optimization tools, we have shown that the continuous equations of motion coupled to the macroscopic nonsmooth Drucker-Prager rheology can be interpreted as the exact analogous of the solid frictional contact problem at the heart of Discrete Element Methods (DEM), extended to the tensorial space. Combined with a carefully chosen finite-element discretization, this new framework allowed us to avoid regularizing the continuum rheology while benefiting from the efficiency of nonsmooth optimization solvers, mainly leveraged by DEM methods so far. Our numerical results were successfully compared to analytic solutions on model problems, such as the silo discharge, and we retrieved qualitative flow features commonly observed in reported experiments of the literature. This work, published at the Journal of Non Newtonian Fluid Mechanics [24], has been extended the approach to account for flows with a varying density, leveraging the Material Point Method to discretize the Drucker Prager yield criterion without linearization. We have also included the handling of anisotropic flow, as well as the coupling of the flow with rigid bodies. These extensions led to a publication at ACM SIGGRAPH 2016 [23].

6.11. Robust Model Predictive Control for biped walking motion generation

Participants: Pierre-Brice Wieber, Diana Serra, Alexander Sherikov, Dimitar Dimitrov.

One of the main sources of nonlinearity in the Newton and Euler equations of motion of biped walking robots lie in the vertical motion of the Center of Mass. We proposed last year an approach that considers this nonlinearity as an uncertainty, in what would else be a linear system. We proposed then to use a robust linear MPC approach accordingly. The use of a linear approach allows fast computations to generate walking motions online. This year, we further developed this approach, by adapting the bounds on the uncertainty at each iteration of a Newton scheme, when solving the original nonlinear problem [35]. By using a robust approach within a Newton scheme, every iteration can be ensured to satisfy all dynamic constraints, so that we can limit the number iterations depending on the available computing power and always obtain a feasible solution. We also developed this year an application of this MPC approach to cases of collaborative carrying of heavy objects with a human partner [32].

6.12. Lexicographic Model Predictive Control for collision avoidance in dynamic environments

Participants: Pierre-Brice Wieber, Nestor Alonso Bohorquez Dorante, Alexander Sherikov, Dimitar Dimitrov.

Collision avoidance may not always be feasible in dynamic environments, when new obstacles can appear too late and move too fast with respect to the dynamic limitations of the system. A typical situation is with a biped robot walking in a compact and uncooperative crowd, with limited field of view. This year, we have investigated and compared 3 different relaxations of the collision avoidance constraint in this setting [33]. In the first case, collisions are accepted if the robot first comes to a stop, what corresponds to standard ISO norms for the safety of robots. In the second case, collisions are actively minimised by the robot, what gives significantly better results. In the third case, for the sake of completeness, the robot is allowed to fall in order to further avoid collisions. All three options were implemented with different formulations of lexicographic relaxation of the constraints in a standard MPC scheme for biped walking motion generation. This work raises important issues regarding safety norms for robots in human environments and how they are implemented.

6.13. Lexicographic Programming

Participants: Pierre-Brice Wieber, Alexander Sherikov, Dimitar Dimitrov, Adrien Escande.

Lexicographic Programming has proved to be a very valuable tool in the last few years for relaxing selectively various constraints and objectives in the control of complex systems such as biped humanoid robots. A major difficulty however is that solutions to such problems very often lie at singular points, making the convergence of standard Newton schemes difficult. We have shown this year how a trust region with filter method can help improve convergence, at least in simple situations [40].

MISTIS Project-Team

7. New Results

7.1. Mixture models

7.1.1. High dimensional Kullback-Leilbler divergence for supervised clustering

Participant: Stephane Girard.

Joint work with: C. Bouveyron (Univ. Paris 5), M. Fauvel and M. Lopes (ENSAT Toulouse))

In the PhD work of Charles Bouveyron [74], we proposed new Gaussian models of high dimensional data for classification purposes. We assume that the data live in several groups located in subspaces of lower dimensions. Two different strategies arise:

- the introduction in the model of a dimension reduction constraint for each group
- the use of parsimonious models obtained by imposing to different groups to share the same values of some parameters

This modelling yielded a supervised classification method called High Dimensional Discriminant Analysis (HDDA) [4]. Some versions of this method have been tested on the supervised classification of objects in images. This approach has been adapted to the unsupervised classification framework, and the related method is named High Dimensional Data Clustering (HDDC) [3]. In the framework of Mailys Lopes PhD, our recent work [50], consists in adapting this work to the classification of grassland management practices using satellite image time series with high spatial resolution. The study area is located in southern France where 52 parcels with three management types were selected. The spectral variability inside the grasslands was taken into account considering that the pixels signal can be modeled by a Gaussian distribution. A parsimonious model is discussed to deal with the high dimension of the data and the small sample size. A high dimensional symmetrized Kullback-Leibler divergence (KLD) is introduced to compute the similarity between each pair of grasslands. The model is positively compared to the conventional KLD to construct a positive definite kernel used in SVM for supervised classification.

7.1.2. Single-run model selection in mixtures

Participants: Florence Forbes, Alexis Arnaud.

Joint work with: Russel Steele, McGill University, Montreal, Canada.

A number of criteria exist to select the number of components in a mixture automatically based on penalized likelihood criteria (eg. AIC, BIC, ICL etc.) but they usually require to run several models for different number of components to choose the best one. In this work, the goal was to investigate existing alternatives that can select the component number from a single run and to develop such a procedure for our MRI analysis. These objectives were achieved for the main part as 1) different single run methods have been implemented and tested for Gaussian and Standard mixture models, 2) a Bayesian version of Generalized Student mixtures have been designed that allows the use of the methods in 1), and 3) we also proposed a new heuristic based on this Bayesian model that shows good performance and lower computational times. A more complete validation on simulated data and tests on real MRI data need still to be performed. The single run methods studied are based on a fully Bayesian approach involving therefore specification of appropriate priors and choice of hyperparameters. To estimate our Bayesian mixture model, we use a Variational Expectation-Maximization algorithm (VEM). For the heuristic, we add an additional step inside VEM in order to compute in parallel the corresponding VEM step with one less component. If the lower-bound of the model likelihood is higher with one less component, then we delete this component and go to the next VEM step, until convergence of the algorithm. As regards software development, the Rcpp package has been used to bridge pure R code with more efficient C++ code. This project has been initiated with Alexis Arnaud's visit to McGill University in Montreal in the context of his Mitacs award.

7.1.3. Sequential Quasi Monte Carlo for Dirichlet Process Mixture Models

Participant: Julyan Arbel.

Joint work with: Jean-Bernard Salomond (Université Paris-Est).

In mixture models, latent variables known as allocation variables play an essential role by indicating, at each iteration, to which component of the mixture observations are linked. In sequential algorithms, these latent variables take on the interpretation of particles. We investigate the use of quasi Monte Carlo within sequential Monte Carlo methods (a technique known as sequential quasi Monte Carlo) in nonparametric mixtures for density estimation. We compare them to sequential and non sequential Monte Carlo algorithms. We highlight a critical distinction of the allocation variables exploration of the latent space under each of the three sampling approaches. This work has been presented at the *Practical Bayesian Nonparametrics* NIPS workshop [48].

7.1.4. Truncation error of a superposed gamma process in a decreasing order representation

Participant: Julyan Arbel.

Joint work with: Igor Prünster (University Bocconi, Milan).

Completely random measures (CRM) represent a key ingredient of a wealth of stochastic models, in particular in Bayesian Nonparametrics for defining prior distributions. CRMs can be represented as infinite random series of weighted point masses. A constructive representation due to Ferguson and Klass provides the jumps of the series in decreasing order. This feature is of primary interest when it comes to sampling since it minimizes the truncation error for a fixed truncation level of the series. We quantify the quality of the approximation in two ways. First, we derive a bound in probability for the truncation error. Second, we study a moment-matching criterion which consists in evaluating a measure of discrepancy between actual moments of the CRM and moments based on the simulation output. This work focuses on a general class of CRMs, namely the superposed gamma process, which suitably transformed have already been successfully implemented in Bayesian Nonparametrics. To this end, we show that the moments of this class of processes can be obtained analytically. This work has been presented at the *Advances in Approximate Bayesian Inference* NIPS workshop [47].

7.1.5. Non linear mapping by mixture of regressions with structured covariance matrix

Participant: Emeline Perthame.

Joint work with: Emilie Devijver (KU Leuven, Belgium) and Mélina Gallopin (Université Paris Sud).

In genomics, the relation between phenotypical responses and genes are complex and potentially non linear. Therefore, it could be interesting to provide biologists with statistical models that mimic and approximate these relations. In this paper, we focus on a dataset that relates genes expression to the sensitivity to alcohol of drosophila. In this framework of non linear regression, GLLiM (Gaussian Locally Linear Mapping) is an efficient tool to handle non linear mappings in high dimension. Indeed, this model based on a joint modeling of both responses and covariates by Gaussian mixture of regressions has demonstrated its performance in non linear prediction for multivariate responses when the number of covariates is large. This model also allows the addition of latent factors which have led to interesting interpretation of the latent factors in image analysis. Nevertheless, in genomics, biologists are more interested in graphical models, representing gene regulatory networks. For this reason, we developed an extension of GLLiM in which covariance matrices modeling the dependence structure of genes in each clusters are blocks diagonal, using tools derived for graphical models. This extension provides a new class of interpretable models that are suitable to genomics application fields while keeping interesting prediction properties.

7.1.6. Extended GLLiM model for a subclustering effect: Mixture of Gaussian Locally Linear Mapping (MoGLLiM)

Participant: Florence Forbes.

Joint work with: Naisyin Wang and Chun-Chen Tu from University of Michigan, Ann Arbor, USA.

The work of Chun-Chen Tu and Naisyin Wang pointed out a problem with the original GLLiM model that they propose to solve with a divide-remerge method. The proposal seems to be efficient on test data but the resulting procedure does not anymore correspond to the optimization of a single statistical model. The idea of this work is then to discuss the possibility to change the original GLLiM model in order to account for sub-clusters directly. A small change in the definition seems to have such an effect while remaining tractable. However, we will probably have to be careful with potential non-identifiability issue when dealing with clusters and sub-clusters.

7.2. Semi and non-parametric methods

7.2.1. Robust estimation for extremes

Participants: Clement Albert, Stephane Girard.

Joint work with: M. Stehlik (Johannes Kepler Universitat Linz, Austria and Universidad de Valparaiso, Chile) and A. Dutfoy (EDF R&D).

In the PhD thesis of Clément Albert (funded by EDF), we study the sensitivity of extreme-value methods to small changes in the data [46]. To reduce this sensitivity, robust methods are needed and, in [21], we proposed a novel method of heavy tails estimation based on a transformed score (the t-score). Based on a new score moment method, we derive the t-Hill estimator, which estimates the extreme value index of a distribution function with regularly varying tail. t-Hill estimator is distribution sensitive, thus it differs in e.g. Pareto and log-gamma case. Here, we study both forms of the estimator, i.e. t-Hill and t-lgHill. For both estimators we prove weak consistency in moving average settings as well as the asymptotic normality of t-lgHill estimator in the i.i.d. setting. In cases of contamination with heavier tails than the tail of original sample, t-Hill outperforms several robust tail estimators, especially in small sample situations. A simulation study emphasizes the fact that the level of contamination is playing a crucial role. We illustrate the developed methodology on a small sample data set of stake measurements from Guanaco glacier in Chile. This methodology is adapted to bounded distribution tails in [26] with an application to extreme snow loads in Slovakia.

7.2.2. Conditional extremal events

Participant: Stephane Girard.

Joint work with: L. Gardes (Univ. Strasbourg) and J. Elmethni (Univ. Paris 5)

The goal of the PhD theses of Alexandre Lekina and Jonathan El Methni was to contribute to the development of theoretical and algorithmic models to tackle conditional extreme value analysis, *ie* the situation where some covariate information X is recorded simultaneously with a quantity of interest Y. In such a case, the tail heaviness of Y depends on X, and thus the tail index as well as the extreme quantiles are also functions of the covariate. We combine nonparametric smoothing techniques [77] with extreme-value methods in order to obtain efficient estimators of the conditional tail index and conditional extreme quantiles. When the covariate is functional and random (random design) we focus on kernel methods [18].

Conditional extremes are studied in climatology where one is interested in how climate change over years might affect extreme temperatures or rainfalls. In this case, the covariate is univariate (time). Bivariate examples include the study of extreme rainfalls as a function of the geographical location. The application part of the study is joint work with the LTHE (Laboratoire d'étude des Transferts en Hydrologie et Environnement) located in Grenoble [31], [32].

7.2.3. Estimation of extreme risk measures

Participant: Stephane Girard.

Joint work with: A. Daouia (Univ. Toulouse), L. Gardes (Univ. Strasbourg) and G. Stupfler (Univ. Aix-Marseille).

One of the most popular risk measures is the Value-at-Risk (VaR) introduced in the 1990's. In statistical terms, the VaR at level $\alpha \in (0,1)$ corresponds to the upper α-quantile of the loss distribution. The Value-at-Risk however suffers from several weaknesses. First, it provides us only with a pointwise information: VaR(α) does not take into consideration what the loss will be beyond this quantile. Second, random loss variables with light-tailed distributions or heavy-tailed distributions may have the same Value-at-Risk. Finally, Value-at-Risk is not a coherent risk measure since it is not subadditive in general. A first coherent alternative risk measure is the Conditional Tail Expectation (CTE), also known as Tail-Value-at-Risk, Tail Conditional Expectation or Expected Shortfall in case of a continuous loss distribution. The CTE is defined as the expected loss given that the loss lies above the upper α-quantile of the loss distribution. This risk measure thus takes into account the whole information contained in the upper tail of the distribution. In [64], we investigate the extreme properties of a new risk measure (called the Conditional Tail Moment) which encompasses various risk measures, such as the CTE, as particular cases. We study the situation where some covariate information is available under some general conditions on the distribution tail. We thus has to deal with conditional extremes (see paragraph 7.2.2).

A second possible coherent alternative risk measure is based on expectiles [63]. Compared to quantiles, the family of expectiles is based on squared rather than absolute error loss minimization. The flexibility and virtues of these least squares analogues of quantiles are now well established in actuarial science, econometrics and statistical finance. Both quantiles and expectiles were embedded in the more general class of M-quantiles as the minimizers of a generic asymmetric convex loss function. It has been proved very recently that the only M-quantiles that are coherent risk measures are the expectiles.

7.2.4. Multivariate extremal events

Participants: Stephane Girard, Florence Forbes.

Joint work with: F. Durante (Univ. Bolzen-Bolzano, Italy) and G. Mazo (Univ. Catholique de Louvain, Belgique).

Copulas are a useful tool to model multivariate distributions [83]. However, while there exist various families of bivariate copulas, much fewer has been done when the dimension is higher. To this aim an interesting class of copulas based on products of transformed copulas has been proposed in the literature. The use of this class for practical high dimensional problems remains challenging. Constraints on the parameters and the product form render inference, and in particular the likelihood computation, difficult. As an alternative, we proposed a new class of copulas constructed by introducing a latent factor. Conditional independence with respect to this factor and the use of a nonparametric class of bivariate copulas lead to interesting properties like explicitness, flexibility and parsimony. In particular, various tail behaviours are exhibited, making possible the modeling of various extreme situations [17], [22].

7.2.5. Level sets estimation

Participant: Stephane Girard.

Joint work with: G. Stupfler (Univ. Aix-Marseille).

The boundary bounding the set of points is viewed as the larger level set of the points distribution. This is then an extreme quantile curve estimation problem. We proposed estimators based on projection as well as on kernel regression methods applied on the extreme values set, for particular set of points [10]. We also investigate the asymptotic properties of existing estimators when used in extreme situations. For instance, we have established in collaboration with G. Stupfler that the so-called geometric quantiles have very counter-intuitive properties in such situations [20] and thus should not be used to detect outliers.

7.2.6. Robust Sliced Inverse Regression.

Participants: Stephane Girard, Alessandro Chiancone, Florence Forbes.

This research theme was supported by a LabEx PERSYVAL-Lab project-team grant.

Sliced Inverse Regression (SIR) has been extensively used to reduce the dimension of the predictor space before performing regression. Recently it has been shown that this technique is, not surprisingly, sensitive to noise. Different approaches have thus been proposed to robustify SIR. In [14], we start considering an inverse problem proposed by R.D. Cook and we show that the framework can be extended to take into account a non-Gaussian noise. Generalized Student distributions are considered and all parameters are estimated via an EM algorithm. The algorithm is outlined and tested comparing the results with different approaches on simulated data. Results on a real dataset show the interest of this technique in presence of outliers.

7.2.7. Collaborative Sliced Inverse Regression.

Participants: Stephane Girard, Alessandro Chiancone.

This research theme was supported by a LabEx PERSYVAL-Lab project-team grant.

Joint work with: J. Chanussot (Gipsa-lab and Grenoble-INP).

In his PhD thesis work, Alessandro Chiancone studies the extension of the SIR method to different sub-populations. The idea is to assume that the dimension reduction subspace may not be the same for different clusters of the data [15]. One of the difficulty is that standard Sliced Inverse Regression (SIR) has requirements on the distribution of the predictors that are hard to check since they depend on unobserved variables. It has been shown that, if the distribution of the predictors is elliptical, then these requirements are satisfied. In case of mixture models, the ellipticity is violated and in addition there is no assurance of a single underlying regression model among the different components. Our approach clusterizes the predictors space to force the condition to hold on each cluster and includes a merging technique to look for different underlying models in the data. A study on simulated data as well as two real applications are provided. It appears that SIR, unsurprisingly, is not able to deal with a mixture of Gaussians involving different underlying models whereas our approach is able to correctly investigate the mixture.

7.2.8. Hapke's model parameter estimation from photometric measurements

Participants: Florence Forbes, Emeline Perthame.

Joint work with: Sylvain Douté (IPAG, Grenoble).

The Hapke's model is a widely used analytical model in planetology to describe the spectro-photometry of granular materials. It is a non linear model F that links a set of parameters x to a "theoretical" Bidirectional Reflectance Diffusion Function (BRDF). In practice, we assume that the observed BRDF Y is a noisy version of the "theoretical" one

$$Y = F(x) + \epsilon \tag{1}$$

where ϵ is a centered Gaussian noise with diagonal covariance matrix Σ. Then x is also assumed to be random with some prior distribution to be specified, e.g. uniform on the parameters range in [84]. The overall goal is to estimate the posterior distribution $p(x|y)$ for some observed BRDF y. Equation (5) defines the likelihood of the model which is $p(y|x) = \mathcal{N}(y; F(x), \Sigma)$. Then since F is non linear, it is not possible to obtain an analytical expression for $p(x|y)$. However, it is easy to simulate parameters x that follows the posterior distribution $p(x|y) \propto p(y|x)\ p(x)$ for instance using MCMC techniques [84]. If only point estimate are desired, the MAP can be used and evolutionary algorithms can then be used also using $p(y|x)\ p(x)$ as a fitness function. But obtaining such simulations is time consuming and has to be done for each observed value of y. In this work, we propose to use a locally linear mapping approximation and an inverse regression strategy to provide an analytical expression of $p(x|y)$. The idea is that the non linear F can be approximated by a number K of locally linear functions and that each of this function is easy to inverse. It follows that the inverse of F is also approximated as locally linear. Preliminary results were presented at the MultiPlaNet workshop in Orsay, December 14, 2016. They show that the proposed method does not fully reproduce the previous results obtained using MCMC techniques. Further investigations are required to understand the origin of the difference. Also ABC (approximate Bayes computation) methods will be considered as a subsequent step that may improved the current procedure while remaining computationally efficient.

7.2.9. Prediction intervals for inverse regression models in high dimension

Participant: Emeline Perthame.

Joint work with: Emilie Devijver (KU Leuven, Belgium).

Inverse regression, as a dimension reduction technique, is a reliable and efficient approach to handle large regression issues in high dimension, when the number of features exceeds the number of observations. Indeed, under some conditions, dealing with the inverse regression problem associated to a forward regression problem drastically reduces the number of parameters to estimate and make the problem tractable. However, regression models are often used to predict a new response from a new observed profile of covariates, and we may be interested in deriving confidence bands for the prediction to quantify the uncertainty around a predicted response. Theoretical results have already been derived for the well-known linear model, but recently, the curse of dimensionality has increased the interest of practitioners and theoreticians into generalization of those results on a high-dimension context. When both the responses and the covariates are multivariate, we derive in this work theoretical prediction bands for the inverse regression linear model and propose an analytical expression of these intervals. The feasibility, the confidence level and the accuracy of the proposed intervals are also analyzed through a simulation study.

7.2.10. Multi sensor fusion for acoustic surveillance and monitoring

Participants: Florence Forbes, Jean-Michel Becu.

Joint work with: Pascal Vouagner and Christophe Thirard from ACOEM company.

In the context of the DGA-rapid WIFUZ project with the ACOEM company, we addressed the issue of determining the localization of shots from multiple measurements coming from multiple sensors. We used Bayesian inversion and simulation techniques to recover multiple sources mimicking collaborative interaction between several vehicles. This project is at the intersection of data fusion, statistics, machine learning and acoustic signal processing. The general context is the surveillance and monitoring of a zone acoustic state from data acquired at a continuous rate by a set of sensors that are potentially mobile and of different nature. The overall objective is to develop a prototype for surveillance and monitoring that is able to combine multi sensor data coming from acoustic sensors (microphones and antennas) and optical sensors (infrared cameras) and to distribute the processing to multiple algorithmic blocs.

7.3. Graphical and Markov models

7.3.1. Conditional independence properties in compound multinomial distributions

Participant: Jean-Baptiste Durand.

Joint work with: Pierre Fernique (Inria, Virtual Plants) and Jean Peyhardi (Université de Montpellier).

We developed a unifying view of two families of multinomial distributions: the singular – for modeling univariate categorical data – and the non-singular – for modeling multivariate count data. In the latter model, we introduced sum-compound multinomial distributions that encompass re-parameterization of non-singular multinomial and negative multinomial distributions. The estimation properties within these compound distributions were obtained, thus generalizing know results in univariate distributions to the multivariate case. These distributions were used to address the inference of discrete-state models for tree-structured data. In particular, they were used to introduce parametric generation distributions in Markov-tree models [66].

7.3.2. Change-point models for tree-structured data

Participant: Jean-Baptiste Durand.

Joint work with: Pierre Fernique (Inria) and Yann Guédon (CIRAD), Inria Virtual Plants.

In the context of plant growth modelling, methods to identify subtrees of a tree or forest with similar attributes have been developed. They rely either on hidden Markov modelling or multiple change-point approaches. The latter are well-developed in the context of sequence analysis, but their extensions to tree-structured data are not straightforward. Their advantage on hidden Markov models is to relax the strong constraints regarding dependencies induced by parametric distributions and local parent-children dependencies. Heuristic approaches for change-point detection in trees were proposed and applied to the analysis of patchiness patterns (consisting of canopies made of clumps of either vegetative or flowering botanical units) in mango trees [43].

7.3.3. Hidden Markov models for the analysis of eye movements

Participants: Jean-Baptiste Durand, Brice Olivier.

This research theme is supported by a LabEx PERSYVAL-Lab project-team grant.

Joint work with: Marianne Clausel (LJK) Anne Guérin-Dugué (GIPSA-lab) and Benoit Lemaire (Laboratoire de Psychologie et Neurocognition)

In the last years, GIPSA-lab has developed computational models of information search in web-like materials, using data from both eye-tracking and electroencephalograms (EEGs). These data were obtained from experiments, in which subjects had to make some kinds of press reviews. In such tasks, reading process and decision making are closely related. Statistical analysis of such data aims at deciphering underlying dependency structures in these processes. Hidden Markov models (HMMs) have been used on eye movement series to infer phases in the reading process that can be interpreted as steps in the cognitive processes leading to decision. In HMMs, each phase is associated with a state of the Markov chain. The states are observed indirectly though eye-movements. Our approach was inspired by Simola et al. (2008), but we used hidden semi-Markov models for better characterization of phase length distributions. The estimated HMM highlighted contrasted reading strategies (ie, state transitions), with both individual and document-related variability. However, the characteristics of eye movements within each phase tended to be poorly discriminated. As a result, high uncertainty in the phase changes arose, and it could be difficult to relate phases to known patterns in EEGs.

This is why, as part of Brice Olivier's PhD thesis, we are developing integrated models coupling EEG and eye movements within one single HMM for better identification of the phases. Here, the coupling should incorporate some delay between the transitions in both (EEG and eye-movement) chains, since EEG patterns associated to cognitive processes occur lately with respect to eye-movement phases. Moreover, EEGs and scanpaths were recorded with different time resolutions, so that some resampling scheme must be added into the model, for the sake of synchronizing both processes.

To begin with, we first proved why HMM would be the best option in order to conduct this analysis and what could be the alternatives. A brief state of the art was made on models similar to HMMs. However, since our data is very specific, we needed to make use of unsupervised graphical generative models for the analysis of sequences which would keep a deep meaning. It resulted that Hidden semi-Markov model (HSMM) was the most powerful tool satisfying all our needs. Indeed, a HSMM is characterized by meaningful parameters such as an initial distribution, transition distributions, emission distributions and sojourn distributions, which allows us to directly characterize a reading strategy. Second, we found and improved an existing implementation of such a model. After searching for libraries to make inference in HSMM, the Vplants library embedded in the OpenAlea software turned out to be the most viable solution regarding the functionalities, though it was still incomplete. Consequently, we proposed improvements to this library and added functions in order to boost the likelihood of the data. This lead us to also propose a new library included in that software which is specific at the analysis of eye movements. Third, in order to improve and validate the interpretation of the reading strategies, we calculated indicators specific to each reading strategy. Fourth, since the parameters obtained from the model suggested individual and text variability, we first investigated text clustering to reduce the variability of the model. In order to do this, we supervised a group of 6 students to explore the text clustering component with the mission of clustering the texts by evolution of the semantic similarity throughout text. We therefore explored different methods for time series clustering and we retained the usage of Ascendant Hierarchical Clustering (AHC) using the Dynamic Time Warping (DTW) metric, which allows

global dynamics of the time series to be captured, but not local dynamics. Plus, we preferred the simplicity and good understanding of the results using that method. Therefore, we deduced three text profiles giving meaning to the evolution of the semantic similarity: a step profile, a ramp profile, and a saw profile. With that new information in hand, we are now able to decompose our model over text profiles and hence, reduce its variability.

As discussed in the previous section, our work is focused on the standalone analysis of the eye-movements. We are currently polishing this phase of work. The common work and the goal for this coming year is to develop and implement a model for jointly analyzing eye-movements and EEGs in order to improve the discrimination of the reading strategies.

7.3.4. Lossy compression of tree structures

Participant: Jean-Baptiste Durand.

Joint work with: Christophe Godin (Inria, Virtual Plants) and Romain Azais (Inria BIGS)

In a previous work [79], a method to compress tree structures and to quantify their degree of self-nestedness was developed. This method is based on the detection of isomorphic subtrees in a given tree and on the construction of a DAG (Directed Acyclic Graph), equivalent to the original tree, where a given subtree class is represented only once (compression is based on the suppression of structural redundancies in the original tree). In the lossless compressed graph, every node representing a particular subtree in the original tree has exactly the same height as its corresponding node in the original tree. A lossy version of the algorithm consists in coding the nearest self-nested tree embedded in the initial tree. Indeed, finding the nearest self-nested tree of a structure without more assumptions is conjectured to be an NP-complete or NP-hard problem. We obtained new theoretical results on the combinatorics of self-nested structures [60]. We improved this lossy compression method by computing a self-nested reduction of a tree that better approximates the initial tree. The algorithm has polynomial time complexity for trees with bounded outdegree. This approximation relies on an indel edit distance that allows (recursive) insertion and deletion of leaf vertices only. We showed using a simulated dataset that the error rate of this lossy compression method is always better than the loss based on the nearest embedded self-nestedness tree [79] while the compression rates are equivalent. This procedure is also a keystone in our new topological clustering algorithm for trees. Perspectives of improving the time complexity of our algorithm include taking profit from one of its byproduct, which could be used as an indicator of both the number of potential candidates to explore and of the proximity of the tree to the nearest self-nested tree.

7.3.5. Learning the inherent probabilistic graphical structure of metadata

Participants: Thibaud Rahier, Stephane Girard, Florence Forbes.

Joint work with: Sylvain Marié, Schneider Electric.

The quality of prediction and inference on temporal data can be significantly improved by taking advantage of the associated metadata. However, metadata are often only partially structured and may contain missing values. In the context of T. Rahier's PhD with Schneider Electric, we first considered the problem of learning the inherent probabilistic graphical structure of metadata, which has two main benefits: (i) graphical models are very flexible and therefore enable the fusion of different types of data together (ii) the learned graphical model can be interrogated to perform tasks on metadata alone: variable clustering, conditional independence discovery or missing data replenishment. Bayesian Network (and more generally Probabilistic Graphical Model) structure learning is a tremendous mathematical challenge, that involves a NP-Hard optimisation problem. In the past year, we have explored many approaches to tackle this issue, and begun to develop a tailor-made algorithm, that exploits dependencies typically present in metadata, and that significantly speeds up the structure learning task and increases the chance of finding the optimal structure.

7.3.6. Robust Graph estimation

Participants: Karina Ashurbekova, Florence Forbes.

Joint work with: Sophie Achard, CNRS, Gipsa-lab.

In the face of increasingly high dimensional data and of trying to understand the dependency/association present in the data the literature on graphical modelling is growing rapidly and covers a range of applications (from bioinformatics e.g gene expression data to document modelling). A major limitation of recent work on using the (standard) Student t distribution for robust graphical modelling is the lack of independence and conditional independence of the Student t distribution, and estimation in this context (with the standard student t) is very difficult. We propose to develop and assess a generalized Student t from a new family (which has independence and conditional independence as special properties) for the general purpose of graphical modelling in high dimensional settings. Its main characteristic is to include multivariate heavy-tailed distributions with variable marginal amounts of tailweight that allow more complex dependencies than the standard case. We target an application to brain connectivity data for which standard Gaussian graphical models have been applied. Brain connectivity analysis consists in the study of multivariate time series representing local dynamics at each of multiple sites or sources throughout the whole human brain while functioning using for example functional magnetic resonance imaging (fMRI). The acquisition is difficult and often spikes are observed due to the movement of the subjects inside the scanner. In the case of identifying Gaussian graphical models, the glasso technique has been developed for estimating sparse graphs. However, this method can be severely impacted by the inclusion of only a few contaminated values, such as spikes that commonly occur in fMRI time series, and the resulting graph has the potential to contain false positive edges. Therefore, our goal was to assess the performance of more robust methods on such data.

7.4. Robust non Gaussian models

7.4.1. Robust Locally linear mapping with mixtures of Student distributions

Participants: Florence Forbes, Emeline Perthame, Brice Olivier.

The standard GLLiM model [6] for high dimensional regression assumes Gaussian noise models and is in its unconstrained version equivalent to a joint GMM. The fact that response and independent variables (X, Y) are jointly a mixture of Gaussian distribution is the key for all derivations in the model. In this work, we show that similar developments are possible based on a joint Student Mixture model, joint SMM. It follows a new model referred to as SLLiM for Student Locally linear mapping for which we investigate the robustness to outlying data in a high dimensional regression context [71]. The corresponding code is available on the CRAN in the *xLLiM* package.

7.4.2. Rectified binaural ratio: A complex T-distributed feature for robust sound localization

Participant: Florence Forbes.

Joint work with: Antoine Deleforge, Inria PANAMA team in Rennes.

Most existing methods in binaural sound source localization rely on some kind of aggregation of phase-and level-difference cues in the time-frequency plane. While different aggregation schemes exist, they are often heuristic and suffer in adverse noise conditions. In this work, we introduce the rectified binaural ratio as a new feature for sound source localization. We show that for Gaussian-process point source signals corrupted by stationary Gaussian noise, this ratio follows a complex t-distribution with explicit parameters. This new formulation provides a principled and statistically sound way to aggregate binaural features in the presence of noise. We subsequently derive two simple and efficient methods for robust relative transfer function and time-delay estimation. Experiments on heavily corrupted simulated and speech signals demonstrate the robustness of the proposed scheme. This work has been presented at the Eusipco conference in 2016 [30].

7.4.3. Statistical reconstruction methods for multi-energy tomography

Participants: Florence Forbes, Pierre-Antoine Rodesch.

Joint work with: Veronique Rebuffel from CEA Grenoble.

In the context of Pierre-Antoine Rodesh's PhD thesis, we investigate new statistical and optimization methods for tomographic reconstruction from non standard detectors providing multiple energy signals.

7.5. Statistical models for Neuroscience

7.5.1. Advanced statistical analysis of functional Arterial Spin Labelling data

Participants: Florence Forbes, Aina Frau Pascual.

Joint work with: Philippe Ciuciu from Team PARIETAL and Neurospin, CEA Saclay.

Arterial Spin Labelling (ASL) is a non-invasive perfusion MR imaging technique that can be also used to measure brain function (fASL for functional ASL). In contrast to BOLD fMRI, it gives a quantitative and absolute measure of cerebral blood flow (CBF), making this modality appealing for clinical neuroscience and patient's follow-up over longitudinal studies. However, its limited signal-to-noise ratio makes the analysis of fASL data challenging. In this work, we compared different approaches (GLM vs JDE) in the analysis of functional ASL data for the detection of evoked brain activity at the group level during visual and motor task performance. Our dataset has been collected at Neurospin on a 3T Tim Trio Siemens scanner (CEA Saclay, France), during the HEROES project (Inria Grant). It contains BOLD data (165 scans, TR=2.5s, TE=30ms, 3x3x3mm3) and functional pulsed ASL data (Q2TIPS PICORE scheme [Luh,00], 165 scans, TR=2.5s, TE=11ms, 3x3x7.5 mm3) of 13 right-handed subjects (7 men and 6 women) of age between 20 and 29. The experimental design consists of a mini-block paradigm of visual, motor and auditory tasks with 16 blocks of 15s each followed by 10s of rest. Data have been scaled, realigned, and normalized. For univariate analysis, the images have also been spatially smoothed with a Gaussian kernel of 5 mm full width half at maximum. Three data analysis approaches have been compared: (a) univariate General Linear Model (GLM) that considers canonical shapes for the perfusion and hemodynamic responses; (b) physiologically informed joint detection estimation (PI-JDE) [4] that jointly estimates effect maps and response functions in a multivariate manner in a Bayesian framework; (c) A restricted version of PI-JDE that considers fixed canonical shapes for the perfusion and hemodynamic responses (PRF and HRF, respectively), defining an intermediate approach between the first two. Since methods (b)-(c) embed adaptive spatial regularization, they do not require a preliminary smoothing of the data. Our results demonstrate that the PI-JDE multivariate approach is a competing alternative to GLM for the analysis of fASL: it recovers more localized and stronger effects. Our findings also replicate the state-of-the-art by showing more localized activation patterns in perfusion as compared to hemodynamics.

7.5.2. BOLD VEM multi session extension of the JDE approach

Participants: Florence Forbes, Aina Frau Pascual.

Joint work with: Philippe Ciuciu from Team PARIETAL and Neurospin, CEA Saclay.

The fast solution of the JDE approach for BOLD fMRI presented in [5] uses a variational expectation maximization (VEM) algorithm and considers a single session of BOLD data. This paper shows the faster performance of this algorithm with respect to the Markov Chain Monte Carlo (MCMC) approach presented in earlier work, with similar results. In fMRI, usually several sessions are acquired for the same subject to be able to compare them or combine them. In [73], a multiple-session extension of the JDE approach has been proposed to analyze several sessions together. The solution proposed uses MCMC and considers that the response levels have a mean value per condition and a common variance between sessions. In the context of Aina Frau's PhD, a VEM solution of this extension has been implemented. Experimental results have shown that the solution of the multiple-session VEM is not very different from the average of the results computed with single session VEM. For this reason, we proposed a heteroscedastic version of the multiple-session VEM. It amounts to considering session-specific variances. The goal is to be able to weight the importance of the different sessions so as to diminish the contribution of any potential noisy session to the final parameter estimates.

7.5.3. Estimating biophysical parameters from multimodal fMRI data

Participants: Florence Forbes, Pablo Mesejo Santiago.

Joint work with: Jan Warnking from Grenoble Institute of Neuroscience.

Functional Magnetic Resonance Imaging (fMRI) indirectly studies brain function. With Jan M. Warnking (Grenoble Institute of Neurosciences) we worked on the estimation of biophysical parameters from fMRI signals. We first used only BOLD signals, using a stochastic population-based optimization method to estimate 15 parameters without neither providing initial estimates nor computing gradients. Initial results were published at MICCAI 2015 and in the IEEE JSTSP journal [81], [82]. Also a MATLAB toolbox was released (see software section). The current ongoing work is to study the impact of the combination of different fMRI modalities in the estimation of this biophysical parameters. We can use 3 fMRI modalities (BOLD, ASL and MION) and 13 rats. We ran our optimizer with all possible combinations of modalities. The initial hypothesis was that as long as we introduce more fMRI modalities we would like to see more consistent estimates but we need to assess possible limits due to potential lack of data: only 13 rats, 6 of them without MION, and potential outliers among the rats that would better be excluded from the analysis.

7.5.4. Multi-subject joint parcelation detection estimation in functional MRI

Participant: Florence Forbes.

Joint work with: Lotfi Chaari, Mohanad Albughdadi, Jean-Yves Tourneret from IRIT-ENSEEIHT in Toulouse and Philippe Ciuciu from Neurospin, CEA Saclay.

fMRI experiments are usually conducted over a population of interest for investigating brain activity across different regions, stimuli and subjects. Multi-subject analysis usually proceeds in two steps: an intra-subject analysis is performed sequentially on each individual and then a group-level analysis is carried out to report significant results at the population level. This work considers an existing Joint Parcellation Detection Estimation (JPDE) model which performs joint hemodynamic parcellation, brain dynamics estimation and evoked activity detection. The hierarchy of the JPDE model is extended for multi-subject analysis in order to perform group-level parcellation. Then, the corresponding underlying dynamics is estimated in each parcel while the detection and estimation steps are iterated over each individual. Validation on synthetic and real fMRI data shows its robustness in inferring group-level parcellation and the corresponding hemodynamic profiles. This work has been presented at ISBI 2016 [42].

7.5.5. Automatic segmentation and characterization of brain tumors using robust multivariate clustering of multiparametric MRI

Participants: Florence Forbes, Alexis Arnaud.

Joint work with: Emmanuel Barbier and Benjamin Lemasson from Grenoble Institute of Neuroscience.

Brain tumor segmentation is a difficult task in the field of multiparametric MRI analysis because of the number of maps that are available. Furthermore, the characterization of brain tumors can be time-consuming, even for medical experts, and the reference method is biopsy which is a local and invasive technique. Because of this, it is important to develop automatic and non-invasive approaches in order to help the medical expert with these issues. In this study we use a robust statistical model-based method to classify multiparametric MRI of rat brains. The voxels are gather into classes resulting from multivariate multi-scaled Student distributions, which can accommodate outliers. First we adjust a mixture model on a reference group of rats to learn the MRI characteristics of healthy tissues. Second we use this model to delineate the brain tumors as atypical voxels in the data set of unhealthy rats. Third we adjust a new mixture model only on the atypical voxels to learn the MRI characteristics of tumorous tissues. Finally, we extract a fingerprint for each tumor type to make a tumor dictionary.

Our data set is composed of healthy rats (n=8 rats) and 4 groups of rats bearing a brain tumor model (n=8 per group). For each rat, we acquired 5 quantitative MRI parameters along 5 slices. And the proposed tumor dictionary reaches a rate of 75% of accurate prediction with a leave-one-out procedure.

7.5.6. Monitoring brain tumor evolution using multiparametric MRI

Participants: Florence Forbes, Alexis Arnaud.

Joint work with: Emmanuel Barbier, Nora Collomb and Benjamin Lemasson from Grenoble Institute of Neuroscience.

Analyzing brain tumor tissue composition can improve the handling of tumor growth and resistance to therapies. We showed on a 6 time point dataset of 8 rats that multiparametric MRI could be exploited via statistical clustering to quantify intra-lesional heterogeneity in space and time. More specifically, MRI can be used to map structural, eg diffusion, as well as functional, eg volume (BVf), vessel size (VSI), oxygen saturation of the tissue (StO2), characteristics. In previous work, these parameters were analyzed to show the great potential of multiparametric MRI (mpMRI) to monitor combined radio- and chemo-therapies. However, to exploit all the information contained in mpMRI while preserving information about tumor heterogeneity, new methods need to be developed. We demonstrated the ability of clustering analysis applied to longitudinal mpMRI to summarize and quantify intra-lesional heterogeneity during tumor growth. This study showed the interest of a clustering analysis on mpMRI data to monitor the evolution of brain tumor heterogeneity. It highlighted the type of tissue that mostly contributes to tumor development and could be used to refine the evaluation of therapies and to improve tumor prognosis.

7.5.7. Assessment of tissue injury in severe brain trauma

Participant: Florence Forbes.

Joint work with: Michel Dojat and Christophe Maggia from Grenoble Institute of Neuroscience and Senan Doyle from Pixyl.

Traumatic brain injury (TBI) remains a leading cause of death and disability among young people worldwide and current methods to predict long-term outcome are not strong. TBI initiates a cascade of events that can lead to secondary brain damage or exacerbate the primary injury, and these develop hours to days after the initial accident. The concept of secondary brain damage is the focus of modern TBI management in Intensive Care Units. The imbalance between oxygen supply to the brain tissue and utilization, i.e. brain tissue hypoxia, is considered the major cause for the development of secondary brain damage, and hence poor neurological outcome. Monitoring brain tissue oxygenation after TBI using brain tissue O_2 pressure (PbtO_2) probes surgically inserted into the parenchyma, may help clinicians to initiate adequate actions when episodes of brain ischemia/hypoxia are identified. The aggressive treatment of low PbtO_2 values (< 15mmHg for more than 30 minutes) was associated with better outcome compared to standard therapy in some cohort studies of severe head-injury patients. However, another study was unable to find similar benefits to patient outcome. MRI is an excellent modality for estimating global and regional alterations in TBI and for following their longitudinal evolution. To assess the complexity of TBI, several morphological sequences are required for assessing volume loss. Moreover, diffusion tensor imaging (DTI) offers the most sensitive modality for the detection of changes in the acute phase of TBI and increases the accuracy of long-term outcome prediction compared to the available clinical/radiographic prognostic score. Mean Diffusivity (MD) or Apparent Diffusion Coefficient (ADC) have been widely used to determine the volume of ischemic tissue, and assess intra- and extracellular conditions. A reduction of MD is related to cytotoxic edema (intracellular) while an increase of MD indicates a vasogenic edema (extracellular). Changes of MD are expected with severe TBI. The volume of lesions on DTI shows a strong correlation with neurological outcome at patient discharge. We consider a clinically relevant criterion to be the volume of vulnerable brain lesions after TBI, as previously suggested. In consequence, we need an automatic segmentation method to assess the tissue damage in severe trauma, acute phase i.e. before 10 days after the event. Skull deformation, the presence of blood in the acute phase, the high variability of brain damage that excludes the use of anatomical *a priori* information, and the diffuse aspect of brain injury affecting potentially all brain structures, render TBI segmentation particularly demanding. The methods proposed in the literature are mainly concerned with volumetric changes following TBI and scarcely report lesion load. In this work, we report our methodological developments to assess lesion load in severe brain trauma in the entire brain. We use P-LOCUS to perform brain tissue segmentation and exclude voxels labeled as CSF, ventricle and

hemorrhagic lesion. We propose a fusion of several atlases to parcel cortical, subcortical and WM structures into well identified regions where MD values can be expected to be homogenous. Abnormal voxels are detected in these regions by comparing MD values with normative values computed from healthy volunteers. The preliminary results, evaluated in a single center, are a first step in defining a robust methodology intended to be used in multi-center studies. This work has been published in [58].

7.5.8. Automatic multiple sclerosis lesion segmentation with P-Locus

Participant: Florence Forbes.

Joint work with: Michel Dojat from Grenoble Institute of Neuroscience and Senan Doyle from Pixyl.

P-LOCUS provides automatic quantitative neuroimaging biomarker extraction tools to aid diagnosis, prognosis and follow-up in multiple sclerosis studies. The software performs accurate and precise segmentation of multiple sclerosis lesions in a multi-stage process. In the first step, a weighted Gaussian tissue model is used to perform a robust segmentation. The algorithm avails of complementary information from multiple MR sequences, and includes additional estimated weight variables to account for the relative importance of each voxel. These estimated weights are used to define candidate lesion voxels that are not well described by a normal tissue model. In the second step, the candidate lesion regions are used to populate the weighted Gaussian model and guide convergence to an optimal solution. The segmentation is unsupervised, removing the need for a training dataset, and providing independence from specific scanner type and MRI scanner protocol. The procedure was applied to participate to the MSSEG Challenge at Miccai 2016 in Athen: Multiple Sclerosis Lesions Segmentation Challenge Using a Data Management and Processing Infrastructure [55].

NANO-D Project-Team

6. New Results

6.1. Development of a novel minimization method

Participants: Clement Beitone, Stephane Redon.

Finding the optimized configuration of a system of particles so that it minimizes the energy of the system is a very common task in the field of particles simulation. More precisely, we are interested in finding the closest atomic structure located at a minima on the Potential Energy Surface (PES) starting from a given initial configuration. Achieving faster but reliable minimizations of such systems help to enhance a wide range of applications in molecular dynamics. To improve the efficiency of the convergence some authors have proposed alternative methods to the steepest descent algorithm; for example, the conjugate gradient technique or the Fast Inertial Relaxation Engine (FIRE).

In this work, we are developing a novel method that helps to increase the efficiency and the reliability of existing optimizers, *e.g.* FIRE and Interactive Modelling (IM).

We have implemented the modified versions of these algorithms along with others optimization algorithms like L-BFGS and Conjugate Gradient as state updaters in SAMSON. To assess the efficiency of the proposed approaches we have developed an App in SAMSON that allows us to reliably and conveniently probe several criteria during the minimization process (Figure 3).

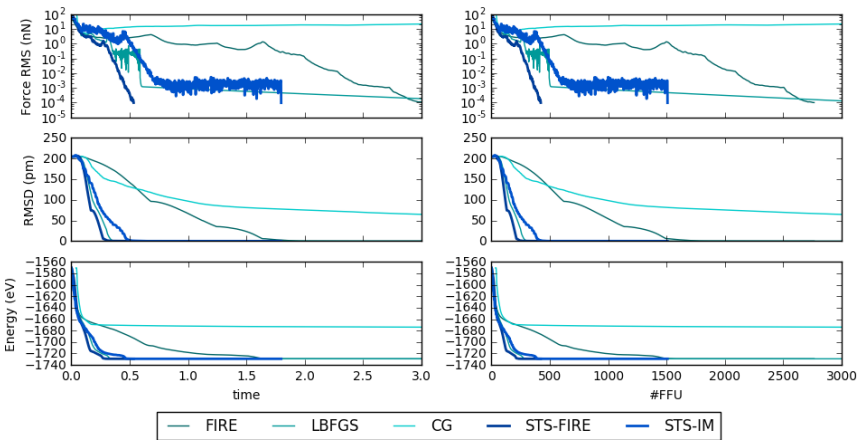


Figure 3. Comparison of different optimizers with the proposed methods on the fullerene C240. For this experiment the force field used to model the interactions between the atoms is the Brenner potential.

6.2. Parallel algorithms for adaptive molecular dynamics simulations

Participants: Dmitriy Marin, Stephane Redon.

We have developed a parallel implementation of Adaptively Restrained Particle Simulations (ARPS) in LAMMPS Molecular Dynamics Simulator with the usage of Kokkos [0] package. The main idea of the ARPS method [22] is to speed up particle simulations by adaptively switching on and off positional degrees of freedom, while letting momenta evolve; this is done by using adaptively restrained Hamiltonian. The developed parallel implementation allows us to run LAMMPS with ARPS integrator on central processing units (CPU), graphics processing units (GPU), or many integrated core architecture (MIC). We modified the ARPS algorithm for efficient usage of GPU and many-core CPU, e.g. all computations were parallelized for efficient calculations on computational device; communications between host and device were decreased.

To measure speed up of the developed parallel implementation we used several benchmarks and heterogeneous computational systems with next parameters: 2x CPU Intel Xeon E5-2680 v3 (24 cores in total), GPU Nvidia Quadro K4200, GPU Nvidia Tesla K20c. Results on the speed up in comparison with serial ARPS code for one of the benchmarks (Lennard–Jones liquid, 515K atoms, $\sim$1% of particles switches their state at each timestep from active to restrained or from restrained to active) are shown in Figure 4 . It can be seen, that for small number of CPU cores the speed up is almost constant for all the percentage of active atoms in the system. But for large number of CPU cores and for GPUs the speed up is decreasing with decreasing percentage of active atoms, because of divergence of threads and limited occupancy. The achieved speed up on 20 CPU-cores is up to 14 times, on GPU Nvidia Tesla K20c is up to 24 times.

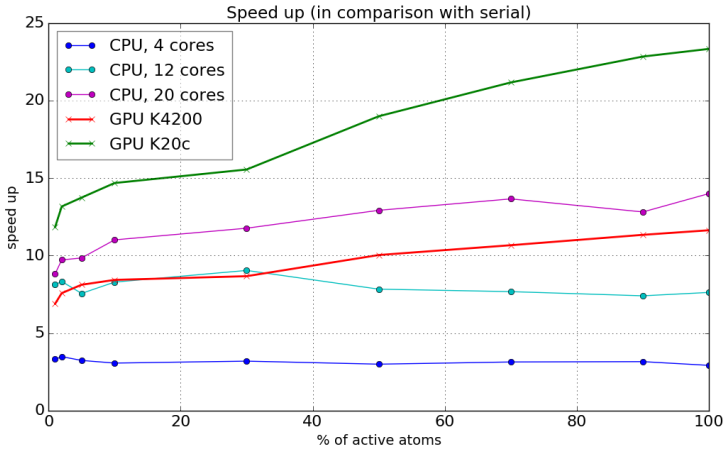


Figure 4. The parallel ARPS results

6.3. Adaptive Algorithms for Orbital-Free Density Functional Theory

Participants: Francois Rousse, Stephane Redon.

The SAMSON App developed to simulate molecular systems with an adaptive version of OF-DFT has been continued. It has been tested on several small systems : atoms, dimers, etc. The errors found on the energies and the bond length found were coherent with the predictive characteristics of OF-DFT and with other OF-DFT softwares like PROFESS.

[0] The Kokkos package is based on Kokkos library, which is a templated C++ library that provides two key abstractions: it allows a single implementation of an application kernel to run efficiently on different hardware, such as a many-core CPU, GPU, or MIC; it provides data abstractions to adjust (at compile time) the memory layout of basic data structures — like 2d and 3d arrays — for performance optimization on different platforms. These abstractions are set at build time (during compilation of LAMMPS).

The pseudopotentials computed by the Carter Group of Princeton (who developed PROFESS) have been implemented in the SAMSON App. The electronic densities became smoother and the predictions were improved, but it restricted the applicability of the SAMSON App since the pseudopotentials were computed only for the elements of the columns III (like aluminum) and V (like Potassium) of the periodic table.

Several optimization algorithms have been tried : projected gradient, Primal-Dual, Lagrangian multiplier improved with a penalization, different nonlinear conjugate gradient minimization algorithms ... None of them showed a clear superiority on the other in both stability and speed. Currently, we use the projected gradient since it is the most stable.

We have implemented an interaction model in SAMSON based on the OF-DFT code and tested its ability to predict the geometry of system on a small crystal of aluminum. The crystal contracted itself, which is coherent with the OF-DFT theory, since it tends to underestimate bond lengths, and with the surface tension, since it tends to minimize the surface of the system. The next step will be to make this interaction model adaptive and measure how much time is gained.

6.4. A crystal creator app

Participants: Francois Rousse, Stephane Redon.

We developed a new SAMSON Element able to generate models of crystals. The user can either write its own unit cell or load it from a CIF file ("Crystallographic Information File"). Once written or imported, this unit cell can be replicated again in every direction to generate a whole crystal. As the important characteristics of crystals often comes from the defects, the replacements and the insertions, these repetitions of unit cells are not mere copies but are whole new unit cells generated again each time. Thus a crystal with enough unit cells shall have the right proportion of elements, with the right amount of defects, replacements and insertions, randomly disposed. In the document view, the unit cells are separated to ease the manipulation of the crystal. Last, it allows the user to cut the crystal on the planes given by Miller indexes.

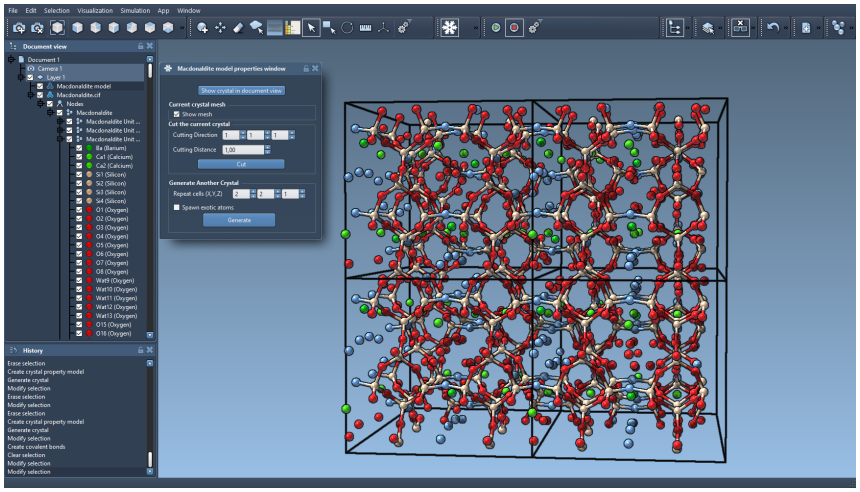


Figure 5. A Macdonaldite crystal generated in SAMSON

6.5. Software development process improvements

Participants: Jocelyn Gate, Stephane Redon.

We set up a Jenkins server on a virtual machine at Inria. The server is accessible to the team and is able to build and generate everything related to SAMSON. This Jenkins server is linked to diferents slaves, located in our offices:

- Window 7 / Windows 10
- Fedora 21 / Fedora 25 / Ubuntu 16.04
- MacOs 10.10.5

Slave machines are used by the Jenkins server to build the specified version of SAMSON, generate the associated SDK, build all SAMSON elements that are specified on Jenkins and upload everything to our private version of SAMSON Connect. Thanks to this, the team has access each day to the latest developments.

In order to efficiently upload everything from slaves nodes, Jenkins uses a private helpers that is able to communicate with SAMSON-Connect, and that knows every SAMSON files format.

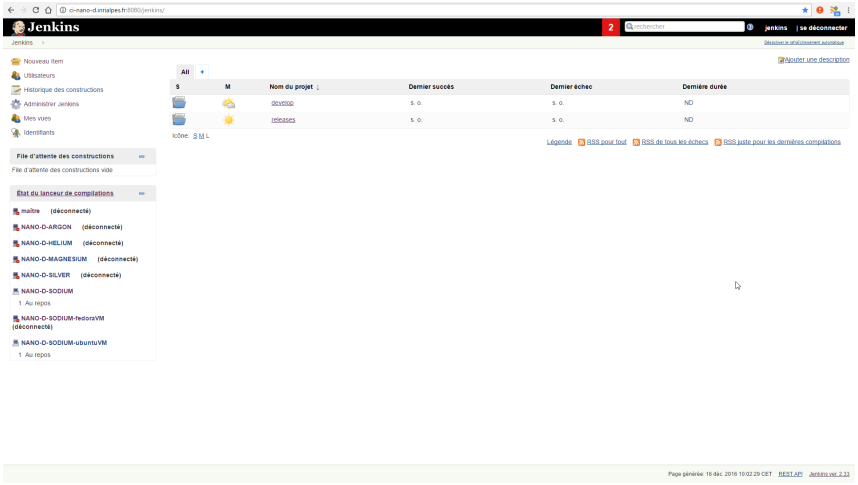


Figure 6. The jenkins interface

We developed a private, command line SAMSON helper that is able to do everything concerning the packing and the uploading of new versions of SAMSON, the SAMSON SDK and the installer to SAMSON Connect. It can:

- Upload the SAMSON or SAMSON-SDK packaged file to SAMSON-Connect (adding a new version of SAMSON/SAMSON-SDK).
- Upload the SAMSON or SAMSON-SDK Setup executable to SAMSON-Connect.
- Package the SAMSON elements of a developer to .element files.
- Upload .element files to SAMSON Connect.

6.6. Updates to SAMSON and SAMSON Connect

Participants: Jocelyn Gate, Stephane Redon.

To be able to know if SAMSON works well on users computers, we added some logging features inside SAMSON, SAMSON installers and SAMSON Helpers. Thanks to this functionality, users may accept to send logs when bugs are found. For example, if SAMSON crashes on a user computer, a log is generated, anonymized, and automatically sent to the SAMSON Connect web service. If SAMSON crashes because of a SAMSON Element, an email is sent to the author of the SAMSON Element. If a new user tries to install SAMSON or the SAMSON SDK, a log is sent to the SAMSON Connect web service.

We also added the possibility for users to configure proxy access to SAMSON Connect.
These functionalities will be part of the upcoming 0.6.0 release of SAMSON.

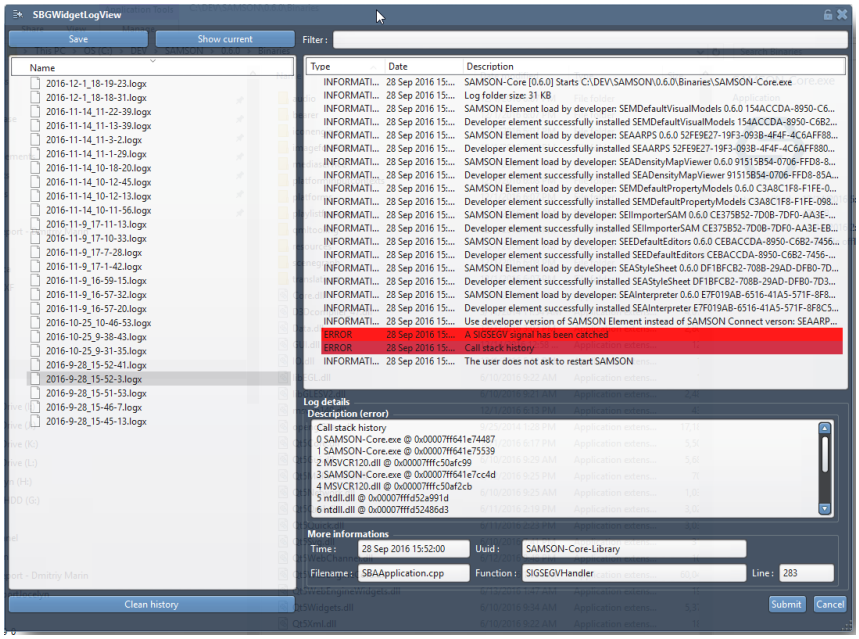


Figure 7. The SAMSON log interface

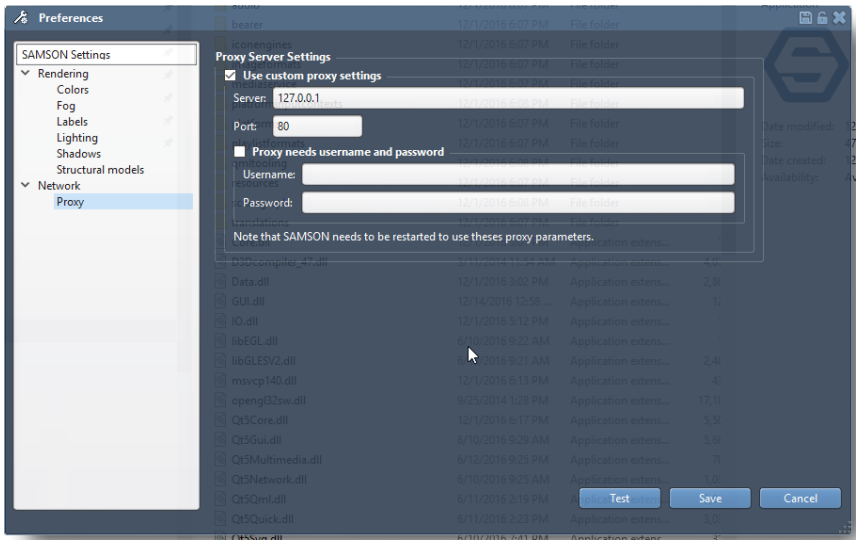


Figure 8. The proxy setting interface

6.7. As-Rigid-As-Possible molecular interpolation paths

Participants: Minh Khoa Nguyen, Leonard Jaillet, Stephane Redon.

We submitted a paper describing a new method to generate interpolation paths between two given molecular conformations. It applies the As-Rigid-As-Possible (ARAP) from the field of computer graphics to manipulate complex meshes while preserving their essential structural characteristics. The adaptation of ARAP interpolation approach to the case of molecular systems was presented. Experiments were conducted on a large set of benchmarks and the performance was compared between ARAP interpolation and linear interpolation. They show that ARAP interpolation generates more relevant paths, that preserve bond lengths and bond angles better.

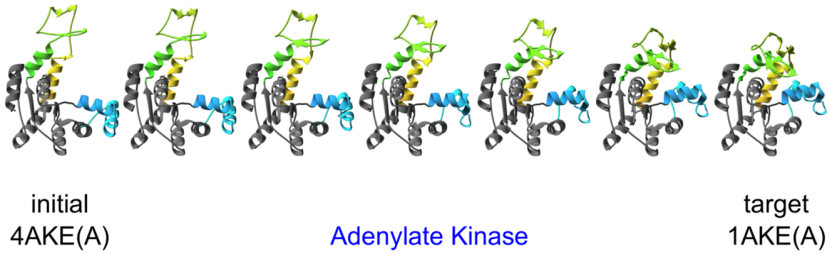


Figure 9. The morphing path for Adenylate Kinase from 4AKE (chain A) to 1AKE (chain A) by ARAP interpolation:

6.8. As-Rigid-As-Possible molecular interpolation paths

Participants: Krishna Kant Singh, Stephane Redon.

We have continued our work on the development of parallel adaptively restrained particle simulations. We proposed new algorithms to compute forces involving active particles faster. These algorithms involved construction of the Active Neighbor List (ANL) and incremental force computations. These algorithms have advantages over the state-of-the-art methods for simulating a system using Adaptively Restrained Molecular Dynamics (ARMD). Previously proposed algorithms required at-least 60% restrained particles in order to achieve speed up. In new algorithms, we overcome this limitation and speed up can be achieve with 10% retrained particles. We implemented our algorithm in the popular molecular dynamics package LAMMPS and submitted our results in the *Computer Physics Communications* Journal [0]. Figure 11 show that speed-up can be achieved for more than 10% of the particles are restrained. We also achieved significant speed up in constructing the ANL (figure 12).

6.9. Refining the energy landscape sampling of protein-protein associations

Participants: Dmytro Kozakov, Leonard Jaillet.

PIPER is a FFT-based protein docking program with pairwise potentials. It combines a systematic sampling procedure with an original pairwise potential that provides an energy landscape representation through a set of samples [48].

[0] K.K. Singh, S. Redon, Adaptively Restrained Molecular Dynamics in LAMMPS, Submitted to *Computer Physics Communications*.

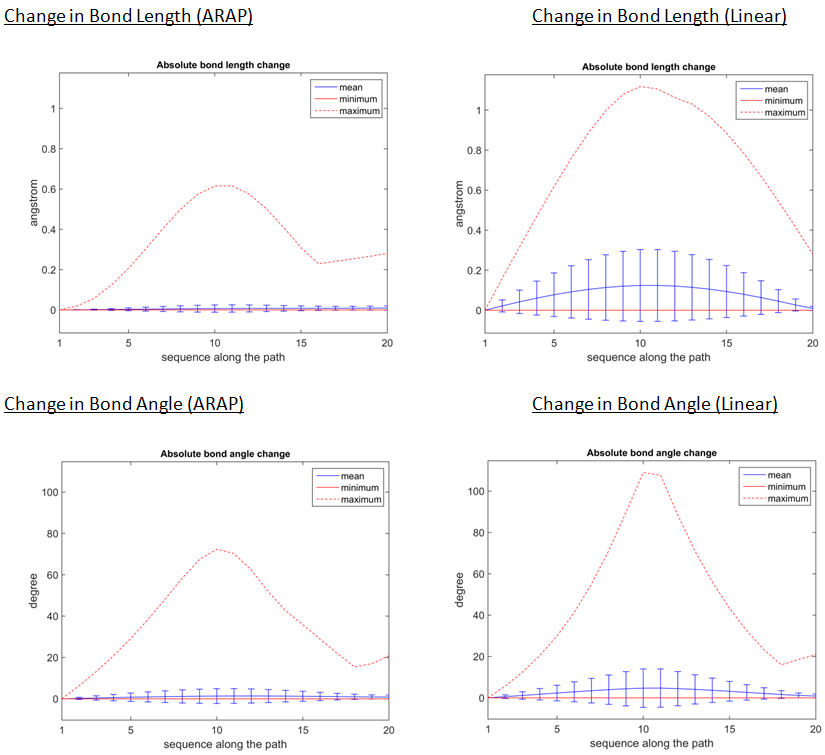


Figure 10. Comparison of ARAP and linear interpolation for preserving structural characteristics of adenylate kinase

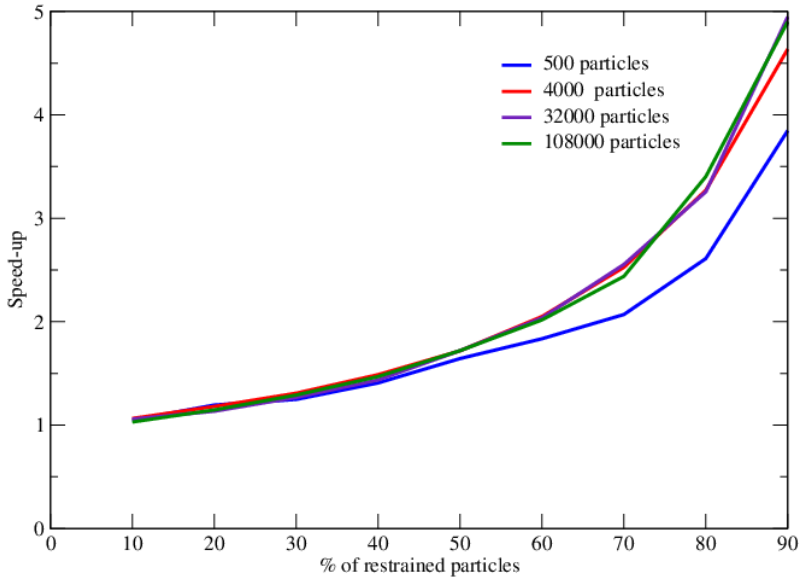


Figure 11. speed up using ARMD on different benchmark

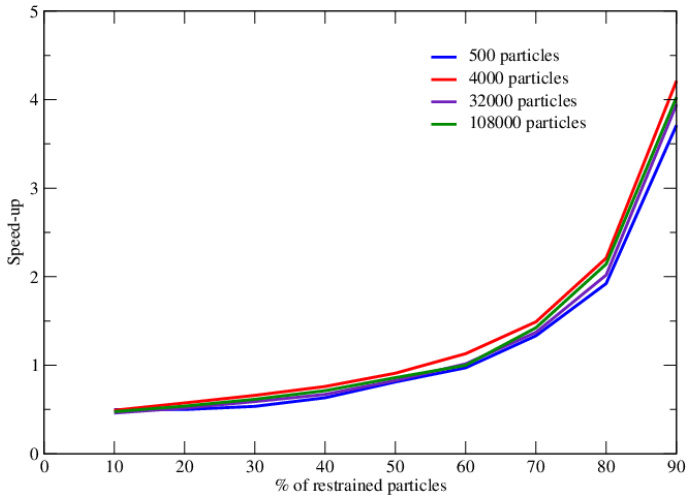


Figure 12. Obtained speed up in constructing the ANL.

In [49], an experimental validation of the complexes obtained with PIPER, has been made possible thanks to the PRE method [31]. PRE (NMR paramagnetic relaxation enhancement) is an experimental technique used to characterize the states present for a given system. Hence, it characterizes the accessible region of the energy landscape corresponding to a given protein. For this, it introduces paramagnetic labels (tags) one at a time at few sites on one protein. The method then relies on measures of the transverse paramagnetic relaxation enhancement rates of the backbone amide protons (HN) of the partner protein. These value correspond to the weighted averages of the values for the various states present. One advantage of PRE is that it is nicely sensitive to lowly populated states.

In [49] the values measured obtained from a set of PIPER output have been compared to those obtained when using only the native state. It appears that using all the PIPER states give a better correlation respect to experimental results than when using only the native state.

In this context, our objective is to refine the energy landscape description by filtering some of the PIPER output complexes in order to improve even further the correlation with experimental measures. The method is developed as a module of the SAMSON software package (http://www.samson-connect.net/).

We have proposed a refinement from process of PIPER complexes based on two criterions: a RMSD-based filtering and an energy-based filtering.

The RMSD-filtering first creates a graph of connected component by connecting a pair of complexes if their distance is lower than a given RMSD threshold. Such a process forms clusters. Then, only the complexes that are in the cluster where belongs the native state are conserved. Since only rigid transforms are applied, RMSD are computed thanks to the fast RMSD computation method previously proposed in the team [56].

The energy-based filtering compares the energy of the complexes to the native state energy. The states for which the difference of energy is higher than a given threshold are discarded.

We have evaluated the results obtained when using our filtering scheme, for a distance threshold ranging form 3 to 9 and for an energy threshold ranging from 70 to $240 kJ.mol^{-1}$. Some setting of the filtering are able to improve the correlation (see figure 13), but the gain around 0.3% remains limited (e.g. the correlation rising from 0.770 to 0.773). We are currently working on a more sophisticated state selection process to filter more precisely the PIPER states and hence to further improve the correlation.

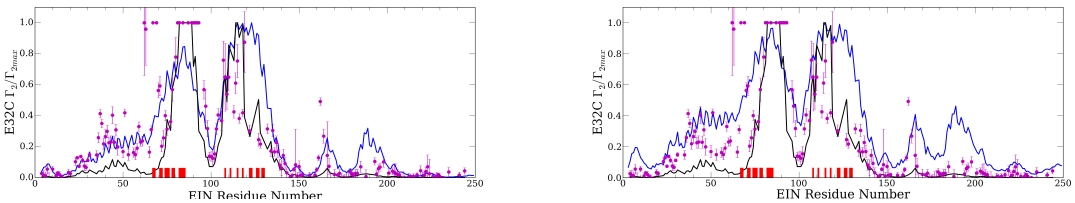


Figure 13. The experimental PRE rates ($\Gamma 2$) are displayed as filled-in magenta circles. Theoretical intermolecular PREs, calculated only from the coordinates of the specific EIN/HPr complex, are shown as black lines. Calculated PRE values from PIPER output are shown as blue lines. The calculated PRE value obtained from the filtered complexes (left) gives a higher correlation with experimental ($c = 0.773$) than the correlation obtained from all the complexes generated with PIPER (right) ($c = 0.770$).

6.10. CREST: Chemical Reactivity Exploration with Stochastic Trees

Participants: Leonard Jaillet, Stephane Redon.

We have proposed the CREST method (Chemical Reactivity Exploration with Stochastic Trees), a new simulation tool to assess the chemical reaction paths of molecular systems. First, it builds stochastic trees based on motion planning principles to search for relevant pathways inside a system's state space. This generates low energy paths transforming a reactant to a given product. Then, a nudged elastic band optimization step locally improves the quality of the initial solutions. The consistency of our approach has been evaluated through tests in various scenarios. It shows that CREST allows to appropriately describe conformational changes as well as covalent bond breaking and formations present in chemical reactions (see figure 14).

This contribution appears in continuity of our previous work regarding the development of a geenric Motion planning architecture for nanosystems. Important features have been added to specifically treat the case of chemical reaction, such as structure alignment, exploration based on multiple trees, automatic resizing of the sampling volume, etc.

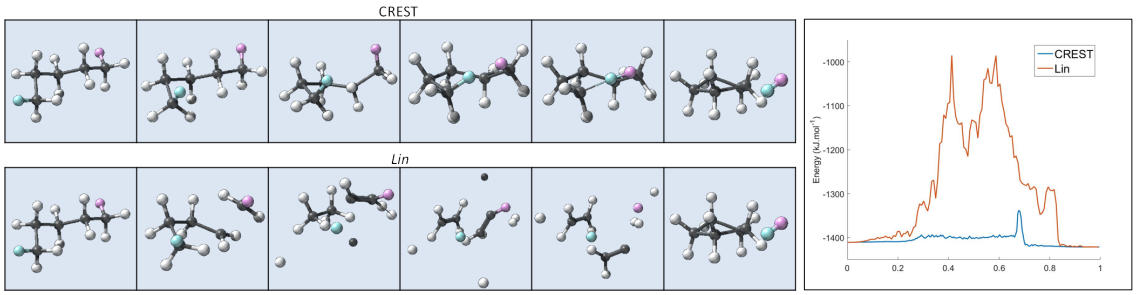


Figure 14. Fictive chemical reaction transforming a pentane into a cyclopentane with a H2 molecule. Hydrogen atoms leading to the H2 molecule are colored. The path obtained with CREST (top) is able to capture the CH3 internal rotations that approaches the two H2 Hydrogens and thus, lead to a low energy barrier. By comparison, a method based on linear interpolation (Lin) gives intermediate broken structures after local path optimization (down). The plot on the right shows the respective energies along the paths. This represents Scenario 3 described in our Results section.

6.11. IM-UFF: extending the Universal Force Field for interactive molecular modeling

Participants: Leonard Jaillet, Svetlana Artemova, Stephane Redon.

We have completed the development of IM-UFF (Interactive Modeling - UFF), an extension of UFF that combines the possibility to significantly modify molecular structures (as with reactive force fields) with a broad diversity of supported systems thanks to the universality of UFF. Such an extension lets the user easily build and edit molecular systems interactively while being guided by physically-based inter-atomic forces. This approach introduces weighted atom types and weighted bonds, used to update topologies and atom parameterizations at every time step of a simulation. IM-UFF has been evaluated on a large set of benchmarks and is proposed as a self-contained implementation integrated in a new module for the SAMSON software platform for computational nanoscience.

This contribution has been submitted to the Journal of Molecular Modeling.

6.12. Incremental methods for long range interactions

Participants: Semeho Edorh, Stephane Redon.

Adaptively Restrained Particles Simulations (**ARPS**) were recently proposed with the purpose of speeding up molecular simulations. The main idea is to modify the Hamiltonian such that the kinetic energy is set to zero for low velocities, which allows to save computational time since particles do not move and forces need not be updated.

We continued our work on developing an extension of **ARPS** to electrostatic simulations.

We have decided to compute the electrostatic contribution by using Multigrid method. This choice have been made because of its $\mathcal{O}(N)$ behavior and its good scalability. In systems containing point charges, Multigrid can't be applied directly because of the discontinuous distribution created by these charges. To overcome this problem, one can replace this distribution by a smooth charge distribution. This charge distribution will be the source term of a Poisson equation which will be solved by Multigrid method. By doing so we retrieve an approximative electrostatic contribution which can be corrected by a near field correction. Concretely each charge will be smeared by a smooth density function.This function is chosen with a compact support. The accuracy of the method is related to the degree of smoothness and the size of the support r_{cut} of the chosen function Fig(15). The bottleneck of this method is often the time spent building the smooth charge distribution. To overcome this issue, We've introduced an interpolation scheme in the near field correction. This leads to a significant reduction of the support required to achieve a specified accuracy. The time spent building the smooth charge distribution is also reduced. Conversely the near correction is slowed down. Nevertheless, the introduction of the interpolation scheme speeds up the method in most of cases Fig(16).

Finally we modified our algorithm to take advantage of ARPS dynamics. This leads to a speed up related to the amount of restrained particles. According to our benchmarks our method can challenge Particle Particle Particle Mesh(**PPPM**), the traditional fast method to compute electrostatics Fig(17). Our algorithm is implemented in LAMMPS.

6.13. Error Analysis of Modified Langevin Dynamics

Participants: Zofia Trstanova, Gabriel Stoltz, Stephane Redon.

We performed a mathematical analysis of modified Langevin dynamics. The aim of this work was first to prove the ergodicity of the modified Langevin dynamics (which fails to be hypoelliptic), and next to analyze how the asymptotic variance on ergodic averages depends on the parameters of the modified kinetic energy. Numerical results illustrated the approach, both for low-dimensional systems where we resorted to a Galerkin approximation of the generator, and for more realistic systems using Monte Carlo simulations.

6.14. Estimating the speed-up of Adaptively Restrained Langevin Dynamics

Participants: Zofia Trstanova, Stephane Redon.

We performed a computational analysis of Adaptively Restrained Langevin dynamics, in which the kinetic energy function vanishes for small velocities. Properly parameterized, this dynamics makes it possible to reduce the computational complexity of updating inter-particle forces, and to accelerate the computation of ergodic averages of molecular simulations. We analyzed the influence of the method parameters on the total achievable speed-up. In particular, we estimated both the algorithmic speed-up, resulting from incremental force updates, and the influence of the change of the dynamics on the asymptotic variance. This allowed us to propose a practical strategy for the parametrization of the method. We validated these theoretical results by representative numerical experiments on the system of a dimer surrounded by a solvent.

6.15. Stable and accurate schemes for Langevin dynamics with general kinetic energies

Participants: Zofia Trstanova, Gabriel Stoltz.

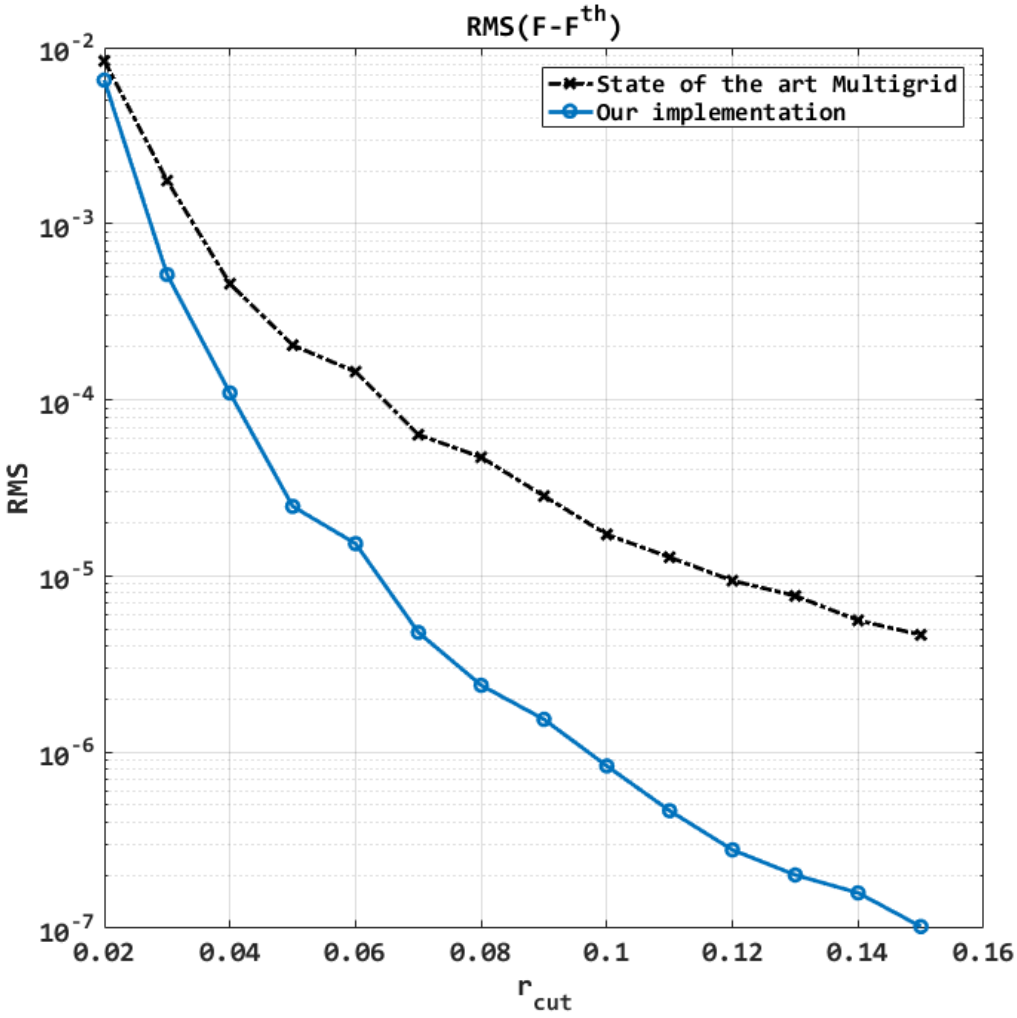


Figure 15. Accuracy in forces for the state of the art multigrid and our implementation : 125000 charged particles randomly distributed in a cubic box. r_{cut} represents the width of the chosen function.

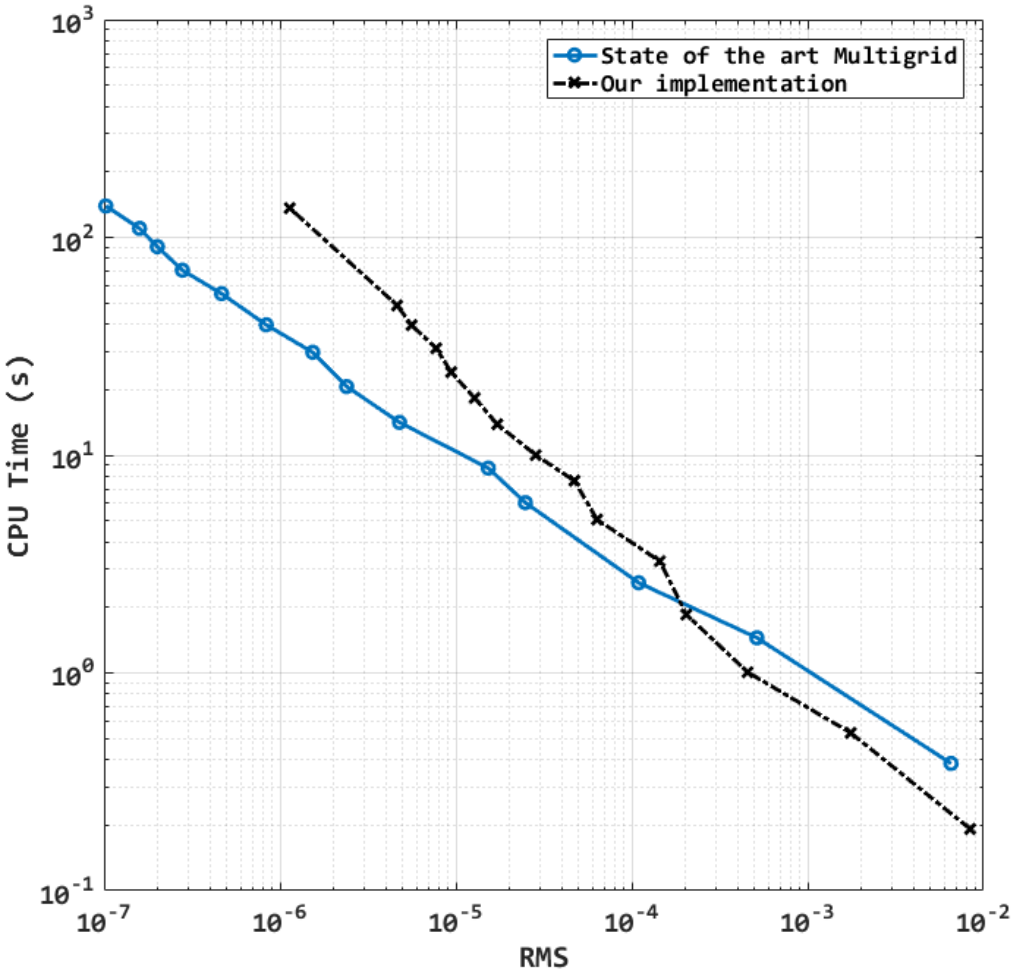


Figure 16. Comparison in terms of CPU time between the state of the art multigrid and our implementation : 125000 charged particles randomly distributed in a cubic box. r_{cut} represents the width of the chosen function.

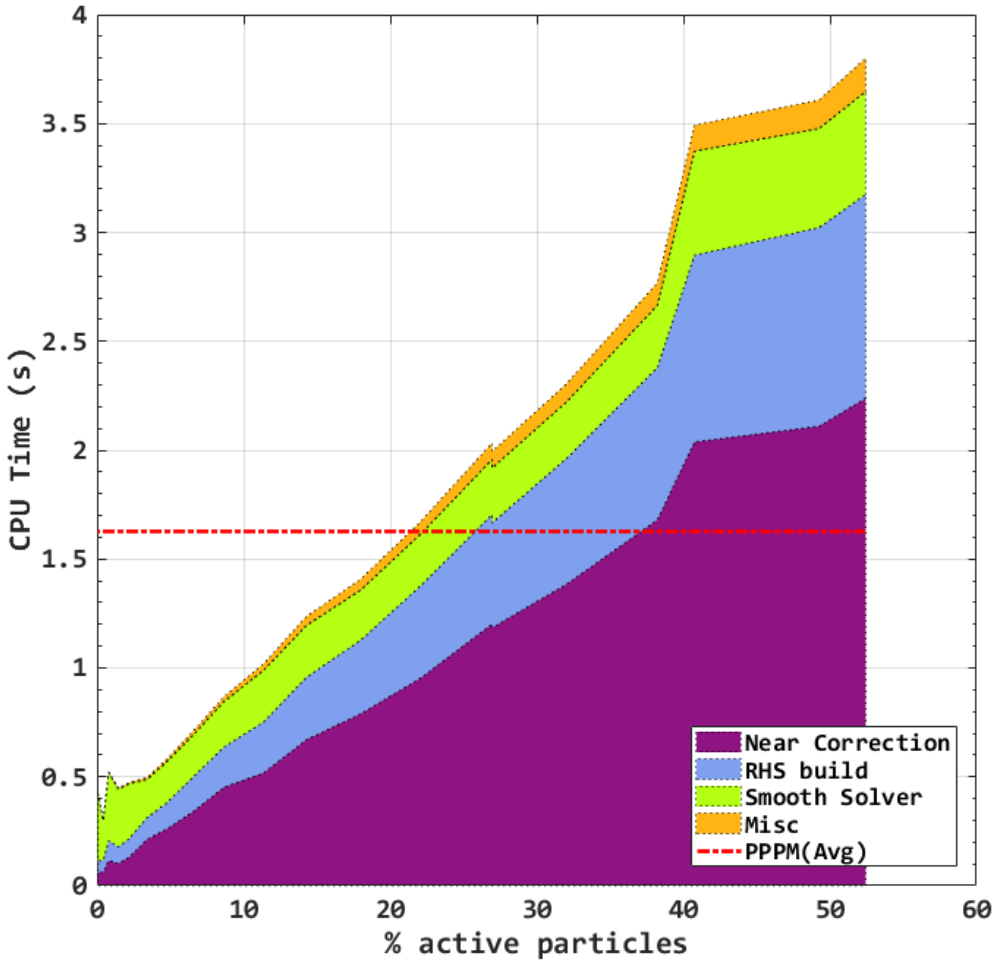


Figure 17. Comparison in terms of CPU time between PPPM and our implementation for a fixed accuracy : 64000 charged particles randomly distributed in a cubic box. Some particles are in restrained dynamics. Colored areas show the associated contribution of each part of our multigrid algorithm. Red dash-dot line represents CPU Time of Particle Particle Particle Mesh needed for this system.

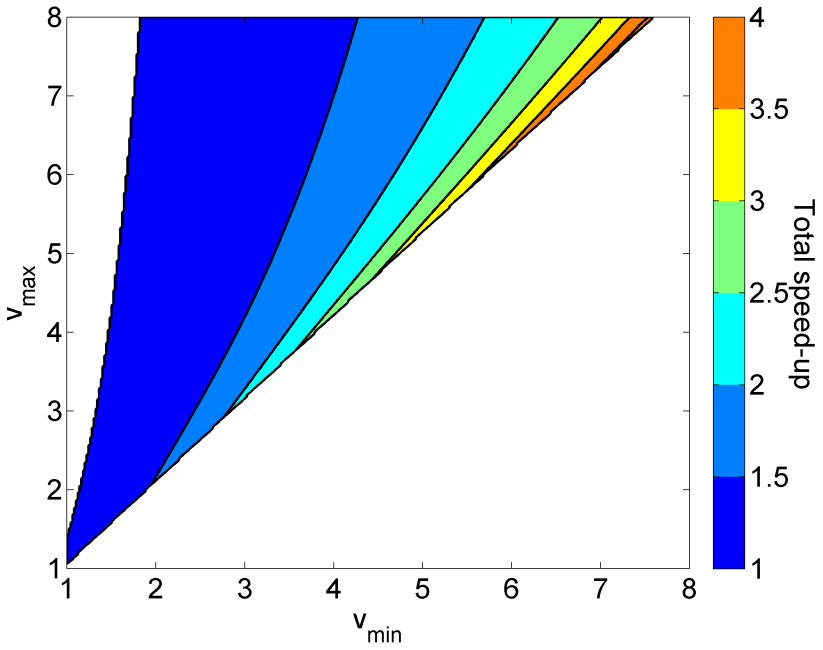


Figure 18. Analytical estimation of the total speed-up of the 3D simulation of the dimer in solvent. Only the solvent particles are restrained by the AR-method. We estimated the expected total speed-up S_{total} for the observable dimer distance A_D with respect to the restraining parameters v_{min} and v_{max} ($v_{\mathrm{max}} \leq 0.95 v_{\mathrm{max}}$). The variance was estimated from three points as a linear function of v_{min} and v_{max} and we used the analytical estimation of the algorithmic speed-up S_{a}. Only $S_{\mathrm{total}} > 1$ is plotted.

We studied integration schemes for Langevin dynamics with a kinetic energy different from the standard, quadratic one in order to accelerate the sampling of the Boltzmann–Gibbs distribution. We considered two cases: kinetic energies which are local perturbations of the standard kinetic energy around the origin, where they vanish (this corresponds to the so-called adaptively restrained Langevin dynamics); and more general non-globally Lipschitz energies. We developed numerical schemes which are stable and of weak order two, by considering splitting strategies where the discretizations of the fluctuation/dissipation are corrected by a Metropolis procedure. We used the newly developed schemes for two applications: optimizing the shape of the kinetic energy for the adaptively restrained Langevin dynamics, and reducing the metastability of some toy models with non-globally Lipschitz kinetic energies.

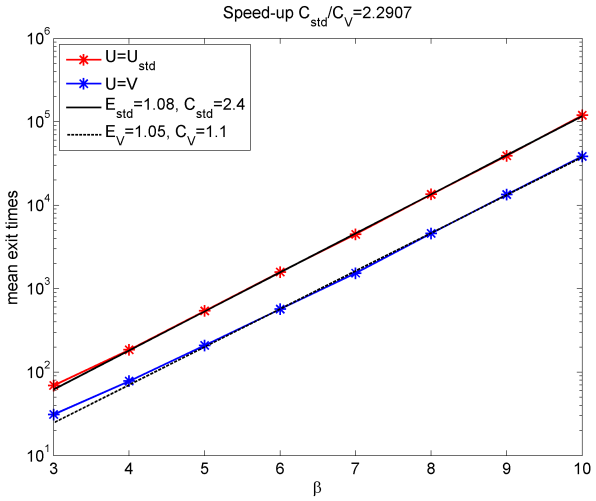


Figure 19. Comparison of the mean exit times for 2D double-well potential with the standard and the modified kinetic energy function (2000 realizations) as a function of the inverse temperature $\beta \in \{3, 4, 5, 6, 7, 8, 9, 10\}$. Thanks to the modified kinetic energy, the transition between two metastable states occurs on average three times faster.

6.16. Quadratic Programming Approach to Fit Protein Complexes into Electron Density Maps

Participants: Alexander Katrutsa, Sergei Grudinin.

We investigated the problem of simultaneous fitting protein complexes into electron density maps of their assemblies. These are represented by high-resolution cryo-EM density maps converted into overlapping matrices and partly show a structure of a complex. The general purpose is to define positions of all proteins inside it. This problem is known to be NP-hard, since it lays in the field of combinatorial optimization over a set of discrete states of the complex. We introduced quadratic programming approaches to the problem. To find an approximate solution, we converted a density map into an overlapping matrix, which is generally indefinite. Since the matrix is indefinite, the optimization problem for the corresponding quadratic form is non-convex. To treat non-convexity of the optimization problem, we use different convex relaxations to find which set of proteins minimizes the quadratic form best.

6.17. Inverse Protein Folding Problem via Quadratic Programming

Participants: Mikhail Karasikov, Sergei Grudinin.

We presented a method of reconstruction a primary structure of a protein that folds into a given geometrical shape. This method predicts the primary structure of a protein and restores its linear sequence of amino acids in the polypeptide chain using the tertiary structure of a molecule. Unknown amino acids are determined according to the principle of energy minimization. This study represents inverse folding problem as a quadratic optimization problem and uses different relaxation techniques to reduce it to the problem of convex optimizations. Computational experiment compares the quality of these approaches on real protein structures.

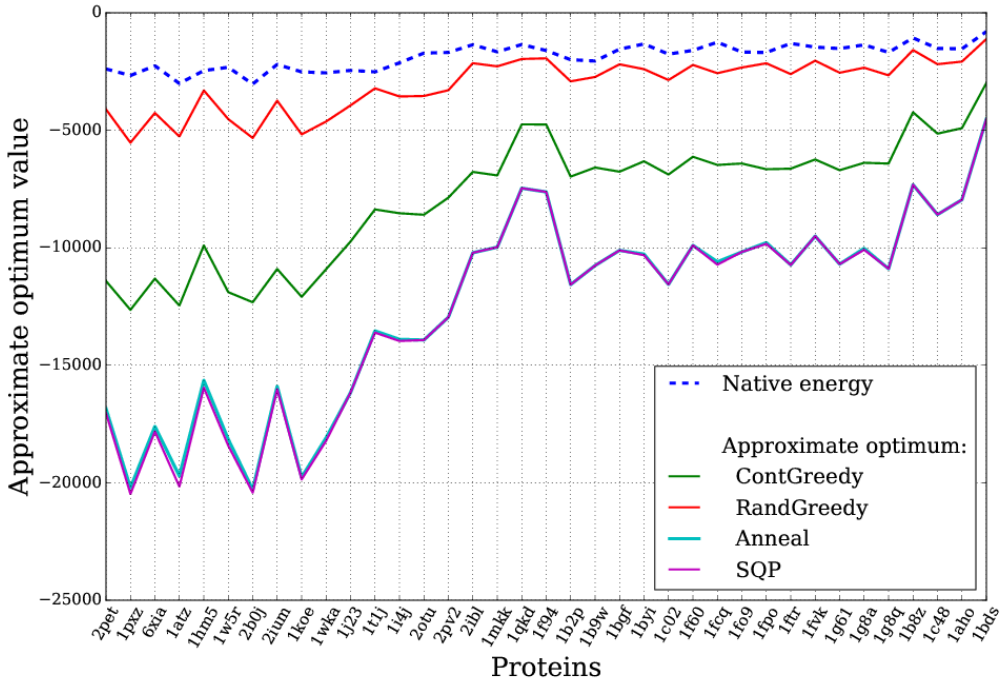


Figure 20. Approximate energy optimum for different relaxations computed on the test set

6.18. Coarse-Grained Protein Scoring Based on Geometrical Features

Participants: Mikhail Karasikov, Sergei Grudinin.

We learnt a scoring function to score protein structures with application to highly important problems in structural biology, namely, protein design, side-chain prediction, and selection of mutations increasing protein stability. For each native structure P_0 a set of ordered decoy structures $\mathcal{D}$ is given:

$$\mathcal{D} = \{P_1, \cdots, P_m\} \subset \mathcal{P},$$

$$(i_1, \cdots, i_m): \; P_{i_m} \preceq \cdots \preceq P_{i_1} \prec P_0.$$

The problem is to train protein scoring function

$$S: \mathcal{P} \to \mathbb{R},$$

such that

$$S(P_0) < S(P_{i_1}) \leq \cdots \leq S(P_{i_m}).$$

We proposed a residue-based scoring function, which uses not the positions of protein's atoms separately, but configurations of the entire residues. The proposed method requires artificially generated decoy structures for the training process and provides high quality scoring functions, which are efficient to compute. Several types of scoring functions are considered according to restrictions imposed by the specific application. For the prediction problems where the whole domain should be searched for the best prediction, we use functions that allow the reduction of emerging optimization problem

$$\sum_{k=1}^{m}\sum_{l=1}^{m} E_{kl}(a_k, a_l) \to \min_{(a_1,\cdots,a_m)\in\mathcal{A}^m} \tag{2}$$

to quadratic binary constrained optimization

$$\begin{array}{ll} \underset{\overrightarrow{x}\in\{0,1\}^n}{\text{minimize}} & \overrightarrow{x}^{\mathsf{T}}\mathbf{Q}\overrightarrow{x} \\ \text{subject to} & \mathbf{A}\overrightarrow{x} = \overrightarrow{1}_m. \end{array} \tag{3}$$

6.19. Development of a Normal Modes Analysis element for SAMSON platform

Participants: Yassine Naimi, Alexandre Hoffmann, Sergei Grudinin, Stephane Redon.

We are currently developing an element for the SAMSON platform for the calculation of normal modes based on the Normal Modes Analysis method. This element will be based on the program developed by Alexandre Hoffmann and Sergei Grudinin on Linux and Mac operating systems. First, we have ported the initial program from Linux and Mac operating systems to Windows and linked the program to the libraries needed for the calculations. These libraries consist in: an optimized version of BLAS (Basic Linear Algebra Subprograms) library called OpenBLAS for basic vector and matrix operations; LAPACK (Linear Algebra PACKage) library for solving systems of simultaneous linear equations, least-squares solutions of linear systems of equations, eigenvalue problems, and singular value problems; ARPACK library for solving large scale eigenvalue problems and ARMADILLO library which is a linear algebra library for the C++ language. We will also compare the performances of our program using these libraries to the Intel MKL (Math Kernel Library) libraries. The ultimate goal is to develop the interface for the SAMSON platform using the SAMSON SDK and Qt software.

6.20. Pairwise distance potential for protein folding

Participants: Maria Kadukova, Guillaume Pages, Alisa Patotskaya, Sergei Grudinin.

We have developed a new knowledge-based pairwise distance-dependent potential using convex optimization. This method uses histogram of distances repartition between each different pair of atom types as feature to feed an SVM-like algorithm. We then obtained a potential for each pair of atom types that can be used to score protein conformations. This method have been extensively used during the CASP12 blind assesement.

6.21. Knowledge-based scoring function for protein-ligand interactions

Participants: Maria Kadukova, Sergei Grudinin.

We have developed a knowledge-based pairwise distance-dependent scoring function based on the similar physical principles, as the protein folding potentials. It was trained on a set of protein-ligand complexes taken from the PDBBindCN database and validated on the CASF 2013 benchmark [50]. The corresponding paper submitted to Journal of Chemical Information and Modeling is currently under revision. We used this scoring function while participating in the 2015-2016 D3R Challenge.

6.22. Updates for the atomic typization software

Participants: Maria Kadukova, Sergei Grudinin.

We have additionally validated Knodle – our atomic typization software – on an extensive set of more than 300,000 small molecules based on the LigandExpo database. Knodle workflow involves machine-learning based "models" for different atoms, this year we retrained several of them on the updated version of PDBBindCN database. These results were published in Journal of Chemical Information and Modeling [45]. We also added functions that add missing hydrogen atoms to the molecules. Knodle was used to classify ligand atoms into different types in our protein-ligand interactions scoring function.

6.23. FFT-accelerated methods for fitting molecular structures into Cryo-EM maps

Participants: Alexandre Hoffmann, Sergei Grudinin.

We have developed a set of new methods for fitting molecular structures into Cryo-EM maps. The problem can be formally written as follows, We are given two proteins $\mathcal{P}_1$ and $\mathcal{P}_2$, and we also have $d_1 : \mathbb{R}^3 \to \mathbb{R}$, the electron density of $\mathcal{P}_1$ and $(Y_k)_{k=0\cdots N_{atoms}-1}$, the starting positions of the atoms of $\mathcal{P}_2$. Assuming we can generate an artificial electron density $d_2 : \mathbb{R}^3 \to \mathbb{R}$ from $(Y_k)_{k=0\cdots N_{atoms}-1}$, our problem is to find a transformation of the atoms $T : \mathbb{R}^3 \to \mathbb{R}^3$ that minimize the L^2 distance between d_1 and d_2.

In image processing this problem is usually solved using the optimal transport theory, but this method assumes that both of the densities have the same L^2 norm which is not necessarily the case for the fitting problem. To solve this problem, one instead starts by splitting T into a rigid transformation T_{rigid} (which is a combination of translation and rotation) and a flexible transformation $T_{flexible}$. Two classes of methods have been developed to find T_{rigid} :

- the first one uses optimization techniques such as gradient descent, and
- the second one uses Fast Fourier Transform (FFT) to compute the Cross Correlation Function (CCF) of d_1 and d_2.

The NANO-D team has already developed some algorithms based on the FFT to find T_{rigid} and we have been developing an efficient extension of these to find $T_{flexible}$.

6.24. Protein sequence and structure aligner for SAMSON

Participants: Guillaume Pages, Sergei Grudinin.

Aligning sequences and structures of proteins is important to understand both the homologies and differences between them. We developed a SAMSON element for this purpose, that can perform both sequence and structure alignment. The sequence alignment is done thanks to the software MUSCLE [36]. The structural alignment is done by finding the transformation that minimize the RMSD between corresponding backbones atoms in both structures. We used the algorithm presented by Kearsley [47].

6.25. Implementation of an Interactive Ramachandran Plot Element for SAMSON

Participants: Guillaume Pages, Sergei Grudinin.

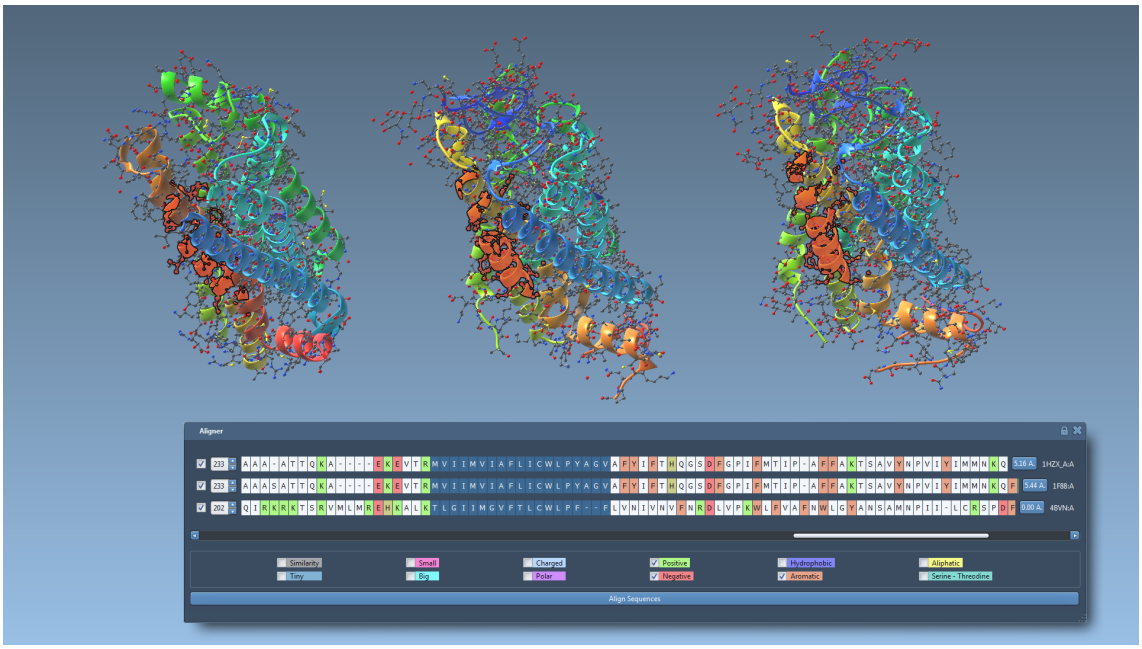


Figure 21. The protein aligner element

Each residue of a protein have two degrees of conformational freedom, described by the two dihedral angles of the backbone ϕ and ψ. Those two angles are crucial to visualize since they determine most of the protein backbone's overall conformation. A very useful way to represent them has been proposed by Ramachandran, Sasisekharan, and Ramakrishnan in 1963 [63].

We have developed a SAMSON element for displaying and editing the Ramachandran Plot of a protein. The favoured regions of the plot have been determined by analysing a database of high quality solved protein structure, provided on Richardson Lab's website (http://kinemage.biochem.duke.edu/databases/top8000.php).

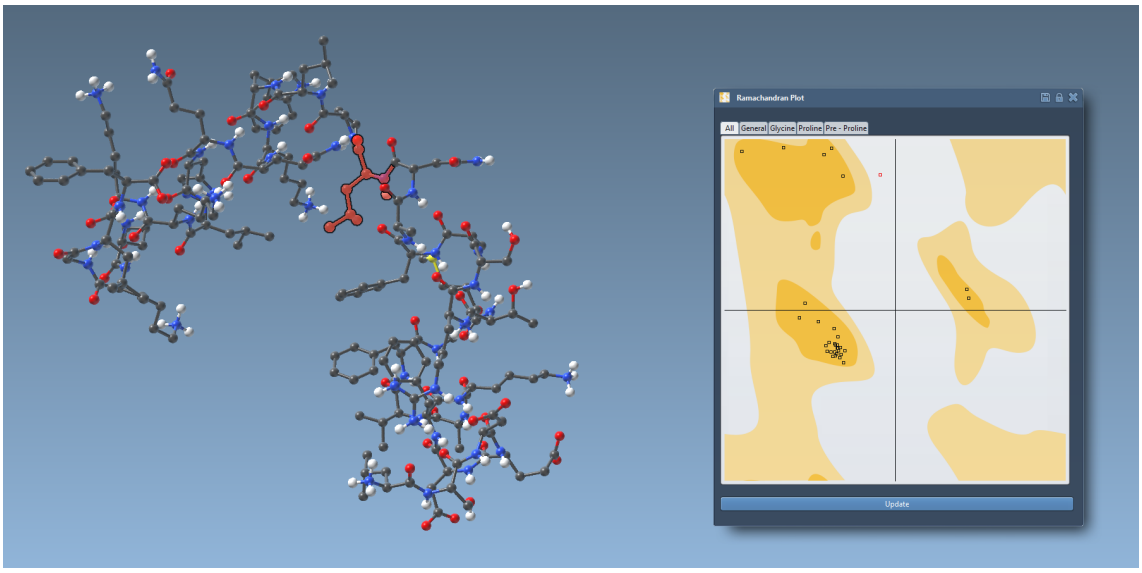


Figure 22. The Ramachandran plot element

NECS Project-Team

7. New Results

7.1. Networked and multi-agent systems: modeling, analysis, and estimation

7.1.1. Modeling of animal groups

Participants: P. Frasca [Contact person], A. Aydogdu [Rutgers University at Camden], C. d'Apice [Univ. Salerno], R. Manzo [Univ. Salerno], W. Saidel [Rutgers University at Camden], B. Piccoli [Rutgers University at Camden].

The paper [13] introduces a mathematical model to study the group dynamics of birds resting on wires. The model is agent-based and postulates attraction-repulsion forces between the interacting birds: the interactions are "topological", in the sense that they involve a given number of neighbors irrespective of their distance. The main properties of the model are investigated by combining rigorous mathematical analysis and simulations. This analysis gives indications about the total length of a group and the inter-animal spacings within it: in particular, the model predicts birds to be more widely spaced near the borders of each group. We compare these insights from the model with new experimental data, derived from the analysis of pictures of pigeons and starlings taken by the team in New Jersey. We have used two different image elaboration protocols to derive the data for the statistical analysis, which allowed us to establish a good agreement with the model and to quantify its main parameters. Our data also seem to indicate potential handedness of the birds: we investigated this issue by analyzing the group organization features and the group dynamics at the arrival of new birds. However, data are still insufficient to draw a definite conclusion on this matter. Finally, arrivals and departures of birds from the group are included in a refined version of the model, by means of suitable stochastic processes.

7.1.2. Cyber-Physical Systems: a control-theoretic approach to privacy and security

Participants: A. Kibangou [Contact person], F. Garin, S. Gracy, H. Nouasse.

Cyber-physical systems are composed of many simple components (agents) with interconnections giving rise to a global complex behaviour. Interesting recent research has been exploring how the graph describing interactions affects control-theoretic properties such as controllability or observability, namely answering the question whether a small group of agents would be able to drive the whole system to a desired state, or to retrieve the state of all agents from the observed local states only. A related problem is observability in the presence of an unknown input, where the input can represent a failure or a malicious attack, aiming at disrupting the normal system functioning while staying undetected. In our work [24], we study linear network systems affected by a single unknown input. The main result is a characterization of input and state observability, namely the conditions under which both the whole network state and the unknown input can be reconstructed from some measured local states. This characterization is in terms of observability of a suitably-defined subsystem, which allows the use of known graphical characterizations of observability of cyber-physical systems, leading to structural results (true for almost all interaction weights) or strong structural results (true for all non-zero interaction weights). Observability is also related to privacy issues. In the ProCyPhyS project, started recently (October 2016), we are studying privacy-preserving properties of cyber-physical systems, by analyzing observability properties of such systems, in order to derive privacy-preserving policies for applications related to smart mobility.

7.1.3. Sensor networks: Multisensor data fusion for attitude estimation

Participants: H. Fourati [Contact person], A. Kibangou, A. Makni, T. Michel, P. Geneves [Tyrex, Inria], N. Layaida [Tyrex, Inria], J. Wu [University of Electronic Science and Technology of China, Chengdu], Z. Zhou [University of Electronic Science and Technology of China, Chengdu], D. Belkhiat [University Ferhat Abbas, Setif, Algeria].

Attitude estimation consists in the determination of rigid body orientation in 3D space (principally in terms of Euler angles, rotation matrix, or quaternion). This research area is a multilevel, multifaceted process involving the automatic association, correlation, estimation, and combination of data and information from several sources. Another interest consists in the fact that redundant and complementary sensor data can be fused and integrated using multisensor fusion techniques to enhance system capability and reliability. Data fusion for attitude estimation is therefore a research area that borrows ideas from diverse fields, such as signal processing, sensor fusion, and estimation theory, where enhancements are involved in point-of-view data authenticity or availability. Data fusion for attitude estimation is motivated by several issues and problems, such as data imperfection, data multimodality, data dimensionality, and processing framework. As a majority of these problems have been identified and heavily investigated, no single data fusion algorithm is capable of addressing all the aforementioned challenges. Consequently, a variety of methods in the literature focuses on a subset of these issues. These concepts and ideas are treated in the book [28], as a response to the great interest and strong activities in the field of multisensor attitude estimation during the last few years, both in theoretical and practical aspects. In the team, we have carried out works related to attitude estimation evaluation for pedestrian navigation purpose. In [18], we focused on two main challenges. The first one concerns the attitude estimation during dynamic cases, in which external acceleration occurs. In order to compensate for such external acceleration, we design a quaternion-based adaptive Kalman filter q-AKF. Precisely, a smart detector is designed to decide whether the body is in static or dynamic case. Then, the covariance matrix of the external acceleration is estimated to tune the filter gain. The second challenge is related to the energy consumption issue of gyroscope. In order to ensure a longer battery life for the Inertial Measurement Units, we study the way to reduce the gyro measurements acquisition by switching on/off the sensor while maintaining an acceptable attitude estimation. The switching policy is based on the designed detector. The efficiency of the proposed scheme is evaluated by means of numerical simulations and experimental tests. In [31], we investigate the precision of attitude estimation algorithms in the particular context of pedestrian navigation with commodity smartphones and their inertial/magnetic sensors. We report on an extensive comparison and experimental analysis of existing algorithms. We focus on typical motions of smartphones when carried by pedestrians. We use a precise ground truth obtained from a motion capture system. We test state-of-the-art attitude estimation techniques with several smartphones, in the presence of magnetic perturbations typically found in buildings. We discuss the obtained results, analyze advantages and limits of current technologies for attitude estimation in this context. Furthermore, we propose a new technique for limiting the impact of magnetic perturbations with any attitude estimation algorithm used in this context. We show how our technique compares and improves over previous works. A novel quaternion-based attitude estimator with magnetic, angular rate, and gravity (MARG) sensor arrays is proposed in [20] within the framework of collaboration with Prof. Zhou from University of Electronic Science and Technology of China, Chengdu. A new structure of a fixed-gain complementary filter is designed fusing related sensors. To avoid using iterative algorithms, the accelerometer-based attitude determination is transformed into a linear system. Stable solution to this system is obtained via control theory. With only one matrix multiplication, the solution can be computed. Using the increment of the solution, we design a complementary filter that fuses gyroscope and accelerometer together. The proposed filter is fast, since it is free of iteration. We name the proposed filter the fast complementary filter (FCF). To decrease significant effects of unknown magnetic distortion imposing on the magnetometer, a stepwise filtering architecture is designed. The magnetic output is fused with the estimated gravity from gyroscope and accelerometer using a second complementary filter when there is no significant magnetic distortion. Several experiments are carried out on real hardware to show the performance and some comparisons. Results show that the proposed FCF can reach the accuracy of Kalman filter. It successfully finds a balance between estimation accuracy and time consumption. Compared with iterative methods, the proposed FCF has much less convergence speed. Besides, it is shown that the magnetic distortion would not affect the estimated Euler angles.

7.2. Control design and networked systems

7.2.1. Control design for hydro-electric power-plants

Participants: C. Canudas de Wit [Contact person], S. Gerwig [Feb 2014–Mar 2016], F. Garin, B. Sari [Alstom].

This is the study of collaborative and resilient control of hydro-electric power-plants, in collaboration with Alstom. The goal is to improve performance of a hydro-electric power-plant outside its design operation conditions, by cancellation of oscillations that occur in such an operation range. Indeed, current operation of power-plants requires to operate on a variety of conditions, often different from the ones initially considered when designing the plant. At off-design operation pressure, the hydraulic turbine exhibits a vortex rope below the runner. This vortex generates pressure fluctuations after the turbine and can excite the hydraulic pipes. Indeed the water is compressible and the pipe walls elastic, so the system can oscillate. The goal is to damp these pressure oscillations as they create vibrations in the system and can lead to damages. Our first contribution [23] has been to model the effect of the vortex rope on the hydraulic system as an external perturbation source acting on pipes. The pipes themselves are described with equations taking into account water compressibility and pipe-wall elasticity. The resulting model is nonlinear with hyperbolic functions in the equations (analogous to high-frequency transmission lines), from which we obtain a suitably linearized model. This model can then be used for control design.

7.2.2. Collaborative source seeking

Participants: F. Garin [Contact person], C. Canudas de Wit, R. Fabbiano.

The problem of source localization consists in finding the point or the spatial region from which a quantity of interest is being emitted. We consider collaborative source seeking, where various moving devices, each equipped with a sensor, share information to coordinate their motion towards the source. We focus on the case where information can only be shared locally (with neighbor agents) and where the the agents have no global position information, and only limited relative information (bearing angle of neighbor agents). This setup is relevant when GPS navigation is not available, as in underwater navigation or in cave exploration, and when relative position of neighbors is vision-based, making it easier to measure angles than distances. In [16] we propose and analize a control law, which is able to bring and keep the agents on a circular equispaced formation, and to steer the circular formation towards the source via a gradient-ascent technique; the circular equispaced formation is beneficial to a good approximation of the gradient from local pointwise measurements. This algorithm is different from the ones present in the literature, because it can cope with our above-described restrictive assumptions on the available position information.

7.2.3. Distributed control and game theory: self-optimizing systems

Participants: F. Garin [Contact person], B. Gaujal [POLARIS], S. Durand.

The design of distributed algorithms for a networked control system composed of multiple interacting agents, in order to drive the global system towards a desired optimal functioning, can benefit from tools and algorithms from game theory. This is the motivation of the Ph.D. thesis of Stéphane Durand, a collaboration between POLARIS and NECS teams. The first results of this thesis concern the complexity of a classical algorithm in game theory, the Best Response Algorithm, an iterative algorithm to find a Nash Equilibrium. For potential games, Best Response Algorithm converges in finite time to a pure Nash Equilibrium. The worse-case convergence time is known to be exponential in the number of players, but surprisingly it turns out that on average (over the possible values of the potentials) the complexity is much smaller, only linearly growing, see [27], [26], [22].

7.3. Transportation networks and vehicular systems

7.3.1. Travel time prediction

Participants: A. Kibangou [Contact person], C. Canudas de Wit, H. Fourati, A. Ladino.

One of the regular performance metrics for qualifying the level of congestion in traffic networks is the travel time. Precision in the estimation or measurement of this variable is one of the most desired features for traffic management. The computation of the travel time is regularly performed based on instantaneous information so called instantaneous travel time (ITT), but regularly traffic changes on time and spaces and the computation depends dynamically on the speeds of the system and the notion of dynamic travel time (DTT) is required. Here the computation requires future information of speed so a short term forecast is required. First in [25] we have presented a framework for instantaneous travel time predictions for multiple origins and destinations in a highway. Secondly in [32], a detailed real time application to compute predictions of dynamic travel time (DTT) is presented. Speed measurements describing a spatio-temporal distribution are captured, from there the DTT is constructed. Definitions, computational details and properties in the construction of DTT are provided. DTT is dynamically clustered using a K-means algorithm and then information on the level and the trend of the centroid of the clusters is used to devise predictors computationally simple to be implemented. To take into account lack of information of cluster assignment of the data to be predicted, a fusion strategy based on the best linear unbiased estimator principle is proposed to combine the predictions of each model. The algorithm is deployed in a real time application and the performance is evaluated using real traffic data from the South Ring of the Grenoble city in France.

7.3.2. Urban traffic control

Participants: C. Canudas de Wit [Contact person], F. Garin, P. Grandinetti.

This work deals with optimal or near-optimal operation of traffic lights in an urban area, e.g., a town or a neighborhood. The goal is on-line optimization of traffic lights schedule in real time, so as to take into account variable traffic demands, with the objective of obtaining a better use of the road infrastructure. More precisely, we aim at maximizing total travel distance within the network, while also ensuring good servicing of demands of incoming cars in the network from other areas. The complexity of optimization over a large area is addressed both in the formulation of the optimization problem, with a suitable choice of the traffic model, and in a distributed solution, which not only parallelizes computations, but also respects the geometry of the town, i.e., it is suitable for an implementation in a smart infrastructure where each intersection can compute its optimal traffic lights by local computations combined with exchanges of information with neighbor intersections.

7.3.3. Optimal control of freeway access

Participants: C. Canudas de Wit [Contact person], D. Pisarski.

The work [19] contains Dominik Pisarski's major results which he obtained during the realization of his Ph.D. thesis at Inria-Rhone Alpes. In concerns the problem of optimal control for balancing traffic density in freeway traffic. The control is realized by ramp metering. The balancing of traffic was proposed as a new objective to improve the vehicular flow on freeways and ring-roads. It was demonstrated that the balancing may result in significantly shortened travel delays and reduced pollution. It may also be beneficial for safety and comfort during a travel. For the controller, a novel modular decentralized structure was proposed where each of the modules computes its optimal decision by using local traffic state and supplementary information arriving from the neighboring controllers. For such a structure, the optimal control problem was formulated as a Nash game, where each player (controller's module) optimizes its local subsystem with respect to decisions of the other players. In comparison to the existing solutions, this new approach significantly reduces the computational burden needed for optimal traffic control, allowing for on-line implementation over long freeway segments. In the paper, the proposed control method was tested via numerical examples with the use of Cell Transmission Model. Later, the performance of the designed method was validated by employing a micro-simulator and real traffic data collected from the south ring of Grenoble. The designed distributed controller resulted in 5% reduction of total time spent on the ring road, 18% reduction of total time spent in the on-ramp queues, 2reduction of the average fuel consumption, and 4% reduction of the traffic density.

AIRSEA Project-Team

7. New Results

7.1. Modeling for Oceanic and Atmospheric flows

7.1.1. Numerical Schemes for Ocean Modelling

Participants: Eric Blayo, Laurent Debreu, Florian Lemarié.

The increase of model resolution naturally leads to the representation of a wider energy spectrum. As a result, in recent years, the understanding of oceanic submesoscale dynamics has significantly improved. However, dissipation in submesoscale models remains dominated by numerical constraints rather than physical ones. Effective resolution is limited by the numerical dissipation range, which is a function of the model numerical filters (assuming that dispersive numerical modes are efficiently removed). In [20], we present a Baroclinic Jet test case set in a zonally reentrant channel that provides a controllable test of a model capacity at resolving submesoscale dynamics. We compare simulations from two models, ROMS and NEMO, at different mesh sizes (from 20 to 2 km). Through a spectral decomposition of kinetic energy and its budget terms, we identify the characteristics of numerical dissipation and effective resolution. It shows that numerical dissipation appears in different parts of a model, especially in spatial advection-diffusion schemes for momentum equations (KE dissipation) and tracer equations (APE dissipation) and in the time stepping algorithms. Effective resolution, defined by scale-selective dissipation, is inadequate to qualify traditional ocean models with low-order spatial and temporal filters, even at high grid resolution. High- order methods are better suited to the concept and probably unavoidable. Fourth-order filters are suited only for grid resolutions less than a few kilometers and momentum advection schemes of even higher-order may be justified. The upgrade of time stepping algorithms (from filtered Leapfrog), a cumbersome task in a model, appears critical from our results, not just as a matter of model solution quality but also of computational efficiency (extended stability range of predictor-corrector schemes). Effective resolution is also shaken by the need for non scale-selective barotropic mode filters and requires carefully addressing the issue of mode splitting errors. Possibly the most surprising result is that submesoscale energy production is largely affected by spurious diapycnal mixing (APE dissipation). This result justifies renewed efforts in reducing tracer mixing errors and poses again the question of how much vertical diffusion is at work in the real ocean.

7.1.2. Coupling Methods for Oceanic and Atmospheric Models

Participants: Eric Blayo, Mehdi-Pierre Daou, Laurent Debreu, Florian Lemarié, Charles Pelletier, Antoine Rousseau.

7.1.2.1. Coupling heterogeneous models in hydrodynamics

The coupling of models of different kinds is gaining more and more attention, due in particular to a need for more global modeling systems encompassing different disciplines (e.g. multi-physics) and different approaches (e.g. multi-scale, nesting). In order to develop such complex systems, it is generally more pragmatic to assemble different modeling units inside a user friendly modelling software platform rather than to develop new complex global models.

In the context of hydrodynamics, global modeling systems have to couple models of different dimensions (1D, 2D or 3D) and representing different physics (Navier-Stokes, hydrostatic Navier-Stokes, shallow water...). We have been developing coupling approaches for several years, based on so-called Schwarz algorithms. Our recent contributions address the development of absorbing boundary conditions for Navier-Stokes equations [4], and of interface conditions for coupling hydrostatic and nonhydrostatic Navier-Stokes flows [5]. In the context of our partnership with ARTELIA Group (PhD thesis of Medhi Pierre Daou), implementations of Schwarz coupling algorithms have been performed for hydrodynamics industrial codes (Mascaret, Telemac and OpenFoam), using the PALM coupling software. A first series of experiments was realized in an academic test case, and a second one in the much more realistic context of the Rusumo hydropower plant, coupling Telemac-3D (Navier-Stokes equations) with OpenFoam (diphasic solver) - see Figure 1 . M.-P. Daou defended his PhD on September 27, 2016 [1].

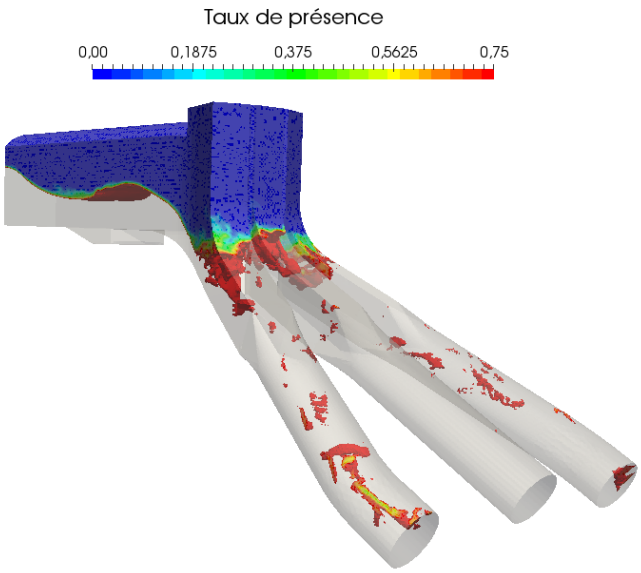


Figure 1. Biphasic simulation of the air/water flow in the Rusumo hydropower plant (PhD of M. P. Daou)

7.1.2.2. Ocean-atmosphere coupling

Coupling methods routinely used in regional and global climate models do not provide the exact solution to the ocean-atmosphere problem, but an approximation of one [63]. For the last few years we have been actively working on the analysis of Schwarz waveform relaxation to apply this type of iterative coupling method to air-sea coupling [65], [66], [64]. In the context of the simulation of tropical cyclone, sensitivity tests to the coupling method have been carried out using ensemble simulations (through perturbations of the coupling frequency and initial conditions). We showed that the use of the Schwarz iterative coupling methods leads to a significantly reduced spread in the ensemble results (in terms of cyclone trajectory and intensity), thus suggesting that a source of error is removed w.r.t coupling methods en vogue in existing coupled models [68].

Motivated by this encouraging result, our activities over the last few years can be divided into four general topics

1. *Stability and consistency analysis of existing coupling methods*: in [63] we showed that the usual methods used in the context of ocean-atmosphere coupling are prone to splitting errors because they correspond to only one iteration of an iterative process without reaching convergence. Moreover, those methods have an additional condition for the coupling to be stable even if unconditionally stable time stepping algorithms are used. This last remark was further studied last year in [3] and it turned out to be a major source of instability in atmosphere-snow coupling.
2. *Study of physics-dynamics coupling*: during the PhD-thesis of Charles Pelletier (funded by Inria) the scope is on including the formulation of physical parameterizations in the theoretical analysis of the coupling, in particular the parameterization schemes to compute air-sea fluxes. To do so, a metamodel representative of the behavior of the full parameterization but with a continuous form easier to manipulate has been derived thanks to a sensitivity analysis based on Sobol' indexes This metamodel has the advantage to be more adequate to conduct the mathematical analysis of the coupling while being physically satisfactory. A publication is currently in preparation for the Quarterly Journal of the Royal Meteorological Society. In parallel we have contributed to a general review gathering the main international specialists on the topic [38].

3. *Design of a coupled single column model*: in order to focus on specific problems of ocean-atmosphere coupling, a work on simplified equation sets has been started. The aim is to implement a one-dimensional (in the vertical direction) coupled model with physical parameterizations representative of those used in realistic models. Thanks to this simplified coupled model the objective is to develop a benchmark suite for coupled models evaluation. Last year the single column oceanic and atmospheric components have been developed in the framework of the SIMBAD project and should be coupled in early 2017 (collaboration with Mercator-océan).
4. *Analysis of air-sea interactions in realistic high-resolution realistic simulations*: part of our activity has been in collaboration with atmosphericists and physical oceanographers to study the impact on some modeling assumptions (e.g. [67]) in large-scale realistic ocean-atmosphere coupled simulations [16], [15].

These four topics are addressed through strong collaborations between the applied mathematics and the climate community.

Moreover a PPR (*Projet à partenariat renforcé*) called SIMBAD (SIMplified Boundary Atmospheric layer moDel for ocean modeling purposes) is funded by Mercator-Ocean for the next three years (from march 2015 to march 2018). The aim of this project in collaboration with Meteo-France, Ifremer, LMD, and LOCEAN is to derive a metamodel to force high-resolution oceanic operational models for which the use of a full atmospheric model is not possible due to a prohibitive computational cost. Another industrial contract named ALBATROS is also funded by (from June 2016 to June 2019) to couple SIMBAD with the NEMO global ocean model and a wave model called WW3.

An ANR project COCOA (COmprehensive Coupling approach for the Ocean and the Atmosphere, P.I.: E. Blayo) has been funded in 2016 and will officially start in January 2017.

7.1.2.3. Data assimilation for coupled models

In the context of operational meteorology and oceanography, forecast skills heavily rely on proper combination of model prediction and available observations via data assimilation techniques. Historically, numerical weather prediction is made separately for the ocean and the atmosphere in an uncoupled way. However, in recent years, fully coupled ocean-atmosphere models are increasingly used in operational centers to improve the reliability of seasonal forecasts and tropical cyclones predictions. For coupled problems, the use of separated data assimilation schemes in each medium is not satisfactory since the result of such assimilation process is generally inconsistent across the interface, thus leading to unacceptable artefacts. Hence, there is a strong need for adapting existing data assimilation techniques to the coupled framework. As part of our ERACLIM2 contribution, R. Pellerej started a PhD on that topic late 2014. So far, three general data assimilation algorithms, based on variational data assimilation techniques, have been developed and applied to a simple coupled problem. The dynamical equations of the considered problem are coupled using an iterative Schwarz domain decomposition method. The aim is to properly take into account the coupling in the assimilation process in order to obtain a coupled solution close to the observations while satisfying the physical conditions across the air-sea interface. Preliminary results shows significant improvement compared to the usual approach on this simple system [28].

The aforementioned system has been recoded within the OOPS framework (Object Oriented Prediction System) in order to ease the transfer to more complex/realistic models.

7.1.3. Parameterizing subgrid scale eddy effects

Participant: Eugene Kazantsev.

Basing on the maximum entropy production principle, the influence of subgrid scales on the flow is presented as the harmonic dissipation accompanied by the backscattering of the dissipated energy. This parametrization is tested on the shallow water model in a square box. Two possible solutions of the closure problem are compared basing on the analysis of the energy dissipation-backscattering balance. Results of this model on the coarse resolution grid are compared with the reference simulation at four times higher resolution. It is shown that the mean flow is correctly recovered, as well as variability properties, such as eddy kinetic energy fields and its spectrum [40].

7.2. Model reduction / multiscale algorithms

7.2.1. Multigrid Methods for Variational Data Assimilation.

Participants: Laurent Debreu, François-Xavier Le Dimet, Arthur Vidard.

In order to lower the computational cost of the variational data assimilation process, we investigate the use of multigrid methods to solve the associated optimal control system. On a linear advection equation, we study the impact of the regularization term on the optimal control and the impact of discretization errors on the efficiency of the coarse grid correction step. We show that even if the optimal control problem leads to the solution of an elliptic system, numerical errors introduced by the discretization can alter the success of the multigrid methods. The view of the multigrid iteration as a preconditioner for a Krylov optimization method leads to a more robust algorithm. A scale dependent weighting of the multigrid preconditioner and the usual background error covariance matrix based preconditioner is proposed and brings significant improvements. This work is summarized in ([7]).

7.2.2. Intrusive sensitivity analysis, reduced models

Participants: Maëlle Nodet, Clémentine Prieur.

Another point developed in the team for sensitivity analysis is model reduction. To be more precise regarding model reduction, the aim is to reduce the number of unknown variables (to be computed by the model), using a well chosen basis. Instead of discretizing the model over a huge grid (with millions of points), the state vector of the model is projected on the subspace spanned by this basis (of a far lesser dimension). The choice of the basis is of course crucial and implies the success or failure of the reduced model. Various model reduction methods offer various choices of basis functions. A well-known method is called “proper orthogonal decomposition" or “principal component analysis". More recent and sophisticated methods also exist and may be studied, depending on the needs raised by the theoretical study. Model reduction is a natural way to overcome difficulties due to huge computational times due to discretizations on fine grids. In [61], the authors present a reduced basis offline/online procedure for viscous Burgers initial boundary value problem, enabling efficient approximate computation of the solutions of this equation for parametrized viscosity and initial and boundary value data. This procedure comes with a fast-evaluated rigorous error bound certifying the approximation procedure. The numerical experiments in the paper show significant computational savings, as well as efficiency of the error bound.

When a metamodel is used (for example reduced basis metamodel, but also kriging, regression, ...) for estimating sensitivity indices by Monte Carlo type estimation, a twofold error appears: a sampling error and a metamodel error. Deriving confidence intervals taking into account these two sources of uncertainties is of great interest. We obtained results particularly well fitted for reduced basis metamodels [62]. In [60], the authors provide asymptotic confidence intervals in the double limit where the sample size goes to infinity and the metamodel converges to the true model. These results were also adapted to problems related to more general models such as Shallow-Water equations, in the context of the control of an open channel [11].

When considering parameter-dependent PDE, it happens that the quantity of interest is not the PDE’s solution but a linear functional of it. In [10], we have proposed a probabilistic error bound for the reduced output of interest (goal-oriented error bound). By probabilistic we mean that this bound may be violated with small probability. The bound is efficiently and explicitly computable, and we show on different examples that this error bound is sharper than existing ones.

A collaboration has been started with Christophe Prieur (Gipsa-Lab) on the very challenging issue of sensitivity of a controlled system to its control parameters [11]. In [32], we propose a generalization of the probabilistic goal-oriented error estimation in [10] to parameter-dependent nonlinear problems. One aims at applying such results in the previous context of sensitivity of a controlled system.

7.3. Dealing with uncertainties

7.3.1. Sensitivity Analysis for Forecasting Ocean Models

Participants: Eric Blayo, Laurent Gilquin, Céline Helbert, François-Xavier Le Dimet, Elise Arnaud, Simon Nanty, Maëlle Nodet, Clémentine Prieur, Laurence Viry, Federico Zertuche.

7.3.1.1. Scientific context

Forecasting geophysical systems require complex models, which sometimes need to be coupled, and which make use of data assimilation. The objective of this project is, for a given output of such a system, to identify the most influential parameters, and to evaluate the effect of uncertainty in input parameters on model output. Existing stochastic tools are not well suited for high dimension problems (in particular time-dependent problems), while deterministic tools are fully applicable but only provide limited information. So the challenge is to gather expertise on one hand on numerical approximation and control of Partial Differential Equations, and on the other hand on stochastic methods for sensitivity analysis, in order to develop and design innovative stochastic solutions to study high dimension models and to propose new hybrid approaches combining the stochastic and deterministic methods.

7.3.1.2. Data assimilation and second order sensitivity analysis

Sensitivity Analysis is defined by some scalar response function giving an evaluation of the state of a system with respect to parameters. By definition, sensitivity is the gradient of this response function. In the case of Variational Data Assimilation, sensitivity analysis have to be carried out on the optimality system because this is the only system in which all the information is located. An important application is the sensitivity, for instance, of the prediction with respect to observations. It's necessary to derive the optimality system and to introduce a second order adjoint.We have applied it to a simulated pollution transport problem and in the case of an oceanic model [18], [19]. More applications to water pollution using a complex hydrological model are under development.

7.3.2. Estimating variance-based sensitivity indices

Participants: Elise Arnaud, Laurent Gilquin, Clémentine Prieur, Simon Nanty, Céline Helbert, Laurence Viry.

In variance-based sensitivity analysis, a classical tool is the method of Sobol' [74] which allows to compute Sobol' indices using Monte Carlo integration. One of the main drawbacks of this approach is that the estimation of Sobol' indices requires the use of several samples. For example, in a d-dimensional space, the estimation of all the first-order Sobol' indices requires $d + 1$ samples. Some interesting combinatorial results have been introduced to weaken this defect, in particular by Saltelli [72] and more recently by Owen [70] but the quantities they estimate still require $O(d)$ samples.

In a recent work [80] we introduce a new approach to estimate all first-order Sobol' indices by using only two samples based on replicated latin hypercubes and all second-order Sobol' indices by using only two samples based on replicated randomized orthogonal arrays. This method is referred as the replicated method. We establish theoretical properties of such a method for the first-order Sobol' indices and discuss the generalization to higher-order indices. As an illustration, we propose to apply this new approach to a marine ecosystem model of the Ligurian sea (northwestern Mediterranean) in order to study the relative importance of its several parameters. The calibration process of this kind of chemical simulators is well-known to be quite intricate, and a rigorous and robust — i.e. valid without strong regularity assumptions — sensitivity analysis, as the method of Sobol' provides, could be of great help. The computations are performed by using CIGRI, the middleware used on the grid of the Grenoble University High Performance Computing (HPC) center. We are also applying these estimates to calibrate integrated land use transport models. As for these models, some groups of inputs are correlated, Laurent Gilquin extended the approach based on replicated designs for the estimation of grouped Sobol' indices [58].

We can now wonder what are the asymptotic properties of these new estimators, or also of more classical ones. In [60], the authors deal with asymptotic properties of the estimators. In [57], the authors establish also a multivariate central limit theorem and non asymptotic properties.

The use of replicated designs to estimate first-order Sobol' indices has the major advantage of reducing drastically the estimation cost as the number of runs n becomes independent of the input space dimension. The generalization to closed second-order Sobol' indices relies on the replication of randomized orthogonal arrays. However, if the input space is not properly explored, that is if n is too small, the Sobol' indices estimates may not be accurate enough.

To address this challenge, we propose approaches to render the replication method recursive, enabling the required number of evaluations to be controlled. With these approaches, more accurate Sobol' estimates are obtained while recycling previous sets of model evaluations. The estimation procedure is therefore stopped when the convergence of estimates is considered reached. One of these approaches corresponds to a recursive version of the replication method and is based on the iterative construction of stratified designs, latin hypercubes and orthogonal arrays [36]. A second approach combines the use of quasi-Monte Carlo sampling and the construction of a new stopping criterion [9], [39].

Extension of the replication method has also been proposed to face constraints arising in an application on the land use and transport model Tranus, such as the presence of dependency among the model inputs, as well as multivariate outputs [37].

7.3.2.1. Sensitivity analysis with dependent inputs

An important challenge for stochastic sensitivity analysis is to develop methodologies which work for dependent inputs. For the moment, there does not exist conclusive results in that direction. Our aim is to define an analogue of Hoeffding decomposition [59] in the case where input parameters are correlated. Clémentine Prieur supervised Gaëlle Chastaing's PhD thesis on the topic (defended in September 2013) [49]. We obtained first results [50], deriving a general functional ANOVA for dependent inputs, allowing defining new variance based sensitivity indices for correlated inputs. We then adapted various algorithms for the estimation of these new indices. These algorithms make the assumption that among the potential interactions, only few are significant. Two papers have been recently accepted [48], [51]. We also considered (see the paragraph 7.3.2) the estimation of groups Sobol' indices, with a procedure based on replicated designs. These indices provide information at the level of groups, and not at a finer level, but their interpretation is still rigorous.

Céline Helbert and Clémentine Prieur supervised the PhD thesis of Simon Nanty (funded by CEA Cadarache, and defended in October, 2015). The subject of the thesis is the analysis of uncertainties for numerical codes with temporal and spatio-temporal input variables, with application to safety and impact calculation studies. This study implied functional dependent inputs. A first step was the modeling of these inputs [14]. The whole methodology proposed during the PhD is presented in [13].

More recently, the Shapley value, from econometrics, was proposed as an alternative to quantify the importance of random input variables to a function. Owen [71] derived Shapley value importance for independent inputs and showed that it is bracketed between two different Sobol' indices. Song et al. [75] recently advocated the use of Shapley value for the case of dependent inputs. In a very recent work [42], in collaboration with Art Owen (Standford's University), we show that Shapley value removes the conceptual problems of functional ANOVA for dependent inputs. We do this with some simple examples where Shapley value leads to intuitively reasonable nearly closed form values.

7.3.3. Optimal Control of Boundary Conditions

Participants: Eric Blayo, Eugene Kazantsev, Florian Lemarié.

A variational data assimilation technique is applied to the identification of the optimal boundary conditions for a simplified configuration of the NEMO model. A rectangular box model placed in mid-latitudes, and subject to the classical single or double gyre wind forcing, is studied. The model grid can be rotated on a desired angle around the center of the rectangle in order to simulate the boundary approximated by a staircase-like coastlines. The solution of the model on the grid aligned with the box borders was used as a reference solution and as artificial observational data. It is shown in [24], [25] that optimal boundary has a rather complicated geometry which is neither a staircase, nor a straight line. The boundary conditions found in the data assimilation procedure bring the solution toward the reference solution allowing to correct the influence of the rotated grid.

Adjoint models, necessary to variational data assimilation, have been produced by the TAPENADE software, developed by the SCIPORT team. This software is shown to be able to produce the adjoint code that can be used in data assimilation after a memory usage optimization.

7.3.4. Non-Parametric Estimation for Kinetic Diffusions

Participants: Clémentine Prieur, Jose Raphael Leon Ramos.

This research is the subject of a collaboration with Venezuela and is partly funded by an ECOS Nord project.

We are focusing our attention on models derived from the linear Fokker-Planck equation. From a probabilistic viewpoint, these models have received particular attention in recent years, since they are a basic example for hypercoercivity. In fact, even though completely degenerated, these models are hypoelliptic and still verify some properties of coercivity, in a broad sense of the word. Such models often appear in the fields of mechanics, finance and even biology. For such models we believe it appropriate to build statistical non-parametric estimation tools. Initial results have been obtained for the estimation of invariant density, in conditions guaranteeing its existence and unicity [45] and when only partial observational data are available. A paper on the non parametric estimation of the drift has been accepted recently [46] (see Samson et al., 2012, for results for parametric models). As far as the estimation of the diffusion term is concerned, a paper has been accepted [46], in collaboration with J.R. Leon (Caracas, Venezuela) and P. Cattiaux (Toulouse). Recursive estimators have been also proposed by the same authors in [47], also recently accepted. In a recent collaboration with Adeline Samson from the statistics department in the Lab, we considered adaptive estimation, that is we proposed a data-driven procedure for the choice of the bandwidth parameters. A paper has been submitted.

In [6], we focused on damping Hamiltonian systems under the so-called fluctuation-dissipation condition.

Note that Professor Jose R. Leon (Caracas, Venezuela) is now funded by an international Inria Chair, allowing to collaborate further on parameter estimation.

We recently proposed a paper on the use of the Euler scheme for inference purposes, considering reflected diffusions. This paper could be extended to the hypoelliptic framework.

7.3.5. Multivariate Risk Indicators

Participants: Clémentine Prieur, Patricia Tencaliec.

Studying risks in a spatio-temporal context is a very broad field of research and one that lies at the heart of current concerns at a number of levels (hydrological risk, nuclear risk, financial risk etc.). Stochastic tools for risk analysis must be able to provide a means of determining both the intensity and probability of occurrence of damaging events such as e.g. extreme floods, earthquakes or avalanches. It is important to be able to develop effective methodologies to prevent natural hazards, including e.g. the construction of barrages.

Different risk measures have been proposed in the one-dimensional framework . The most classical ones are the return level (equivalent to the Value at Risk in finance), or the mean excess function (equivalent to the Conditional Tail Expectation CTE). However, most of the time there are multiple risk factors, whose dependence structure has to be taken into account when designing suitable risk estimators. Relatively recent regulation (such as Basel II for banks or Solvency II for insurance) has been a strong driver for the development of realistic spatio-temporal dependence models, as well as for the development of multivariate risk measurements that effectively account for these dependencies.

We refer to [52] for a review of recent extensions of the notion of return level to the multivariate framework. In the context of environmental risk, [73] proposed a generalization of the concept of return period in dimension greater than or equal to two. Michele et al. proposed in a recent study [53] to take into account the duration and not only the intensity of an event for designing what they call the dynamic return period. However, few studies address the issues of statistical inference in the multivariate context. In [54], [56], we proposed non parametric estimators of a multivariate extension of the CTE. As might be expected, the properties of these estimators deteriorate when considering extreme risk levels. In collaboration with Elena Di Bernardino (CNAM, Paris), Clémentine Prieur is working on the extrapolation of the above results to extreme risk levels.

Elena Di Bernardino, Véronique Maume-Deschamps (Univ. Lyon 1) and Clémentine Prieur also derived an estimator for bivariate tail [55]. The study of tail behavior is of great importance to assess risk.

With Anne-Catherine Favre (LTHE, Grenoble), Clémentine Prieur supervises the PhD thesis of Patricia Tencaliec. We are working on risk assessment, concerning flood data for the Durance drainage basin (France). The PhD thesis started in October 2013 and will be defended in next February. A first paper on data reconstruction has been accepted [79]. It was a necessary step as the initial series contained many missing data. A second paper is in preparation, considering the modeling of precipitation amount with semi-parametric sparse mixtures.

7.4. Assimilation of Images

Participants: Elise Arnaud, François-Xavier Le Dimet, Maëlle Nodet, Arthur Vidard, Nelson Feyeux.

7.4.1. Direct assimilation of image sequences

At the present time the observation of Earth from space is done by more than thirty satellites. These platforms provide two kinds of observational information:

- Eulerian information as radiance measurements: the radiative properties of the earth and its fluid envelops. These data can be plugged into numerical models by solving some inverse problems.
- Lagrangian information: the movement of fronts and vortices give information on the dynamics of the fluid. Presently this information is scarcely used in meteorology by following small cumulus clouds and using them as Lagrangian tracers, but the selection of these clouds must be done by hand and the altitude of the selected clouds must be known. This is done by using the temperature of the top of the cloud.

MOISE was the leader of the ANR ADDISA project dedicated to the assimilation of images, and is a member of its follow-up GeoFluids (along with EPI FLUMINANCE and CLIME, and LMD, IFREMER and Météo-France) that ended in 2013.

During the ADDISA project we developed Direct Image Sequences Assimilation (DISA) and proposed a new scheme for the regularization of optical flow problems [77], which was recently extended [76]. Thanks to the nonlinear brightness assumption, we proposed an algorithm to estimate the motion between two images, based on the minimization of a nonlinear cost function. We proved its efficiency and robustness on simulated and experimental geophysical flows. As part of the ANR project GeoFluids, we are investigating new ways to define distance between a couple of images. One idea is to compare the gradient of the images rather than the actual value of the pixels. This leads to promising results. Another idea, currently under investigation, consists in comparing main structures within each image. This can be done using, for example, a wavelet representation of images. Both approaches have been compared, in particular their relative merits in dealing with observation errors. This work has been extended to the progressive assimilation of different scales contained in the observations [22]

In recent developments we have also used "Level Sets" methods to describe the evolution of the images. The advantage of this approach is that it permits, thanks to the level sets function, to consider the images as a state variable of the problem. We have derived an Optimality System including the level sets of the images. This approach is being applied to the tracking of oceanic oil spills [41]

7.4.2. Optimal transport for image assimilation

We investigate the use of optimal transport based distances for data assimilation, and in particular for assimilating dense data such as images. The PhD thesis of N. Feyeux studied the impact of using the Wasserstein distance in place of the classical Euclidean distance (pixel to pixel comparison). In a simplified one dimensional framework, we showed that the Wasserstein distance is indeed promising. Figure 2 illustrates the advantage of using Wasserstein over L^2 : imagine that the density ρ_0 represents the observation, ρ_1 the background, and that we wish to find the "best" interpolation of the two. The "middle point" between them in the sense of the L^2 distance does not have the correct characteristics: its amplitude is smaller, and its shape is not correct as well. On the contrary, the W^2 middle point presents a similar structure and is indeed physically a better candidate for the interpolation of ρ_0 and ρ_1. Data assimilation experiments with the Shallow Water model have been realised and confirm the interest of the Wasserstein distance. Results have been presented at ISDA conference [35] and a paper has been submitted [34]. N. Feyeux will defend his PhD thesis on Dec. 8th.

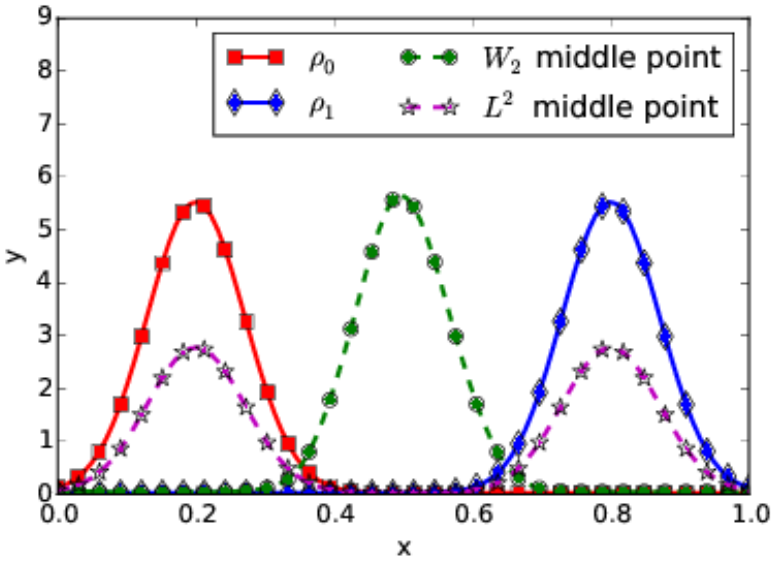


Figure 2. Illustration of the impact of using the Wasserstein distance in Data Assimilation instead of the classical L^2 distance.

7.5. Tracking of Mesoscale Convective Systems

Participant: Clémentine Prieur.

We are interested in the tracking of mesoscale convective systems.A particular region of interest is West Africa. Data and hydrological expertise is provided by T. Vischel and T. Lebel (LTHE, Grenoble).

A first approach involves adapting the multiple hypothesis tracking (MHT) model originally designed by the NCAR (National Centre for Atmospheric Research) for tracking storms [78] to the data for West Africa. With A. Makris (working on a post-doctoral position), we proposed a Bayesian approach [69], which consists in considering that the state at time t is composed on one hand by the events (birth, death, splitting, merging) and on the other hand by the targets' attributes (positions, velocities, sizes, ...). The model decomposes the state into two sub-states: the events and the targets positions/attributes. The events are updated first and are conditioned to the previous targets sub-state. Then given the new events the target substate is updated. A simulation study allowed to verify that this approach improves the frequentist approach by Storlie et al. (2009). It has been tested on simulations [69] and investigated in the specific context of real data on West Africa [12]. Using PHD (probability hypothesis density) filters adapted to our problem, generalizing recent developments in particle filtering for spatio-temporal branching processes (e.g. [44]) could be an interesting alternative to explore. The idea of a dynamic, stochastic tracking model should then provide the base for generating rainfall scenarios over a relatively vast area of West Africa in order to identify the main sources of variability in the monsoon phenomenon.

7.6. Land Use and Transport Models Calibration

Participants: Thomas Capelle, Laurent Gilquin, Clémentine Prieur, Arthur Vidard, Peter Sturm, Elise Arnaud.

Given the complexity of modern urban areas, designing sustainable policies calls for more than sheer expert knowledge. This is especially true of transport or land use policies, because of the strong interplay between the land use and the transportation systems. Land use and transport integrated (LUTI) modelling offers invaluable analysis tools for planners working on transportation and urban projects. Yet, very few local authorities in charge of planning make use of these strategic models. The explanation lies first in the difficulty to calibrate these models, second in the lack of confidence in their results, which itself stems from the absence of any well-defined validation procedure. Our expertise in such matters will probably be valuable for improving the reliability of these models. To that purpose we participated to the building up of the ANR project CITiES led by the STEEP EPI. This project started early 2013 and two PhD about sensitivity analysis and calibration were launched late 2013. Laurent Gilquin defended his PhD in October 2016 [2] and Thomas Capelle will defend his in February 2017.

On top of the development on calibration procedure and sensitivity analysis for LUTI models, a study was conducted to understand in what extend modelling is or could be more integrated into urban planning [17].

BEAGLE Project-Team

6. New Results

6.1. Open-Ended Novelty: Requirements, Guidelines, and Challenges

Participants: G. Beslon

We started in 2014 a collective reflexion on the concept of "Open-Endedness". This reflexion led to a collective paper published this year in "Theory in Biosciences" [12]. The open-endedness of a system is often defined as a continual production of novelty. In this paper we pin down this concept more fully by defining several types of novelty that a system may exhibit, classified as variation, innovation, and emergence. We then provide a meta-model for including levels of structure in a system's model. From there, we define an architecture suitable for building simulations of open-ended novelty-generating systems and discuss how previously proposed systems fit into this framework. We discuss the design principles applicable to those systems and close with some challenges for the community.

6.2. Endocannabinoid dynamics gate spike timing dependent depression and potentiation

Participants: I. Prokin and H. Berry, in collaboration with L. Venance lab, CIRB, Collège de France, Paris.

Learning and memory depend on processes that alter the connections – or synapses – between neurons in the brain. For example, molecules called endocannabinoids can alter synapses to decrease the influence that one neuron has on another neuron's activity. This "synaptic depression" is an important mechanism through which the brain can adapt to an experience. However, recent research also suggests that endocannabinoids might also increase the influence one neuron has on another neuron's activity by strengthening the synaptic connection between neurons. This opposite process is known as synaptic potentiation, and is also important for learning from experience. But how do endocannabinoids manage to produce opposing effects? Using a combination of electrophysiological recording experiments from our experimental collaborator lab and mathematical modeling, we have deciphered the molecular mechanisms that govern the action of endocannabinoids at key synapses in rat brain slices. This revealed that both the levels and timing of endocannabinoid release control changes in the strength of the synaptic connections. Electrical stimulations that produced moderate amounts of endocannabinoids over a prolonged period led to synaptic depression. However, stimulation that produced short but large endocannabinoid peaks caused synaptic potentiation. The enzymes that control endocannabinoid levels thus play a crucial role in determining whether a given stimulation leads to the strengthening or weakening of a synaptic connection. In the type of synapses studied, changes to synaptic strength also depend on another chemical called dopamine. Abnormal dopamine production is implicated in a number of disorders, including Parkinson's disease and addiction. These results have been published in eLife [16].

6.3. Quantitative convergence towards a self similar profile in an age-structured renewal equation for subdiffusion

Participants: A. Mateos Gonzalez and H. Berry, in collaboration with T. Lepoutre, EPI Dracula, Inria.

Continuous-time random walks are generalisations of random walks frequently used to account for the consistent observations that many molecules in living cells undergo anomalous diffusion, i.e. subdiffusion. We described the subdiffusive continuous-time random walk using age-structured partial differential equations with age renewal upon each walker jump, where the age of a walker is the time elapsed since its last jump. In the spatially-homogeneous (zero-dimensional) case, we followed the evolution in time of the age distribution. An approach inspired by relative entropy techniques allows us to obtain quantitative explicit rates for the convergence of the age distribution to a self-similar profile, which corresponds to convergence to a stationnary profile for the rescaled variables. An important difficulty arises from the fact that the equation in self-similar variables is not autonomous and we do not have a specific analyitcal solution. Therefore, in order to quantify the latter convergence, we estimate attraction to a time-dependent "pseudo-equilibrium", which in turn converges to the stationnary profile. These results have been published in Acta Applicandae Mathematicae [13].

6.4. Modulation of Synaptic Plasticity by Glutamatergic Gliotransmission

Participants: M. De Pittà in collaboration with N. Brunel, Dept of Neuroscience and Statistics, University of Chicago, USA.

Glutamatergic gliotransmission, that is the release of glutamate from perisynaptic astrocyte processes in an activity-dependent manner, has emerged as a potentially crucial signal-ing pathway for regulation of synaptic plasticity, yet its modes of expression and function in vivo remain unclear. We focused on two experimentally well-identified gliotransmitter patwhays: (i) modulations of synaptic release and (ii) postynaptic slow inward currents mediated by glutamate released from astrocytes, and investigate their possible functional relevance on synaptic plasticity in a biophysical model of an astrocyte-regulated synapse. Our model predicts that both pathways could profoundly affect both short-and long-term plasticity. In particular, activity-dependent glutamate release from astrocytes, could dramatically change spike-timing–dependent plasticity, turning potentiation into depression (and vice versa) for the same protocol. These results have been published in Neural plasticity [17] and in a review targetting a biologist audience in the journal Neuroscience [18].

6.5. Comparative Genomics and artificial life

Participants: P Biller, C Knibbe, G Beslon, E Tannier

Molecular evolutionary methods and tools are difficult to validate as we have almost no direct access to ancient molecules. Inference methods may be tested with simulated data, producing full scenarios they can be compared with. But often simulations design is concomitant with the design of a particular method, developed by a same team, based on the same assumptions, when both should be blind to each other. In silico experimental evolution consists in evolving digital organisms with the aim of testing or discovering complex evolutionary processes. Models were not designed with a particular inference method in mind, only with basic biological principles. As such they provide a unique opportunity to blind test the behavior of inference methods. We give a proof of this concept on a comparative genomics problem: inferring the number of inversions separating two genomes. We use Aevol, an in silico experimental evolution platform, to produce benchmarks, and show that most combinatorial or statistical estimators of the number of inversions fail on this dataset while they were behaving perfectly on ad-hoc simulations. We argue that biological data is probably closer to the difficult situation.

This work has been published in the article [23] and presented at the Jobim conference [25] and provided the inspiration for a new estimator of the evolutionary distance between two genomes (see below).

6.6. Breaking good

Participants: P Biller, C Knibbe, E Tannier, in collaboration with L Guéguen, University of Lyon 1.

Models of evolution by genome rearrangements are prone to two types of flaws: one is to ignore the diversity of susceptibility to breakage across genomic regions, the other is to suppose that susceptibility values are given. Without necessarily supposing their precise localization, we call "solid" the regions that are improbably broken by rearrangements and "fragile" the regions outside solid ones. We propose a model of evolution by inversions where breakage probabilities vary across fragile regions and over time. It contains as a particular case the uniform breakage model on the nucleotidic sequence, where breakage probabilities are proportional to fragile region lengths. This is very different from the frequently used pseudo-uniform model where all fragile regions have the same probability to break. Estimations of rearrangement distances based on the pseudo-uniform model completely fail on simulations with the truly uniform model. On pairs of amniote genomes, we show that identifying coding genes with solid regions yields incoherent distance estimations, especially with the pseudo-uniform model, and to a lesser extent with the truly uniform model. This incoherence is solved when we co-estimate the number of fragile regions with the rearrangement distance. The estimated number of fragile regions is surprisingly small, suggesting that a minority of regions are recurrently used by rearrangements. Estimations for several pairs of genomes at different divergence times are in agreement with a slowly evolvable co-localization of active genomic regions in the cell.

This work has been published in an article for a reference biology journal [14].

6.7. Subspace clustering

Participants: S Peignier, C Rigotti

We developed an algorithm to tackle the subspace clustering problem over a data stream containing clusters than change over time. Very few subspace clustering algorithms can handle such streams. Our starting point was the work made in the team on evolution of evolution mechanisms and on a preliminary bio-inspired algorithm that we have proposed last year. This previous algorithm included many bio-like features like variable genome length and organization, functional and non-functional elements, and variation operators including chromosomal rearrangements. It achieved satisfying results on standard benchmark data sets but was not designed to process dynamic streams. The new algorithm finds and adapts changing clusters over such streams, while preserving high cluster quality. It has been successfully used to build the evolving music generation system EvoMove.

DRACULA Project-Team

6. New Results

6.1. Mathematical modeling of memory CD8 T cell ontogeny and quantitative predictions

Primary immune responses generate both short-term effector and long-term protective memory cells from naive CD8 T cells. The delineation of the genealogy linking those cell types has been complicated by the lack of molecular markers allowing to discriminate effector from memory cells at the peak of the response. Coupling transcriptomics and phenotypic analyses, and in collaboration with immunologists from Lyon (Jacqueline Marvel's team, Centre International de Recherche en Infectiologie), we identified a novel marker combination that allows to track nascent memory cells within the effector phase [13]. We then used mathematical models based upon our previous description of the dynamics of T cell immune response ([35], [45]) to investigate potential differentiation pathways. We thereby could describe the dynamics of population-size evolutions to test potential progeny links and we could demonstrate that most cells follow a linear naive-early effector-late effector-memory pathway. Of interest for vaccine design, our mathematical model also allows long-term prediction of memory cell numbers from early experimental measurements. Alltogether, our work thus provides a phenotypic means to identify effector and memory cells, as well as a mathematical framework to investigate the ontology of their generation and to predict the outcome of immunization regimens (vaccines) in terms of memory cell numbers generated.

6.2. Multiscale model of the CD8 T cell immune response

We presented in [43] the first multiscale model of CD8 T cell activation in a lymph node. We now described in [14] an update of this modeling approach. CD8 T cell dynamics are described using a cellular Potts model (hence cells are discrete interacting objects), whereas intracellular regulation is associated with a continuous system of nonlinear ordinary differential equations. We focused our study on describing the role of Interleukin 2 (IL2) secretion. One major result was the demonstration of the full relevance of a bona fide multiscale description of the process: the observed (all or none) emergent behavior at the cell population scale could not have been straightforwardly deduced from the simple examination of (seemingly tenuous) differences in the cellular or molecular levels in separation.

6.3. Moving the Boundaries of Granulopoiesis Modelling

The human blood cell production system usually remains extremely robust, in terms of cell number or function, with little signs of decline in old age. To achieve robustness, circulating blood cells rely on a formidable production machinery, the hematopoietic system, located in the bone marrow. All circulating blood cells-red blood cells, white blood cells and platelets-are renewed on a daily basis. The hematopoietic system produces an estimated $1e12$ cells per day. This is a significant fraction of the $3.7e13$ cells in an adult. Robustness is partly due to the short time scales at which cell populations are able to return to equilibrium, combined with large cell numbers and renewal rates. White blood cells (WBCs), among which neutrophils are most prevalent, are the body's first line, innate immune system. Upon infection, WBCs are mobilized from the bone marrow, to increase their number in circulation and fight off pathogen within hours. The 26 billion circulating neutrophils in human have a mean residence time of only 11h in the blood. After their release from the bone marrow, they quickly disappear in the peripheral tissues and are destroyed in the spleen, liver and bone marrow. In addition to the high renewal rate of circulating blood cells, a large number of mature neutrophils, ten times or more the circulating number, is kept in a bone marrow reserve, ready for entering circulation. This high renewal rate and mobilization capability, however, come at a cost. The blood system is an easy target for chemotherapeutic drugs, whose main way of acting is by killing proliferating

cells. White blood cells and end especially neutrophils, with their fast turnover, are particularly vulnerable to chemotherapy. Chemotherapy can induce neutropenia-a state of low absolute neutrophil count (ANC)-in cancer patients, which puts them at risk of infection. Homeostatic regulation of white blood cells is mainly controlled by the cytokine Granulocye-Colony Stimulating Factor (G-CSF). G-CSF promotes survival of white blood cell precursors and their differentiation into mature cells. The identification of this protein in the 1980's, and the subsequent development of human recombinant forms of G-CSF paved the way to the treatment of chemotherapy-induced neutropenia. G-CSF therapy as also been successful at treating congenital and other forms of neutropenia. Today, G-CSF is used as an adjuvant in several anti-cancer treatment protocols. The aim of the adjuvant therapy is to minimize the length of the neutropenic episodes. However, exogenous G-CSF administration interferes with white blood cell production regulation. What should be a straightforward effect–administer G-CSF to cause the ANC to increase–turns to be more complicated than that. For instance, it was observed that early timing of G-CSF administration could lead to prolonged neutropenic phase. Thus, in order to take advantage of the full potential of G-CSF, a detailed understanding of the physiological interaction between neutrophils and exogenous G-CSF is necessary. In this issue of the Bulletin (see [7]), Craig and colleagues present a physiological model of neutrophil production that includes a detailed modelling of the kinetics of G-CSF.

6.4. Bone marrow infiltration by multiple myeloma causes anemia by reversible disruption of erythropoiesis

Multiple myeloma (MM) infiltrates bone marrow and causes anemia by disrupting erythropoiesis, but the effects of marrow infiltration on anemia are difficult to quantify. Marrow biopsies of newly diagnosed MM patients were analyzed before and after four 28-day cycles of nonerythrotoxic remission induction chemotherapy. Complete blood cell counts and serum paraprotein concentrations were measured at diagnosis and before each chemotherapy cycle. At diagnosis, marrow area infiltrated by myeloma correlated negatively with hemoglobin, erythrocytes, and marrow erythroid cells. After successful chemotherapy, patients with less than 30% myeloma infiltration at diagnosis had no change in these parameters, whereas patients with more than 30% myeloma infiltration at diagnosis increased all three parameters. Clinical data were used to develop mathematical models of the effects of myeloma infiltration on the marrow niches of terminal erythropoiesis, the erythroblastic islands (EBIs) (see [12]). A hybrid discrete-continuous model of erythropoiesis based on EBI structure/function was extended to sections of marrow containing multiple EBIs. In the model, myeloma cells can kill erythroid cells by physically destroying EBIs and by producing proapoptotic cytokines. Following chemotherapy, changes in serum paraproteins as measures of myeloma cells and changes in erythrocyte numbers as measures of marrow erythroid cells allowed modeling of myeloma cell death and erythroid cell recovery, respectively. Simulations of marrow infiltration by myeloma and treatment with nonerythrotoxic chemotherapy demonstrate that myeloma-mediated destruction and subsequent reestablishment of EBIs and expansion of erythroid cell populations in EBIs following chemotherapy provide explanations for anemia development and its therapy-mediated recovery in MM patients.

6.5. Mathematical modelling of hematopoiesis dynamics with growth factor-dependent coefficients

In [4] and [5], we propose and analyze an age-structured partial differential model for hematopoietic stem cell dynamics, in which proliferation, differentiation and apoptosis are regulated by growth factor concentrations. By integrating the age-structured system over the age and using the characteristics method, we reduce it to a delay differential system (with discret delay [4] and distribute delay [5]). We investigate the existence and stability of the steady states of the reduced delay differential system. By constructing a Lyapunov function, the trivial steady state, describing cell's dying out, is proven to be globally asymptotically stable when it is the only equilibrium of the system. The asymptotic stability of the positive steady state, the most biologically meaningful one, is analyzed using the characteristic equation. This study may be helpful in understanding the uncontrolled proliferation of blood cells in some hematological disorders. This study may be helpful in understanding the behavior of hematopoietic cells in some hematological disorders.

6.6. Mathematical modelling of Chronic Myleoid Leukemia (CML)

Firstly, an analysis of a reduced version of our model has been performed by A. Besse et al. (manuscript in revision). It allows to analyze the structure of the steady states and their stability. Typically, the situation is as follows. There are 4 steady states: 0 (unstable) a low one (stable) an intermediate (unstable) and a high (stable).

Secondly, considering another framework of modelling [37], it was observed by A. Besse et al. (see also the thesis of A. Besse) that, under the assumptions of the models, the long term response might be non monotonous with respect to the dose. In words, when the disease load has been reduced enough, it might be more efficient (it is not a question of toxicity) to reduce the dose. This comes from a balance between quiescence induction and apoptosis effects of the drug.

6.7. Hybrid Modelling in Biology

The paper [19] presents a general review on hybrid modelling which is about to become ubiquitous in biological and medical modelling. Hybrid modelling is classically defined as the coupling of a continuous approach with a discrete one, in order to model a complex phenomenon that cannot be described in a standard homogeneous way mainly due to its inherent multiscale nature. In fact, hybrid modelling can be more than that since any types of coupled formalisms qualify as being hybrid. The paper [19], first presents the evolution and current context of this modelling approach. It then proposes a classification of the models through three different types that relate to the nature and level of coupling of the formalisms used.

6.8. Design and study of a new model describing the effect of radiotherapy on healthy cells

This new project started in January 2016 between a start up Neolys Diagnostics, an Inserm team from Lyon and some members of the Dracula team (Léon Matar Tine and Laurent Pujo-Menjouet) (see [11]). We recruited a student to start a PhD (Aurélien Canet) paid for one half by Neolys and the other half by the labex Milyon. The objective of this collaboration is to use deterministic models (as a first step) to describe the dynamics of ATM proteins in the cytoplasm moving to the nucleus. Once there, they recognize and repair damaged DNA (due to nuclear radiations) and to give solid mechanistic explanations of the phenomenological linear quadratic model used until now by biologists and clinicians. Next step is then to use data provided by the Inserm team to calibrate our model and use it for clinical tests by Neolys (to detect radiosensitive persons (3 different groups) and prevent individual from creating cancer induced by nuclear radiations).

6.9. Contribution to the interaction between Alzheimer's disease and prion with the analysis of a mathematical model arising from in vitro experiments

Alzheimer's disease (AD) is a fatal incurable disease leading to progressive neuron destruction. AD is caused in part by the accumulation of $A\beta$ monomers inside the brain, which have the faculty to aggregate into oligomers and fibrils. Oligomers are the most toxic structures as they can interact with neurons via membrane receptors, including PrPc proteins. This interaction leads to the misconformation of $PrPc$ into pathogenic oligomeric prions, $PrPol$. The objective of our collaboration with the Inra team lead by Human Rezaei (Jouy en Josas), is to design and study a brand new model describing in vitro $A\beta$ polymerization process (see [25]). We include interactions between oligomers and $PrPc$ that induces the misconformation of PrPc into PrPol. The model consists of nine equations, including size structured transport equations, ordinary differential equations and delayed differential equations. Our collaboration is only at its beginning and we applied for an ANR grant highlighting this interdisciplinary work.

6.10. Methods of Blood Flow Modelling

The paper [9] is devoted to recent developments in blood flow modelling. It begins with the discussion of blood rheology and its non-Newtonian properties. After that it presents some modelling methods where blood is considered as a heterogeneous fluid composed of plasma and blood cells. Namely, it describes the method of Dissipative Particle Dynamics and presents some results of blood flow modelling. The last part of this paper deals with one-dimensional global models of blood circulation. It explains the main ideas of this approach and presents some examples of its application.

6.11. Anomalous diffusion as an age-structured renewal process

Continuous-time random walks (CTRW) are one of the main mechanisms that are recurrently evoked to explain the emergence of subdiffusion in cells. CTRW were introduced fifty years ago as a generalisation of random walks, where the residence time (the time between two consecutive jumps) is a random variable. If the expectation of the residence time is defined, for instance when it is dirac-distributed or decays exponentially fast, one recovers "normal" Brownian motion. However, when the residence time expectation diverges, the CTRW describes a subdiffusive behavior. The classical approach to CTRW yields a non-Markovian (mean-field) transport equation, which is a serious obstacle when one wants to couple subdiffusion with (bio)chemical reaction. In [8], we took an alternative approach to CTRW that maintains the Markovian property of the transport equation at the price of a supplementary independent variable. We associate each random walker with an age a, that is the time elapsed since its last jump and describe the subdiffusive CTRW using an age-structured partial differential equations with age renewal upon each walker jump. In the spatially-homogeneous (zero-dimensional) case, we follow the evolution in time of the age distribution. An approach inspired by relative entropy techniques allows us to obtain quantitative explicit rates for the convergence of the age distribution to a self-similar profile, which corresponds to convergence to a stationary profile for the rescaled variables. An important difficulty arises from the fact that the equation in self-similar variables is not autonomous and we do not have a specific analytical solution. Therefore, in order to quantify the latter convergence, we estimate attraction to a time-dependent "pseudo-equilibrium", which in turn converges to the stationary profile.

6.12. Doubly nonlocal reaction-diffusion equations and the emergence of species

The paper [6] is devoted to a reaction-diffusion equation with doubly nonlocal nonlinearity arising in various applications in population dynamics. One of the integral terms corresponds to the nonlocal consumption of resources while another one describes reproduction with different phenotypes. Linear stability analysis of the homogeneous in space stationary solution is carried out. Existence of travelling waves is proved in the case of narrow kernels of the integrals. Periodic travelling waves are observed in numerical simulations. Existence of stationary solutions in the form of pulses is shown, and transition from periodic waves to pulses is studied. In the applications to the speciation theory, the results of this work signify that new species can emerge only if they do not have common offsprings. Thus, it is shown how Darwin's definition of species as groups of morphologically similar individuals is related to Mayr's definition as groups of individuals that can breed only among themselves.

6.13. Existence of very weak global solutions to cross diffusion models

The entropy structure has been used in [26] to derived a very general theorem for existence for cross diffusion models. The theory is based on the interplay between the entropy structure which gives some compactness in space (gradient control) and the duality strucutre identified by Michel Pierre for general parabolic systems, which gives integrability. We derive a very general results under very general structural hypothesis (exsitence of an entropy which is compatible with reaction terms and relevance of the duality strucutre). The key is the construction of implicit solutions of the semi discete version (time is discretized) which happens to verify all the strucutres and are very regular. Moreover, we give a simple condition for mutliple case (more than 3

species) for building examples with an entropy structure based on the detailed balance structure proposed in [34].

ERABLE Project-Team

6. New Results

6.1. General comments

We present in this section the main results obtained in 2016. Some were already in preparation or submitted at the end of 2015. This will be indicated whenever it is the case.

We tried to organise the results following the five main axes of research of the team. Clearly, in some cases, a result obtained overlaps more than one axis. We chose the one that could be seen as the main one concerned by such results.

We did not indicate here the results on more theoretical aspects of computer science if it did not seem for now that they could be relevant in contexts related to computational biology. Actually, we do believe those on rumour spreading (by Pierluigi Crescenzi) [9] or on general network analysis (by Pierluigi Crescenzi or Roberto Grossi) [31], [36], [40], [39], [37], [38], [10], [42] could in the future become relevant for life sciences (biology or ecology). In the other direction, algorithmic ideas that were developed in the context of a problem in life sciences could prove useful for solving more general problems (possibly with other applications). This was the case of some of the ideas explored in previous years to deal with de Bruijn graphs in the context of NGS analysis that led to the team fruitfully collaborating with a group of researchers at the ETH in Switzerland on a problem related to transport systems [34].

Below however, we preferred to only indicate the theoretical results related to problems closely resembling questions that have already been addressed by us in computational biology. Notice that such CS results concern not only cross-fertilising issues among different computational approaches, and we therefore extended the title of this axis for the purpose of presenting such results, for now purely theoretical.

A few other results are not mentioned either in this report, not because the corresponding work is not important, but because it was likewise more specialised, or the work represented a survey.

6.2. Identifying the molecular elements

RNA-seq NGS algorithms and data analysis
SNPs (Single Nucleotide Polymorphisms) are genetic markers whose precise identification is a prerequisite for association studies. Methods to identify them are currently well developed for model species, but rely on the availability of a (good) reference genome, and therefore cannot be applied to non-model species. They are also mostly tailored for whole genome (re-)sequencing experiments, whereas in many cases, transcriptome sequencing can be used as a cheaper alternative which already enables to identify SNPs located in transcribed regions. In a paper accepted this year [18], we proposed the use of a previously developed method, KISSPLICE, that identifies, quantifies and annotates SNPs without any reference genome, using RNA-seq data only. Individuals can be pooled prior to sequencing if not enough material is available from one individual. Using pooled human RNA-seq data, we clarified the precision and recall of our method and discussed them with respect to other methods which use a reference genome or an assembled transcriptome. We then validated experimentally the predictions of our method using RNA-seq data from two non-model species. KISSPLICE can be used for any species to annotate SNPs and predict their impact on the protein sequence. We further enable to test for the association of the identified SNPs with a phenotype of interest.

We participated also in two other works, one computational and the other biological, on alternative splicing in Human.

The first is associated to the ANR Colib'read project in which we were one of the partners. A Colib'read Galaxy tools suite was developed that should enable a broad range of life science researchers to analyse raw NGS data, allows the maximum biological information to be retained in the data, and uses a very low memory footprint [17]. The algorithms implemented in the tools are based on the use of a de Bruijn graph and of a bloom filter. The analyses can be performed in a few hours, using small amounts of memory. Applications using real data further demonstrate the good accuracy of these tools compared to classical approaches.

KISSPLICE was also used in the context of myotonic dystrophy (DM), which is caused by the expression of mutant RNAs containing expanded CUG repeats that sequester muscleblind-like (MBNL) proteins, leading to alternative splicing changes. Cardiac alterations, characterised by conduction delays and arrhythmia, are the second most common cause of death in DM. Using RNA sequencing, the authors of [14] identified novel splicing alterations in DM heart samples, including a switch from adult exon 6B towards fetal exon 6A in the cardiac sodium channel, SCN5A. They found that MBNL1 regulates alternative splicing of SCN5A mRNA and that the splicing variant of SCN5A produced in DM presents a reduced excitability compared to the control adult isoform. Importantly, reproducing splicing alteration of Scn5a in mice is sufficient to promote heart arrhythmia and cardiac-conduction delay, two predominant features of myotonic dystrophy. Misregulation of the alternative splicing of SCN5A may therefore contribute to a subset of the cardiac dysfunctions observed in myotonic dystrophy.

We introduced CIDANE, a novel framework for genome-based transcript reconstruction and quantification from RNA-seq reads [8]. CIDANE assembles transcripts efficiently with significantly higher sensitivity and precision than existing tools. Its algorithmic core not only reconstructs transcripts *ab initio*, but also allows the use of the growing annotation of known splice sites, transcription start and end sites, or full-length transcripts, which are available for most model organisms. CIDANE supports the integrated analysis of RNA-seq and additional gene-boundary data and recovers splice junctions that are invisible to other methods.

Landscape of somatic mutations in breast cancer whole-genome sequences
In the context of the International Cancer Genome Consortium (ICGC), we conducted a whole-genome, exome, RNASeq and methylome characterisation of 560 breast cancers. The results were published this year in three main papers.

The first one describes the general landscape of somatic mutations and rearrangements in all subtypes of breast cancers [21]. This allowed to extend our current repertoire of probable breast cancer drivers to 93 genes. The mutational signature analysis was extended to genome rearrangements as well and revealed six typical rearrangement signatures. Three of them, characterised by tandem duplications or deletions, appear associated with defective homologous- recombination-based DNA repair (BRCA1/2). This analysis highlighted the repertoire of cancer genes and mutational processes operating in human, and represented a progress towards obtaining a comprehensive account of the somatic genetic basis of breast cancer.

This first analysis was then used to link known and novel drivers and mutational signatures to gene expression (transcriptome) of 266 cases [28]. One important and still debated question is to know to what extend somatic aberrations could trigger an immune-response. Our data suggested that substitutions of a particular type could be more effective in doing so than others.

Finally, in the context of ICGC, France was in charge of the analysis of a clinically specific subgroup of breast cancers, called HER2-positive, characterised by the HER2/ERBB2 amplification and over-expression. This is a subgroup for which several efficient targeted therapies (trastuzumab) are now available. However, resistance to treatment has been observed, revealing the underlying diversity of these cancers. An in-depth genomic and transcriptomic characterisation of 64 HER2-positive breast tumour was carried out. We delineated four subgroups, based on the expression data, each of them with distinctive genomic features in terms of somatic mutations, copy-number changes or structural variations [12]. The results suggested that, despite being clinically delineated by a specific gene amplification, HER2-positive tumours actually melt into the luminal-basal breast cancer spectrum rather, probably following their "cell-of-origin" fate and suggesting that the ERBB2 amplification is an embedded event in the natural history of these tumours. Finally, WGS data allowed us to gain more information about the amplification process itself and brought some indications about

how (and maybe when) it arose. Whole genome paired-end sequencing provides two important experimental clues to this purpose: a) high dynamics and resolution analysis of copy numbers, and b) ability to pinpoint large scale structural rearrangements by using clipping and abnormal mapping of read pairs. We could show that, in several cases, the observed sequence of copy numbers as well as the orientation of clipped reads was consistent with a breakage-fusion-bridge folding mechanism (BFB). However, the observation of long distance and inter-chromosomal rearrangements further showed that the amplification is a complex event (or sequence of events), likely involving several amplicons on the same or different chromosomes and several intertwined mechanisms. Indeed one of the features of HER2+ tumours is the ubiquitous presence of firestorms, corresponding to multiple closely spaced amplicons on highly rearranged chromosomal arms. It is therefore tempting to combine two mechanisms to explain the complex amplification patterns observed: chromothripsis, which will generate a mosaic of fragments (but no amplification per se), followed by a BFB amplification of chromosomal arm(s). This work was done at the "Plateforme Bioinformatique Gilles Thomas" located at Centre Léon Bérard (Lyon).

Sequence comparison
Sequence comparison is a fundamental step in many important computational biology tasks, in particular the reconstruction of genomes, a first key step before being able to identify the molecular elements present in them.

Traditional algorithms for measuring approximation in sequence comparison are based on the notions of distance or similarity, and are generally computed through sequence alignment techniques. As circular molecular structures are a common phenomenon in nature, a caveat of the adaptation of alignment techniques for circular sequence comparison is that they are computationally expensive, requiring from super-quadratic to cubic time in the length of the sequences. We introduced a new distance measure based on q-grams, and showed how it can be applied effectively and computed efficiently for circular sequence comparison [15]. Experimental results, using real DNA, RNA, and protein sequences as well as synthetic data, demonstrated orders-of-magnitude superiority of our approach in terms of efficiency, while maintaining an accuracy very competitive in relation to the state of the art.

Data structures for text indexing and string (sequence) comparison
Suffix trees are important data structures for text indexing and string algorithms. For any given string w of length $n = |w|$, a suffix tree for w takes $O(n)$ vertices and links. It is often presented as a compacted version of a suffix trie for w, where the latter is the trie (or digital search tree) built on the suffixes of w. The compaction process replaces each maximal chain of unary vertices with a single arc. For this, the suffix tree requires that the labels of its arcs are substrings encoded as pointers to w (or equivalent information). On the contrary, the arcs of the suffix trie are labeled by single symbols but there can be $\Theta(n^2)$ vertices and links for suffix tries in the worst case because of their unary vertices. It was an interesting question if the suffix trie can be stored using $O(n)$ vertices. We addressed it and thus presented the linear-size suffix trie, which guarantees $O(n)$ vertices [11]. We used a new technique for reducing the number of unary vertices to $O(n)$, that stems from some results on anti-dictionaries. For instance, by using the linear-size suffix trie, we are able to check whether a pattern p of length $m = |p|$ occurs in w in $O(m \log |\Sigma|)$ time and we can find the longest common substring of two strings w_1 and w_2 in $O((|w_1| + |w_2|) \log |\Sigma||)$ time for an alphabet $\Sigma|$.

Haplotype assembly
Haplotype assembly is the computational problem of reconstructing haplotypes in diploid organisms and is of fundamental importance for characterising the effects of single-nucleotide polymorphisms on the expression of phenotypic traits. Haplotype assembly highly benefits from the advent of "future-generation" sequencing technologies and their capability to produce long reads at increasing coverage. Existing methods are not able to deal with such data in a fully satisfactory way, either because accuracy or performances degrade as read length and sequencing coverage increase or because they are based on restrictive assumptions.

By exploiting a feature of future-generation technologies – the uniform distribution of sequencing errors – we designed an exact algorithm, called HAPCOL, that is exponential in the maximum number of corrections for each single-nucleotide polymorphism position and that minimises the overall error-correction score [22]. We performed an experimental analysis, comparing HAPCOL to the current state-of-the-art combinatorial methods both on real and simulated data. On a standard benchmark of real data, we showed that HAPCOL is competitive with state-of-the-art methods, improving the accuracy and the number of phased positions. Furthermore, experiments on realistically simulated datasets revealed that HAPCOL requires significantly less computing resources, especially memory. Thanks to its computational efficiency, HAPCOL can overcome the limits of previous approaches, allowing to phase datasets with higher coverage and without the traditional all-heterozygous assumption.

HAPCOL is based on MEC (Minimum error correction) which is computationally hard to solve. However, some approximation-based or fixed-parameter approaches have been proved capable of obtaining accurate results on real data. In another work [5], we then attempted to expand the current characterisation of the computational complexity of MEC from such approximation and fixed-parameter tractability points of view. We showed that MEC is not approximable within a constant factor, whereas it is approximable within a logarithmic factor in the size of the input. Furthermore, we answered open questions on the fixed-parameter tractability for parameters of classical or practical interest: the total number of corrections and the fragment length. In addition, we presented a direct 2-approximation algorithm for a variant of the problem that has also been applied in the framework of clustering data. Finally, since polyploid genomes, such as those of plants and fishes, are composed of more than two copies of the chromosomes, we introduced a novel formulation of MEC, namely the k-ploid MEC problem, that extends the traditional problem to deal with polyploid genomes. We showed that the novel formulation remains both computationally hard and hard to approximate. Nonetheless, from the parameterised point of view, we proved that the problem is tractable for parameters of practical interest such as the number of haplotypes and the coverage, or the number of haplotypes and the fragment length.

6.3. Inferring and analysing the networks of molecular elements

Metamodules in transcriptomic analysis

The human microbiome plays a key role in health and disease. Thanks to comparative metatranscriptomics, the cellular functions that are deregulated by the microbiome in disease can now be computationally explored. Unlike gene-centric approaches, pathway-based methods provide a systemic view of such functions; however, they typically consider each pathway in isolation and in its entirety. They can therefore overlook the key differences that (i) span multiple pathways, (ii) contain bidirectionally deregulated components, (iii) are confined to a pathway region. To capture these properties, computational methods that reach beyond the scope of predefined pathways are needed.

By integrating an existing module discovery algorithm into comparative metatranscriptomic analysis, we developed METAMODULES, a novel computational framework for automated identification of the key functional differences between health- and disease-associated communities [20]. Using this framework, we recovered significantly deregulated subnetworks that were indeed recognised to be involved in two well-studied, microbiome-mediated oral diseases, such as butanoate production in periodontal disease and metabolism of sugar alcohols in dental caries. More importantly, our results indicated that our method can be used for hypothesis generation based on automated discovery of novel, disease-related functional subnetworks, which would otherwise require extensive and laborious manual assessment.

Metabolic environmental dialog

What an organism needs at least from its environment to produce a set of metabolites, *e.g.* target(s) of interest and/or biomass, has been called a minimal precursor set. Early approaches to enumerate all minimal precursor sets took into account only the topology of the metabolic network (topological precursor sets). Due to cycles and the stoichiometric values of the reactions, it is often not possible to produce the target(s) from a topological precursor set in the sense that there is no feasible flux. Although considering the stoichiometry makes the problem harder, it enables to obtain biologically reasonable precursor sets that we call stoichiometric. Recently a method to enumerate all minimal stoichiometric precursor sets was proposed in the literature. The relationship between topological and stoichiometric precursor sets had however not yet been studied.

Such relationship was explored in a recently accepted paper [3]. In there, we also presented two algorithms that enumerate all minimal stoichiometric precursor sets. The first one is of theoretical interest only and is based on the above mentioned relationship. The second approach solves a series of mixed integer linear programming (MILP) problems. We compared the computed minimal precursor sets to experimentally obtained growth media of several *Escherichia coli* strains using genome-scale metabolic networks.

The results showed that the second approach, called SASITA, efficiently enumerates minimal precursor sets taking stoichiometry into account, and allows for broad *in silico* studies of strains or species interactions that may help to understand *e.g.* pathotype and niche-specific metabolic capabilities.

This work was also part of the PhD of Martin Wannagat, defended in June 2016 [2].

Metabolic hyperstories

In the context of a PhD in the team (whose defence took place in Dec 8, 2016) [1] and using metabolomics data, we focused on inferring the metabolic behaviour of an organism when it is subjected to a change in conditions. In this case, one can infer the reactions impacted when the changes are controlled and known (*e.g.* exposition to toxic compounds, changes in culture conditions). However, understanding how the metabolism of an organism changes of equilibrium is also of interest to infer the processes related for example to a transition between a commensal or beneficial bacterium to a pathogenic one. This question led to two different methods. The first, that we called TOTORO (for TOpological analysis of Transient metabOlic RespOnse), is based on the topology of metabolic networks to infer the reactions involved in a transient state, when an organism goes from one state of growth to another. We proposed a novel definition using the directed hypergraph representation and discussed its application on a dataset of Yeast exposed to cadmium. We showed that this method suggests more complete solutions of the reactions impacted during the metabolic shift. The second method, called KOTOURA (for Kantitative analysis Of Transient metabOlic and regUlatory Response And control), offers a constraint-based perspective in a more quantitative approach. We applied it to a simulated dataset and we are currently trying to infer the possible quantitative responses to mutations with a more complete kinetic model. An image previously used is that condition-specific models provide a snapshot of the metabolism of an organism, whether it is at the evolutionary-time scale or at the scale of a specific environmental condition describing a physiological process. Our idea here is thus to infer the transitions between those snapshots.

Besides the PhD manuscript, two papers are in preparation to present this work. They should be submitted in early 2017. A prototype for the two methods is available at: http://hyperstories.gforge.inria.fr/.

6.4. Modelling and analysing a network of individuals, or a network of individuals' networks

Robustness of the parsimonious reconciliation method in cophylogeny

The currently most used method in cophylogenetic studies is the so-called *phylogenetic tree reconciliation*. In this model, we are given the phylogenetic tree of the hosts H, the one of the symbionts S, and a mapping ϕ from the leaves of S to the leaves of H indicating the known symbiotic relationships among present-day organisms. The common evolutionary history of the hosts and of their symbionts is then explained through a number of macroevolutionary events (four in general). A reconciliation is then a function λ which is an extension of the mapping ϕ between leaves to a mapping that includes all internal nodes and that can be constructed using the different types of events considered. An optimal reconciliation is usually defined in a parsimonious way: a cost is associated to each event and a solution of minimum total cost is searched for.

An important issue in this model is that it makes strong assumptions on the input data which may not be verified in practice. We examine two cases where this situation happens. The first is related to a limitation in the currently available methods for tree reconciliation where the association ϕ of the leaves is for now required to be a function. This is not realistic as a single symbiont species can infect more than one host. For each present-day symbiont involved in a multiple association, one is currently forced to choose a single one. The second case addresses a different type of problem related to the phylogenetic trees of hosts and symbionts. These indeed are assumed to be correct, which may not be the case. In this work, we addressed the problem of correctly rooting a phylogenetic tree.

We thus explored the robustness of the parsimonious tree reconciliation method under "editing" (multiple associations) or "small perturbations" of the input (rooting problem) [29].

An extended version of this paper has been submitted to *IEEE/ACM Transactions on Computational Biology and Bioinformatics*.

Insights on the virulence of swine respiratory tract mycoplasmas through genome-scale metabolic modelling

The respiratory tract of swines is colonised by several bacteria among which are three Mycoplasma species: *Mycoplasma flocculare*, *Mycoplasma hyopneumoniae* and *Mycoplasma hyorhinis*. While colonisation by *M. flocculare* was shown to be virtually asymptomatic, *M. hyopneumoniae* is known to be the causative agent of enzootic pneumonia and *M. hyorhinis* to be present in cases of pneumonia, polyserositis and arthritis. Nonetheless, the elevated genomic resemblance among these three mycoplasmas combined with their different levels of pathogenicity is an indication that they have unknown mechanisms of virulence and differential expression. In 2015, we performed whole-genome metabolic network reconstructions for these three mycoplasmas. The results obtained were then submitted for publication to *BMC Genomics*. The paper has since been published [13].

Maximal chain subgraphs and covers of bipartite graphs motivated by analysis of cytoplasmic incompatibility

In a previous work of the team (Nor *et al.American Naturalist*, 182(1):15-24, 2013; Nor*et al.Information and Computation*, 213:23-32, 2012), we showed that a minimum chain subgraph cover of a given bipartite graph provides a good model for identifying the minimum genetic architecture enabling to explain one type of manipulation, called *cytoplasmic incompatibility*, by some parasite bacteria on their hosts. This phenomenon results in the death of embryos produced in crosses between males carrying the infection and uninfected females. The observed cytoplasmic compatibility relationships, can then be represented by a bipartite graph with males and females in different classes. Moreover, as different minimum (resp. minimal) covers may correspond to solutions that differ in terms of their biological interpretation, the capacity to enumerate all such minimal chain covers becomes crucial.

We recently addressed three related problems that bear some interest for the above problem besides raising interesting theoretical questions [35]. One is the enumeration of all the maximal *edge induced* chain subgraphs of a bipartite graph, for which we provided a polynomial delay algorithm. We gave bounds on the number of maximal chain subgraphs for a bipartite graph and used them to establish the input-sensitive complexity of the enumeration problem. The second problem we treated was the one of finding the minimum number of chain subgraphs needed to cover all the edges a bipartite graph. For this, we provided an exact exponential algorithm with a non trivial complexity. Finally, we approached the problem of enumerating all minimal chain subgraph covers of a bipartite graph and showed that it can be solved in quasi-polynomial time.

An extended version of the conference paper has been submitted to a journal in December 2016.

6.5. Cross-fertilising different computational approaches and other theoretical results

On the Complexity of Quadratic-Time Solvable Problems

Quadratic-time solvable problems may be classified into two classes: problems that are solvable in *truly subquadratic* time (that is, in time $(n^{2-\epsilon})$ for some $\epsilon > 0$) and problems that are not, unless the well known Strong Exponential Time Hypothesis (in short, SETH) is false. We proved that some quadratic-time solvable problems are indeed easier than expected [6]. We provided an algorithm that computes the transitive closure of a directed graph in time $(mn^{\frac{\omega+1}{4}})$, where m denotes the number of edges in the transitive closure and ω is the exponent for matrix multiplication. As a side effect of our analysis, we were able to prove that our algorithm runs in time $(n^{\frac{5}{3}})$ if the transitive closure of the graph is sparse. The same time bounds hold if we want to check whether a graph is transitive, by replacing m with the number of edges in the graph itself. As far as we know, this gives us the fastest algorithm for checking whether a sparse graph is transitive. Finally, we applied our algorithm to the comparability graph recognition problem (which dates back to 1941): also in this case, we obtained the first truly subquadratic algorithm. We then dealt with some hardness results. In particular, we started from an artificial quadratic-time solvable variation of the k-SAT problem and constructed a graph of Karp reductions, proving that a truly subquadratic-time algorithm for any of the problems in the graph falsifies SETH. More specifically, the analysed problems were the following: computing the subset graph, finding dominating sets, computing the betweenness centrality of a vertex, computing the minimum closeness centrality, and computing the hyperbolicity of a pair of vertices. We were also able to include in our framework three proofs that had already appeared in the literature, concerning the problems of distinguishing between split graphs of diameter 2 and diameter 3, of solving the local alignment of strings, and of finding two orthogonal binary vectors inside a collection.

Enumeration of solutions produced by closure operations
In enumeration problems, we are interested in listing a set of elements, which can be of exponential cardinality in the size of the input. The complexity of such problems is thus measured in terms of their input and output sizes. An enumeration algorithm with a complexity polynomial in both sizes is called output polynomial or total polynomial time. Another more precise notion of complexity is related to the *delay*, that is to the time between the production of two consecutive solutions. We are especially interested in problems solvable with a delay polynomial in the input size. These are considered as the tractable problems in enumeration complexity.

We addressed the problem of generating all elements obtained by the saturation of an initial set by some operations [41]. More precisely, we proved that we can generate the closure by polymorphisms of a boolean relation with a polynomial delay. This implies for instance that we can compute with polynomial delay the closure of a family of sets by any set of "set operations" (*e.g.* union, intersection, difference, symmetric difference, etc.). To do so, we proved that for any set of operations $\mathcal{F}$, one can decide in polynomial time whether an element belongs to the closure by $\mathcal{F}$ of a family of sets. When the relation is over a domain larger than two elements, our generic enumeration method fails for some cases since the associated decision problem is NP-hard, and we then provide an alternative algorithm.

6.6. Going towards control

Combinatorial approach for microbial consortia synthetic design
Synthetic biology has boomed since the early 2000s when it started being shown that it was possible to efficiently synthetise compounds of interest in a much more rapid and effective way by using other organisms than those naturally producing them. However, to thus engineer a single organism, often a microbe, to optimise one or a collection of metabolic tasks may lead to difficulties when attempting to obtain a production system that is efficient, or to avoid toxic effects for the recruited microorganism. The idea of using instead a microbial consortium has thus started being developed in the last decade. This was motivated by the fact that such consortia may perform more complicated functions than could single populations and be more robust to environmental fluctuations. Success is however not always guaranteed. In particular, establishing which consortium is best for the production of a given compound or set thereof remains a great challenge. This is the problem we addressed in a paper accepted this year [16].

We thus introduced an initial model and a method, called MULTIPUS, that enable to propose a consortium to synthetically produce compounds that are either exogenous to it, or are endogenous but where the interaction among the species in the consortium could improve the production line. In mathematical terms, given a weighted directed hypergraph $\mathcal{H}$, the problem is to enumerate all directed sub-hypergraphs whose sets of vertices and of hyperarcs are included in those of $\mathcal{H}$, enable to produce the set of targets of interest from a subset of the sources of $\mathcal{H}$, and are of minimum weight. We called this the Directed Steiner Hypertree (DSH) problem.

We showed that the main issue in terms of the complexity of the problem comes from the hyperarcs with multiple source vertices (we called those the *tentacular hyperarcs*), not from the possible multiplicity of the target vertices. This is not the only issue though, and we thus further demonstrated that even when there is only one target that needs to be reached, the problem remains NP-hard. When both parameters, number of tentacular hyperarcs and of targets, are fixed, the problem becomes tractable. We then explored two methods for addressing it. One is a dynamic programming approach, and the other logic programming using ASP (Answer Set Programming). The second was more efficient for now, and the software MULTIPUS is thus based on it.

As initial validations of the model and of the method, we applied it to two case-studies taken from the literature.

This work was also part of the PhD of Alice Julien-Laferrière defended in December 2016 [1].

IBIS Project-Team

6. New Results

6.1. Qualitative modeling of gene regulatory networks in food-borne pathogens

Bacteria are able to respond to a variety of environmental stresses, which poses food safety problems when these bacteria are food-borne pathogens. Addition of salt, one of the most ancient and common way of preserving food, subjects the bacteria to an osmotic stress to which some may survive. However, the molecular mechanisms of adaptation in food-born pathogens are largely unknown. As a first step towards better understanding these adaptation processes on the molecular level, Delphine Ropers and Aline Métris from the Institute for Food Research in Norwich (UK), invited researcher in IBIS this year, have developed a qualitative model of the osmotic stress response in the model bacterium *Escherichia coli* for which more information is available in the literature. The model has allowed to reproduce the behavior of *E. coli* cells adapting to an osmotic stress by including the regulatory mechanisms involved in the process. This work has been published in the *International Journal of Food Microbiology* [15] and in *Data in Brief* [16]. It paves the way to modelling stress responses of other foodborne pathogens like *Salmonella* to stresses relevant for the food industry, for which much less is known.

The tool used for the qualitative modeling and simulation of the regulatory mechanism underlying osmotic stress is GENETIC NETWORK ANALYZER (GNA). This tool describes the dynamics of gene regulatory networks by means of PLDE models, as described in Section 5.1 . GNA has been integrated with the other bioinformatics tools distributed by Genostar (http://www.genostar.com/). Version 8.7.2 of GNA was released by IBIS and Genostar this year and has been deposited at the Agence pour la Protection des Programmes (APP). Some bugs have been corrected in the new version and the program has been adapted to the latest versions of Java and the software platform of Genostar. Version 8.7.2 supports the SBML standard and is also capable of exporting its models to the newly-developed standard for qualitative models, SBML Qual. This standard has been elaborated by the community of developers of logical and related modeling tools (CoLoMoTo), in which the GNA developers participate.

6.2. Analysis of fluorescent reporter gene data

The use of fluorescent and luminescent reporter genes allows real-time monitoring of gene expression, both at the level of individual cells and cell populations (Section 3.2). In order to fully exploit this technology, we need methods to rapidly construct reporter genes, both on plasmids and on the chromosome, mathematical models to infer biologically relevant quantities from the primary data, and computer tools to achieve this in an efficient and user-friendly manner. For instance, in a typical microplate experiment, 96 cultures are followed in parallel, over several hours, resulting in 10,000-100,000 measurements of absorbance and fluorescence and luminescence intensities.

Valentin Zulkower, former PhD student in IBIS, developed novel methods for the analysis of reporter gene data obtained in microplate experiments, based on the use of regularized linear inversion. This allows a range of estimation problems in the analysis of reporter gene data, notably the inference of growth rate, promoter activity, and protein concentration profiles, to be solved in a mathematically sound and practical manner. This work was presented at the major bioinformatics conference ISMB/ECCB and published in the special issue of *Bioinformatics* associated with the conference last year. The linear inversion methods have been implemented in the Python package WELLFARE and integrated in the web application WELLINVERTER (Section 5.3). Funded by the Institut Français de Bioinformatique (IFB), Yannick Martin is currently extending WellInverter into a scalable and user-friendly web service providing a guaranteed quality of service, in terms of availability and response time. This web service will be deployed on the IFB platform and accompanied by extensive user documentation, online help, and a tutorial.

While the use of microplate readers results in population-level measurements of gene expression, for many applications it is mandatory to monitor gene expression over time on the level of individual cells. Several developments in the past decade have enormously extended the capabilities to achieve this, in particular the combination of fluorescence time-lapse microscopy for precisely quantifying gene expression in single cells and microfluidics technology for cultivating bacteria in confined spatial compartments and under well-controlled experimental conditions. One of the most wide-spread microfluidics devices is the so-called mother machine shown in Figure 5 . A major problem is that software for image analysis (segmentation, tracking, lineage reconstruction, ...) adapted to the requirements of mother machine applications are still missing. IBIS therefore collaborates with the BEAGLE project-team for the adaptation of their tool FLUOBACTRACKER to the analysis of time-lapse movies of fluorescent reporter expression and bacterial growth in microfluidics devices. This collaboration is supported by the Technology Transfer and Innovation department of Inria, in the framework of the Inria Hub program, and has allowed the hiring of Cyril Dutrieux as a software engineer in IBIS.

6.3. Models of carbon metabolism in bacteria

Adaptation of bacterial growth to changes in environmental conditions, such as the availability of specific carbon sources, is triggered at the molecular level by the reorganization of metabolism and gene expression: the concentration of metabolites is adjusted, as well as the concentration and activities of enzymes, the rate of metabolic reactions, the transcription and translation rates, and the stability of proteins and RNAs. This reprogramming of the bacterial cell is carried out by i) specific interactions involving regulatory proteins or RNAs that specifically respond to the change of environmental conditions and ii) global regulation involving changes in the concentration of RNA polymerase, ribosomes, and metabolite pools that globally affect the rates of transcription, translation, and degradation of all RNAs and proteins. While these phenomena have been well studied in steady-state growth conditions, much less is known about adaptation during growth transitions. In particular, only very few data are available on changes in the concentration and activity of the transcription and translation machineries and almost no data exost for the dynamic response of the degradation machinery.

In the framework of the PhD thesis of Manon Morin, supported by a Contrat Jeune Scientifique INRA-Inria (2012-2015), the collaboration of Delphine Ropers with Muriel Cocaign-Bousquet and Brice Enjalbert at INRA/INSA de Toulouse has allowed to disentangle the role of post-transcriptional regulation from other regulatory interactions in the dynamic adaptation of central carbon metabolism in *E. coli*. In a multi-scale analysis of a wild-type strain and its isogenic mutant attenuated for the protein CsrA, a variety of experimental data have been acquired in relevant conditions, including growth parameters, gene expression levels, metabolite pools, enzyme activities and metabolic fluxes. Data integration, metabolic flux analysis and regulation analysis revealed the pivotal role of post-transcriptional regulation for shaping carbon metabolism. In particular, the work has shed light on *csrA* essentiality and has provided an explanation for the glucose-phosphate stress observed in the mutant strain. A paper summarizing the work has been published in *Molecular Microbiology* this year [14]. A follow-up study conducted with various mutant strains of the carbon storage regulator system has elucidated the role of post-transcriptional regulation in the dynamics of glycogen storage and consumption, as well as the key role of the latter compound for bacterial fitness. A paper summarizing the work is being prepared for publication.

The collaboration with INRA/INSA de Toulouse is continued in the context of the PhD thesis of Thibault Etienne, funded by an INRA-Inria PhD grant, with the objective of developing models able to explain how cells coordinate their physiology and the functioning of the transcription, translation, and degradation machineries following changes in the availability of carbon sources in the environment.

6.4. Stochastic modeling and identification of gene regulatory networks in bacteria

At the single-cell level, the processes that govern single-cell dynamics in general and gene expression in particular are better described by stochastic models. Modern techniques for the real-time monitoring of gene

expression in single cells enable one to apply stochastic modelling to study the origins and consequences of random noise in response to various environmental stresses, and the emergence of phenotypic variability. The potential impact of single-cell stochastic analysis and modelling ranges from a better comprehension of the biochemical regulatory mechanisms underlying cellular phenotypes to the development of new strategies for the (computer assisted or genetically engineered) control of cell populations and even of single cells.

Work in IBIS on gene expression and interaction dynamics at the level of individual cells is addressed in terms of identification of intrinsic noise models from population snapshot data, on the one hand, and the inference of models focusing on cellular variability within isogenic populations from fluorescence microscopy gene expression profiles, on the other hand. Along with modelling and inference comes analysis of the inferred models in various respects, notably in terms of identifiability, single-cell state estimation and control. Other problems related with single-cell modelling and extracellular variability are considered in eukaryotic cells through external collaborations.

In the context of the response of yeast cells to osmotic shocks, in collaboration with the LIFEWARE project team and colleagues from Université Paris Descartes and University of Pavia (Italy), Eugenio Cinquemani has investigated the use of mixed effects-modelling and identification techniques to characterize individual cell dynamics in isogenic cell populations. Mixed-effects models are hierarchical models where parametric response profiles of individuals is subject to inter-individual parameter variability following a common population distribution. Starting from identification approaches in pharmacokinetics, we have developed and applied inference methods to microfluidics data, with a focus on the response of budding yeast to osmotic shocks. Results were described in a publication in *PLoS Computational Biology* [13]. A study of statistical validation methods for mixed-effects and alternative stochastic modelling paradigms has been presented at the *IFAC Conference on Foundations of Systems Biology in Engineering (FOSBE)* in Magdeburg [19]. In collaboration with the project-team BIOCORE at Inria Sophia-Antipolis - Méditerranée, the approach is now being investigated for the joint modelling of growth and gene expression in *E. coli*, based on single-cell microfluidics data from growth arrest-and-restart experiments. Further challenges stemming from this activity toward modelling and identification of extrinsic noise in individual cells are part of the recently started ANR project MEMIP (Section 8.2).

Work on identification and state estimation for single-cell gene network dynamics has been focused on the reconstruction of promoter activity profiles from fluorescent reporter data. In a stochastic, intrinsic noise modelling context, Eugenio Cinquemani addressed the problem of inferring promoter activity statistics over a cell population, such as mean and variance, from analogous statistics of the reporter output, as obtained from so-called population snapshot data. This nontrivial extension of the deterministic promoter activity deconvolution problem from population-average data is the first, crucial step toward reconstruction of promoter activity regulation and inference of stochastic network models. Earlier results, concerning parameter identifiability of stochastic promoter activity models and reconstruction of promoter activity distributions in the special case of single-switch systems, were further developed in a contribution to the HSB conference this year [18]. The relationship between the spectrum of the promoter process (cross-correlation function) and the mean-variance profiles of fluorescent reporter readouts was derived and demonstrated on examples, laying down the bases for a full-blown observability analysis and the development of spectrum estimation methods.

The collaboration of Eugenio Cinquemani with Marianna Rapsomaniki (IBM Zurich Research Lab, Switzerland), Zoi Lygerou (University of Patras, Greece) and John Lygeros (ETH Zurich, Switzerland) is moving on to applications of joint work published in *Bioinformatics* last year. Deployment of the methods developed into an efficient cluster-based software for the inference of protein kinetics in single cells from Fluorescence Recovery After Photobleaching (FRAP) experiments is under study. Exploitation of the same methods for the simulation and analysis of more general biochemical processes in single cells is part of the ongoing research efforts.

6.5. Growth control in bacteria and biotechnological applications

The ability to experimentally control the growth rate is crucial for studying bacterial physiology. It is also of central importance for applications in biotechnology, where often the goal is to limit or even arrest

growth. Growth-arrested cells with a functional metabolism open the possibility to channel resources into the production of a desired metabolite, instead of wasting nutrients on biomass production. The objective of the RESET project, supported in the framework of the Programme d'Investissements d'Avenir (Section 8.2), is to develop novel strategies to limit or completely stop microbial growth and to explore biotechnological applications of these approaches.

A foundation result for growth control in bacteria was published in the journal *Molecular Systems Biology* last year. In that publication, we described an engineered *E. coli* strain where the transcription of a key component of the gene expression machinery, RNA polymerase, is under the control of an inducible promoter. By changing the inducer concentration in the medium, we can adjust the RNA polymerase concentration and thereby switch bacterial growth between zero and the maximal growth rate supported by the medium. The publication also presented a biotechnological application of the synthetic growth switch in which both the wild-type *E. coli* strain and our modified strain were endowed with the capacity to produce glycerol when growing on glucose. Cells in which growth has been switched off continue to be metabolically active and harness the energy gain to produce glycerol at a twofold higher yield than in cells with natural control of RNA polymerase expression. Remarkably, without any further optimization, the improved yield is close to the theoretical maximum computed from a flux balance model of *E. coli* metabolism. This work is being continued in several directions in the context of the RESET project by Célia Boyat. In order to further explore the possibility of transferring this technology to biotechnology companies, we participated in the Challenge Out of Labs (http://www.linksium.fr/lancez-vous/resultat-challenge-out-of-labs/) organized by Linksium, the local incubator for technology transfer and start-up building. The presentation by Hans Geiselmann was selected for further development by Linksium.

In a review recently accepted for publication in *Trends in Microbiology* [11], we have put the scientific results mentioned above in a broader context. As illustrated by the synthetic growth switch, reengineering the gene expression machinery allows modifying naturally evolved regulatory networks and thereby profoundly reorganizing the manner in which bacteria allocate resources to different cellular functions. This opens new opportunities for our fundamental understanding of microbial physiology and for a variety of applications. We describe how recent breakthroughs in genome engineering and the miniaturization and automation of culturing methods have offered new perspectives for the reengineering of the transcription and translation machinery in bacteria as well as the development of novel *in vitro* and *in vivo* gene expression systems. In our paper, we review different examples from the unifying perspective of resource reallocation, and discuss the impact of these approaches for microbial systems biology and biotechnological applications.

Whereas the synthetic growth switch has been designed for biotechnological purposes, the question can be asked how resource allocation is organized in wild-type strains that have naturally evolved. Recent work has shown that coarse-grained models of resource allocation can account for a number of empirical regularities relating the the macromolecular composition of the cell to the growth rate. Some of these models hypothesize control strategies enabling microorganisms to optimize growth. While these studies focus on steady-state growth, such conditions are rarely found in natural habitats, where microorganisms are continually challenged by environmental fluctuations. The aim of the PhD thesis of Nils Giordano is to extend the study of microbial growth strategies to dynamical environments, using a self-replicator model. In collaboration with the BIOCORE project-team, we formulate dynamical growth maximization as an optimal control problem that can be solved using Pontryagin's Maximum Principle. We compare this theoretical gold standard with different possible implementations of growth control in bacterial cells. We find that simple control strategies enabling growth-rate maximization at steady state are suboptimal for transitions from one growth regime to another, for example when shifting bacterial cells to a medium supporting a higher growth rate. A near-optimal control strategy in dynamical conditions is shown to require information on several, rather than a single physiological variable. Interestingly, this strategy has structural analogies with the regulation of ribosomal protein synthesis by ppGpp in *E. coli*. It involves sensing a mismatch between precursor and ribosome concentrations, as well as the adjustment of ribosome synthesis in a switch-like manner. Our results show how the capability of regulatory systems to integrate information about several physiological variables is critical for optimizing growth in a changing environment. A paper describing the above results was published in *PLoS Computational Biology* this year [12].

NUMED Project-Team (section vide)

STEEP Project-Team

7. New Results

7.1. Ecological accounting

Besides the publication of the article [2] on environmental pressures in supply chains in the leading journal in the field (*Journal of Industrial Ecology*), the most important result obtained on this front this year bears on the quantification of the errors associated with the national road freight transport database (SITRAM). This database is informed year by year through a dedicated sampling campaign, but the errors associated with the various types of material goods transported have never been quantified. This was achieved by our team through the use of appropriate error estimators. This result is eagerly awaited by a number of scientific teeams and public territorial agencies. Furthermore, the methodology that we have developed can easily be transposed to other countries. This result constitutes an important piece in the overall effort that the team has devoted to the question of the quantification of uncertainties in material flow analyses.

7.2. Modeling of human-mediated dispersal via road network in invasive spreads

In the case of ecosystem invasions, human-mediated dispersal often acts as a vector for many exotic species, both at the introduction and secondary spread stages. The introduction stage is mainly a consequence of human-mediated long distance dispersal and is known to happen at continental or global scales. Secondary spread, however occurs at smaller spatial and time scales (e.g. landscape), and can result from natural or human-mediated dispersal. Despite the importance of local goods and materials transportation (e.g. for landscaping, construction, or road-building) potentially promoting the spreading of invasive species, few studies have investigated short distance human-mediated dispersal. This lack of consideration seems to be the consequence of multiple factors:

- human-mediated dispersal is generally considered as a long distance dispersal process, more important for invasive species introduction than for secondary spread;
- it is difficult to qualify and quantify this mode of dispersal because of the multiplicity of potentially involved human activities;
- for organisms that can disperse naturally, it is complicated to distinguish between natural and human-mediated dispersal, as they may occur at similar scales.

Even though a range of methodologies are available for describing population spread by natural dispersal, only few models have been developed to describe and predict human-mediated dispersal consequences at small scales, and none of them take into account the topology of the transport infrastructure (roads, waterways). In this result, and in order to fill this gap and provide new insights into how invasion dynamics impact ecosystem services, we combined ecological (invasive species occurrence data) and geographical (transportationnetwork topology) data in a computer model to provide estimated frequencies and distances of materials transportations through the landscape. In this study (cf. [7]), we investigated the spreading pattern of Lasius neglectus, an invasive ant species originating from Turkey, which spread into Europe in the last decades. In this species, no mating or dispersal flights are performed, and its spread is therefore solely ensured by the transport of soil materials in which individuals are present. We built a numerical model enabling the estimation of multiple human-mediated dispersal parameters based on ground-truth sampling and a priori minimizing. After having built a model of the landscape-level spreading process that takes explicitly into account the topology of the road network, we localized the most probable sites of introduction, the number of jump events, as well as parameters of jump distances linked to the road network. Our model was also able to compute presence probability map, and can be used to calibrate sampling campaigns, explore invasion scenarios, and more generally perform invasion spread predictions. It could be applied to all the species that can be disseminated at local to regional scales by human activities through transportation networks.

7.3. A computer framework for measuring urban land-use mix

The number of people living in cities has been increasing considerably since 1950, from 746 million to 3.9 billion in 2014, and more than 66% of the world's population are projected to live in urban areas by 2050. As this continuing population growth and urbanization are projected to add 2.5 billion people to the world's urban population in 30 years, this situation brings new challenges on how to conceive cities that host such amounts of population in a sustainable way. This sustainability question should address several aspects, ranging from economical to social and environmental matters among others. In this work, we focus on the formalization of a measure of mixed use development or land use mix in a city, i.e. how the structure of the city can help to provide a car-free sustainable living. Such type of land use mix has been largely proven to contain beneficial outcomes in terms of sustainability and to positively contribute to societal outcome, health, and public transportation among others. We developped a framework to compute mixed uses development index. A main caracteristic of our approach is to use only crowd-sourcing data (from OpenStreetMap) to extract the geo-localized land uses. Due to the universality of this data source, we are able to process any geographical area in the world, as long as sufficient data are available in OSM. A Kernel Density Estimation is performed for each of the land uses, outputing the spatial distribution of the different land uses. Based on this representation, a measure of land use mix is then calculated using the Entropy Index. The resulting GIS output shows enriched information for urban planners, supporting and aiding the decision-making procedure.

The framework, still in the phase of validation, was applied on the cities of London and Grenoble [9]. Future work includes integrating the LUM output for measuring the urban sprawl phenomenon and performing numerical interpretations of desirable mixed use values. We will also study the potential integration to transportation models, where land use mix correlation with the activities and residential uses can help to improve demand estimation. In addition, further investigation can be done by means of analyzing in detail the different types of activities. Finally, the estimation of LUM can be refined by taking into account, besides their location, the accessibility between different land uses, which is partly conditioned by the transportation infrastructure.

7.4. Calibration and sensitivity analysis for LUTI models

This year, we have consolidated our previous works on calibration of LUTI models, in particular of the Tranus model [6]. The developed approaches are currently applied to instantiate a complete Tranus model for the Grenoble catchment area, in collaboration with AURG (Urban Planning Agency of the Grenoble area) and Brian Morton (U North Carolina).

We have also collaborated with the AIRSEA project-team towards applying novel sensitivity analysis tools to study the influence of the different parameter sets of a Tranus model [13]. The rationale is to then apply optimization methods to the most influential parameters. As a result, we were able to calibrate a real-life Tranus model such that results were of higher quality than with the baseline ad hoc approach, while reducing calibration time significantly.

AVALON Project-Team

6. New Results

6.1. Energy Efficiency of Large Scale Distributed Systems

Participants: Laurent Lefevre, Daniel Balouek-Thomert, Eddy Caron, Radu Carpa, Marcos Dias de Assunção, Jean-Patrick Gelas, Olivier Glück, Jean-Christophe Mignot, Violaine Villebonnet.

6.1.1. Energy Efficient Core Networks with SDN

This work [14], [15] seeks to improve the energy efficiency of backbone networks by providing an intra-domain Software Defined Network (SDN) approach to selectively and dynamically turn off and on a subset of links. We proposed the STREETE framework (SegmenT Routing based Energy Efficient Traffic Engineering) that represents an online method to switch some links off/on dynamically according to the network load. We have implemented a working prototype in the OMNET++ simulator and design a validation platform [15] based on NetFPGA and Raspberry equipment with SDN frameworks (ONOS).

6.1.2. Energy Proportionality in HPC Systems

Energy savings are among the most important topics concerning Cloud and HPC infrastructures nowadays. Servers consume a large amount of energy, even when their computing power is not fully utilized. These static costs represent quite a concern, mostly because many datacenter managers are over-provisioning their infrastructures compared to the actual needs. This results in a high part of wasted power consumption. In this work [25], [24], [23] , we proposed the BML ("Big, Medium, Little") infrastructure, composed of heterogeneous architectures, and a scheduling framework dealing with energy proportionality. We introduce heterogeneous power processors inside datacenters as a way to reduce energy consumption when processing variable workloads. Our framework brings an intelligent utilization of the infrastructure by dynamically executing applications on the architecture that suits their needs, while minimizing energy consumption. Our first validation process focuses on distributed stateless web servers scenario and we analyze the energy savings achieved through energy proportionality. This research activity is performed with the collaboration of Sepia Team (IRIT, Toulouse) through the co-advising of Violaine Villebonnet.

6.1.3. Energy-Aware Server Provisioning

Several approaches to reduce the power consumption of datacenters have been described in the literature, most of which aim to improve energy efficiency by trading off performance for reducing power consumption. However, these approaches do not always provide means for administrators and users to specify how they want to explore such trade-offs. This work [11] provides techniques for assigning jobs to distributed resources, exploring energy efficient resource provisioning. We use middleware-level mechanisms to adapt resource allocation according to energy-related events and user-defined rules. A proposed framework enables developers, users and system administrators to specify and explore energy efficiency and performance trade-offs without detailed knowledge of the underlying hardware platform. Evaluation of the proposed solution under three scheduling policies shows gains of 25% in energy-efficiency with minimal impact on the overall application performance. We also evaluate reactivity in the adaptive resource provisioning This research activity is performed with the collaboration of NewGen SR society through the co-advising of Daniel Balouek-Thomert.

6.1.4. Improving Energy Re-Usage of Large Scale Computing Systems

The heat induced by computing resources is generally a waste of energy in supercomputers. This is especially true in very large scale supercomputers, where the produced heat has to be compensated with expensive and energy consuming cooling systems. Energy is a critical point for future supercomputing trends that currently try to achieve exascale, without having its energy consumption reaching an important fraction of a nuclear power plant. Thus, new ways of generating or recovering energy have to be explored. Energy harvesting consists in recovering wasted energy. ThermoElectric Generators (TEGs) aim to recover energy by converting wasted dissipated energy into usable electricity. By combining computing units (CU) and TEGs at very large scale, we spotted a potential way to recover energy from wasted heat generated by computations on supercomputers. In this work [30], [20], we study the potential gains in combining TEGs with computational units at petascale and exascale. We explored the technology behind TEGs, and finally our results concerning binding TEGs and computational units in a petascale and exascale system. With the available technology, we demonstrate that the use of TEGs in a supercomputer environment could be realistic and quickly profitable, and hence have a positive environmental impact.

6.2. MPI Application Simulation

Participant: Frédéric Suter.

6.2.1. The SMPI approach

In [37], we summarized our recent work and developments on SMPI, a flexible simulator of MPI applications. In this tool, we took a particular care to ensure our simulator could be used to produce fast and accurate predictions in a wide variety of situations. Although we did build SMPI on SimGrid whose speed and accuracy had already been assessed in other contexts, moving such techniques to a HPC workload required significant additional effort. Obviously, an accurate modeling of communications and network topology was one of the key to such achievements. Another less obvious key was the choice to combine in a single tool the possibility to do both offline and online simulation.

6.3. MapReduce Computations on Hybrid Distributed Computations Infrastructures

Participant: Gilles Fedak.

6.3.1. Availability and Network-Aware MapReduce Task Scheduling over the Internet

MapReduce offers an easy-to-use programming paradigm for processing large datasets. In our previous work, we have designed a MapReduce framework called BitDew-MapReduce for desktop grid and volunteer computing environment, that allows non-expert users to run data-intensive MapReduce jobs on top of volunteer resources over the Internet. However, network distance and resource availability have great impact on MapReduce applications running over the Internet. To address this, an availability and network-aware MapReduce framework over the Internet is proposed in [9]. Simulation results show that the MapReduce job response time could be decreased by 27.15%, thanks to Naive Bayes Classifier-based availability prediction and landmark-based network estimation.

6.4. Managing Big Data Life Cycle

Participants: Gilles Fedak, Valentin Lorentz, Laurent Lefevre.

6.4.1. Data Energy Traceability

In this work, we have opened a new research topic around the energy traceability of data. The objective is to answer the question of how many energy has been consumed to produce a particular data. This work is partially based on the concept of data life cycle, that is extended to record each step of the data life cycle.

6.5. Desktop Grid Computing

Participant: Gilles Fedak.

6.5.1. Multi-Criteria and Satisfaction Oriented Scheduling for Hybrid Distributed Computing Infrastructures

Assembling and simultaneously using different types of distributed computing infrastructures (DCI) like Grids and Clouds is an increasingly common situation. Because infrastructures are characterized by different attributes such as price, performance, trust, greenness, the task scheduling problem becomes more complex and challenging. In [7], we presented the design for a fault-tolerant and trust-aware scheduler, which allows to execute Bag-of-Tasks applications on elastic and hybrid DCI, following user-defined scheduling strategies. Our approach, named Promethee scheduler, combines a pull-based scheduler with multi-criteria Promethee decision making algorithm. Because multi-criteria scheduling leads to the multiplication of the possible scheduling strategies, we proposed SOFT, a methodology that allows to find the optimal scheduling strategies given a set of application requirements. The validation of this method is performed with a simulator that fully implements the Promethee scheduler and recreates an hybrid DCI environment including Internet Desktop Grid, Cloud and Best Effort Grid based on real failure traces. A set of experiments shows that the Promethee scheduler is able to maximize user satisfaction expressed accordingly to three distinct criteria: price, expected completion time and trust, while maximizing the infrastructure useful employment from the resources owner point of view. Finally, we present an optimization which bounds the computation time of the Promethee algorithm, making realistic the possible integration of the scheduler to a wide range of resource management software.

6.6. HPC Component Models and OpenMP

Participants: Hélène Coullon, Vincent Lanore, Christian Perez, Jérôme Richard, Thierry Gautier.

6.6.1. Combining Both a Component Model and a Task-based Model

We have studied the feasibility of efficiently combining both a software component model and a task-based model. Task based models are known to enable efficient executions on recent HPC computing nodes while component models ease the separation of concerns of application and thus improve their modularity and adaptability. We have designed a prototype version of the COMET programming model combining concepts of task-based and component models, and a preliminary version of the COMET runtime built on top of StarPU and L2C. Evaluations of the approach have been conducted on a real-world use-case analysis of a sub-part of the production application GYSELA. Results show that the approach is feasible and that it enables easy composition of independent software codes without introducing overheads. Performance results are equivalent to those obtained with a plain OpenMP based implementation. Part of this work is described in [38].

6.6.2. OpenMP Scheduling Heuristic for NUMA Architecture

The recent addition of data dependencies to the OpenMP 4.0 standard provides the application programmer with a more flexible way of synchronizing tasks. Using such an approach allows both the compiler and the runtime system to know exactly which data are read or written by a given task, and how these data will be used through the program lifetime. Data placement and task scheduling strategies have a significant impact on performances when considering NUMA architectures. While numerous papers focus on these topics, none of them has made extensive use of the information available through dependencies. One can use this information to modify the behavior of the application at several levels: during initialization to control data placement and during the application execution to dynamically control both the task placement and the tasks stealing strategy, depending on the topology. This paper [26] introduces several heuristics for these strategies and their implementations in our OpenMP runtime XKAAPI. We also evaluate their performances on linear algebra applications executed on a 192-core NUMA machine, reporting noticeable performance improvement when considering both the architecture topology and the tasks data dependencies. We finally compare them to strategies presented previously by related works.

6.6.3. Extending OpenMP with Affinity Clause: Design and Implementation

OpenMP 4.0 introduced dependent tasks, which give the programmer a way to express fine grain parallelism. Using appropriate OS support (such as NUMA libraries), the runtime can rely on the information in the depend clause to dynamically map the tasks to the architecture topology. Controlling data locality is one of the key factors to reach a high level of performance when targeting NUMA architectures. On this topic, OpenMP does not provide a lot of flexibility to the programmer yet, which lets the runtime decide where a task should be executed. In [27], we present a class of applications which would benefit from having such a control and flexibility over tasks and data placement. We also propose our own interpretation of the new affinity clause for the task directive, which is being discussed by the OpenMP Architecture Review Board. This clause enables the programmer to give hints to the runtime about tasks placement during the program execution, which can be used to control the data mapping on the architecture. In our proposal, the programmer can express affinity between a task and the following resources: a thread, a NUMA node, and a data. We then present an implementation of this proposal in the Clang-3.8 compiler, and an implementation of the corresponding extensions in our OpenMP runtime LIBKOMP. Finally , we present a preliminary evaluation of this work running two task-based OpenMP kernels on a 192-core NUMA architecture, that shows noticeable improvements both in terms of performance and scalability.

6.6.4. Support of High Task Throughput for Complex OpenMP Application

In [4], we present block algorithms and their implementation for the parallelization of sub-cubic Gaussian elimination on shared memory architectures using OpenMP standard. Contrarily to the classical cubic algorithms in parallel numerical linear algebra, we focus here on recursive algorithms and coarse grain parallelization. Indeed, sub-cubic matrix arithmetic can only be achieved through recursive algorithms making coarse grain block algorithms perform more efficiently than fine grain ones. This work is motivated by the design and implementation of dense linear algebra over a finite field, where fast matrix multiplication is used extensively and where costly modular reductions also advocate for coarse grain block decomposition. We incrementally build efficient kernels, for matrix multiplication first, then triangular system solving, on top of which a recursive PLUQ decomposition algorithm is built. We study the parallelization of these kernels using several algorithmic variants: either iterative or recursive and using different splitting strategies. Experiments show that recursive adaptive methods for matrix multiplication, hybrid recursive-iterative methods for triangular system solve and tile recursive versions of the PLUQ decomposition, together with various data mapping policies, provide the best performance on a 32 cores NUMA architecture. Overall, we show that the overhead of modular reductions is more than compensated by the fast linear algebra algorithms and that exact dense linear algebra matches the performance of full rank reference numerical software even in the presence of rank deficiencies.

6.7. Security for Virtualization and Clouds

Participants: Eddy Caron, Arnaud Lefray.

6.7.1. Secured Systems in Clouds with Model-Driven Orchestration

As its complexity grows, securing a system is harder than it looks. Even with efficient security mechanisms, their configuration remains a complex task. Indeed, the current practice is the hand-made configuration of these mechanisms to protect systems about which we generally lack information. Cloud computing brings its share of new security concerns but it may also be considered as leverage to overcome these issues. In [13] we adressed the key challenge of achieving global security of Cloud systems and advocate for a new approach: Model-Driven Orchestration. We have designed an implementation of this new approach called Security-Aware Models for Clouds. With this approach an industrial use-case has been deployed and secured using the Sam4C software.

6.8. Large Scale Cloud Deployment

Participants: Eddy Caron, Marcos Dias de Assunção, Christian Perez, Pedro de Souza Bento Da Silva.

6.8.1. Efficient Heuristics for Placing Large-Scale Distributed Applications on Multiple Clouds

With the fast growth of the demand for Cloud computing services, the Cloud has become a very popular platform to develop distributed applications. Features that in the past were available only to big corporations, like fast scalability, availability, and reliability, are now accessible to any customer, including individuals and small companies, thanks to Cloud computing. In order to place an application, a designer must choose among VM types, from private and public cloud providers, those that are capable of hosting her application or its parts using as criteria application requirements, VM prices, and VM resources. This procedure becomes more complicated when the objective is to place large component based applications on multiple clouds. In this case, the number of possible configurations explodes making necessary the automation of the placement. In this context, scalability has a central role since the placement problem is a generalization of the NP-Hard multi-dimensional bin packing problem.

In this work [22], we propose efficient greedy heuristics based on first fit decreasing and best fit algorithms, which are capable of computing near optimal solutions for very large applications, with the objective of minimizing costs and meeting application performance requirements. Through a meticulous evaluation, we show that the greedy heuristics took a few seconds to calculate near optimal solutions to placements that would require hours or even days when calculated using state of the art solutions, namely exact algorithms or meta-heuristics.

6.8.2. Multi-Criteria Malleable Task Management for Hybrid-Cloud Platforms

The use of large distributed computing infrastructure is a mean to address the ever increasing resource demands of scientific and commercial applications. The scale of current large-scale computing infrastructures and their heterogeneity make scheduling applications an increasingly complex task. Cloud computing minimises the heterogeneity by using virtualization mechanisms, but poses new challenges to middleware developers, such as the management of virtualization, elasticity and economic models. In this context, we proposed algorithms for efficient scheduling and execution of malleable computing tasks with high granularity while taking into account multiple optimisation criteria such as resource cost and computation time. We focused on hybrid platforms that comprise both clusters and cloud providers. In [12] we defined and formalized the main aspects of the problem, introduced the difference between local and global scheduling algorithms and evaluate their efficiency using discrete-event simulation.

6.9. Workflow management on Cloud environment

Participants: Daniel Balouek-Thomert, Eddy Caron, Laurent Lefevre.

6.9.1. Multi-objective workflow placements in Clouds

The recent rapid expansion of Cloud computing facilities triggers an attendant challenge to facility providers and users for methods for optimal placement of workflows on distributed resources, under the often-contradictory impulses of minimizing makespan, energy consumption, and other metrics. Evolutionary Optimization techniques that from theoretical principles are guaranteed to provide globally optimum solutions, are among the most powerful tools to achieve such optimal placements. Multi-Objective Evolutionary algorithms by design work upon contradictory objectives, gradually evolving across generations towards a converged Pareto front representing optimal decision variables – in this case the mapping of tasks to resources on clusters. However the computation time taken by such algorithms for convergence makes them prohibitive for real time placements because of the adverse impact on makespan. In [11], we described parallelization, on the same cluster, of a Multi-objective Differential Evolution method (NSDE-2) for optimization of workflow placement, and the attendant speedups that bring the implicit accuracy of the method into the realm of practical utility. We did experimental validation on a reallife testbed using diverse Cloud traces. The solutions under different scheduling policies demonstrate significant reduction in energy consumption with some improvement in makespan. We designed, implementation and evaluation of an energy-efficient resource management system that builds upon DIET, an open source middleware and NSDE-divisible tasks with precedence constraints. Real-life experiment of this approach on the Grid'5000 testbed demonstrates its effectiveness in a dynamic environment.

CTRL-A Team

7. New Results

7.1. Design and programming

7.1.1. Component-based approaches

Participants: Gwenaël Delaval, Eric Rutten.

Architecting in the context of variability has become a real need in today's software development. Modern software systems and their architecture must adapt dynamically to events coming from the environment (e.g., workload requested by users, changes in functionality) and the execution platform (e.g., resource availability). Component-based architectures have shown to be very suited for self-adaptation especially with their dynamical reconfiguration capabilities. However, existing solutions for reconfiguration often rely on low level, imperative, and non formal languages. We have defined Ctrl-F, a domain-specific language whose objective is to provide high-level support for describing adaptation behaviors and policies in component-based architectures. It relies on reactive programming for formal verification and control of reconfigurations. We integrate Ctrl-F with the FraSCAti Service Component Architecture middleware platform, and apply it to the Znn.com self-adaptive case study

We have obtained new results in the application of modular controller synthesis and BZR compilation integrated in Ctrl-F, in order to attack issues in scalability, and reusability. We are also considering integration at the DSL level of expressivity extensions, for which the compilation and controller synhesis is relying on the ReaX tool developed at Inria Rennes, in the Sumo team.

7.1.2. Rule-based systems

Participants: Adja Sylla, Eric Rutten.

We are starting a cooperation with CEA LETI/DACLE on the topic of a high-level language for safe rule-based programming in the LINC platform. The general context is that of the runtime redeployment of distributed applications, for example managing smart buildings. Motivations for redeployment can be diverse: load balancing, energy saving, upgrading, or fault tolerance. Redeployment involves changing the set of components in presence, or migrating them. The basic functionalities enabling to start, stop, migrate, or clone components, and the control managing their safe coordination, will have to be designed in the LINC middleware developed at CEA.

Rule based middlewares such as LINC enable high level programming of distributed adaptive systems behaviours. LINC also provides the systems with transactional guarantees and hence ensures their reliability at runtime. However, the set of rules may contain design errors (e.g. conflicts, violations of constraints) that can bring the system in unsafe safe or undesirables states, despite the guarantees provided by LINC. On the other hand, automata based languages such as Heptagon/BZR enable formal verification and especially synthesis of discrete controllers to deal with design errors. Our work studies these two languages and combines their execution mechanisms, from a technical perspective. A case study taken in the field of building automation is treated to illustrate the proposed approach [18].

The PhD of Adja Sylla at CEA on this topic is co-advised with F. Pacull and M. Louvel.

7.2. Infrastructure-level support

We apply the results of the previous axes of the team's activity to a range of infrastructures of different natures, but sharing a transversal problem of reconfiguration control design. From this very diversity of validations and experiences, we draw a synthesis of the whole approach, towards a general view of Feedback Control as MAPE-K loop in Autonomic Computing [23], [22].

7.2.1. Autonomic Cloud and Big-Data systems

Participants: Soguy Mak Kare Gueye, Gwenaël Delaval, Eric Rutten.

Complex computing systems are increasingly self-adaptive, with an autonomic computing approach for their administration. Real systems require the co-existence of multiple autonomic management loops, each complex to design. However their uncoordinated co-existence leads to performance degradation and possibly to inconsistency. There is a need for methodological supports facilitating the coordination of multiple autonomic managers. To tackle this problem, we take a global view and underscore that Autonomic Management Systems (AMS) are intrinsically reactive, as they react to flows of monitoring data by emitting flows of reconfiguration actions. Therefore we propose a new approach for the design of AMSs, based on synchronous programming and discrete controller synthesis techniques. They provide us with high-level languages for modeling the system to manage, as well as means for statically guaranteeing the absence of logical coordination problems. Hence, they suit our main contribution, which is to obtain guarantees at design time about the absence of logical inconsistencies in the taken decisions. We detail our approach, illustrate it by designing an AMS for a realistic multi-tier application, and evaluate its practicality with an implementation [16].

We addressed these problems in the context of follow-ups of the ANR project Ctrl-Green, in cooperation with LIG (N. de Palma) in the framework of the PhD of S. Gueye [17] and the post-doc of N. Berthier.

7.2.2. Reconfiguration control in DPR FPGA

Participants: Soguy Mak Kare Gueye, Eric Rutten.

Dynamically reconfigurable hardware has been identified as a promising solution for the design of energy efficient embedded systems. However, its adoption is limited by the costly design effort including verification and validation, which is even more complex than for non dynamically reconfigurable systems. We worked on this topic in the context of a ensign environment, developed in the framework of the ANR project Famous, in cooperation with LabSticc in Lorient and Inria Lille (DaRT team). We proposed a tool-supported formal method to automatically design a correct-by-construction control of the reconfiguration. By representing system behaviors with automata, we exploit automated algorithms to synthesize controllers that safely enforce reconfiguration strategies formulated as properties to be satisfied by control. We design generic modeling patterns for a class of reconfigurable architectures, taking into account both hardware architecture and applications, as well as relevant control objectives. We validate our approach on two case studies implemented on FPGAs [3].

We are currently valorizing results in more publications [15], and extending the use of control techniques by evaluating the new tool ReaX developed at Inria Rennes (Sumo).

We are starting a new ANR project called HPeC, within which some of these topics will be extended, especially regarding hierarchical and modular control, and logico-numeric aspects.

7.2.3. Autonomic memory management in HPC

Participants: Naweiluo Zhou, Gwenaël Delaval, Bogdan Robu, Eric Rutten.

Parallel programs need to manage the time trade-off between synchronization and computation. A high parallelism may decrease computing time but meanwhile increase synchronization cost among threads. Software Transactional Memory (STM) has emerged as a promising technique, which bypasses locks, to address synchronization issues through transactions. A way to reduce conflicts is by adjusting the parallelism, as a suitable parallelism can maximize program performance. However, there is no universal rule to decide the best parallelism for a program from an offline view. Furthermore, an offline tuning is costly and error-prone. Hence, it becomes necessary to adopt a dynamical tuning-configuration strategy to better manage a STM system. Autonomic control techniques begin to receive attention in computing systems recently. Control technologies offer designers a framework of methods and techniques to build autonomic systems with well-mastered behaviors. The key idea of autonomic control is to implement feedback control loops to design safe, efficient and predictable controllers, which enable monitoring and adjusting controlled systems dynamically while keeping overhead low. We propose to design feedback control loops to automate the choice of parallelism

at runtime and diminish program execution time [20], [24], [21]. It is then combined with another objective related to Thread Mapping Control [19]

In the context of the action-team HPES of the Labex Persyval-lab [0] (see 9.1), this work is performed in cooperation with LIG (J.F. Méhaut) in the framework of the PhD of N. Zhou [14].

7.2.4. Control of smart environments

Participants: Adja Sylla, Armando Ochoa, Eric Rutten, Stéphane Mocanu.

7.2.4.1. A service-oriented approach to smart home applications control with reactive programming

The need for adaptability in pervasive computing is growing, driven in part by the increasing number and variety of communication devices. In autonomic applications, however, the control architecture frequently becomes itself a complex system that needs to be adapted. Autonomic applications are often composed of multiple control loops ? each addressing a specific aspect ? whose execution needs to be coordinated for efficient and correct administration. We therefore propose to investigate the use of reactive control models with events and states to coordinate autonomic loops in service-oriented architectures. In this work, we illustrate our approach by integrating a controller based on discrete controller synthesis in an autonomic pervasive environment. The role of the controller is to influence the service-binding criteria of multiple control loops, while respecting logical constraints. In particular, we consider reconfiguration operations of known and dynamic service sets. This work constituted the M2R internship of Armando Ochoa, and was performed in cooperation with the Adele team at LIG, co-advised by E. Rutten and V. Lestideau, in the framework of the Labex Persyval-lab project CASE.

Another activity in this topic was the M2R internship of Ronak Feizimirkhani, co-advised by S. Mocanu and V. Lestideau, The context is the development of an application for a smart home in which automation devices are connected through a wireless communication protocol, Z-Wave, and controlled by a central controller, USB plug in. This involves methods and tools to design fail-safe controllers for autonomic, adaptive, reconfigurable computing systems by combining Computer Science and Control Theory techniques. For this purpose, it is necessary to access required information over the network, derive out a simplified model of the physical network, and then link it to the User interface application. According to the information achieved, there will be an estimation of the network diagnostics to find some probable solutions for. The final application is in a user media to do installing, maintaining or even optimizing the network and devices.

7.2.4.2. Rule-based specification of smart environments control

In the context of IoT applications like mart home environments, the rules for programming in the LINC framework are used as a flexible tool to govern the relations between sensors and actuators. Runtime coordination and formal analysis becomes a necessity to avoid side effects mainly when applications are critical. In cooperation with CEA LETI/DACLE, we are working on a case study for safe applications development in IoT and smart home environments.

New results from Section 7.1.2 are applied in case studies regarding smart environments (offices or homes) [18].

[0] https://persyval-lab.org/en/sites/hpes

DANTE Project-Team

7. New Results

7.1. Graph & Signal Processing

Participants: Sarra Ben Alaya, Éric Fleury, Paulo Gonçalves Andrade.

7.1.1. Isometric graph shift operator]

Following up the PhD work of Benjamin Girault [57], we demonstrated in [26] that the isometric graph shift operator we originally proposed, does have a vertex-domain interpretation as a diffusion operator using a polynomial approximation. We showed that its impulse response exhibits an exponential decay of the energy way from the impulse, demonstrating localisation preservation. Additionally, we formalised several techniques that can be used to study other graph signal operators.

7.2. Performance analysis and networks protocols

Participants: Mohammed Amer, Thomas Begin, Anthony Busson, Éric Fleury, Paulo Gonçalves Andrade, Yannick Léo, Isabelle Guérin Lassous, Philippe Nain, Huu Nghi Nguyen, Laurent Reynaud.

7.2.1. Use of large scale CDR for protocol performance evaluation and modelling

In [11] we use large scale CDR (Call Data Records) coming from a nationwide cellular telecommunication operator during a two month period to validate several DTN approaches for conveying SMS traffic in dense urban areas taking benefits of the density of users and the mobility of the users. We study a mobile dataset including 8 Million users living in large urban area. This gives us a precise estimation of the average transmission time and the global performance of our approach. Our analysis shows that after 30 min, half of the SMS are delivered successfully to destination. In [10], we study the temporal activity of a user and the user movements. At the user scale, the usage is not only defined by the amount of calls but also by the userâs mobility. At a higher level, the base stations have a key role on the quality of service. From a very large Call Detail Records (CDR) we first study call duration and inter-arrival time parameters. Then, we assess user movements between consecutive calls (switching from a station to another one). Our study suggests that user mobility is pretty dependent on user activity. Furthermore, we show properties of the inter-call mobility by making an analysis of the call distribution.

7.2.2. End-to-end delay

Because of the growing complexity of computer networks, a new paradigm has been introduced to ease their design and management, namely, the SDN (Software-defined Networking). In particular, SDN defines a new entity, the controller that is in charge of controlling the devices belonging to the data plane In order to let the controller take its decisions, it must have a global view on the network. This includes the topology of the network and its links capacity, along with other possible performance metrics such delays, loss rates, and available bandwidths. This knowledge can enable a multi-class routing, or help guarantee levels of Quality of Service. In [33], [20], [42], we proposed new algorithms that allow a centralised entity, such as the controller in an SDN network, to accurately estimate the end-to-end delay for a given flow in its network. The proposed methods are passive in the sense that they do not require any additional traffic to be run. Through extensive simulations, we show that these methods are able to accurately estimate the expectation and the standard deviation of end-to-end delays.

In [14] we investigated the traversal time of a file across N communication links subject to stochastic changes in the sending rate of each link. Each link's sending rate is modelled by a finite-state Markov process. Two cases, one where links evolve independently of one another (N mutually independent Markov processes), and the second where their behaviours are dependent (these N Markov processes are not mutually independent) were considered. A particular instance where the above is

7.2.3. Circumventing the complexity of multi-server queues

Many real-life systems can be viewed as instances of multi-server queues. However, when the number of servers is high (say more than 16) and the arrival or/and service process exhibit high variability, current state-of-the-art solutions often become intractable due to the combinatorial growth of the underlying state space of the Markov chain. We proposed two efficient, fast and easy-to-implement approximate solutions to deal with *G/G/c*-like queues in [4], [2]. Our solutions rely the use of an original, though incomplete, state description that heavily breaks the complexity of multi-server queues. We have extensively validated our approximations against discrete-event simulation for several QoS performance metrics such as mean sojourn time and blocking probability with excellent results.

7.2.4. Wi-Fi networks optimization

Densification of Wi-Fi networks has led to the possibility for a station to choose between several access points (APs). On the other hand, the densification of APs generates interference, contention and decreases the global throughput as APs have to share a limited number of channels. Optimizing the association step between APs and stations can alleviate this problem and increase the overall throughput and fairness between stations. We proposed original solutions [23], [22] to this optimization problem based on two contributions. First, we modeled the association optimization problem assuming a realistic share of the medium between APs and stations and among APs when using the 802.11 DCF (Distributed Coordination Function) mode. Then, we introduced a local search algorithm to solve this problem through a suitable neighborhood structure. We show that the classical approaches in the literature, based on a time based fairness scheme, is less efficient than our solution when the number of orthogonal channels is limited. Also, we show through a large set of simulations and scenarios that our models are able to capture the real throughputs of Wi-Fi networks.

7.2.5. Controlled mobility in wireless networks

In this work, we have investigated the application of an adapted controlled mobility strategy on self-propelling nodes, which could efficiently provide network resource to users scattered on a designated area. In [7], we describe an adapted controlled mobility strategy and detail the design of our Virtual Force Protocol (VFP) which allows a swarm of vehicles to track and follow hornets to their nests, while maintaining connectivity through a wireless multi-hop communication route with a remote ground station used to store applicative data such as hornet trajectory and vehicle telemetry. In [43], we design a physics-based controlled mobility strategy, which we name the extended Virtual Force Protocol (VFPe), allowing self-propelled nodes, and in particular here unmanned aerial vehicles, to fly autonomously and cooperatively. In this way, ground devices scattered on the operation site may establish communications through the wireless multi-hop communication routes formed by the network of aerial nodes. In [28], we design a virtual force-based controlled mobility scheme, named VFPc, and evaluate its ability to be jointly used with a dual packet-forwarding and epidemic routing protocol. In particular, we study the possibility for end-users to achieve synchronous communications at given times of the considered scenarios.

7.3. Modeling of Dynamics of Complex Networks

Participants: Christophe Crespelle, Éric Fleury, Márton Karsai, Yannick Léo, Philippe Nain, Matteo Morini.

7.3.1. Data Driven studies on socioeconomic data and communication networks

The study of correlations between the social network and economic status of individuals is difficult due to the lack of large-scale multimodal data disclosing both the social ties and economic indicators of the same population. Thanks to our collaboration with GranData, we close this gap through the analysis of coupled datasets recording the mobile phone communications and bank transaction history of one million anonymised individuals living in a Latin American country. From this large scale data set based on a representative, society-large population we empirically demonstrate some long-lasting hypotheses on socioeconomic correlations, which potentially lay behind social segregation, and induce differences in human mobility. More precisely, in [12] we show that wealth and debt are unevenly distributed among people in agreement with the Pareto principle; the observed social structure is strongly stratified, with people being better connected to others

of their own socioeconomic class rather than to others of different classes; the social network appears to have assortative socioeconomic correlations and tightly connected ârich clubsâ; and that individuals from the same class live closer to each other but commute further if they are wealthier. In [41], we show that typical consumption patterns are strongly correlated with identified socioeconomic classes leading to patterns of stratification in the social structure. In addition we measure correlations between merchant categories and introduce a correlation network, which emerges with a meaningful community structure. We detect multivariate relations between merchant categories and show correlations in purchasing habits of individuals. Our work provides novel and detailed insight into the relations between social and consuming behaviour with potential applications in recommendation system design. In [36] we provide insight about the effects of marking events on the structure and the dynamics of egocentric networks. More precisely, we study the impact of university admission on the composition and evolution of the egocentric networks of freshmen. In other words, we study whether university helps to build connections between egos from different socioeconomic classes, or new social ties emerge via homophilic effects between students of similar economic status. Finally, in [44],

7.3.2. Generalisation of multilayer and temporal graphs

In [16] we introduce the concept of MultiAspect Graph (MAG) as a graph generalisation that we prove to be isomorphic to a directed graph, and also capable of representing all previous generalisations of multilayer and temporal networks. In our proposal, the set of vertices, layers, time instants, or any other independent features are considered as an aspect of the MAG. For instance, a MAG is able to represent multilayer or time-varying networks, while both concepts can also be combined to represent a multilayer time-varying network and even other higher-order networks. Since the MAG structure admits an arbitrary (finite) number of aspects, it hence introduces a powerful modelling abstraction for networked complex systems. In [17] we develop the algebraic representation and basic algorithms for MultiAspect Graphs (MAGs). In particular, we show that, as a consequence of the properties associated with the MAG structure, a MAG can be represented in matrix form. Moreover, we also show that any possible MAG function (algorithm) can be obtained from this matrix-based representation. This is an important theoretical result since it paves the way for adapting well-known graph algorithms for application in MAGs. We present a set of basic MAG algorithms, constructed from well-known graph algorithms, such as degree computing, Breadth First Search (BFS), and Depth First Search (DFS).

Multilayer networks arise in scenarios when a common set of nodes form multiple networks via different co-existing, and sometimes interdependent means of connectivity. In [6] we studied the threshold on the occupation density in the individual network layers for long-range connectivity to emerge in a large multilayer network. For a multilayer network formed via merging M random instances of a graph G with site-occupation probability q in each layer, we showed that when q exceeds a threshold $q_c(M)$, a giant connected component appears in the M-layer network. We showed that $q_c(M) \lesssim \sqrt{-\ln(1-p_c)}/\sqrt{M}$, where p_c is the bond percolation threshold of G, and $q_c(1) \equiv q_c$ is by definition the site percolation threshold of G. We found $q_c(M)$ exactly for when G is a large random graph with any given node-degree distribution. We calculated $q_c(M)$ numerically for various regular lattices, and obtained an exact lower bound for the kagome lattice. Finally, we established an intriguing close connection between the aforesaid multilayer percolation model and the well-studied problem of site-bond (or, mixed) percolation, in the sense that both models provide a bridge between the traditional independent site and independent bond percolation models. Using this connection, and leveraging some analytical approximations to the site-bond critical region developed in the 1990s, we derived an excellent general approximation to the multilayer threshold $q_c(M)$ for regular lattices, which are not only functions solely of the p_c and q_c of the respective lattices, but also closely match the true values of $q_c(M)$ for a large class of lattices, even for small (single-digit) vales of M.

7.3.3. User-based representation of dynamical multimodal public transportation networks

In this project published as an invited paper [9], we provide a novel user-based representation of public transportation systems, which combines representations, accounting for the presence of multiple lines and reducing the effect of spatial embeddedness, while considering the total travel time, its variability across the schedule, and taking into account the number of transfers necessary. After the adjustment of earlier

techniques to the novel representation framework, we analyse the public transportation systems of several French municipal areas and identify hidden patterns of privileged connections. Furthermore, we study their efficiency as compared to the commuting flow. The proposed representation could help to enhance resilience of local transportation systems to provide better design policies for future developments.

7.3.4. Local cascades induced global contagion

In this paper [8] we analyse and model product adoption dynamics in the world's largest voice over internet service, the social network of Skype. We provide empirical evidence about the heterogeneous distribution of fractional behavioural thresholds, which appears to be independent of the degree of adopting egos. We show that the structure of real-world adoption clusters is radically different from previous theoretical expectations, since vulnerable adoptions induced by a single adopting neighbour appear to be important only locally, while spontaneous adopters arriving at a constant rate and the involvement of unconcerned individuals govern the global emergence of social spreading.

7.3.5. Asymptotic theory of time-varying social networks with heterogeneous activity and tie allocation

In this work [15] we empirically characterise social activity and memory in seven real networks describing temporal human interactions in three different settings: scientific collaborations, Twitter mentions, and mobile phone calls. We find that the individuals' social activity and their strategy in choosing ties where to allocate their social interactions can be quantitatively described and encoded in a simple stochastic network modelling framework. The Master Equation of the model can be solved in the asymptotic limit. The analytical solutions provide an explicit description of both the system dynamic and the dynamical scaling laws characterising crucial aspects about the evolution of the networks. The analytical predictions match with accuracy the empirical observations, thus validating the theoretical approach. Our results provide a rigorous dynamical system framework that can be extended to include other processes shaping social dynamics and to generate data driven predictions for the asymptotic behaviour of social networks.

7.3.6. Link prediction in the Twitter mention network

In this project [35] we analyse a large Twitter data corpus and quantify similarities between people by considering the set of their common friends and the set of their commonly shared hashtags in order to predict mention links among them. We show that these similarity measures are correlated among connected people and that the combination of contextual and local structural features provides better predictions as compared to cases where they are considered separately.

DATAMOVE Team

6. New Results

6.1. In Situ Statistical Analysis for Parametric Studies

In situ processing proposes to reduce storage needs and I/O traffic by processing results of parallel simulations as soon as they are available in the memory of the compute processes. We focus in this paper [11] on computing in situ statistics on the results of N simulations from a parametric study. The classical approach consists in running various instances of the same simulation with different values of input parameters. Results are then saved to disks and statistics are computed post mortem, leading to very I/O intensive applications. Our solution is to develop Melissa, an in situ library running on staging nodes as a parallel server. When starting, simulations connect to Melissa and send the results of each time step to Melissa as soon as they are available. Melissa implements iterative versions of classical statistical operations, enabling to update results as soon as a new time step from a simulation is available. Once all statistics ar updated, the time step can be discarded. We also discuss two different approaches for scheduling simulation runs: the jobs-in-job and the multi-jobs approaches. Experiments run instances of the Computational Fluid Dynamics Open Source solver Code_Saturne. They confirm that our approach enables one to avoid storing simulation results to disk or in memory.

6.2. Online Non-preemptive Scheduling in a Resource Augmentation Model based on Duality

Resource augmentation is a well-established model for analyzing algorithms, particularly in the online setting. It has been successfully used for providing theoretical evidence for several heuristics in scheduling with good performance in practice. According to this model, the algorithm is applied to a more powerful environment than that of the adversary. Several types of resource augmentation for scheduling problems have been proposed up to now, including speed augmentation, machine augmentation and more recently rejection. In this paper [7], we present a framework that unifies the various types of resource augmentation. Moreover, it allows generalize the notion of resource augmentation for other types of resources. Our framework is based on mathematical programming and it consists of extending the domain of feasible solutions for the algorithm with respect to the domain of the adversary. This, in turn allows the natural concept of duality for mathematical programming to be used as a tool for the analysis of the algorithm's performance. As an illustration of the above ideas, we apply this framework and we propose a primal-dual algorithm for the online scheduling problem of minimizing the total weighted flow time of jobs on unrelated machines when the preemption of jobs is not allowed. This is a well representative problem for which no online algorithm with performance guarantee is known. Specifically, a strong lower bound of $\Omega(\sqrt{n})$ exists even for the offline unweighted version of the problem on a single machine. In this paper, we first show a strong negative result even when speed augmentation is used in the online setting. Then, using the generalized framework for resource augmentation and by combining speed augmentation and rejection, we present an $(1+\epsilon_s)$-speed $O(\frac{1}{\epsilon_s \epsilon_r})$-competitive algorithm if we are allowed to reject jobs whose total weight is an ϵ_r-fraction of the weights of all jobs, for any $\epsilon_s > 0$ and $\epsilon_r \in (0,1)$. Furthermore, we extend the idea for analysis of the above problem and we propose an $(1+\epsilon_s)$-speed ϵ_r-rejection $O(\frac{k^{\frac{(k+3}{k}}}{\epsilon_r^{1/k} \epsilon_s^{\frac{(k+2)}{k}}})$-competitive algorithm for the more general objective of minimizing the weighted l_k-norm of the flow times of jobs.

6.3. Batsim: a Realistic Language-Independent Resources and Jobs Management Systems Simulator

As large scale computation systems are growing to exascale, Resources and Jobs Management Systems (RJMS) need to evolve to manage this scale modification. However, their study is problematic since they

are critical production systems, where experimenting is extremely costly due to downtime and energy costs. Meanwhile, many scheduling algorithms emerging from theoretical studies have not been transferred to production tools for lack of realistic experimental validation. To tackle these problems we propose Batsim [6], an extendable, language-independent and scalable RJMS simulator. It allows researchers and engineers to test and compare any scheduling algorithm, using a simple event-based communication interface, which allows different levels of realism. In this paper we show that Batsim's behavior matches the one of the real RJMS OAR. Our evaluation process was made with reproducibility in mind and all the experiment material is freely available.

POLARIS Team

6. New Results

6.1. Asymptotic Models

The analysis of a set of n stochastic entities interacting with each others can be particularly difficult. The *mean field approximation* is a very effective technique to characterize the transient probability distribution or steady-state regime of such systems when the number of entities n grows very large. The idea of mean-field approximation is to replace a complex stochastic system by a simpler deterministic dynamical system. This dynamical system is constructed by assuming that the objects are asymptotically independent. Each object is viewed as interacting with an average of the other objects (the *mean*-field). When each object has a finite or countable state-space, this dynamical system is usually a non-linear ordinary differential equation (ODE). An introduction to these techniques is provided in the book chapter [29].

- Mean-field games model the rational behavior of an infinite number of indistinguishable players in interaction [79]. An important assumption of mean-field games is that, as the number of player is infinite, the decisions of an individual player do not affect the dynamics of the mass. Each player plays against the mass. A mean-field equilibrium corresponds to the case when the optimal decisions of a player coincide with the decisions of the mass. This leads to a simpler computation of the equilibrium.

  It has been shown in [72], [96] that for some games with a finite number of players, the Nash equilibria converge to mean-field equilibria as the number of players tends to infinity. Hence, many authors argue that mean-field games are a good approximation of symmetric stochastic games with a large number of players. The classical argument is that the impact of one player becomes negligible when the number of players goes to infinity. In [17], [36], we show that, in general, this convergence does not hold. We construct an example for which the mean-field limit only describes a sub-set of the limiting equilibria. Each finite-player game has an equilibrium with a good social cost, this is not the case for the limit game.
- Computer system and network performance can be significantly improved by caching frequently used information. When the cache size is limited, the cache replacement algorithm has an important impact on the effectiveness of caching. In [21], [3], [20] we introduce approximations to determine the cache hit probability of two classes of cache replacement algorithms: the recently introduced h-LRU and LRU(m). These approximations only require the requests to be generated according to a general Markovian arrival process (MAP). This includes phase-type renewal processes and the IRM model as special cases. We provide both numerical and theoretical support for the claim that the proposed TTL approximations are asymptotically exact. We further show, by using synthetic and trace-based workloads, that h-LRU and LRU(m) perform alike, while the latter requires less work when a hit/miss occurs.
- In [16], we consider stochastic models in presence of uncertainty, originating from lack of knowledge of parameters or by unpredictable effects of the environment. We focus on population processes, encompassing a large class of systems, from queueing networks to epidemic spreading. We set up a formal framework for imprecise stochastic processes, where some parameters are allowed to vary in time within a given domain, but with no further constraint. We then consider the limit behaviour of these systems as the population size goes to infinity. We prove that this limit is given by a differential inclusion that can be constructed from the (imprecise) drift. We also we discuss different numerical algorithms to compute bounds of the so-obtained differential inclusions. We are currently working on an implementation of these algorithms in a numerical toolbox.

- In [37], we develop a fluid-limit approach to compute the expected absorbing time T_n of a n-dimensional discrete time Markov chain. We show that the random absorbing time T_n is well approximated by a deterministic time t_n that is the first time when a fluid approximation of the chain approaches the absorbing state at a distance $1/n$. We show the applicability of this approach with three different problems: the coupon collector, the erasure channel lifetime and the coupling times of random walks in high dimensional spaces.

6.2. Simulation

Simgrid is a toolkit providing core functionalities for the simulation of distributed applications in heterogeneous distributed environments. Although it was initially designed to study large distributed computing environments such as grids, we have recently applied it to performance prediction of HPC configurations.

- Finite difference methods are, in general, well suited to execution on parallel machines and are thus commonplace in High Performance Computing. Yet, despite their apparent regularity, they often exhibit load imbalance that damages their efficiency. In [38], we characterize the spatial and temporal load imbalance of Ondes3D, a seismic wave propagation simulator used to conduct regional scale risk assessment. Our analysis reveals that this imbalance originates from the structure of the input data and from low-level CPU optimizations. We then show that the CHARM++ runtime can effectively dynamically rebalance the load by migrating data and computation at the granularity of an MPI rank. We propose a methodology that leverages the capabilities of the SimGrid simulation framework and allows to conduct an experimental study at low computational cost.
- The article [35] summarizes our recent work and developments on SMPI, a flexible simulator of MPI applications. In this tool, we took a particular care to ensure our simulator could be used to produce fast and accurate predictions in a wide variety of situations. Although we did build SMPI on SimGrid whose speed and accuracy had already been assessed in other contexts, moving such techniques to a HPC workload required significant additional effort. Obviously, an accurate modeling of communications and network topology was one of the key to such achievements. Another less obvious key was the choice to combine in a single tool the possibility to do both offline and online simulation.

6.3. Trace and Statistical Analysis

- In [19], we present visual analysis techniques to evaluate the performance of HPC task-based applications on hybrid architectures. Our approach is based on composing modern data analysis tools (pjdump, R, ggplot2, plotly), enabling an agile and flexible scripting framework with minor development cost. We validate our proposal by analyzing traces from the full-fledged implementation of the Cholesky decomposition available in the MORSE library running on a hybrid (CPU/GPU) platform. The analysis compares two different workloads and three different task schedulers from the StarPU runtime system. Our analysis based on composite views allows to identify allocation mistakes, priority problems in scheduling decisions, GPU tasks anomalies causing bad performance, and critical path issues.
- Media events are an area of major concern for the science of territory, with a combination of empirical, methodological and theoretical fields of research. The paper [22] presents three variations of increasing complexity around the questions of the application of the concepts of “territory”, “territoriality” and “territorialization” to the description of media events. Each variation is illustrated by recent results from the research project ANR Geomedia on a corpus of international RSS flows produced by newspapers of French, English and Spanish language located in various countries of the world.

6.4. Electricity Markets

The increased penetration of renewable energy sources in existing power systems has led to necessary developments in electricity market mechanisms. Most importantly, renewable energy generation is increasingly made accountable for deviations between scheduled and actual energy generation. However, there is no mechanism to enforce accountability for the additional costs induced by power fluctuations. These costs are socialized and eventually supported by electricity customers. In [1], we propose some metrics for assessing the contribution of all market participants to power regulation needs, as well as an attribution mechanism for fairly redistributing related power regulation costs. We discuss the effect of various metrics used by the attribution mechanisms, and we illustrate, in a game-theoretical framework, their consequences on the strategic behavior of market participants. We also illustrate, by using the case of Western Denmark, how these mechanisms may affect revenues and the various market participants.

6.5. Power control in random wireless networks

Ever since the early development stages of wireless networks, the importance of radiated power has made power control an essential component of network design. In [13], we analyzed the problem of power control in large, random wireless networks that are obtained by "erasing" a finite fraction of nodes from a regular d-dimensional lattice of N transmit-receive pairs. Drawing on tools and ideas from statistical physics, we showed that this problem can be mapped to the Anderson impurity model for diffusion in random media; in this way, by employing the so-called *coherent potential approximation* (CPA) method, we calculated the average power in the system (and its variance) for 1-D and 2-D networks. In this regard, even though infinitely large systems are always unstable beyond a critical value of the users' SINR target, finite systems remain stable with high probability even beyond this critical SINR threshold. We calculated this probability by analyzing the density of low lying eigenvalues of an associated random Schrödinger operator, and we showed that the network can exceed this critical SINR threshold by a factor of at least $O((\log N)^{-2/d})$ before undergoing a phase transition to the unstable regime.

6.6. Energy efficiency in wireless networks

[6] The recent increase in the use of wireless networks for video transmission has led to the increase in the use of rate-adaptive protocols to maximize the resource utilization and increase the efficiency in the transmission. However, a number of these protocols lead to interactions among the users that are subjective in nature and affect the overall performance. In [6], we analyzed the interplay between the wireless network and video transmission dynamics in the light of subjective perceptions of the end users in their interactions – specifically, the trade-off between maximizing the quality of service (QoS) or quality of experience (QoE) and minimizing the transmission cost. By using methods from game theory, we derived an optimized transmission scheme that allows the efficient use of traditional protocols by taking into account the subjective interactions that occur in practical scenarios.

6.7. Cognitive radio and beyond

In cognitive radio networks, secondary (unlicensed) users (SUs) can access the spectrum opportunistically, whenever they sense an opening by the network's primary (licensed) users (PUs). In [7], we analyzed the minimization of overall power consumption over several orthogonal frequency bands under constraints on the minimum quality of service (QoS) and maximum peak and average interference to the network's PUs. To that end, we proposed a projected sub-gradient algorithm which quickly converges to an optimal configuration if the users' channels are fast fading.

Despite such benefits, the conventional cognitive radio network (CCRN) paradgim is not particularly attractive for opportunistic spectrum access because the network's PUs can recapture SU channels at will, thus interrupting the transmission of the latter. To address this crucial limitation, we proposed in [24] a semi-cognitive radio network (SCRN) paradigm where PUs are constrained to first use any free channels before being allowed to capture channels that are in use by SUs. These constraints slightly degrade the performance of the network's PUs, but *a*) they offer remarkable performance improvements to the network's SUs; and *b*)

they can be compensated by imposing a monetary (or other) penalty to the network's secondary owners. In [24], we provided a game-theoretic analysis of the performance trade-offs involved for both the PUs and the SUs, and we derived both centralized and distributed learning algorithms that allow the system control process to converge to a stable state.

6.8. Online resource allocation in dynamic wireless networks

The vast majority of works on wireless resource allocation (spectrum, power, etc.) has focused on two limit cases: In the *static regime*, the attributes of the network are assumed effectively static and the system's optimality analysis relies on techniques from (static) optimization. On the other hand, in the so-called *stochastic regime*, the network is assumed to evolve randomly following some fixed probability law, and the allocation of wireless resources is optimized using tools from stochastic optimization and control. In practical wireless networks however, both assumptions fail because of factors that introduce an unpredictable variability to the system (such as user mobility, users going arbitrarily on- and off-line, etc.).

The works [15], [27], [28] treat this problem by providing no-regret learning algorithms for single-user rate maximization and power control in multi-carrier cognitive radio and Internet of Things networks. The extension of these works to multi-antenna systems was carried out in [44], where we derived a matrix exponential learning algorithm for dynamic power allocation and control in time-varying MIMO systems. Building on this, we also showed in [8] that regret minimization techniques can also be applied to the much more challenging problem of energy efficiency maximization in dynamic networks – i.e. the maximization of successfully received bits per Watt of transmitted power in environments that fluctuate unpredictably over time. Finally, as was shown in [39], [23], [9], these unilateral performance gains also extend to large networks comprising hundreds (or even thousands) of users: there, the proposed matrix exponential learning algorithm converges to a stable state within a few iterations, even for very large of antennas and subcarriers.

6.9. Adaptive multi-path routing

Routing plays a crucial part in the efficient operation of packet-switched data networks, especially with regard to latency reduction and energy efficiency. However, in addition to being distributed (so as to cope with the prolific size of today's networks), optimized routing schemes must also be able to adapt to changes in the underlying network (e.g. due to variations in traffic demands, link quality, etc.).

First, to address the issue of latency reduction, we provided in [32] an adaptive multi-flow routing algorithm to select end-to-end paths in packet-switched networks. The algorithm is based only on local information, so it is suitable for distributed implementation; furthermore, it provides guarantees that the network configuration converges to a stable state and exhibits several robustness properties that make it suitable for use in dynamic real-life networks (such as robustness to measurement errors, outdated information and update desynchronization).

Concerning energy efficiency, [41] examines the problem of routing in optical networks with the aim of minimizing traffic-driven power consumption. To tackle this, [41] proposed a pricing scheme which, combined with a distributed learning method based on the Boltzmann distribution of statistical mechanics, exhibits remarkable operation properties even under uncertainty. Specifically, the long-term average of the network's power consumption converges quickly to its minimum value (in practice, within a few iterations of the algorithm), and this convergence remains robust in the face of uncertainty of arbitrarily high magnitude.

6.10. Learning in finite games

One of the most widely used algorithms for learning in finite games is the so-called *best response algorithm* (BRA); nonetheless, even though sevaral worst-case bounds are known for its convergence time, the algorithm's performance in typical game-theoretic scenarios seems to be far better than these worst-case bounds suggest. In [26], [18], [25], [31], we computed the average execution time of the BR algorithm using Markov chain coupling techniques that recast the average execution time of this discrete algorithm as the solution of an ordinary differential equation. In so doing, we showed that the worst-case complexity of the BR algorithm

in a potential game with N players and A actions per player is $AN(N-1)$, while its average complexity over random potential games is $O(N)$, independently of A.

In [34], we also studied the convergence rate of the HEDGE algorithm (which, contrary to the BR algorithm, leads to no regret even in adversarial settings). Motivated by applications to data networks where fast convergence is essential, we analyzed the problem of learning in generic N-person games that admit Nash equilibria in pure strategies. Despite the (unbounded) uncertainty in the players' observations, we show that hedging eliminates dominated strategies (a.s.) and, with high probability, it converges locally to pure Nash equilibria at the exponential rate $O(\exp(-c\sum_{j=1}^{t}\gamma_j))$, where γ_j is the algorithm s step size.

These results are strongly related to the long-term rationality properties (elimination of dominated strategies, convergence to pure Nash equilibria and evolutionarily stable states, etc.) of an underlying class of game dynamics based on regularization and Riemannian geometry. Specifically, in [42], we introduced a class of evolutionary game dynamics whose defining element is a state-dependent geometric structure on the set of population states. When this geometric structure satisfies a certain integrability condition, the resulting dynamics preserve many further properties of the replicator and projection dynamics and are equivalent to a class of reinforcement learning dynamics studied in [10]. Finally, as we showed in [2], these properties also hold even in the presence of noise, i.e. when the players only have noisy observations of their payoff vectors.

6.11. Learning in games with continuous action spaces

A key limitation of existing game-theoretic learning algorithms is that they invariably revolve around games with a finite number of actions per players. However, this assumption is often unrealistic (especially in network-based applications of game theory), a factor which severely limits the applicability of learning techniques in real-life problems.

To address this issue, we studied in [14] a class of control problems that can be formulated as potential games with continuous action sets, and we proposed an actor-critic reinforcement learning algorithm that provably converges to equilibrium in said class. The method employed is to analyse the learning process under study through a mean-field dynamical system that evolves in an infinite-dimensional function space (the space of probability distributions over the players' continuous controls). To do so, we extend the theory of finite-dimensional two-timescale stochastic approximation to an infinite-dimensional, Banach space setting, and we proved that the continuous dynamics of the process converge to equilibrium in the case of potential games. These results combine to give a provably-convergent learning algorithm in which players do not need to keep track of the controls selected by the other agents.

Finally, to address cases where mixing over a continuum of actions is unrealistic, we examined in [40] the convergence properties of a class of learning schemes for N-person games with continuous action spaces based on a continuous optimization technique known as "dual averaging". To study this multi-agent, pure-strategy learning process, we introduced the notion of *variational stability* (VS), and we showed that stable equilibria are locally attracting with high probability whereas globally stable states are globally attracting with probability 1. Finally, we examined the scheme's convergence speed and we showed that if the game admits a strict equilibrium and the players' mirror maps are surjective, then, with high probability, the process converges to equilibrium in a finite number of steps, no matter the level of uncertainty in the players' observations (or payoffs).

6.12. Stochastic optimization

A key feature of modern data networks is their distributed nature and the stochasticity surrounding users and their possible actions. To account for these issues in a general optimization context, we proposed in [4] a distributed, asynchronous algorithm for stochastic semidefinite programming which is a stochastic approximation of the continous-time matrix exponential scheme derived in [9]. This algorithm converges almost surely to an ϵ-approximation of an optimal solution requiring only an unbiased estimate of the gradient of the problem's stochastic objective. When applied to throughput maximization in wireless multiple-input and multiple-output (MIMO) systems, the proposed algorithm retains its convergence properties under a wide array

of mobility impediments such as user update asynchronicities, random delays and/or ergodically changing channels.

More generally, in view of solving convex optimization problems with noisy gradient input, we also analyzed in [43] the asymptotic behavior of gradient-like flows that are subject to stochastic disturbances. For concreteness, we focused on the widely studied class of mirror descent methods for constrained convex programming and we examined the dynamics' convergence and concentration properties in the presence of noise. In the small noise limit, we showed that the dynamics converge to the solution set of the underlying problem with probability 1. Otherwise, in the case of persistent noise, we estimated the measure of the dynamics' long-run concentration around interior solutions and their convergence to boundary solutions that are sufficiently "robust". Finally, we showed that a rectified variant of the method with a decreasing sensitivity parameter converges irrespective of the magnitude of the noise or the structure of the underlying convex program, and we derived an explicit estimate for its rate of convergence.

6.13. Benchmarking

In modern High Performance Computing architectures, the memory subsystem is a common performance bottleneck. When optimizing an application, the developer has to study its memory access patterns and adapt accordingly the algorithms and data structures it uses. The objective is twofold: on one hand, it is necessary to avoid missuses of the memory hierarchy such as false sharing of cache lines or contention in a NUMA interconnect. On the other hand, it is essential to take advantage of the various cache levels and the memory hardware prefetcher. Still, most profiling tools focus on CPU metrics. The few of them able to provide an overview of the memory patterns involved by the execution rely on hardware instrumentation mechanisms and have two drawbacks. The first one is that they are based on sampling which precision is limited by hardware capabilities. The second one is that they trace a subset of all the memory accesses, usually the most frequent, without information ab out the other ones. In [30] we present Moca, an efficient tool for the collection of complete spatio-temporal memory traces. Moca is based on a Linux kernel module and provides a coarse grained trace of a superset of all the memory accesses performed by an application over its addressing space during the time of its execution. The overhead of Moca is reasonable when taking into account the fact that it is able to collect complete traces which are also more precise than the ones collected by comparable tools.

Benchmarking has proven to be crucial for the investigation of the behavior and performances of a system. However, the choice of relevant benchmarks still remains a challenge. To help the process of comparing and choosing among benchmarks, in [33] we propose a solution for automatic benchmark profiling. It computes unified benchmark profiles reflecting benchmarks' duration, function repartition, stability, CPU efficiency, parallelization and memory usage. It identifies the needed system information for profile computation, collects it from execution traces and produces profiles through efficient and reproducible trace analysis treatments. The paper presents the design, implementation and the evaluation of the approach.

ROMA Project-Team

7. New Results

7.1. A backward/forward recovery approach for the preconditioned conjugate gradient method

Participants: Massimiliano Fasi [Univ. Mancherster, UK], Julien Langou [UC Denver, USA], Yves Robert, Bora Uçar.

Several recent papers have introduced a periodic verification mechanism to detect silent errors in iterative solvers. Chen [PPoPP'13, pp. 167-176] has shown how to combine such a verification mechanism (a stability test checking the orthogonality of two vectors and recomputing the residual) with checkpointing: the idea is to verify every d iterations, and to checkpoint every $c \times d$ iterations. When a silent error is detected by the verification mechanism, one can rollback to and re-execute from the last checkpoint. In this work, we also propose to combine checkpointing and verification, but we use algorithm-based fault tolerance (ABFT) rather than stability tests. ABFT can be used for error detection, but also for error detection and correction, allowing a forward recovery (and no rollback nor re-execution) when a single error is detected. We introduce an abstract performance model to compute the performance of all schemes, and we instantiate it using the preconditioned conjugate gradient algorithm. Finally, we validate our new approach through a set of simulations.

This work has been accepted for publication in the *Journal of Computational Science* [13].

7.2. High performance parallel algorithms for the tucker decomposition of sparse tensors

Participants: Oguz Kaya, Bora Uçar.

We investigate an efficient parallelization of a class of algorithms for the well-known Tucker decomposition of general N-dimensional sparse tensors. The targeted algorithms are iterative and use the alternating least squares method. At each iteration, for each dimension of an N-dimensional input tensor, the following operations are performed: (i) the tensor is multiplied with $N-1$ matrices (TTMc step); (ii) the product is then converted to a matrix; and (iii) a few leading left singular vectors of the resulting matrix are computed (TRSVD step) to update one of the matrices for the next TTMc step. We propose an efficient parallelization of these algorithms for the current parallel platforms with multicore nodes. We discuss a set of preprocessing steps which takes all computational decisions out of the main iteration of the algorithm and provides an intuitive shared-memory parallelism for the TTM and TRSVD steps. We propose a coarse and a fine-grain parallel algorithm in a distributed memory environment, investigate data dependencies, and identify efficient communication schemes. We demonstrate how the computation of singular vectors in the TRSVD step can be carried out efficiently following the TTMc step. Finally, we develop a hybrid MPI-OpenMP implementation of the overall algorithm and report scalability results on up to 4096 cores on 256 nodes of an IBM BlueGene/Q supercomputer.

This work has been published at *ICPP'16* [28].

7.3. Preconditioning techniques based on the Birkhoff–von Neumann decomposition

Participants: Michele Benzi [Emory University, Atlanta, USA], Bora Uçar.

We introduce a class of preconditioners for general sparse matrices based on the Birkhoff–von Neumann decomposition of doubly stochastic matrices. These preconditioners are aimed primarily at solving challenging linear systems with highly unstructured and indefinite coefficient matrices. We present some theoretical results and numerical experiments on linear systems from a variety of applications.

This work has been accepted for publication in the journal *Computational Methods in Applied Mathematics* [10].

7.4. Parallel CP decomposition of sparse tensors using dimension trees

Participants: Oguz Kaya, Bora Uçar.

Tensor factorization has been increasingly used to address various problems in many fields such as signal processing, data compression, computer vision, and computational data analysis. CANDECOMP/PARAFAC (CP) decomposition of sparse tensors has successfully been applied to many well-known problems in web search, graph analytics, recommender systems, health care data analytics, and many other domains. In these applications, computing the CP decomposition of sparse tensors efficiently is essential in order to be able to process and analyze data of massive scale. For this purpose, we investigate an efficient computation and parallelization of the CP decomposition for sparse tensors. We provide a novel computational scheme for reducing the cost of a core operation in computing the CP decomposition with the traditional alternating least squares (CP-ALS) based algorithm. We then effectively parallelize this computational scheme in the context of CP-ALS in shared and distributed memory environments, and propose data and task distribution models for better scalability. We implement parallel CP-ALS algorithms and compare our implementations with an efficient tensor factorization library, using tensors formed from real-world and synthetic datasets. With our algorithmic contributions and implementations, we report up to 3.95x, 3.47x, and 3.9x speedups in sequential, shared memory parallel, and distributed memory parallel executions over the state of the art, and up to 1466x overall speedup over the sequential execution using 4096 cores on an IBM BlueGene/Q supercomputer.

This work is described in a technical report [49].

7.5. Scheduling series-parallel task graphs to minimize peak memory

Participants: Enver Kayaaslan, Thomas Lambert, Loris Marchal, Bora Uçar.

We consider a variant of the well-known, NP-complete problem of minimum cut linear arrangement for directed acyclic graphs. In this variant, we are given a directed acyclic graph and asked to find a topological ordering such that the maximum number of cut edges at any point in this ordering is minimum. In our main variant the vertices and edges have weights, and the aim is to minimize the maximum weight of cut edges in addition to the weight of the last vertex before the cut. There is a known, polynomial time algorithm [Liu, SIAM J. Algebra. Discr., 1987] for the cases where the input graph is a rooted tree. We focus on the variant where the input graph is a directed series-parallel graph, and propose a polynomial time algorithm. Directed acyclic graphs are used to model scientific applications where the vertices correspond to the tasks of a given application and the edges represent the dependencies between the tasks. In such models, the problem we address reads as minimizing the peak memory requirement in an execution of the application. Our work, combined with Liu's work on rooted trees addresses this practical problem in two important classes of applications.

This work is described in a technical report [50].

7.6. Matrix symmetrization and sparse direct solvers

Participants: Raluca Portase [Cluj Napoca, Romania], Bora Uçar.

We investigate algorithms for finding column permutations of sparse matrices in order to have large diagonal entries and to have many entries symmetrically positioned around the diagonal. The aim is to improve the memory and running time requirements of a certain class of sparse direct solvers. We propose efficient algorithms for this purpose by combining two existing approaches and demonstrate the effect of our findings in practice using a direct solver. In particular, we show improvements in a number of components of the running time of a sparse direct solver with respect to the state of the art on a diverse set of matrices.

This work is described in a technical report [53].

7.7. Robust Memory-Aware Mapping for Parallel Multifrontal Factorizations

Participants: Emmanuel Agullo [HIEPACS project-team], Patrick Amestoy [INPT-IRIT], Alfredo Buttari [CNRS-IRIT], Abdou Guermouche [HIEPACS project-team], Jean-Yves L'Excellent, François-Henry Rouet [Lawrence Berkeley Laboratory, CA, USA].

In this work, we study the memory scalability of the parallel multifrontal factorization of sparse matrices. In particular, we are interested in controlling the active memory specific to the multifrontal factorization. We illustrate why commonly used mapping strategies (e.g., the proportional mapping) cannot provide a high memory efficiency, which means that they tend to let the memory usage of the factorization grow when the number of processes increases. We propose "memory-aware" algorithms that aim at maximizing the granularity of parallelism while respecting memory constraints. These algorithms provide accurate memory estimates prior to the factorization and can significantly enhance the robustness of a multifrontal code. We illustrate our approach with experiments performed on large matrices.

This work has been published in the *SIAM Journal on Scientific Computing* [1].

7.8. Fast 3D frequency-domain full waveform inversion with a parallel Block Low-Rank multifrontal direct solver: application to OBC data from the North Sea

Participants: Patrick Amestoy [INPT-IRIT], Romain Brossier [ISTerre], Alfredo Buttari [CNRS-IRIT], Jean-Yves L'Excellent, Théo Mary [UPS-IRIT], Ludovic Métivier [CNRS-ISTerre-LJK], Alain Miniussi [Geoazur], Stéphane Operto [Geoazur].

Wide-azimuth long-offset OBC/OBN surveys provide a suitable framework to perform computationally-efficient frequency-domain full waveform inversion (FWI) with a few discrete frequencies. Frequency-domain seismic modeling is performed efficiently with moderate computational resources for a large number of sources with a sparse multifrontal direct solver (Gauss-elimination techniques for sparse matrices). Approximate solutions of the time-harmonic wave equation are computed using a Block Low-Rank (BLR) approximation, leading to a significant reduction in the operation count and in the volume of communication during the LU factorization as well as offering a great potential for reduction in the memory demand. Moreover, the sparsity of the seismic source vectors is exploited to speed up the forward elimination step during the computation of the monochromatic wavefields. The relevance and the computational efficiency of the frequency-domain FWI performed in the visco-acoustic VTI approximation is shown with a real 3D OBC case study from the North Sea. The FWI subsurface models show a dramatic resolution improvement relative to the initial model built by reflection traveltime tomography. The amplitude errors introduced in the modeled wavefields by the BLR approximation for different low-rank thresholds have a negligible footprint in the FWI results. With respect to a standard multifrontal sparse direct factorization, and without compromise on the accuracy of the imaging, the BLR approximation can bring a reduction of the LU factor size by a factor up to three. This reduction is not yet exploited to reduce the effective memory usage (ongoing work). The flop reduction can be larger than a factor of 10 and can bring a factor of time reduction of around three. Moreover, this reduction factor tends to increase with frequency, namely with the matrix size. Frequency-domain visco-acoustic VTI FWI can be viewed as an efficient tool to build an initial model for elastic FWI of 4-C OBC data.

This work has been published in the journal *Geophysics* [2].

7.9. Matching-Based Allocation Strategies for Improving Data Locality of Map Tasks in MapReduce

Participant: Loris Marchal.

MapReduce is a well-know framework for distributing data-processing computations on parallel clusters. In MapReduce, a large computation is broken into small tasks that run in parallel on multiple machines, and scales easily to very large clusters of inexpensive commodity computers. Before the Map phase, the original dataset is first split into chunks, that are replicated (a constant number of times, usually 3) and distributed onto the computing nodes. During the Map phase, nodes request tasks and are allocated first tasks associated to local chunks (if any). Communications take place when requesting nodes do not hold any local chunk anymore. In this work, we provide the first complete theoretical data locality analysis of the Map phase of MapReduce, and more generally, for bag-of-tasks applications that behaves like MapReduce. We show that if tasks are homogeneous (in term of processing time), once the chunks have been replicated randomly on resources with a replication factor larger than 2, it is possible to find a priority mechanism for tasks that achieves a quasi-perfect number of communications using a sophisticated matching algorithm. In the more realistic case of heterogeneous processing times, we prove using an actual trace of a MapReduce server that this priority mechanism enables to complete the Map phase with significantly fewer communications, even on realistic distributions of task durations.

This work is described in a technical report [41].

7.10. Minimizing Rental Cost for Multiple Recipe Applications in the Cloud

Participant: Loris Marchal.

Clouds are more and more becoming a credible alternative to parallel dedicated resources. The pay-per-use pricing policy however highlights the real cost of computing applications. This new criterion, the cost, must then be assessed when scheduling an application in addition to more traditional ones as the completion time or the execution flow. In this work, we tackle the problem of optimizing the cost of renting computing instances to execute an application on the cloud while maintaining a desired performance (throughput). The target application is a stream application based on a DAG pattern, i.e., composed of several tasks with dependencies, and instances of the same execution task graph are continuously executed on the instances. We provide some theoretical results on the problem of optimizing the renting cost for a given throughput then propose some heuristics to solve the more complex parts of the problem, and we compare them to optimal solutions found by linear programming.

This work has been published in *IPDPS Workshops* [27].

7.11. Malleable task-graph scheduling with a practical speed-up model

Participants: Loris Marchal, Bertrand Simon, Oliver Sinnen [Univ. Auckland, New Zealand], Frédéric Vivien.

Scientific workloads are often described by Directed Acyclic task Graphs. Indeed, DAGs represent both a model frequently studied in theoretical literature and the structure employed by dynamic runtime schedulers to handle HPC applications. A natural problem is then to compute a makespan-minimizing schedule of a given graph. In this work, we are motivated by task graphs arising from multifrontal factorizations of sparsematrices and therefore work under the following practical model. We focus on malleable tasks (i.e., a single task can be allotted a time-varying number of processors) and specifically on a simple yet realistic speedup model: each task can be perfectly parallelized, but only up to a limited number of processors. We first prove that the associated decision problem of minimizing the makespan is NP-Complete. Then, we study a widely used algorithm, PropScheduling, under this practical model and propose a new strategy GreedyFilling. Even though both strategies are 2-approximations, experiments on real and synthetic data sets show that GreedyFilling achieves significantly lower makespans.

This work is described in a technical report [52].

7.12. Dynamic memory-aware task-tree scheduling

Participant: Loris Marchal.

Factorizing sparse matrices using direct multifrontal methods generates directed tree-shaped task graphs, where edges represent data dependency between tasks. This work revisits the execution of tree-shaped task graphs using multiple processors that share a bounded memory. A task can only be executed if all its input and output data can fit into the memory. The key difficulty is to manage the order of the task executions so that we can achieve high parallelism while staying below the memory bound. In particular, because input data of unprocessed tasks must be kept in memory, a bad scheduling strategy might compromise the termination of the algorithm. In the single processor case, solutions that are guaranteed to be below a memory bound are known. The multi-processor case (when one tries to minimize the total completion time) has been shown to be NP-complete. We designed in this work a novel heuristic solution that has a low complexity and is guaranteed to complete the tree within a given memory bound. We compared our algorithm to state of the art strategies, and observed that on both actual execution trees and synthetic trees, we always performed better than these solutions, with average speedups between 1.25 and 1.45 on actual assembly trees. Moreover, we showed that the overhead of our algorithm is negligible even on deep trees (10^5), and would allow its runtime execution.

This work is described in a technical report [39].

7.13. Optimal resilience patterns to cope with fail-stop and silent errors

Participants: Anne Benoit, Aurélien Cavelan, Yves Robert, Hongyang Sun.

This work focuses on resilience techniques at extreme scale. Many papers deal with fail-stop errors. Many others deal with silent errors (or silent data corruptions). But very few papers deal with fail-stop and silent errors simultaneously. However, HPC applications will obviously have to cope with both error sources. This work presents a unified framework and optimal algorithmic solutions to this double challenge. Silent errors are handled via verification mechanisms (either partially or fully accurate) and in-memory checkpoints. Fail-stop errors are processed via disk checkpoints. All verification and checkpoint types are combined into computational patterns. We provide a unified model, and a full characterization of the optimal pattern. Our results nicely extend several published solutions and demonstrate how to make use of different techniques to solve the double threat of fail-stop and silent errors. Extensive simulations based on real data confirm the accuracy of the model, and show that patterns that combine all resilience mechanisms are required to provide acceptable overheads.

This work was presented at the *IPDPS'2016* conference [20].

7.14. Two-level checkpointing and partial verifications for linear task graphs

Participants: Anne Benoit, Aurélien Cavelan, Yves Robert, Hongyang Sun.

Fail-stop and silent errors are unavoidable on large-scale platforms. Efficient resilience techniques must accommodate both error sources. A traditional checkpointing and rollback recovery approach can be used, with added verifications to detect silent errors. A fail-stop error leads to the loss of the whole memory content, hence the obligation to checkpoint on a stable storage (e.g., an external disk). On the contrary, it is possible to use in-memory checkpoints for silent errors, which provide a much smaller checkpoint and recovery overhead. Furthermore, recent detectors offer partial verification mechanisms, which are less costly than guaranteed verifications but do not detect all silent errors. In this work, we show how to combine all these techniques for HPC applications whose dependence graph is a chain of tasks, and provide a sophisticated dynamic programming algorithm returning the optimal solution in polynomial time. Simulations demonstrate that the combined use of multi-level checkpointing and partial verifications further improves performance.

This work was presented at the *17th IEEE International Workshop on Parallel and Distributed Scientific and Engineering Computing (PDSEC 2016)* [21].

7.15. Resilient application co-scheduling with processor redistribution

Participants: Anne Benoit, Loïc Pottier, Yves Robert.

Recently, the benefits of co-scheduling several applications have been demonstrated in a fault-free context, both in terms of performance and energy savings. However, large-scale computer systems are confronted to frequent failures, and resilience techniques must be employed to ensure the completion of large applications. Indeed, failures may create severe imbalance between applications, and significantly degrade performance. In this work, we propose to redistribute the resources assigned to each application upon the striking of failures, in order to minimize the expected completion time of a set of co-scheduled applications. First we introduce a formal model and establish complexity results. When no redistribution is allowed, we can minimize the expected completion time in polynomial time, while the problem becomes NP-complete with redistributions, even in a fault-free context. Therefore, we design polynomial-time heuristics that perform redistributions and account for processor failures. A fault simulator is used to perform extensive simulations that demonstrate the usefulness of redistribution and the performance of the proposed heuristics.

This work was presented at the *ICCP'16* conference [22].

7.16. A different re-execution speed can help

Participants: Anne Benoit, Aurélien Cavelan, Valentin Le Fèvre, Yves Robert, Hongyang Sun.

We consider divisible load scientific applications executing on large-scale platforms subject to silent errors. While the goal is usually to complete the execution as fast as possible in expectation, another major concern is energy consumption. The use of dynamic voltage and frequency scaling (DVFS) can help save energy, but at the price of performance degradation. Consider the execution model where a set of K different speeds is given, and whenever a failure occurs, a different re-execution speed may be used. Can this help? We address the following bi-criteria problem: how to compute the optimal checkpointing period to minimize energy consumption while bounding the degradation in performance. We solve this bi-criteria problem by providing a closed-form solution for the checkpointing period, and demonstrate via a comprehensive set of simulations that a different re-execution speed can indeed help.

This work was presented at the *5th International Workshop on Power-aware Algorithms, Systems, and Architectures* [19].

7.17. Coping with recall and precision of soft error detectors

Participants: Anne Benoit, Aurélien Cavelan, Yves Robert, Hongyang Sun.

Many methods are available to detect silent errors in high-performance computing (HPC) applications. Each method comes with a cost, a recall (fraction of all errors that are actually detected, i.e., false negatives), and a precision (fraction of true errors amongst all detected errors, i.e., false positives). The main contribution of this work is to characterize the optimal computing pattern for an application: which detector(s) to use, how many detectors of each type to use, together with the length of the work segment that precedes each of them. We first prove that detectors with imperfect precisions offer limited usefulness. Then we focus on detectors with perfect precision, and we conduct a comprehensive complexity analysis of this optimization problem, showing NP-completeness and designing an FPTAS (Fully Polynomial-Time Approximation Scheme). On the practical side, we provide a greedy algorithm, whose performance is shown to be close to the optimal for a realistic set of evaluation scenarios. Extensive simulations illustrate the usefulness of detectors with false negatives, which are available at a lower cost than the guaranteed detectors.

This work was accepted for publication in the *Journal of Parallel and Distributed Computing* [7].

7.18. Checkpointing strategies for scheduling computational workflows

Participants: Anne Benoit, Yves Robert.

We study the scheduling of computational workflows on compute resources that experience exponentially distributed failures. When a failure occurs, rollback and recovery is used to resume the execution from the last checkpointed state. The scheduling problem is to minimize the expected execution time by deciding in which order to execute the tasks in the workflow and deciding for each task whether to checkpoint it or not after it completes. We give a polynomial-time optimal algorithm for fork DAGs (Directed Acyclic Graphs) and show that the problem is NP-complete with join DAGs. We also investigate the complexity of the simple case in which no task is checkpointed. Our main result is a polynomial-time algorithm to compute the expected execution time of a workflow, with a given task execution order and specified to-be-checkpointed tasks. Using this algorithm as a basis, we propose several heuristics for solving the scheduling problem. We evaluate these heuristics for representative workflow configurations.

This work was published in the *International Journal of Networking and Computing* [4].

7.19. Assessing General-Purpose Algorithms to Cope with Fail-Stop and Silent Errors

Participants: Anne Benoit, Aurélien Cavelan, Yves Robert, Hongyang Sun.

We combine the traditional checkpointing and rollback recovery strategies with verification mechanisms to cope with both fail-stop and silent errors. The objective is to minimize makespan and/or energy consumption. For divisible load applications, we use first-order approximations to find the optimal checkpointing period to minimize execution time, with an additional verification mechanism to detect silent errors before each checkpoint, hence extending the classical formula by Young and Daly for fail-stop errors only. We further extend the approach to include intermediate verifications, and to consider a bi-criteria problem involving both time and energy (linear combination of execution time and energy consumption). Then, we focus on application workflows whose dependence graph is a linear chain of tasks. Here, we determine the optimal checkpointing and verification locations, with or without intermediate verifications, for the bi-criteria problem. Rather than using a single speed during the whole execution, we further introduce a new execution scenario, which allows for changing the execution speed via dynamic voltage and frequency scaling (DVFS). In this latter scenario, we determine the optimal checkpointing and verification locations, as well as the optimal speed pairs for each task segment between any two consecutive checkpoints. Finally, we conduct an extensive set of simulations to support the theoretical study, and to assess the performance of each algorithm, showing that the best overall performance is achieved under the most flexible scenario using intermediate verifications and different speeds.

This work was accepted for publication in the journal *ACM Transactions on Parallel Computing* [8].

7.20. A failure detector for HPC platforms

Participant: Yves Robert.

Building an infrastructure for Exascale applications requires, in addition to many other key components, a stable and efficient failure detector. This work describes the design and evaluation of a robust failure detector, able to maintain and distribute the correct list of alive resources within proven and scalable bounds. The detection and distribution of the fault information follow different overlay topologies that together guarantee minimal disturbance to the applications. A virtual observation ring minimizes the overhead by allowing each node to be observed by another single node, providing an unobtrusive behavior. The propagation stage is using a non-uniform variant of a reliable broadcast over a circulant graph overlay network, and guarantees a logarithmic fault propagation. Extensive simulations, together with experiments on the Titan ORNL supercomputer, show that the algorithm performs extremely well, and exhibits all the desired properties of an Exascale-ready algorithm.

This work was presented at the *SC'16* conference [24].

7.21. Optimal multistage algorithm for adjoint computatio

Participant: Yves Robert.

We reexamine the work of Stumm and Walther on multistage algorithms for adjoint computation. We provide an optimal algorithm for this problem when there are two levels of checkpoints, in memory and on disk. Previously, optimal algorithms for adjoint computations were known only for a single level of checkpoints with no writing and reading costs; a well-known example is the binomial checkpointing algorithm of Griewank and Walther. Stumm and Walther extended that binomial checkpointing algorithm to the case of two levels of checkpoints, but they did not provide any optimality results. We bridge the gap by designing the first optimal algorithm in this context. We experimentally compare our optimal algorithm with that of Stumm and Walther to assess the difference in performance.

This work was accepted for publication in the *SIAM Journal on Scientific Computing* [5].

7.22. Assessing the cost of redistribution followed by a computational kernel: Complexity and performance results

Participant: Yves Robert.

The classical redistribution problem aims at optimally scheduling communications when reshuffling from an initial data distribution to a target data distribution. This target data distribution is usually chosen to optimize some objective for the algorithmic kernel under study (good computational balance or low communication volume or cost), and therefore to provide high efficiency for that kernel. However, the choice of a distribution minimizing the target objective is not unique. This leads to generalizing the redistribution problem as follows: find a re-mapping of data items onto processors such that the data redistribution cost is minimal, and the operation remains as efficient. This work studies the complexity of this generalized problem. We compute optimal solutions and evaluate, through simulations, their gain over classical redistribution. We also show the NP-hardness of the problem to find the optimal data partition and processor permutation (defined by new subsets) that minimize the cost of redistribution followed by a simple computational kernel. Finally, experimental validation of the new redistribution algorithms are conducted on a multicore cluster, for both a 1D-stencil kernel and a more compute-intensive dense linear algebra routine.

This work has been published in the Parallel Computing journal [14].

7.23. When Amdahl Meets Young/Daly

Participants: Aurélien Cavelan, Yves Robert.

This work investigates the optimal number of processors to execute a parallel job, whose speedup profile obeys Amdahl's law, on a large-scale platform subject to fail-stop and silent errors. We combine the traditional checkpointing and rollback recovery strategies with verification mechanisms to cope with both error sources. We provide an exact formula to express the execution overhead incurred by a periodic checkpointing pattern of length T and with P processors, and we give first-order approximations for the optimal values T* and P* as a function of the individual processor MTBF. A striking result is that P* is of the order of the fourth root of the individual MTBF if the checkpointing cost grows linearly with the number of processors, and of the order of its third root if the checkpointing cost stays bounded for any P. We conduct an extensive set of simulations to support the theoretical study. The results confirm the accuracy of first-order approximation under a wide range of parameter settings.

This work was presented at the *Cluster'16* conference [26].

7.24. Computing the expected makespan of task graphs in the presence of silent errors

Participants: Julien Herrmann, Yves Robert.

Applications structured as Directed Acyclic Graphs (DAGs) of tasks correspond to a general model of parallel computation that occurs in many domains, including popular scientific workflows. DAG scheduling has received an enormous amount of attention, and several list-scheduling heuristics have been proposed and shown to be effective in practice. Many of these heuristics make scheduling decisions based on path lengths in the DAG. At large scale, however, compute platforms and thus tasks are subject to various types of failures with no longer negligible probabilities of occurrence. Failures that have recently received increasing attention are silent errors, which cause a task to produce incorrect results even though it ran to completion. Tolerating silent errors is done by checking the validity of the results and re-executing the task from scratch in case of an invalid result. The execution time of a task then becomes a random variable, and so are path lengths. Unfortunately, computing the expected makespan of a DAG (and equivalently computing expected path lengths in a DAG) is a computationally difficult problem. Consequently, designing effective scheduling heuristics is preconditioned on computing accurate approximations of the expected makespan. In this work we propose an algorithm that computes a first order approximation of the expected makespan of a DAG when tasks are subject to silent errors. We compare our proposed approximation to previously proposed such approximations for three classes of application graphs from the field of numerical linear algebra. Our evaluations quantify approximation error with respect to a ground truth computed via a brute-force Monte Carlo method. We find that our proposed approximation outperforms previously proposed approaches, leading to large reductions in approximation error for low (and realistic) failure rates, while executing much faster.

This work was presented at the *Ninth Int. Workshop on Parallel Programming Models and Systems Software for High-End Computing (P2S2)* [25].

7.25. Toward an Optimal Online Checkpoint Solution under a Two-Level HPC Checkpoint Model

Participants: Yves Robert, Frédéric Vivien.

The traditional single-level checkpointing method suffers from significant overhead on large-scale platforms. Hence, multilevel checkpointing protocols have been studied extensively in recent years. The multilevel checkpoint approach allows different levels of checkpoints to be set (each with different checkpoint overheads and recovery abilities), in order to further improve the fault tolerance performance of extreme-scale HPC applications. How to optimize the checkpoint intervals for each level, however, is an extremely difficult problem. In this work, we construct an easy-to-use two-level checkpoint model. Checkpoint level 1 deals with errors with low checkpoint/recovery overheads such as transient memory errors, while checkpoint level 2 deals with hardware crashes such as node failures. Compared with previous optimization work, our new optimal checkpoint solution offers two improvements: (1) it is an online solution without requiring knowledge of the job length in advance, and (2) it shows that periodic patterns are optimal and determines the best pattern. We evaluate the proposed solution and compare it with the most up-to-date related approaches on an extreme-scale simulation testbed constructed based on a real HPC application execution. Simulation results show that our proposed solution outperforms other optimized solutions and can improve the performance significantly in some cases. Specifically, with the new solution the wall-clock time can be reduced by up to 25.3% over that of other state-of-the-art approaches. Finally, a brute-force comparison with all possible patterns shows that our solution is always within 1% of the best pattern in the experiments.

This work has been published in IEEE Transactions on Parallel and Distributed Systems [11].

7.26. Cell morphing: from array programs to array-free Horn clauses

Participants: Laure Gonnord, David Monniaux [(CNRS/Verimag)], Julien Braine [(M2 Student)].

Automatically verifying safety properties of programs is hard. Many approaches exist for verifying programs operating on Boolean and integer values (e.g. abstract interpretation, counterexample-guided abstraction refinement using interpolants), but transposing them to array properties has been fraught with difficulties. Our work addresses that issue with a powerful and flexible abstraction that morphes concrete array cells into a finite set of abstract ones. This abstraction is parametric both in precision and in the back-end analysis used. From

our programs with arrays, we generate nonlinear Horn clauses over scalar variables only, in a common format with clear and unambiguous logical semantics, for which there exist several solvers. We thus avoid the use of solvers operating over arrays, which are still very immature. Experiments with our prototype VAPHOR show that this approach can prove automatically and without user annotations the functional correctness of several classical examples, including *selection sort*, *bubble sort*, *insertion sort*, as well as examples from literature on array analysis.

This work has been published in Static Analysis Symposium [30] for the array part. We are currently deseigning an extension to programs with inductive data structures.

7.27. Symbolic Analyses of pointers

Participants: Laure Gonnord, Maroua Maalej, Fernando Pereira [(UFMG, Brasil)], Leonardo Barbosa [(UFMG, Brasil)], Vitor Paisante [(UFMG, Brasil)], Pedro Ramos [(UFMG, Brasil)].

Alias analysis is one of the most fundamental techniques that compilers use to optimize languages with pointers. However, in spite of all the attention that this topic has received, the current state-of-the-art approaches inside compilers still face challenges regarding precision and speed. In particular, pointer arithmetic, a key feature in C and C++, is yet to be handled satisfactorily.

A first work presents a new range-based alias analysis algorithm to solve this problem. The key insight of our approach is to combine alias analysis with symbolic range analysis. This combination lets us disambiguate fields within arrays and structs, effectively achieving more precision than traditional algorithms. To validate our technique, we have implemented it on top of the LLVM compiler. Tests on a vast suite of benchmarks show that we can disambiguate several kinds of C idioms that current state-of-the-art analyses cannot deal with. In particular, we can disambiguate 1.35x more queries than the alias analysis currently available in LLVM. Furthermore, our analysis is very fast: we can go over one million assembly instructions in 10 seconds.

A second work starts from an obvious, yet unexplored, observation: if a pointer is strictly less than another, they cannot alias. Motivated by this remark, we use the abstract interpretation framework to build strict less-than relations between pointers. To this end, we construct a program representation that bestows the Static Single Information (SSI) property onto our dataflow analysis. SSI gives us an efficient sparse algorithm, which, once seen as a form of abstract interpretation, is correct by construction. We have implemented our static analysis in LLVM. It runs in time linear on the number of program variables, and, depending on the benchmark, it can be as much as six times more precise than the pointer disambiguation techniques already in place in that compiler.

This work has been published in the *International Symposium of Code Generation and Optmization* [31] and at CGO'17 [29].

7.28. High-Level Synthesis of Pipelined FSM from Loop Nests

Participants: Christophe Alias, Fabrice Rastello [(Inria/CORSE)], Alexandru Plesco [(XtremLogic SAS, France)].

Embedded systems raise many challenges in power, space and speed efficiency. The current trend is to build heterogeneous systems on a chip with specialized processors and hardware accelerators. Generating an hardware accelerator from a computational kernel requires a deep reorganization of the code and the data. Typically, parallelism and memory bandwidth are met thanks to fine-grain loop transformations. Unfortunately, the resulting control automaton is often very complex and eventually bound the circuit frequency, which limits the benefits of the optimization. This is a major lock, which strongly limits the power of the code optimizations appliable by high-level synthesis tools.

In this work, we propose an architecture of control automaton and an algorithm of high-level synthesis which translates efficiently the control required by fine-grain loop optimizations. Unlike the previous approaches, our control automaton can be pipelined *at will, without any restriction*. Hence, the frequency of the automaton can be as high as possible. Experimental results on FPGA confirms that our control circuit can reach a high frequency with a reasonable resource consumption.

This work is described in a technical report [36].

7.29. Estimation of Parallel Complexity with Rewriting Techniques

Participants: Christophe Alias, Laure Gonnord, Carsten Fuhs [(Birbeck, UK)].

We show how monotone interpretations - a termination analysis technique for term rewriting systems - can be used to assess the inherent parallelism of recursive programs manipulating inductive data structures. As a side effect, we show how monotone interpretations specify a parallel execution order, and how our approach extends naturally affine scheduling - a powerful analysis used in parallelising compilers - to recursive programs. This preliminary work opens new perspectives in automatic parallelisation.

This work has been published in the *Workshop on Termination,* [15].

SOCRATE Project-Team

6. New Results

6.1. Flexible Radio Front-End

6.1.1. Wake-Up Radio

The last decades have been really hungry in new ways to reduce energy consumption. That is especially true when talking about wireless sensor networks in general and home multimedia networks in particular, since electrical energy consumption is the bottleneck of the network. One of the most energy-consuming functional block of an equipment is the radio front end, and methods to switch it off during the time intervals where it is not active must be implemented. This previous study has proposed a wake-up radio circuit which is capable of both addressing and waking up not only a more efficient but also more energy-consuming radio front end. By using a frequency footprint to differentiate each sensor, awaking all the sensors except for the one of interest is avoided. The particularity of the proposed wake-up receiver [22] is that the decision is taken in the radio-frequency part and no baseband treatment is needed. The global evaluation in theory and in simulation was performed, and a first testbed of this technology was fabricated, demonstrating that this principle actually works in practice [21].

6.1.1.1. Full-Duplex

An important work was done in this axis previously around Full-Duplex systems, in order to enhance throughput, flexibility, and, potentially security of wireless links. A PhD thesis grant from DGA and Inria has allowed us to extend this through a collaboration with axis 2, focusing on Physical layer security mechanisms based on Full-Duplex systems. Starting by a theoretical study of the secrecy capacity in the presence of an eavesdropper, this work studies [13] the duality between wiretap channels and state-dependent channels. This represents a basic framework to extend in a near future this study to Full-Duplex scenarios, where the Full-Duplex capability of a node could increase the secrecy of the wireless communication.

6.1.1.2. SDR for SRDs

The technologies employed in urban sensor networks are permanently evolving, and thus the gateways of these networks have to be regularly upgraded. The existing method to do so is to stack-up receivers dedicated to one communication protocol. However, this implies to have to replace the gateway every time a new protocol is added to the network. A more practical way to do this is to perform a digitization of the full band and to perform digitally the signal processing, as done in Software-Defined Radio (SDR). The main hard point in doing this is the dynamic range of the signals: indeed the signals are emitted with very different features because of the various propagation conditions. It has been proved that the difference of power between two signals can be so important that no existing Analog-to-Digital Converter (ADC) is able to properly digitize the signals. We propose a solution to reduce the dynamic range of signals before digital conversion. In this study [9], the assumption is made that there is one strong signal, and several weak signals. This assumption is made from the existing urban sensor networks topology. A receiver architecture with two branches is proposed with a “Coarse Digitization Path” (CDP) and a “Fine Digitization Path” (FDP). The CDP allows to digitize the strong signal and to get data on it that is used to reconfigure the FDP. The FDP then uses a notch filter to attenuate the strong signal (and then to reduce the dynamic range of the signals) and digitizes the rest of the band.

6.2. Multi-User Communications

6.2.1. Fundamental Limits

6.2.1.1. Approximate Capacity Region of the Gaussian Interference Channel with Feedback

An achievability region and a converse region for the two-user Gaussian interference channel with noisy channel-output feedback (G-IC-NOF) are presented [42], [30], [43], [47]. The achievability region is obtained using a random coding argument and three well-known techniques: rate splitting, superposition coding and backward decoding. The converse region is obtained using some of the existing perfect-output feedback outer-bounds as well as a set of new outer-bounds that are obtained by using genie-aided models of the original G-IC-NOF. Finally, it is shown that the achievability region and the converse region approximate the capacity region of the G-IC-NOF to within a constant gap in bits per channel use.

6.2.1.2. Full Characterization of the Capacity Region of the Linear Deterministic Interference Channel with Feedback

The capacity region of the two-user linear deterministic (LD) interference channel with noisy output feedback (IC-NOF) has been fully characterized [29]. This result allows the identification of several asymmetric scenarios in which implementing channel-output feedback in only one of the transmitter-receiver pairs is as beneficial as implementing it in both links, in terms of achievable individual rate and sum-rate improvements w.r.t. the case without feedback. In other scenarios, the use of channel-output feedback in any of the transmitter-receiver pairs benefits only one of the two pairs in terms of achievable individual rate improvements or simply, it turns out to be useless, i.e., the capacity regions with and without feedback turn out to be identical even in the full absence of noise in the feedback links.

6.2.1.3. Full Characterization of the Information Equilibrium Region of the Multiple Access Channel

The fundamental limits of decentralized information transmission in the K-user Gaussian multiple access channel (G-MAC), with $K \geqslant 2$, are fully characterized [38]. Two scenarios are considered. First, a game in which only the transmitters are players is studied. In this game, the transmitters autonomously and independently tune their own transmit configurations seeking to maximize their own information transmission rates, $R_1, R_2, \cdots, R_K$, respectively. On the other hand, the receiver adopts a fixed receive configuration that is known a priori to the transmitters. The main result consists of the full characterization of the set of rate tuples $(R_1, R_2, ..., R_K)$ that are achievable and stable in the G-MAC when stability is considered in the sense of the η-Nash equilibrium (NE), with $\eta > 0$ arbitrarily small. Second, a sequential game in which the two categories of players (the transmitters and the receiver) play in a given order is presented. For this sequential game, the main result consists of the full characterization of the set of rate tuples $(R_1, R_2, ..., R_K)$ that are stable in the sense of an η-sequential equilibrium, with $\eta > 0$.

6.2.1.4. Full Characterization of the Information-Energy Capacity Region of the Multiple Access Channel with Energy Harvester with and without Feedback

The fundamental limits of simultaneous information and energy transmission in the two-user Gaussian multiple access channel (G-MAC) with and without feedback have been fully characterized [10], [15]. More specifically, all the achievable information and energy transmission rates (in bits per channel use and energy-units per channel use, respectively) are identified. In the case without feedback, an achievability scheme based on power-splitting and successive interference cancelation is shown to be optimal. Alternatively, in the case with feedback (G-MAC-F), a simple yet optimal achievability scheme based on power-splitting and Ozarow's capacity achieving scheme is presented. Two of the most important observations in this work are: (a) The information-energy capacity region of the G-MAC without feedback can be a proper subset of the information-energy capacity region of the G-MAC-F and (b) Feedback can at most double the energy rate when the information transmission rate is kept fixed at the sum-capacity of the G-MAC.

6.2.1.5. Full Characterization of the Information-Energy Equilibrium Region of the Multiple Access Channel with Energy Harvester

The fundamental limits of decentralized simultaneous information and energy transmission in the two-user Gaussian multiple access channel (G-MAC) have been fully characterized for the case in which a minimum energy transmission rate b is required for successful decoding [14], [39]. All the achievable and stable information-energy transmission rate triplets (R_1, R_2, B) are identified. R_1 and R_2 are in bits per channel use measured at the receiver and B is in energy units per channel use measured at an energy-harvester (EH). Stability is considered in the sense of an η-Nash equilibrium (NE), with $\eta > 0$ arbitrarily small. The main

result consists of the full characterization of the η-NE information-energy region, i.e., the set of information-energy rate triplets (R_1, R_2, B) that are achievable and stable in the G-MAC when: (a) both transmitters autonomously and independently tune their own transmit configurations seeking to maximize their own information transmission rates, R_1 and R_2 respectively; (b) both transmitters jointly guarantee an energy transmission rate B at the EH, such that $B > b$. Therefore, any rate triplet outside the η-NE region is not stable as there always exists one transmitter able to increase by at least η bits per channel use its own information transmission rate by updating its own transmit configuration.

6.2.1.6. Duality Between State-Dependent Channels and Wiretap Channels

A duality between wiretap and state-dependent channels with non-causal channel state information at the transmitter has been established [13]. First, a common achievable scheme is described for a certain class of state-dependent and wiretap channels. Further, state-dependent and wiretap channels for which this scheme is capacity (resp. secrecy capacity) achieving are identified. These channels are said to be dual. This duality is used to establish the secrecy capacity of certain state-dependent wiretap channels with non-causal channel state information at the transmitter. Interestingly, combatting the eavesdropper or combatting the lack of state information at the receiver turn out to be two non-concurrent tasks.

6.2.1.7. Energy efficiency - Spectral Efficiency (EE-SE) Tradeoffs in Wireless RANs

Even for a point-to-point communication, the Shannon capacity can be interpreted for a Gaussian channel as a fundamental spectral and energy efficiency (SE-EE) trade-off. Extending this fundamental trade-off in the context of multi-user communications is not straightforward as it may depend on many parameters. We proposed in [8] a simple and effective method to study this trade-off in cellular networks, an issue that has attracted significant recent interest in the wireless community. The proposed theoretical framework is based on an optimal radio resource allocation of transmit power and bandwidth for the downlink direction, applicable for an orthogonal cellular network. The analysis is initially focused on a single cell scenario, for which in addition to the solution of the main SE-EE optimization problem, it is proved that a traffic repartition scheme can also be adopted as a way to simplify this approach. By exploiting this interesting result along with properties of stochastic geometry, this work is extended to a more challenging multi-cell environment, where interference is shown to play an essential role and for this reason several interference reduction techniques are investigated. Special attention is also given to the case of low signal to noise ratio (SNR) and a way to evaluate the upper bound of EE in this regime is provided. This methodology leads to tractable analytical results under certain common channel properties, and thus allows the study of various models without the need for demanding system level simulations.

6.2.1.8. Spatial Continuum Channel Models

In the context of the deployment of Internet of Things (see next section for more details about our protocol developments), it is expected that a unique cell could serve millions of radio nodes transmitting sporadic short packets. In [18] and [41], our objective is to study this problem from an information theory point of view to derive the fundamental limit in terms of maximal information rates that can be transmitted in such a dense cell. This work proposes a new model called spatial continuum asymmetric channels to study the channel capacity region of asymmetric scenarios in which either one source transmits to a spatial density of receivers or a density of transmitters transmit to a unique receiver. This approach is built upon the classical broadcast channel (BC) and multiple access channel (MAC). For the sake of consistency, the study is limited to Gaussian channels with power constraints and is restricted to the asymptotic regime (zero-error capacity). The reference scenario comprises one base station in Tx or Rx mode, a spatial random distribution of nodes (resp. in Rx or Tx mode) characterized by a probability spatial density of users $u(x)$ where each of them requests a quantity of information with no delay constraint, thus leading to a requested rate spatial density $\rho(x)$. This system is modeled as an ‚àû‚àíuser asymmetric channel (BC or MAC). To derive the fundamental limits of this model, a spatial discretization is first proposed to obtain an equivalent BC or MAC. Then, a specific sequence of discretized spaces is defined to refine infinitely the approximation. Achievability and capacity results are obtained in the limit of this sequence while the access capacity region $\mathcal{C}(Pm)$ is defined as the set of requested rates spatial densities $\rho(x)$ that are achievable with a transmission power Pm. The uniform capacity defined as the maximal symmetric achievable rate is also computed.

6.2.1.9. Finite Block-Length Coding in Wireless Networks

In the context of IoT, the information to be transmitted will be divided in very small packets especially when control and commands will be transmitted over the network. The classical asymptotic information theory relies on the statistic properties of channels and information sources, when the coding block-length tends to infinity. Therefore this framework is not appropriate to study the fundamental limits of short packets transmission over wireless networks. Fortunately, information theory is not only about the asymptotic regime. Shannon himself derived the preliminary foundations of a theory for finite block-length. Later, Gallager extended this framework. Recently this question gained interest after the work of Y. Polyanskiy which extended former results on finite block length to Gaussian channels. This fundamental contribution opens a way for studying wireless networks under finite block-length regime. But this relatively new paradigm suffers from strong problems relative to the compexity of the underlying estimation problem. Starting to work on this topic in the framework of the associated team with Princeton, we exploited in [35] the recent results on the non-asymptotic coding rate for fading channels with no channel state information at the transmitter and we analyzed the goodput in additive white Gaussian noise (AWGN) and the energy-efficiency spectral-efficiency (EE-SE) tradeoff where the fundamental relationship between the codeword length and the EE is given. Finally, the true outage probability in Ricean and Nakagami-m block fading channels is investigated and it is proved that the asymptotic outage capacity is the Laplace approximation of the average error probability in finite blocklength regime. This preliminary work constitutes one of the starting point for our future works in the framework of the ANR project ARBURST.

6.2.2. Algorithm and Protocol Design for Multi-User Communication Scenarios

6.2.2.1. Interference Management in OFDM/MIMO Wireless Networks

Modern cellular networks in traditional frequency bands are notoriously interference-limited especially in urban areas, where base stations are deployed in close proximity to one another. The latest releases of Long Term Evolution (LTE) incorporate features for coordinating downlink transmissions as an efficient means of managing interference. In [4], we review recent field trial results and theoretical studies of the performance of joint transmission (JT) coordinated multi-point (CoMP) schemes. These schemes revealed, however, that their gains are not as high as initially expected, despite the large coordination overhead. These schemes are known to be very sensitive to defects in synchronization or information exchange between coordinating bases stations as well as uncoordinated interference. In this article, we review recent advanced coordinated beamforming (CB) schemes as alternatives, requiring less overhead than JT CoMP while achieving good performance in realistic conditions. By stipulating that, in certain LTE scenarios of increasing interest, uncoordinated interference constitutes a major factor in the performance of CoMP techniques at large, we hereby assess the resilience of the state-of-the-art CB to uncoordinated interference. We also describe how these techniques can leverage the latest specifications of current cellular networks, and how they may perform when we consider standardized feedback and coordination. This allows us to identify some key roadblocks and research directions to address as LTE evolves towards the future of mobile communications.

Among the different techniques described above, we studied in [32] an interference Alignment (IA) technique that, in a large sense, makes use of the increasing signal dimensions available in the system through MIMO and OFDM technologies in order to globally reduce the interference suffered by users in a network. In this paper, we addressed the problem of downlink cellular networks, the so-called interfering broadcast channels, where mobile users at cell edges may suffer from high interference and thus, poor performance. Starting from the downlink IA scheme proposed by Suh et al., a new approach is proposed where each user feeds back multiple selected received signal directions with high signal-to-interference gain. A exhaustive search based scheduler selects a subset of users to be served simultaneously, balancing between sum-rate performance and fairness, but becomes untractable in dense network scenarios where many users send simultaneous requests. Therefore, we develop a sub-optimal scheduler that greatly decreases the complexity while preserving a near-optimal data rate gain. More interestingly, our simulations show that the IA scheme becomes valuable only in correlated channels, whereas the matched filtering based scheme performs the best in the uncorrelated scenarios.

6.2.2.2. Performance of Ultra-NarrowBand Techniques for Internet of Things

This section makes echo to the section entitled Spatial Continuum Channel Models where fundamental limits are studied for a similar scenario. In this section, we investigate the scenario for an existing PHY layer technology, Ultra Narrow Band (UNB) technique, proposer by Sigfox. The ALOHA protocol is regaining interest in the context of the Internet of Things (IoT), especially for UNB signals (dedicated to long range and low power transmission in IoT networks). In this case, the classical assumption of channelization is not verified anymore, modifying the ALOHA performances. Indeed, UNB signals suffer from a lack of precision on the actual transmission carrier frequency, leading to a behavior similar to a frequency unslotted random access. More precisely, the channel access is Random-FTMA, where nodes select their time and frequency in a random and continuous way. The frequency randomness prevents from allocating orthogonal resources for transmission, and induces uncontrolled interference.

In [19], the success probability and throughput of ALOHA is generalized to further describe frequency-unslotted systems such as UNB. The main contribution of this work is the derivation of a generalized expression of the throughput for the random time-frequency ALOHA systems, when neglecting channel attenuation. Besides, this study permits to highlight the duality of ALOHA in time and frequency domain.

Besides, in [26] and [27], to introduce diversity, we propose the use of replication mechanism to enhance the reliability of UNB wireless network. Considering the outage probability, we theoretically evaluate the system performance and show that there exists an optimal number of transmissions. Finally, we highlight that this number of repetitions can be easily optimized by considering a unique global parameter.

Finally, in [28], we also take into consideration the channel effect for such specific network. Indeed, the UNB randomness leads to a new behavior of the interference which has not been theoretically analyzed yet, when considering the pathloss of nodes located randomly in an area. In this work, in order to quantify the system performance, we derive and exploit a theoretical expression of the packet error rate in a UNB based IoT network, when taking into account both interference due to the spectral randomness and path loss due to the propagation.

6.2.2.3. Algorithms and Protocols for BANs

Body Area Networks (BANs) represent a challenging area of research for networking design. Indeed, the topology of these networks differe significantly from classical networks. BANs are dynamic, multi-scale, energy limited and require real time protocols for many applications related to localization. Our work is related to the design of dynamic protocols to gather and exploit localization information in dynamic BANs. Our first contribution is related to the context of group navigation and was developed in the framework of the FUI SMACS project dealing with the localisation of runners during bike races. The problem is to develop fast and reliable protocols to dynamically gather mobility information from moving nodes toward moving sinks.

Our second contribution is relative to the mobility of a single BAN and with the objective of improving localization algorithms based on ranging measures between nodes spread on the body. This work was done in the framework of the ANR CORMORAN project with the PhD of Arturo Gimenez-Guizar who defended his PhD in October 2016 [1].

6.2.2.3.1. Information Gathering in a Group of Mobile Users

In [16], we propose an efficient approach to collect data in mobile wireless sensor networks, with the specific application of sensing in bike races. Recent sensor technology permits to track GPS position of each bike. Because of the inherent correlation between bike positions in a bike race, a simple GPS log is inefficient. The idea presented in this work is to aggregate GPS data at sensors using compressive sensing techniques. We enforce, in addition to signal sparsity, a spatial prior on biker motion because of the group behaviour (peloton) in bike races. The spatial prior is modeled by a graphical model and the data aggregation problem is solved, with both the sparsity and the spatial prior, by belief propagation. We validate our approach on a bike race simulator using trajectories of motorbikes in a real bike race.

6.2.2.3.2. MAC Protocols and Algorithms for Localization at the Body Scale

In this work [20], we have considered the positioning success rate for localization applications deployed in Wireless Body Area Networks (WBAN). Localization is performed with Ultra Wide Band (UWB) pulses, which permits to estimate distances as defined by 3 Way Ranging protocol (3WR). Two channels are considered : the empirical channel CM3, and with our model obtained from our measurement campaign. We first evaluate the positioning loss when considering an aggregation and broadcast scheduling strategy (A&B) upon TDMA MAC. We highlight the channel effects depending on the targeted receiver sensitivity. We then improve the performances by proposing a cooperative algorithm based on conditional permutation of anchors.

6.2.3. Cyber-Physical Systems

6.2.3.1. Attacks in the Electricity Grids

Multiple attacker data injection attack construction in electricity grids with minimum-mean-square-error state estimation has been studied for centralized and decentralized scenarios [6], [11]. A performance analysis of the trade-off between the maximum distortion that an attack can introduce and the probability of the attack being detected by the network operator is considered. In this setting, optimal centralized attack construction strategies are studied. The decentralized case is examined in a game-theoretic setting. A novel utility function is proposed to model this trade-off and it is shown that the resulting game is a potential game. The existence and cardinality of the corresponding set of Nash Equilibria (NEs) of the game is analyzed. Interestingly, the attackers can exploit the correlation among the state variables to facilitate the attack construction. It is shown that attackers can agree on a data injection vector construction that achieves the best trade-off between distortion and detection probability by sharing only a limited number of bits offlline. For the particular case of two attackers, numerical results based on IEEE test systems are presented.

6.2.3.2. Recovering Missing Data in Electricity Grids

The performance of matrix completion based recovery of missing data in electricity distribution systems has been analyzed [17]. Under the assumption that the state variables follow a multivariate Gaussian distribution the matrix completion recovery is compared to estimation and information theoretic limits. The assumption about the distribution of the state variables is validated by the data shared by Electricity North West Limited. That being the case, the achievable distortion using minimum mean square error (MMSE) estimation is assessed for both random sampling and optimal linear encoding acquisition schemes. Within this setting, the impact of imperfect second order source statistics is numerically evaluated. The fundamental limit of the recovery process is characterized using Rate-Distortion theory to obtain the optimal performance theoretically attainable. Interestingly, numerical results show that matrix completion based recovery outperforms MMSE estimator when the number of available observations is low and access to perfect source statistics is not available.

6.3. Software Radio Programming Model

6.3.1. Dataflow programming model

The advent of portable software-defined radio (SDR) technology is tightly linked to the resolution of a difficult problem: efficient compilation of signal processing applications on embedded computing devices. Modern wireless communication protocols use packet processing rather than infinite stream processing and also introduce dependencies between data value and computation behavior leading to dynamic dataflow behavior. Recently, parametric dataflow has been proposed to support dynamicity while maintaining the high level of analyzability needed for efficient real-life implementations of signal processing computations. The team developed a new compilation flow [5] that is able to compile parametric dataflow graphs. Built on the LLVM compiler infrastructure, the compiler offers an actor-based C++ programming model to describe parametric graphs, a compilation front end for graph analysis, and a back end that currently matches the Magali platform: a prototype heterogeneous MPSoC dedicated to LTE-Advanced. We also introduce an innovative scheduling technique, called microscheduling, allowing one to adapt the mapping of parametric dataflow programs to the specificities of the different possible MPSoCs targeted. A specific focus on FIFO sizing

on the target architecture is presented. The experimental results show compilation of 3GPP LTE-Advanced demodulation on Magali with tight memory size constraints. The compiled programs achieve performance similar to handwritten code.

The memory subsystem of modern multi-core architectures is becoming more and more complex with the increasing number of cores integrated in a single computer system. This complexity leads to profiling needs to let software developers understand how programs use the memory subsystem. Modern processors come with hardware profiling features to help building tools for these profiling needs. Regarding memory profiling, many processors provide means to monitor memory traffic and to sample read and write memory accesses. Unfortunately, these hardware profiling mechanisms are often very complex to use and are specific to each micro-architecture. The numap library [44], [31] is dedicated to the profiling of the memory subsytem of modern multi-core architectures. numap is portable across many micro-architectures and comes with a clean application programming interface allowing to easily build profiling tools on top of it.

This numap library as been officially integrated into Turnus, a profiler dedicated to dynamic dataflow programs.

6.3.2. Implementation of filters and FFTs on FPGAs

In collaboration with two researchers from Inria AriC, we have worked on a digital filter synthesis flow targeting FPGAs [46]. Based on a novel approach to the filter coefficient quantization problem, this approach produces results which are faithful to a high-level frequency-domain specification. An automated design process is also proposed where user intervention is limited to a very small number of relevant input parameters. Computing the optimal value of the other parameters not only simplifies the user interface: the resulting architectures also outperform those generated by mainstream tools in accuracy, performance, and resource consumption.

In collaboration with researchers from Isfahan, Iran, a multi-precision Fast Fourier Transform (FFT) module with dynamic run-time reconfigurability has been proposed [3] to trade off accuracy with energy efficiency in an SDR-based architecture. To support variable-size FFT, a reconfigurable memory-based architecture is investigated. It is revealed that the radix-4 FFT has the minimum computational complexity in this architecture. Regarding implementation constraints such as fixed-width memory, a noise model is exploited to statistically analyze the proposed architecture. The required FFT word-lengths for different criteria, (bit-error rate (BER), modulation scheme, FFT size, and SNR) are computed analytically and confirmed by simulations in AWGN and Rayleigh fading channels. At run-time, the most energy-efficient word-length is chosen and the FFT is reconfigured until the required application-specific BER is met. Evaluations show that the implementation area and the number of memory accesses are reduced. The results obtained from synthesizing basic operators of the proposed design on an FPGA show energy consumption saving of over 80 %.

6.3.3. Tools for FPGA development

The pipeline infrastructure of the FloPoCo arithmetic core generator has been completely overhauled [34], [23]. From a single description of an operator or datapath, optimized implementations are obtained automatically for a wide range of FPGA targets and a wide range of frequency/latency trade-offs. Compared to previous versions of FloPoCo, the level of abstraction has been raised, enabling easier development, shorter generator code, and better pipeline optimization. The proposed approach is also more flexible than fully automatic pipelining approaches based on retiming: In the proposed technique, the incremental construction of the pipeline along with the circuit graph enables architectural design decisions that depend on the pipeline. These allow pipeline-dependent changes to the circuit graph for finer optimization. This is particularly important for the filter structures already mentioned [46].

In parallel, we also started to study the integration of arithmetic optimizations in high-level synthesis (HLS) tools [48]. HLS is a big step forward in terms of design productivity. However, it restricts data-types and operators to those available in the `C` language supported by the compiler, preventing a designer to fully exploit the FPGA flexibility. To lift this restriction, a source-to-source compiler may rewrite, inside critical loop nests of the input `C` code, selected floating-point additions into sequences of simpler operator using non-standard

arithmetic formats. This enables hoisting floating-point management out the loop. What remains inside the loop is a sequence of fixed-point additions whose size is computed to enforce a user-specified, application-specific accuracy constraint on the result. Evaluation of this method demonstrates significant improvements in the speed/resource usage/accuracy trade-off.

6.3.4. Computer Arithmetic

In collaboration with researchers from Istanbul, Turkey, operators have also been developed for division by a small positive constant [49]. The first problem studied is the Euclidean division of an unsigned integer by a constant, computing a quotient and a remainder. Several new solutions are proposed and compared against the state of the art. As the proposed solutions use small look-up tables, they match well the hardware resources of an FPGA. The article then studies whether the division by the product of two constants is better implemented as two successive dividers or as one atomic divider. It also considers the case when only a quotient or only a remainder are needed. Finally, it addresses the correct rounding of the division of a floating-point number by a small integer constant. All these solutions, and the previous state of the art, are compared in terms of timing, area, and area-timing product. In general, the relevance domains of the various techniques are very different on FPGA and on ASIC.

On the software side, we have also shown, in collaboration with researchers from LIP and the Kalray company, that correctly rounded elementary functions can be implemented more efficiently using only fixed-point arithmetic than when classically using floating-point arithmetic [24]. A purely integer implementation of the correctly rounded double-precision logarithm outperforms the previous state of the art, with the worst-case execution time reduced by a factor 5. This work also introduces variants of the logarithm that input a floating-point number and output the result in fixed-point. These are shown to be both more accurate and more efficient than the traditional floating-point functions for some applications.

URBANET Team

7. New Results

7.1. Network deployment and characterization

Participants: Ahmed Boubrima, Angelo Furno, Walid Bechkit, Khaled Boussetta, Hervé Rivano, Razvan Stanica.

7.1.1. Deployment of Wireless Sensor Networks for Pollution Monitoring

Monitoring air quality has become a major challenge of modern cities, where the majority of population lives, because of industrial emissions and increasing urbanization, along with traffic jams and heating/cooling of buildings. Monitoring urban air quality is therefore required by municipalities and by the civil society. Current monitoring systems rely on reference sensing stations that are precise but massive, costly and therefore seldom. Wireless sensor networks seem to be a good solution to this problem, thanks to sensors' low cost and autonomy, as well as their fine-grained deployment. A careful deployment of sensors is therefore necessary to get better performances, while ensuring a minimal financial cost.

We have tackled the issue of WSN deployment for air pollution monitoring in a series of papers this year. In [10], we tackled the optimization problem of sensor deployment and we proposed an integer programming model, which allows to find the optimal network topology while ensuring air quality monitoring with a high precision and the minimum financial cost. Most of existing deployment models of wireless sensor networks are generic and assume that sensors have a given detection range. This assumption does not fit pollutant concentrations sensing. Our model takes into account interpolation methods to place sensors in such a way that pollution concentration is estimated with a bounded error at locations where no sensor is deployed. This solution was further tested and evaluated on a data set of the Lyon city [9], giving insights on how to establish a good compromise between the deployment budget and the precision of air quality monitoring.

In practice, multiple pollution sources can be present in an area. For this reason, in [11] we propose to apply a spatial clustering algorithm to the air pollution data in order to determine pollution zones that are due to the same pollutant sources and group them together to find candidate sites for the deployment of sensors. This approach was tested on real world data, namely the Paris pollution data, which was recorded in March 2014.

A very important deployment parameter is the height at which the sensor is placed. In [12], we demonstrate the impact of this parameter, usually neglected in the literature. This pushed us to study a 3D deployment model, based on an air pollution dispersion model issued from real experiments, performed in wind tunnels emulating the pollution emitted by a steady state traffic flow in a typical street canyon.

7.1.2. Access Point Deployment

The problem of designing wireless local networks (WLANs) involves deciding where to install the access points (APs), and assigning frequency channels to them with the aim to cover the service area and to guarantee enough capacity to users. In [5], we propose different solutions to the problems related to the WLAN design. In the first part, we focus on the problem of designing a WLAN by treating separately the AP positioning and the channel assignment problems. For the AP positioning issue, we formulate it as a set covering problem. Since the computation complexity limits the exact solution, we propose two heuristics to offer efficient solutions. On the other hand, for the channel assignment, we define this issue as a minimum interference frequency assignment problem and propose three heuristics: two of them aim to minimize the interference at AP locations, and the third one minimizes the interference at the TPs level. In the second part, we treat jointly the two aforementioned issues based on the concept of virtual forces. In this case, we start from an initial solution provided by the separated approach and try to enhance it by adjusting the APs positions and reassigning their operating frequencies.

7.1.3. Mobile Traffic Analysis

The analysis of operator-side mobile traffic data is a recently emerged research field, and, apart a few outliers, relevant works cover the period from 2005 to date, with a sensible densification over the last four years. In [8], we provided a thorough review of the multidisciplinary activities that rely on mobile traffic datasets, identifying major categories and sub-categories in the literature, so as to outline a hierarchical classification of research lines and proposing a complete introductory guide to the research based on mobile traffic analysis.

The usage of these datasets in the design of new networking solutions, in order to achieve the so-called cognitive networking paradigm, is one of the most important applications of these analytics methods. In fact, cognitive networking techniques root in the capability of mining large amounts of mobile traffic data collected in the network, so as to understand the current resource utilization in an automated manner and realize a more dynamic management of network resources, that adapts to the significant spatiotemporal fluctuations of the mobile demand. In [6], we take a first step towards cellular cognitive networks by proposing a framework that analyzes mobile operator data, builds profiles of the typical demand, and identifies unusual situations in network-wide usages. We evaluate our framework on two real-world mobile traffic datasets, and show how it extracts from these a limited number of meaningful mobile demand profiles. In addition, the proposed framework singles out a large number of outlying behaviors in both case studies, which are mapped to social events or technical issues in the network.

7.2. Data Collection in Multi-hop Networks

Participants: Jin Cui, Jad Oueis, Hervé Rivano, Razvan Stanica, Fabrice Valois.

7.2.1. Data Aggregation in Wireless Sensor Networks

Wireless Sensor Networks (WSNs) have been regarded as an emerging and promising field in both academia and industry. Currently, such networks are deployed due to their unique properties, such as self-organization and ease of deployment. However, there are still some technical challenges needed to be addressed, such as energy and network capacity constraints. Data aggregation, as a fundamental solution, processes information at sensor level as a useful digest, and only transmits the digest to the sink. The energy and capacity consumptions are reduced due to less data packets transmission.

As a key category of data aggregation, aggregation function, solving how to aggregate information at sensor level, was investigated in the Ph.D. thesis of Jin Cui [1]. In this work, we make four main contributions: firstly, we propose two new networking-oriented metrics to evaluate the performance of aggregation function: aggregation ratio and packet size coefficient. Aggregation ratio is used to measure the energy saving by data aggregation, and packet size coefficient allows to evaluate the network capacity change due to data aggregation. Using these metrics, we confirm that data aggregation saves energy and capacity whatever the routing or MAC protocol is used. Secondly, to reduce the impact of sensitive raw data, we propose a data-independent aggregation method which benefits from similar data evolution and achieves better recovered fidelity. This solution, named Simba, is detailed in [15] as well. Thirdly, a property-independent aggregation function is proposed to adapt the dynamic data variations. Comparing to other functions, our proposal can fit the latest raw data better and achieve real adaptability without assumption about the application and the network topology. Finally, considering a given application, a target accuracy, we classify the forecasting aggregation functions by their performance. The networking-oriented metrics are used to measure the function performance, and a Markov Decision Process is used to compute them. Dataset characterization and classification framework are also presented to guide researcher and engineer to select an appropriate functions under specific requirements.

7.2.2. Energy Harvesting in Wireless Sensor Networks

Energy harvesting capabilities are challenging our understanding of wireless sensor networks by adding recharging capacity to sensor nodes. This has a significant impact on the communication paradigm, as networking mechanisms can benefit from these potentially infinite renewable energy sources. In [23], we study photovoltaic energy harvesting in wireless sensor networks, by building a harvesting analytical model, linking three components: the environment, the battery, and the application. Given information on two of

the components, limits on the third one can be determined. To test this model, we adopt several use cases with various indoor and outdoor locations, battery types, and application requirements. Results show that, for predefined application parameters, we are able to determine the acceptable node duty cycle given a specific battery, and vice versa. Moreover, the suitability of the deployment environment (outdoor, well lighted indoor, poorly lighted indoor) for different application characteristics and battery types is discussed .

In a second contribution [22], we study the consequences of implementing photovoltaic energy harvesting on the duty cycle of a wireless sensor node, in both outdoor and indoor scenarios. We show that, for the static duty cycle approach in outdoor scenarios, very high duty cycles, in the order of tens of percents, are achieved. This further eliminates the need for additional energy conservation schemes. In the indoor case, our analysis shows that the dynamic duty cycle approach based solely on the battery residual energy does not necessarily achieve better results than the static approach. We identify the main reasons behind this behavior, and test new design considerations by adding information on the battery level variation to the duty cycle computation. We demonstrate that this approach always outperforms static solutions when perfect knowledge of the harvestable energy is assumed, as well as in realistic deployments, where this information is not available.

7.2.3. Data Collection with Moving Nodes

Patrolling with mobile nodes (robots, drones, cars) is mainly used in situations where the need of repeatedly visiting certain places is critical. In [24], we consider a deployment of a wireless sensor network (WSN) that cannot be fully meshed because of the distance or obstacles. Several robots are then in charge of getting close enough to the nodes in order to connect to them, and perform a patrol to collect all the data in time. We discuss the problem of multi-robot patrolling within the constrained wireless networking settings. We show that this is fundamentally a problem of vertex coverage with bounded simple cycles (CBSC). We offer a formalization of the CBSC problem and prove it is NP-hard and at least as hard as the Traveling Salesman Problem (TSP). Then, we provide and analyze heuristics relying on clusterings and geometric techniques. The performances of our solutions are assessed in regards to robot limitations (storage and energy), networking parameters, but also to random and particular graph models.

Also related to data collection, in [3], we advocate the use of conventional vehicles equipped with storage devices as data carriers whilst being driven for daily routine journeys. The road network can be turned into a large-capacity transmission system to offload bulk transfers of delay-tolerant data from the Internet. The challenges we address include how to assign data to flows of vehicles and while coping with the complexity of the road network. We propose an embedding algorithm that computes an offloading overlay where each logical link spans over multiple stretches of road from the underlying road infrastructure. We then formulate the data transfer assignment problem as a novel linear programming model we solve to determine the optimal logical paths matching the performance requirements of a data transfer. We evaluate our road traffic allocation scheme using actual road traffic counts in France. The numerical results show that 20% of vehicles in circulation in France equipped with only one Terabyte of storage can offload Petabyte transfers in a week.

7.2.4. Network Resilience

The notion of Shared Risk Link Groups (SRLG) captures survivability issues when a set of links of a network may fail simultaneously. The theory of survivable network design relies on basic combinatorial objects that are rather easy to compute in the classical graph models: shortest paths, minimum cuts, or pairs of disjoint paths. In the SRLG context, the optimization criterion for these objects is no longer the number of edges they use, but the number of SRLGs involved. Unfortunately, computing these combinatorial objects is NP-hard and hard to approximate with this objective in general. Nevertheless some objects can be computed in polynomial time when the SRLGs satisfy certain structural properties of locality which correspond to practical ones, namely the star property (all links affected by a given SRLG are incident to a unique node) and the span 1 property (the links affected by a given SRLG form a connected component of the network). The star property is defined in a multi-colored model where a link can be affected by several SRLGs while the span property is defined only in a mono-colored model where a link can be affected by at most one SRLG. In [4], we extend these notions to characterize new cases in which these optimization problems can be solved in polynomial time. We also

investigate the computational impact of the transformation from the multi-colored model to the mono-colored one. Experimental results are presented to validate the proposed algorithms and principles.

7.3. Networks in the Internet of Things

Participants: Soukaina Cherkaoui, Alexis Duque, Guillaume Gaillard, Hervé Rivano, Razvan Stanica, Fabrice Valois.

7.3.1. Service Level Agreements in the Internet of Things

With the growing use of distributed wireless technologies for modern services, the deployments of dedicated radio infrastructures do not enable to ensure large-scale, low-cost and reliable communications. The Ph.D. thesis of Guillaume Gaillard [2] aims at enabling an operator to deploy a radio network infrastructure for several client applications, hence forming the Internet of Things (IoT). We evaluate the benefits earned by sharing an architecture among different traffic flows, in order to reduce the costs of deployment, obtaining a wide coverage through efficient use of the capacity on the network nodes. We thus need to ensure a differentiated Quality of Service (QoS) for the flows of each application.

We propose to specify QoS contracts, namely Service Level Agreements (SLAs), in the context of the IoT. SLAs include specific Key Performance Indicators (KPIs), such as the transit time and the delivery ratio, concerning connected devices that are geographically distributed in the environment. The operator agrees with each client on the sources and amount of traffic for which the performance is guaranteed. Secondly, we describe the features needed to implement SLAs on the operated network, and we organize them into an SLA management architecture. We consider the admission of new flows, the analysis of current performance and the configuration of the operator's relays. Based on a robust, multi-hop technology, IEEE Std 802.15.4-2015 TSCH mode, we provide two essential elements to implement the SLAs : a mechanism for the monitoring of the KPIs [19], and KAUSA, a resource allocation algorithm with multi-flow QoS constraints [18]. The former uses existing data frames as a transport medium to reduce the overhead in terms of communication resources. We compare different piggybacking strategies to find a tradeoff between the performance and the efficiency of the monitoring. With the latter, KAUSA, we dedicate adjusted time-frequency resources for each message, hop by hop. KAUSA takes into account the interference, the reliability of radio links and the expected load to improve the distribution of allocated resources and prolong the network lifetime [17]. We show the gains and the validity of our contributions with a simulation based on realistic traffic scenarios and requirements.

7.3.2. Channel Access in Machine-to-Machine Communications

The densification of the urban population and the rise of smart cities applications foster the need for capillary networks collecting data from sensors monitoring the cities. Among the multiple networking technologies considered for this task, cellular networks, such as LTE-A, bring an ubiquitous coverage of most cities. It is therefore necessary to understand how to adapt LTE-A, and what should be the future 5G architecture, in order to provide efficient connectivity to Machine-to-Machine (M2M) devices alongside the main target of mobile networks, Human-to-Human devices. Indeed, cellular random access procedures are known to suffer from congestion in presence of a large number of devices, while smart cities scenarios expect huge density of M2M devices. Several solutions have been investigated for the enhancement of the current LTE-A access management strategy. In [14], we contribute to the modeling and computation of the capacity of the LTE-A Random Access Channel (RACH) in terms of simultaneous successful access. In particular, we investigate the hypothesis of piggybacking the payload of Machine Type Communications from M2M devices within the RACH, and show that M2M densities considered realistic for smart cities applications are difficult to sustain by the current LTE-A architecture.

7.3.3. Visible Light Communications in the Internet of Things

The Internet of Things connects devices, such as everyday consumer objects, enabling information gathering and improved user experience. Also, this growing and dynamic market makes that consumers nowadays expect electronic products, even the cheapest, to include wireless connectivity. However, despite the fact that radio based solutions exist, such as Bluetooth Low Energy, the manufacturing costs introduced by these radio

technologies are non-negligible compared to the initial product price. As most of the home electronics already integrate small light emitting diodes, Visible Light Communication appears as a competitive alternative. However, its broad adoption is suffering from a lack of integration with smartphones, which represent the communication hubs for most of the users. To overcome this issue, in [16], we propose a line of sight LED-to-camera communication system based on a small color LED and a smartphone. We design a cheap prototype as proof of concept of a near communication framework for the Internet of Things. We evaluate the system performance, its reliability and the environment influence on the LED-to-camera communication, highlighting that a throughput of a few kilobits per second is reachable. Finally, we design a real time, efficient LED detection and image processing algorithm to leverage the specific issues encountered in the system.

7.3.4. Radio Frequency Identification in Dense Environments

Radio Frequency Identification (RFID) is another cheap technology shaping the Internet of Things. The rapid development of RFID has allowed its large adoption and led to increasing deployments of RFID solutions in diverse environments under varying scenarios and constraints. The nature of these constraints ranges from the amount to the mobility of the readers deployed, which in turn highly affects the quality of the RFID system, causing reading collisions. However, the technology suffers from a recurring issue: the reader-to-reader collisions. Numerous protocols have been proposed to attempt to reduce them, but remaining reading errors still heavily impact the performance and fairness of dense RFID deployments.

In order to ensure collision-free reading, a scheduling scheme is needed to read tags in the shortest possible time. In [25], we study this scheduling problem in a stationary setting and the reader minimization problem in a mobile setting. We show that the optimal schedule construction problem is NP-complete and provide an approximation algorithm that we evaluate our techniques through simulation. Moving closer to practical solutions, [20] introduces a new Distributed Efficient & Fair Anticollision for RFID (DEFAR) protocol. DEFAR reduces both monochannel and multichannel collisions, as well as interference, by a factor of almost 90% in comparison with the best state of the art protocols. The fairness of the medium access among the readers is improved to a 99% level. Such improvements are achieved by applying a TDMA-based "server-less" approach and assigning different priorities to readers depending on their behavior over precedent rounds. A distributed reservation phase is organized between readers with at least one winning reader afterwards. Then, multiple reading phases occur within a single frame in order to obtain fast coverage and high throughput. The use of different reader priorities based on reading behaviors of previous frames also contributes to improve both fairness and efficiency.

Another type of collisions appears when the RFID tags are not only dense, but also mobile. mDEFAR [21] is an adaptation of DEFAR, while CORA [7] is more of a locally mutual solution where each reader relies on its neighborhood to enable itself or not. Using a beaconing mechanism, each reader is able to identify potential (non-)colliding neighbors in a running frame and as such chooses to read or not. Performance evaluation shows high performance in terms of coverage delay for both proposals quickly achieving 100% coverage depending on the considered use case while always maintaining consistent efficiency levels above 70%. Compared to the state of the art, our solutions proved to be better suited for highly dense and mobile environments, offering both higher throughput and efficiency. The results reveal that depending on the application considered, choosing either mDEFAR or CORA helps improve efficiency and coverage delay.

CHROMA Team

7. New Results

7.1. Bayesian Perception

Participants: Christian Laugier, Lukas Rummelhard, Amaury Nègre [Gipsa Lab since June 2016], Jean-Alix David, Julia Chartre, Jerome Lussereau, Tiana Rakotovao, Nicolas Turro [SED], Jean-François Cuniberto [SED], Diego Puschini [CEA DACLE], Julien Mottin [CEA DACLE].

7.1.1. Conditional Monte Carlo Dense Occupancy Tracker (CMCDOT)

Participants: Lukas Rummelhard, Amaury Nègre, Christian Laugier.

The research work on *Bayesian Perception* has been done as a continuation and an extension of some previous research results obtained in the scope of the former Inria team-project e-Motion and of the more recent developments done in 2015 in the scope of the Chroma team. This work exploits the *Bayesian Occupancy Filter (BOF)* paradigm [42], developed and patented by the team several years ago [0]. It also extends the more recent concept of *Hybrid Sampling BOF (HSBOF)* [76], whose purpose was to adapt the concept to highly dynamic scenes and to analyze the scene through a static-dynamic duality. In this new approach, the static part is represented using an occupancy grid structure, and the dynamic part (motion field) is modeled using moving particles. The *HSBOF* software has been implemented and tested on our experimental platforms (equipped Toyota Lexus and Renault Zoe) in 2014 and 2015; it has also been implemented in 2015 on the experimental autonomous car of Toyota Motor Europe in Brussels.

The objective of the research work performed in the period 2015-16 was to overcome some of the shortcomings of the initial *HSBOF* approach [0] , and to obtain a better understanding of the observed dynamic scenes through the introduction of an additional object level into the model. The new framework, whose development has been continued in 2016, is called *Conditional Monte Carlo Dense Occupancy Tracker (CMCDOT)* [84]. The whole CMCDOT framework and its results are presented and explained on a video posted on Youtube [0]. This work has mainly been performed in the scope of the project *Perfect* of IRT Nanoelec [0] (financially supported by the French ANR agency [0]), and also used in the scope of our long-term collaboration with Toyota.

In 2016, most of the efforts have been focused on the optimization of the implementation of our grid-based Bayesian filtering CMCDOT framework. Since the beginning of the development of this framework, we have chosen to construct models and algorithms specially designed to attain real-time performances on embedded devices, through a massively parallelization of the involved processes. The whole system have been implemented and scrupulously optimized in Cuda, in order to fully benefit from the Nvidia GPUs and technologies. Starting from the use of the Titan X and GTX980 GPUs (the hardware used in our computers and experimental platforms), we have successfully adapted and transferred our whole real-time perception chain on Nvidia dedicated-to-automotive cards Jetson K1 and X1 [0]. A specific optimization has been performed in term of data access and processing, allowing us to obtain real-time results when processing the data from the 8 lidar layers generated by our IBEO sensors, by using a grid containing 1400x600 cells and 65536 dynamic particles (for motion estimation). The observation grid generation and fusion (representing the input of the CMCDOT) is made in 17ms on Jetson K1 and only in 0.7ms on Jetson X1; a CMCDOT filtering update is performed in 70ms on Jetson K1 and only in 17ms on Jetson X1.

[0]The *Bayesian programming formalism* developed in e-Motion, pioneered (together with the contemporary work of Thrun, Burgards and Fox [94]) a systematic effort to formalize robotics problems under Probability theory –an approach that is now pervasive in Robotics.

[0]In the current implementation of the HSBOF algorithm, many particles are still allocated to irrelevant areas, since no specific representation models are associated to dataless areas. Moreover, if the filtered low level representation can directly be used for various applications (for example mapping process, short-term collision risk assessment [47], [85], etc), the retrospective object level analysis by dynamic grid segmentation can be computationally expensive and subjected to some data association errors.

[0]https://www.youtube.com/watch?v=uwIrk1TLFiM

[0]Nanoelec Technological Research Institute (Institut de Recherche Technologique Nanoelec)

[0]National Research Agency (Agence Nationale de la recherche)

[0]These new Nvidia devices are more suited for embedded applications, in term of power consumption and dimensions.

Figure 5. Jetson X1 card, Nvidia device dedicated to automotive applications

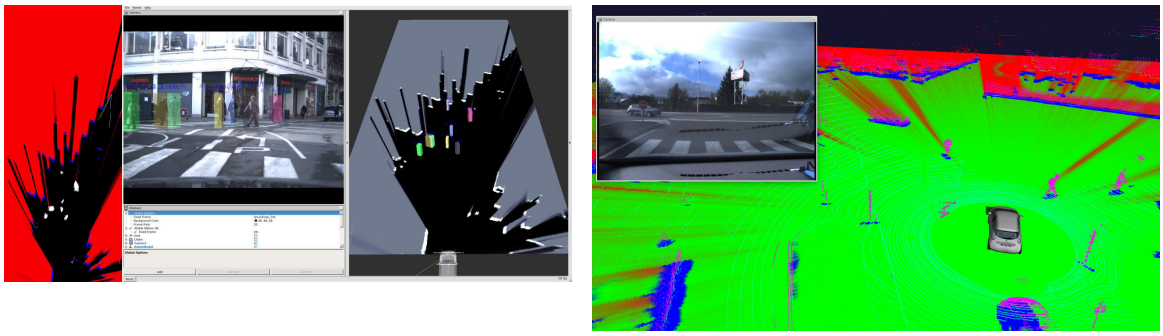


Figure 6. a) CMCDOT results : filtered occupancy grids, enhanced with motion estimations (vectors) and object detection (colored boxes) b) Example of an occupancy grid generated using the classified point cloud and the ground model

7.1.2. A new sensor model for 3D sensors, by Ground Estimation, Data segmentation and adapted Occupancy Grid construction

Participants: Lukas Rummelhard, Amaury Nègre, Anshul Paigwar, Christian Laugier.

As a starting point for the Bayesian perception framework embedded on the vehicles and on the perception boxes, the system generates instantaneous spatial occupancy grids, by interpreting the point clouds generated by the sensors (sensor model). With planar sensors, placed at the level of the wanted occupancy grid, such as the IBEO Lidar on the vehicles or the Hokuyo Lidar on the first developed perception box, a classic sensor model can be used: before the laser impact the space is considered as empty, occupied at the impact point and undefined after the impact. In our previous approach, the angular differences between the 4 laser layers of our IBEO Lidars was taken into account by introducing a *confidence factor* in the data, reducing in this way the effect of the impacts too close to the ground. In this approach the ground is assumed to be flat and the confidence factor is calculated geometrically. Then, given the orientation of these sensors and the environments traversed, such a model was quite satisfactory.

However, this traditional sensor model has to be adapted when using Velodyne or Quanergy sensors mounted on the top of the vehicle and providing dense 3D data with a high horizontal and vertical resolution. Indeed, in this case the laser layers are capable of depicting an obstacle from above, and consequently an impact at a given distance does not certify any more a free area until the impact. Moreover, many impact points are located on the ground and have to be appropriately modeled in order to systematically avoid deceptive obstacle detection. Then, the previous flat-ground assumption doesn't hold anymore, since the actual ground shape is integrated into the data and the correct segmentation of obstacle becomes critical in the process. This is why we have developed the new *Ground Estimator* approach.

The aim of the method is, upstream from the Bayesian filtering step of our current perception system (CMCDOT), to first dynamically *estimate the ground elevation*, to exploit this information for making a *relevant data classification* between actual obstacle impacts and ground impacts, and finally to generate the *relevant occupancy grid using this classified 3D point cloud* (sensor model). The developed method is based on a recursive spatial and temporal filtering of a Bayesian network of elevation nodes, constantly re-estimated and re-evaluated with respect to data and spatial continuity. The construction of the occupancy grid is based, on the one hand, on the location of the laser impacts, and on the other hand on the shape of the ground and the height at which the lasers pass through the different portions of the space.

The approach has been first successfully tested and validated with dense Lidar sensors (Velodyne and Quanergy). The use of the enhanced sensor model is also currently tested with sparser sensors, with the objective to increase their robustness. The obtained results show promising perspectives, offering a robust and efficient ground representation, data segmentation and relevant occupancy grid, and also offering quality inputs for the next steps of the perception framework. A journal paper and a patent are under preparation.

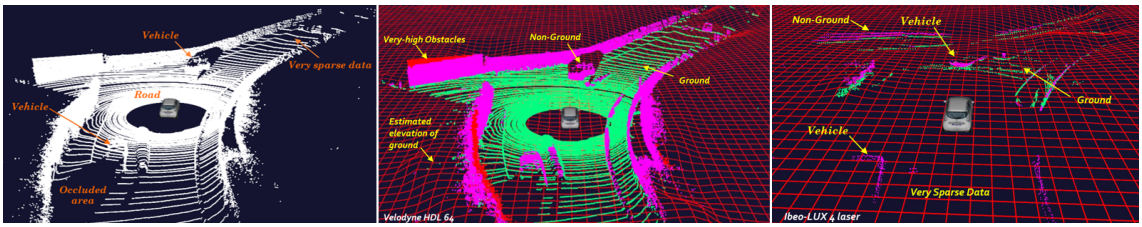


Figure 7. (a) Typical 3D point cloud generated by Velodyne LiDAR, (b) Point cloud segmentation between ground (green points) and non-ground (purple points), and estimated average elevation of the terrain (red grid) (c) Point Cloud Segmentation on 4-Ibeo Lux LiDAR data and estimated elevation of terrain.

7.1.3. Dense & Robust outdoor perception for autonomous vehicles

Participants: Victor Romero-Cano, Christian Laugier.

Robust perception plays a crucial role in the development of autonomous vehicles. While perception in normal and constant environmental conditions has reached a plateau, robust perception in changing and challenging environments has become an active research topic, particularly due to the safety concerns raised by the introduction of self-driving cars to public streets. In collaboration with Toyota Motors Europe and starting in April 2016 we have developed techniques that tackle the robust-perception problem by combining multiple complementary sensor modalities.

Our techniques, similar to those presented in [78], [91] explore the complementary relationships between passive and active sensors at the pixel level. Low-level sensor fusion allows for an effective use of raw data in the fusion process and encourages the development of recognition systems that work directly on multi-modal data rather than higher level estimates. During the last nine months we have developed low-level sensor fusion approaches that, differently from most of the related literature, do not have fixed requirements regarding coverage or density of the active sensors. This provides a competitive advantage due to the elevated costs of dense range sensors such as Velodyne LIDARs.

Our framework outputs a new image-like data representation where each pixel contains not only colour but also other low level features such as depth and regions of interest where generic objects are likely to be. Our approach is generic so it allows for the integration of data coming from any active sensor into the image space. Additionally, it does not aim at tackling the object detection problem directly but it proposes a multi-modal-data representation from which object detection methods may benefit. For evaluation purposes we have tackled the concrete problem of fusing color images and sparse lidar returns, however, as explained before, the framework is amenable for the inclusion of any other range-sensor modality. The framework creates *XDimages* by extrapolating range measurements across the image space in a two-stage procedure. The first stage considers locally homogeneous areas given by a super-pixel segmentation while the second one further expands depth values by performing self-supervised segmentation of areas seeded by the range sensor. The framework's pipeline is illustrated in Figure 8 .

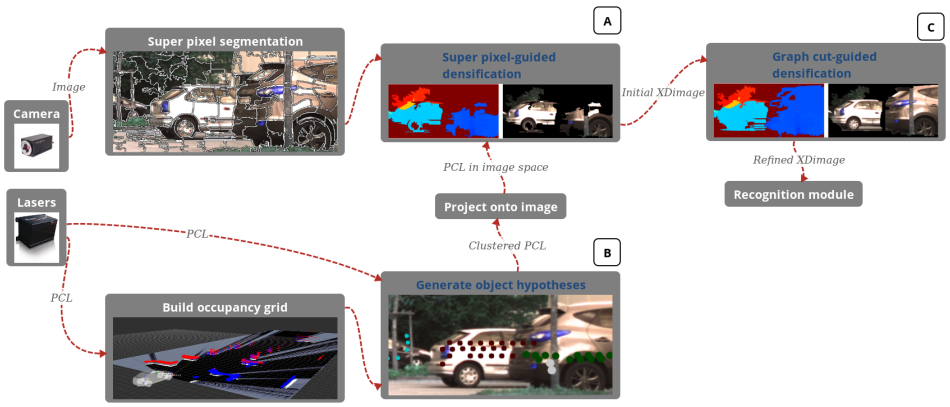


Figure 8. The XDvision framework.

We have named an instance of our data structure an *XDimage*. It corresponds to an augmented camera image where individual pixels contain both appearance and geometric information. The first and more challenging problem to be solved in order to build XDimages is that of densifying sparse point cloud data provided by active range sensors. In our approach we extrapolated depth information using a two-steps procedure as follows:

1. Extend depth values projected onto individual pixels to neighbouring pixels that have similar appearance.
2. Obtain geometry-based object hypothesis.
3. For each geometry-based object hypothesis, extrapolate range measurements in order to account for entire objects.

The results of this work have resulted in a patent application [82] and a paper submission to ICRA 2017 [83].

7.1.4. Integration of Bayesian Perception System on Embedded Platforms

Participants: Tiana Rakotovao, Christian Laugier, Diego Puschini [CEA DACLE], Julien Mottin [CEA DACLE].

Perception is a primary task for an autonomous car where safety is of utmost importance. A perception system builds a model of the driving environment by fusing measurements from multiple perceptual sensors including LIDARs, radars, vision sensors, etc. The fusion based on occupancy grids builds a probabilistic environment model by taking into account sensor uncertainties. Our objective is to integrate the computation of occupancy grids into embedded low-cost and low-power platforms. Occupancy Grids perform though intensive probability calculus that can be hardly processed in real-time on embedded hardware.

As a solution, we introduced a new sensor fusion framework called *Integer Occupancy Grid* [80]. Integer Occupancy Grids rely on a proven mathematical foundation that enables to process probabilistic fusion through simple addition of integers. The hardware/software integration of integer occupancy grids is safe and reliable. The involved numerical errors are bounded and parameterized by the user. Integer Occupancy Grids enable a real-time computation of multi-sensor fusion on embedded low-cost and low-power processing platforms dedicated for automotive applications. This research work has been conducted in the scope of the PhD thesis of Tiana Rakotovao, which will be defended in February 2017.

Figure 9. Fusion of three front LIDARs and one rear LIDAR on the ZOE platform

Experiences showed that Integer Occupancy Grids enable the real-time fusion of the four ibeo LUX LIDARs mounted on the ZOE experimental platform of IRT Nano-Elec [79]. The LIDARs produces about 80,000 points at 25Hz. These points are fused in real-time through a hardware/software integration of the Integer Occupancy Grid framework on an embedded CPU based on ARM A9@1GHz. The platform respects the low-cost and low-power constraints of processing hardware used for automotive. The fusion produces an occupancy grid at more than 25 Hz as illustrated on figure 9 .

7.1.5. Embedded and Distributed Perception

Participants: Christian Laugier, Julia Chartes, Amaury Nègre, Lukas Rummelhard, Jean-Alix David, Jerome Lussereau, Nicolas Turro [SED], Jean-Francçois Cuniberto [SED].

7.1.5.1. Embedded Perception in an Experimental Vehicle Renault ZOE

In the scope of the *Perfect* project of the IRT nanoelec, we have started to build an experimental platform using a Renault Zoe equipped with several types of sensors (see 2014 and 2015 annual activity reports). The platform includes multiple sensors, an embedded perception system based on the CMCDOT, and a collision risk component, figure 10 (a) illustrates.

Figure 10. (a) Display of the HMI (b) Collision simulation with a mannequin (c) On left: picture of the smartbox, on right: picture of the cone.

In 2016, we have continued to develop and to improve the platform using the latest version of the CMCDOT, some optimized perception and localization components, and new V2X communication functions for distributed perception.

In particular, we have developed an improved the localization function using maps and V2X communication devices. We also developed a new embedded component for sharing data between the infrastructure perception boxes and the vehicle; this component is based on the use of V2X communication and GPS time synchronization. This is a first step towards a fully distributed perception system. The development of this system will be continued in 2017 (see next section).

During the year 2016, experiments have been pushed forward on testing the perception algorithms, the collision risk alert and the localization components using a fabric mannequin as shown on figure 10 (b). The mannequin has been motorized for easier and more realistic tests and demos. More details are given in the team publications at MCG 2016 [29] and at GTC Europe 2016. The work of the team is also explained on youtube videos "Irt Nanoelec Perfect Project" [55] and for the technical side "Bayesian Embedded Perception" [54].

New experiments have also been performed on some road intersections and highways, in order to collect new data on driver's behaviors. These experiments have been conducted on mountain roads with changing slopes and on highway (to study the lane changing behaviors). They have been performed in the scope of our cooperation with Renault and with Toyota. The way these experimental data have been used is described in the section "Situation Awareness". More recently, we have also started to work on the development of the automatic driving controls on the Zoe vehicle. For that purpose, we have recently signed a cooperation agreement with Ecole Centrale de Nantes. The basic functions for automatic driving will be implemented on the Zoe at the beginning of 2017. For that purpose, a physical model of the Zoe is currently in construction under ROS Gazebo simulator. This should allow us to implement and to test the required control laws.

7.1.5.2. Distributed Perception

In 2015, we have developed a first *Connected Perception Box* including a GPS, a V2X communication device, a cheap Lidar sensor, and an Nvidia Tegra K1 board. The box was powered using a battery, and the objective was to reduce as far as possible the required energy consumption. Within the box, the perception process is performed using the CMCDOT algorithm. In 2016, we have continued to develop this concept of distributed perception. We have developed a second generation of the perception box, using a Quanergy M8 360° Lidar, a TX1 Nvidia Tegra board, an ITRI V2X communication device and the last version of the CMCDOT system.

This new box is more efficient and powerful than the previous one. It allows the real-time exchange of objects positions and velocities, through a V2X communication between the perception box and the connected vehicle. This leads to the extension of the vehicle perception area to some hidden areas, and to generate some alerts in case of a high collision risk, cf. fig. 11 . In this approach, time synchronization has been performed using GPS time and NTP protocol.

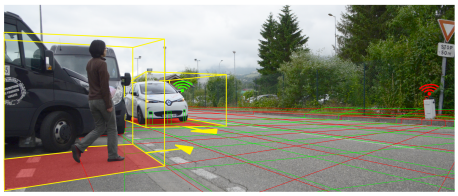


Figure 11. (a) Shared perception between car and perception box

7.1.5.3. Public demonstrations and Technological Transfer

2016 has been a year with many scientific events and public demos. Several public demonstrations of our experimental vehicle have performed, some of them in presence of local government officials during a GIANT show at CEA.

The collaboration with Nvidia on Embedded Perception for autonomous driving has been extended to 2017, and the "GPU research center" label has been renewed.

Toyota Motor Europe (TME) is strongly interested in the CMCDOT technology, and Inria is currently negotiating with them the conditions of a first licence for R&D purpose. A first implementation of the executable code of CMCDOT has successfully been implemented on the TME experimental vehicle in Brussels. We are currently discussing with TME an extension of the license to several other experimental vehicles located in some other places in the world.

At the end of 2016, we also started to transfer the CMCDOT technology to two industrial companies in the fields of industrial mobile robots and automatic shuttles. Confidential contracts for the joint development of proofs of concepts are under signing. The work will be performed at the beginning of 2017.

7.2. Situation Awareness

Participants: Christian Laugier, Olivier Simonin, Jilles Dibangoye, Alejandro Dizan Vasquez Govea [Apple since January 2016], Stephanie Lefevre [Mercedes-Benz North America], David Sierra-Gonzalez, Mathieu Barbier, Victor Romero-Cano.

7.2.1. Framework for Motion Prediction and Collision Risk Assessment

Participants: Christian Laugier, Alejandro Dizan Vasquez Govea [Apple since January 2016], Stephanie Lefevre [Mercedes-Benz North America], Lukas Rummelhard.

For several years, the challenging scientific problem of Motion Prediction, Risk Assessment and Decision-Making in open and dynamic environments has been one of our main research topics (see activity reports of the former e-Motion Inria team-project and Chroma team 2015 activity report).

Throughout 2016, we have continued this line of work by developing and experimentally testing new frameworks for Motion Prediction and Collision Risk Assessment in complex dynamic scenes involving multiple moving agents having various behaviors. This work has been conducted in the scope of three main scenarios: Short-term prediction in crowded urban environments (see approach and results in sections 7.1.1 and 7.1.5), Autonomous driving in highway environments (see section 7.2.2), and Safe Intersection crossing.

The main underlying concepts of the developed framework have recently been published in the second edition of the Handbook of Robotics [31]. They have also been presented into a Mooc course on "Autonomous Mobiles Robots and Vehicles", dedicated to graduate and undergraduate students and to engineers in Robotics [57]. This Mooc has been published twice in 2015 and in 2016.

The recent results have been published at ICRA 2016 [27] and also presented by C. Laugier in several invited talks : Asprom2016 [16], ICIT2016 [14], CUHK2016 [15], GTC-Europe2016 [24] and ARSO2016 [17].

Another contribution relies in the implementation of some the proposed models on two experimental vehicles (Lexus and Zoe experimental platforms). As it has been mentioned in the section 7.1.5 , several experiments on short-term collision risk assessment have successfully been conducted with these platforms in 2015 and 2016 (c.f. [84], [67]).

This work will be continued in 2017, in the scope of our ongoing collaborative projects with Toyota, Renault and IRT nanoelec. It will also be used as a support for the planned technological transfers with two industrial companies in the fields of industrial mobile robots and automatic shuttles (see section 7.1.5).

7.2.2. Planning-based motion prediction for collision risk estimation in autonomous driving scenarios

Participants: David Sierra-Gonzalez, Christian Laugier, Jilles Dibangoye, Alejandro Dizan Vasquez Govea [Apple since January 2016].

The objective is to develop a collision risk estimation system capable of reliably finding the risk of collision associated to the different feasible trajectories of the ego-vehicle. This research work is done in the scope of the Inria-Toyota long-term cooperation and of the PhD thesis work of David Sierra-Gonzáles.

One key factor, and probably the biggest challenge in order to produce robust collision risk estimation in traffic scenes, is the motion prediction of the dynamic obstacles (i.e. the other drivers for highway scenarios). The difficulty stems from the fact that human behavior is determined by a complex set of interdependent factors, which are very hard to model (e.g. intentions, perception, emotions). As a consequence, most existing systems are based on simple short-term motion models assuming constant velocity or acceleration.

We opt here for a planning-based approach, which assumes that drivers instinctively act as to minimize a cost (or equivalently, maximize a reward). This cost function encodes the preferences of the driver to, for instance, keep a minimum distance with the vehicle in front, drive in the right lane in the highway, or respect the speed limits. By using Inverse Reinforcement Learning (IRL) algorithms, we can obtain such cost function directly from expert demonstrations (i.e. simply observing how people drive).

Two well-known IRL algorithms [35], [101] have been implemented and used to obtain driver models from human demonstrations. The algorithms have been adapted to work with simulated demonstrations obtained on a highway simulator, and with real-world data from the Lexus and Renault Zoe platforms. Figure 12 .a shows a slice of one such cost function in the context of a real highway scene.

A novel framework has been developed in order to exploit the predictive potential of these models for the task of highway scene prediction [26]. The ability of these dynamic models to capture the risk-aversive behavior of drivers leads to an interaction-aware prediction. In contrast to other state-of-the-art interaction-aware approaches [59], the complexity of our prediction framework does not grow exponentially in the number of vehicles in the scene, but only quadratically. Figure 12 .b shows the prediction produced by our framework in two prototypical simulated highway scenarios. The figure shows the result of summing up across the occupancy distributions over a discretization of the road for all the vehicles in the scene, at different timesteps (note that the result is no longer a probability distribution, but it is convenient to visualize the prediction).

This framework has been patented by Inria and Toyota Motor Europe in October 2016.

7.2.3. Functional space representation for automated road intersection crossing

Participants: Mathieu Barbier, Christian Laugier, Olivier Simonin.

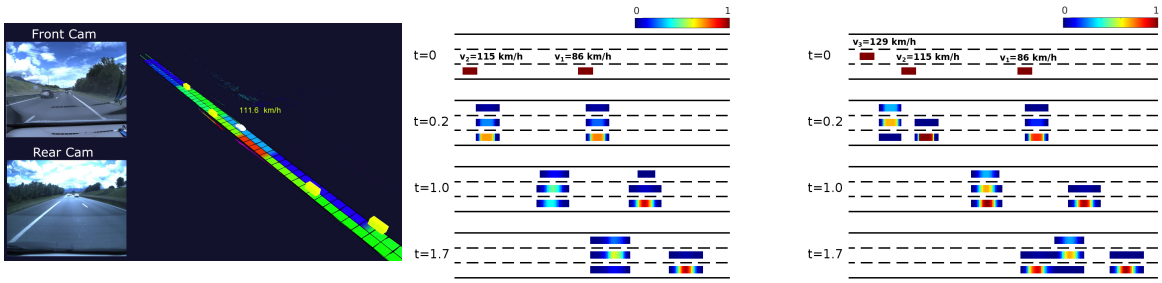


Figure 12. a) Slice of a cost function obtained from real demonstrated data superimposed on the road on a highway scene. Red indicates high cost, green intermediate, and blue low. b) and c) Prediction of two prototypical simulated traffic scenes with the framework from [26]. We show the predicted occupancy probabilities over a discretization of the road for different timesteps.

The objective is to develop a framework for modeling road intersections using relevant functional areas, which can be exploited by an autonomous vehicle for safely crossing the intersection. These functional areas try to capture various characteristics such as crossing, merging or approaching areas, the car dynamics when moving in such areas, or the related uncertainty. We made the assumption that such a functional space representation can be stored in a map and can be constructed using observed trajectories. This work is performed in the scope of the PhD thesis of Mathieu Barbier, which is supported by a CIFRE fellowship with Renault.

In a preliminary work done with map by Renault, it has been observed that the information stored in a map does not fully match with the real motions executed by cars within a given intersection. The differences between the stored and the real path might be important. This difference is not due to error during the map creation, but rather to other constraints related to the driving action (e.g. visibility, dynamics). Such a difference leads to serious difficulties at the level of the autonomous driving decision-making process.

Constructing a functional model, requires to first analyze the topological and dynamics structure of an intersection, and in a second step to imagine how it would be possible to extract such type of information from maps and observed trajectories. Two main structures seem to emerge from this study:

- Merging and Crossing points, areas where two cars are the most likely to collide.
- Approaching areas, areas where drivers are most likely to show constant driving behaviors.

In order to learn motions patterns of multiple cars, we have chosen to train Gaussian processes [81] [93] using simulated data set generated using the SCANER™system. The resulting distribution is segmented using different threshold, in order to find approaching areas and to combine pairs of such areas for constructing overlapping areas. The correlation between this discretization and both real and simulated velocity profiles has been shown by the experimental results, see Fig. 13 . The approach has been published at IEEE ITSC 2016 [18].

We recently started to make use of a Random Forest classifier to connect features of trajectories with an intended maneuver (stop, pass, yield) and to take advantage of the discretization. This research work will be continued in 2017.

7.3. Robust state estimation (Sensor fusion)

7.3.1. Visual-inertial structure from motion: observability properties and state estimation

Participant: Agostino Martinelli.

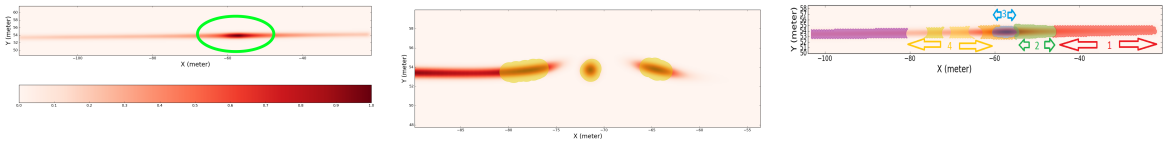


Figure 13. Different step of the framework to discretize the space : a) Map created with prediction from set of GPs, the highlighted area has a high mean probability b) Segmentation of crossing and merging areas, in red the probability of two cars being in the same position and in yellow where this probability is higher than the threshold c) Discretization of approaching area

This research is the follow up of our investigations carried out during the last three years. The main results obtained this year regard the following three topics:

1. Exploitation of the closed form solution introduced in [70] in the framework of Micro Aerial Vehicle (MAV) navigation;
2. Introduction of a new method for simultaneous localization and Gyroscope calibration;
3. Analytic solution of the Unknown Input Observability problem (UIO problem) in the nonlinear case.

Regarding the first two topics, we successfully implemented a new method for MAV localization and mapping, on the aerial vehicles available at the Vision and Perception lab at the university of Zurich [0]. This method is based on our previous closed form solution recently introduced in [70]. The practical advantage of this solution is that it is able to determine several physical quantities (e.g, speed, orientation, absolute scale) by only using the measurements provided by a monocular camera and an Inertial Measurement Unit (IMU) during a short interval of time (about 3 seconds). In other words, an initialization is not requested to determine the aforementioned physical quantities. This fact has a fundamental importance in robotics and it is novel with respect to all the state of the art approaches for visual-inertial sensor fusion, which use filter-based or optimization-based algorithms. Due to the nonlinearity of the system, a poor initialization can have a dramatic impact on the performance of these estimation methods.

Finally, by further studying the impact of noisy sensors on the performance of the closed-form solution introduced in [70], we found that this performance is very sensitive to the gyroscope bias. For, we developed a powerful and simple optimization approach to remove this bias. This method has been tested in collaboration with the vision and perception team in Zurich (in the framework of the ANR-VIMAD) and published on the IEEE Robotics and Automation Letters [12]. Additionally, these results have been presented at the International Conference on Robotics and Automation [21].

Regarding the third topic, we still considered the problem of deriving the observability properties of the visual-inertial structure from motion problem when the number of inertial sensors is reduced. This case corresponds to solve a problem that in control theory is known as the Unknown Input Observability (UIO). This problem was still unsolved in the nonlinear case. In [71] we introduced a new method able to provide sufficient conditions for the state observability. On the other hand, this method is based on a state augmentation. Specifically, the new extended state includes the original state together with the unknown inputs and their time-derivatives up to a given order. Then, the method introduced in [71] is based on the computation of a codistribution defined in the augmented space. This makes the computation necessary to derive the observability properties dependent on the dimension of the augmented state. Our effort to deal with this fundamental issue, was devoted to separate the information on the original state from the information on its extension. Last year, we fully solved this problem in the case of a single unknown input [73], [72]. This year we solved the problem for any number of unknown inputs. We presented this solution at the university of Pisa in June and at the university of Rome, Tor Vergata, in December.

[0]This is the partner of the ANR project VIMAD, in charge of the experiments

7.4. Motion-planning in human-populated environment

We explore motion planning algorithms to allow robots/vehicles to navigate in human populated environment, and to predict human motions.

We have proposed a novel planning-based motion prediction approach [27] which addresses the weaknesses of the previous state-of-the-art motion prediction technique [56], namely *High computational complexity* and *Limited ability to model the temporal evolution along the predicted path*. In 2016, this work has evolved in two new directions, which are prediction of pedestrian behaviors in urban environments and mapping of human flows. We also started to investigate the navigation of a telepresence robot in collaboration with the GIPSA Lab. These work are presented here after.

7.4.1. Urban Behavioral Modeling

Participants: Pavan Vasishta, Raphael Frisch, Anne Spalanzani.

The objective of modeling urban behavior is to predict the trajectories of pedestrians in towns and around car or platoons. We aim to integrate various entities of urban environments such as crosswalks, sidwalks, points of interest, but also characteristics of mobile obstacles (such cars and platoons) and proxemics in order to build a costmap that will show how pedestrians are driven around the ego-vehicle. This work is in the scope of the VALET project and the PhD of Pavan Vasishta (in collaboration with the Inria team Pervasive Interaction). It started in february 2016. Furthermore, a collaboration with the Laboratory of Psychology and NeuroCognition of Grenoble has been initiated to ground interaction and personal space models in experimental data from psychology.

Figure 14. Illustration of (a) Flow-grid mapping in a two-corridor environment where 200 moving pedestrians turns (b) A path-planning computed with different cost functions in this populated environment*

7.4.2. Modeling human-flows from robot(s) perception

Participants: Olivier Simonin, Jacques Saraydaryan, Fabrice Jumel.

In order to deal with robot navigation in dense human populated environments, eg. in flows of humans, we started to investigate the problem of mapping these flows. The challenge is to build such a map from robot perceptions while robots move autonomously to perform some tasks. We developped a counting-based grid model which computes likelihoods of human presence and motion direction in each cell, see red vectors in Fig. 14 .a (this is a statistical learning of repetitive human motions). We extended the flow grid model with a human motion predictive model based on the Von Misses motion patern, allowing to "accelerate" the flow grid mapping, see blue vectors in Fig. 14 .a.

Then we explored how path-planning can benefit of such a flow grid, that is taking into account the risk for a robot to encounter humans in opposite direction. We first implement the Flow-Grid model in a simulator built upon PedSim and ROS tools, allowing to simulate mobile robots, crowd of pedestrians and sensors to detect their motion. Then, we compared three A*-based path-planning models using different levels of information about human presence: non-informed, a grid of human presence likelihood proposed by Tiplaldi [95] and our grid of human motion likelihood. Simulations involving 200 moving persons and 4 collaborative robots allowed to test simultaneously the mapping of human motions and the related path-planning. The different kind of paths obtained are illustrated in Fig. 14 .b, showing the ability of the flow-grid based A* to avoid to cross areas with a possible opposite human flow. These results have been presented at RSS workshop DEMUR [30].

We also started some experiments with our mobile indoor robots (incl. the Pepper) in human populated environments, see [30]. We plan to demonstrate the efficiency of the approach by participating to the new Pepper-league of the Robocup@Home competition, over the future period 2017-2020.

7.4.3. Navigation of telepresence robots

Participants: Olivier Simonin, Anne Spalanzani, Gerard Bailly [GIPSA, CNRS, Grenoble], Rémi Cambuzat.

In 2016 we obtained with the team of Gérard Bailly, from GIPSA/CNRS Grenoble, a regional support for the TENSIVE project. It funds the PhD thesis of Remi Cambuzat on immersive teleoperation of telepresence robots for verbal interaction and social navigation, started in October 2016. Chroma is focusing on the problem of social navigation, and more paricularly on the balance between human commands and autonomous navigation. Two issues are adressed : how to understand the expected direction given by the pilot to the telepresence robot, in order to ease the workload of the pilot ? how to assist the navigation, from embedded processes and sensors on the robot, while following the expected behavior given by the remote pilot ?

First months of the thesis concerned the building of a specific state-of-the-art, the formalization of some experimental scenarios, and the study of the Pepper robot capabilities in this scientific challenge.

7.5. Decision Making in Multi-Robot Systems

7.5.1. Multi-robot path-planning and patrolling

7.5.1.1. Patrolling under connectivity constraints

Participants: Olivier Simonin, Anne Spalanzani, Mihai Popescu, Fabrice Valois [Inria, Agora (ex Urbanet) team].

Patrolling is mainly used in situations where the need of repeatedly visiting certain places is critical. In this work, we consider a deployment of fixed targets, eg. wireless sensors, that several robots are in charge of patrolling while they have to maintain their (periodic) connectivity in order to collect and bring data up to a sink node. We have shown that this is fundamentally a problem of vertex coverage with bounded simple cycles (CBSC). We offered a formalization of the CBSC problem and proved it is NP-hard and at least as hard as the Traveling Salesman Problem (TSP). Then, we provided and analyzed heuristics relying on clusterings and geometric techniques. The proposed approach relies on two steps: the first one partitions the vertices, the second one computes hamiltonian cycles on each partition. We implemented two classic hamiltonian cycle heuristics, one is based on Minimum Spanning Trees computations and the other on Christofides algorithm. Comparisons on randomly-generated graphs showed that the Christofides algorithm computes shorter cycles. This work, started in the Master internship of Mihai-Ioan Popescu, now continuing as PhD student in Chroma, has been published in 2016 in [25]. Work is now focusing on the problem of synchronizing robots to meet at intersection nodes between the cycles.

Another important element of this work is the construction of a new collaboration with the team of Gabriela Czibula in Babes-Bolyai University at Cluj-Napoca (Romania). It will focus on optimization and online adaptation of the multi-cycle patrolling with machine learning (RL) techniques in order to deal with the arrival of new targets in the environment. We obtained, in the end of 2016, a french-romanian PHC [0] bilateral project, called DRONEM, funding students and researchers exchanges during two years.

[0] Hubert Curien Partnership

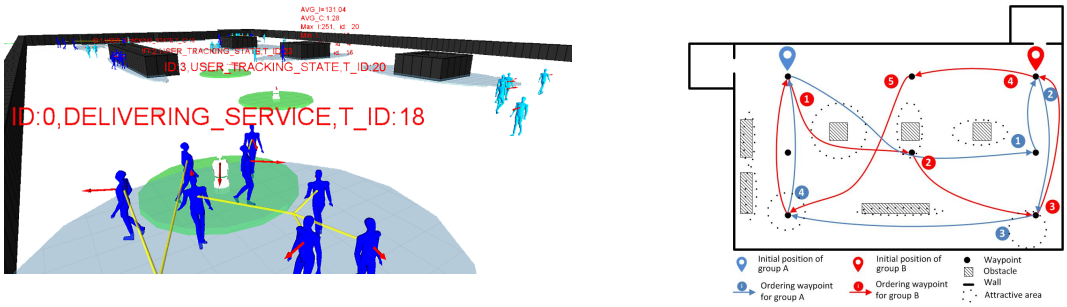


Figure 15. a) Simulator for dynamic patroling of people based on PedSim. b) Scenario of the 200 pedestrians moving along two predefined paths plus local attractors and randoms moves.

7.5.1.2. Patrolling moving people (dynamic patrolling)

Participants: Jacques Saraydaryan, Fabrice Jumel, Olivier Simonin.

In the context of service robotics, we address the problem of serving people by a set of collaborating robots, that is to delever regularly services to moving people. We showed that the problem can be formally expressed as a dynamic patrolling task. We call it the robot-waiters problem, where robots have to regularly visit all the moving persons (to deliver objects/information). In the publication [87], we proposed different criteria and metrics suitable to this problem, by considering not only the time to patrol all the people but also the equity of the delivery. We proposed and compared four algorithms, two are based on standard solutions to the static patrolling problem and two are defined according the specificity of patrolling moving entities, in particular greedy-based solutions on distance and idleness people information. In order to limit robot traveled distances, the last approach introduces a clustering heuristic to identify groups among people. To compare algorithms and to prepare real experiments we evaluated performances by using our simulator (combining PedSim and ROS). The simulator and the scenrio test - paths followed by humans - are illustrated in figure 15 .a and 15 .b. Experimental results show the efficiency of the specific new approaches over standard (static patrolling) approaches. We also analysed the influence of the number of robots on the performances, showing a convergence of performances when it grows. See [87] and extensions in 2016 [28].

We are currently developping new algorithms using the mapping and prediction of human flows based on the work presented in section 7.4.2 to allow robots to predict where human (groups) will move (under hypothesis of repetitive behaviors).

7.5.1.3. Global-local optimization in autonomous multi-vehicles systems

Participants: Olivier Simonin, Jilles Dibangoye, Laetitia Matignon, Florian Peyreron [VOLVO Group, Lyon], Guillaume Bono, Olivier Buffet [Inria Nancy Grand Est], Mohamed Tlig [IRT-Systemx, Paris].

We address transport and traffic management problems with driverless vehicles. We mainly study how local decisions can improve complexity of global (planning) solutions.

A previous work carried in the PhD of M. Tlig [96] concerned stop-free crossing roads with driverless vehicles. We explored distributed algorithms to optimize the global traffic in the road network (time and energy), based on Hill-Climbing techniques, so as to go from a synchronization within each intersection to the synchronization of a network. Experiments in simulation showed that proposed algorithms can efficiently optimize the initial decentralized solution, while keeping its properties (only the temporal phase for crossing in each intersection is modified). In 2016 we extended the experimental study, which was published in the RIA revue [13] and submited to an international journal.

In 2016, we started a new cooperation with the VOLVO Group, in the context of the INSA-VOLVO Chair. It funds the PhD thesis of G. Bono which deals with global-local optimization under uncertainty for goods distribution using a fleet of autonomous vehicles. First months of the thesis focused on building i) a state of the art about online pick-up and delivery solutions with a fleet of autonomous vehicles and ii) defining formally the scenario and hypothesis of the considered problem.

7.5.2. Anytime algorithms for multi-robot cooperation

7.5.2.1. Complex scenes observation

Participants: Olivier Simonin, Laetitia Matignon, Christian Wolf [LIRIS, INSA Lyon], Simon Bultmann [internship], Stefan Chitic.

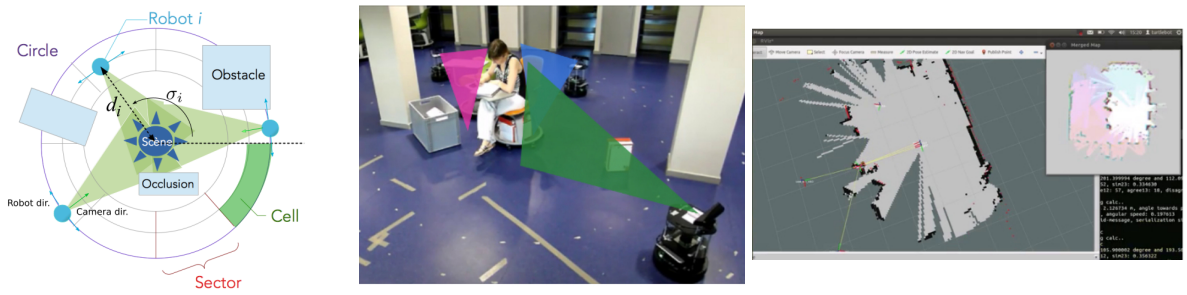


Figure 16. Illustrations (a) Concentric navigation model. (b) Experimental setup and multi-robot mapping with Turtlebot 2.

Solving complex tasks with a fleet of robots requires to develop general strategies that can decide in real time (or time-bounded) efficient and cooperative actions. This is particularly challenging in complex real environments. To this end, we explore anytime algorihms and adaptive/learning techniques.

The "Crome" [0] project [0] is motivated by the exploration of the joint-observation of complex (dynamic) scenes by a fleet of mobile robots. In our current work, the considered scenes are defined as a sequence of activities, performed by a person in a same place. Then, mobile robots have to cooperate to find a spatial configuration around the scene that maximizes the joint observation of the human pose skeleton. It is assumed that the robots can communicate but have no map of the environment and no external localization.

To attack the problem, in cooperation with colleagues from vision (C. Wolf, Liris), we proposed an original concentric navigation model allowing to keep easily each robot camera towards the scene (see fig. 16 .a). This model is combined with an incremental mapping of the environment and exploration guided by meta-heuristics in order to limit the complexity of the exploration state space. We developped a simulator that uses real data from real human pose captures to simulate dynamic scene and noise in sensor information. A video presenting the simulator interface and showing the incremental exploration and mapping can be found at . Results have been published in 2016 in [20] (ICTAI). It compares the variants of the approach and shows its features such as adaptation to the dynamic of the scene and robustness to the noise in the observations.

We have also developped an experimental framework for the circular navigation of several Turtlebot2 robots around a scene, presented in figure 16 .b. Especially, given that we assume in our work that robots have no map of the environment, we implemented a cooperative multi-robot mapping based on the merging of occupancy grid maps. Robots are individually building and communicating to other robots their local maps. Each robot tries to align these local maps to compute a joint, global representation of the environment. We carried out

[0]Coordination d'une flottille de robots mobiles pour l'analyse multi-vue de scènes complexes
[0]Funded by an INSA BQR in 2014-2015 (led by O. Simonin) and a LIRIS transversal project in 2016-2017 (led by L. Matignon)

the map-merging by adapting several methods known in literature [86] to our specific topology, i.e. the single hypothesis of a common center point (the scene) shared by robots. We compared the methods in real-world multi-robot scenarios (see Simon Bultmann's internship report).

7.5.2.2. Middleware for open multi-robot systems

Participants: Stefan Chitic, Julien Ponge [CITI, Dynamid], Olivier Simonin.

Multi-robots systems (MRS) require dedicated tools and models to face the complexity of their design and deployment (there is no or very limited tools/middleware for MRS). In the context of the PhD work of S. Chitic, we address the problem of neighbors and service discovery in an ad-hoc network formed by a fleet of robots. Robots needs a protocol that is able to constantly discover new robots in their coverage area. This led us to propose a robotic middleware, SDfR, that is able to provide service discovery. This protocol is an extension of the Simple Service Discovery Protocol (SSDP) used in Universal Plug and Play (UPnP) to dynamic networks generated by the mobility of the robots. Even if SDfR is platform independent, we proposed a ROS integration in order to facilitate the usage. We evaluated a series of overhead benchmarking across static and dynamic scenarios. Eventualy, we experimented some use-cases where our proposal was successfully tested with Turtlebot 2 robots. Results have been pubished in [19]. In 2016, the work continued by the definition of a timed automata based design and validation tool-set, that offers a framewok to formalize and implement multi-robot behaviors and to check some (temporal) properties.

7.5.3. Sequential decision-making under uncertainty

The holy grail of Artificial Intelligence (AI)—creating an agent (e.g., software, robot or machine) that comes close to mimicking and (possibly) exceeding human intelligence—remains far off. But past years have seen breakthroughs in agents that can gain abilities from experience with the environment, providing significant advances in the society and the industries including: health care, autonomous driving, recommender systems, etc. These advances are partly due to single-agent planning and (deep) reinforcement learning, that is, AI research subfields in which the agent can describe its world as a Markov decision process. Some stand-alone planning and reinforcement learning (RL) algorithms (e.g., Policy and Value Iteration, Q-learning) are guaranteed to converge to the optimal behavior, as long as the environment the agent is experiencing is Markovian. Although Markov decision processes provide a solid mathematical framework for single-agent planning and RL, they do not offer the same theoretical grounding in multi-agent systems, that is, groups of autonomous, interacting agents sharing a common environment, which they perceive through sensors and upon which they act with actuators. Multi-agent systems are finding applications everywhere today. At home, in cities, and almost everywhere, we are surrounded by a growing number of sensing and acting machines, sometimes visibly (e.g., robots, drones, cars, power generators) but often imperceptibly (e.g., smartphones, televisions, vacuum cleaners, wash- ing machines). Before long, through the emergence of a new generation of communication networks, most of these machines will be interacting with one another through the internet of things. In contrast to single-agent systems, when multiple agents interact with one another, how the environment evolves depends not only upon the action of one agent but also on the actions taken by the other agents, rendering the Markov property invalid since the environment is no longer stationary. In addition, a centralized (single-agent) control authority is often inadequate, because agents cannot (e.g., due to communication cost, latency or noise) or do not want (e.g., in competitive or strategic settings) to share all their information all the time. This raises a simple fundamental question: how to design a general algorithm for efficiently computing rational policies for a group of cooperating or competing agents in spite of stochasticity, limited information and computational resources? The remainder of this section points out some of the main results of the year to this question as well as ongoing projects.

7.5.3.1. Optimally solving cooperative games as continuous Markov decision processes

Participants: Jilles S. Dibangoye, Olivier Buffet [Inria Nancy], Christopher Amato [Univ. New Hampshire], François Charpillet [Inria Nancy, Larsen team].

Decentralized partially observable Markov decision processes (Dec-POMDPs) provide a general model for decision-making under uncertainty in decentralized settings, but are difficult to solve optimally (NEXP-Complete). As a new way of solving these problems, we introduce the idea of transforming a Dec-POMDP into a continuous-state deterministic MDP with a piecewise-linear and convex value function. This approach makes use of the fact that planning can be accomplished in a centralized offline manner, while execution can still be decentralized. This new Dec-POMDP formulation, which we call an occupancy MDP, allows powerful POMDP and continuous-state MDP methods to be used for the first time. To provide scalability, we refine this approach by combining heuristic search and compact representations that exploit the structure present in multi-agent domains, without losing the ability to converge to an optimal solution. In particular, we introduce a feature-based heuristic search value iteration (FB-HSVI) algorithm that relies on feature-based compact representations, point-based updates and efficient action selection. A theoretical analysis demonstrates that FB-HSVI terminates in finite time with an optimal solution. We include an extensive empirical analysis using well-known benchmarks, thereby demonstrating that our approach provides significant scalability improvements compared to the state of the art. This work has been published in JAIR journal [11].

7.5.3.2. Optimally solving two-person zero-sum partially observable stochastic games: a convex optimization approach

Participants: Jilles S. Dibangoye, Olivier Buffet [Inria Nancy], Mihai Indricean [INSA Lyon internship].

This work proposes a novel theory and algorithms to optimally solving a two-person zero-sum POSGs (zs-POSGs). That is a general framework for modeling and solving two-person zero-sum games (zs-Games) with imperfect information. Our theory builds upon the result demonstrating that the original problem is reducible to a zs-Game—but now with perfect information. In this form, we show that the dynamic programming theory applies. In particular, we extended Bellman equations [39] for zs-POSGs, and coined them maximin (resp. minimax) equations. Even more importantly, we demonstrated Von Neumann & Morgenstern's minimax theorem [99] [100] holds in zs-POSGs. We further proved that value functions—solutions of maximin (resp. minimax) equations—yield special structures. More specifically, the maximin value functions are convex whereas the minimax value functions are concave. We also showed how our results apply to more restrictive settings, essentially leading to more concise information. Together these findings allow us to introduce a key algorithm avoiding exhaustive enumeration of doubly exponentially many pure strategies, as suggested so far. We further illustrate the use of our algorithm through numerical examples.

7.5.3.3. Decentralized Markov decision processes in open systems: models and first algorithms

Participants: Jilles S. Dibangoye, Abdel-Illah Mouaddib [Univ. Caen Basse-Normandie], Jonathan Cohen [Univ. Caen Basse-Normandie].

Many real-world multiagent applications, e.g., rescue operations, require to dynamically assemble or disassemble teams needed to provide a service based on agents entering or quitting the system. While Decentralized Partially Observable Markov Decision Processes (Dec-POMDPs) formalize decision-making by multiple agents, they fail to exploit the team flexibility. Queueing models can formalize birth-death processes by which agents enter or exit a team, but they fail to capture multiagent planning under uncertainty. This work, in the context of the PhD work of J. Cohen, introduces a new model synthesized from Dec-POMDPs and birth-death processes, called open Dec-POMDPs. The primary result is the proof that the latter is NEXP-Complete. Exploiting the team flexibility, enables us to present a best-response dynamics' algorithm, which can dynamically adapt to agents entering or quitting a team and computes local optimum solutions.

7.5.3.4. Does randomization makes cooperative multi-agent planning easier?

Participant: Jilles S. Dibangoye.

These recent years have seen significant progress in multi-agent planning problems formulated as decentralized partially observable Markov decision processes (Dec-POMDPs). In state-of-the-art algorithms, agents use policies that do not employ random devices, i.e., deterministic policies, which are simple to handle and to implement, and yet are good candidates to be optimal. Integer linear programming (ILP) or constraint optimization programming (COP) can formalize the search for deterministic policies, unfortunately their worst case complexity (NP-Complete) suggest to be little hope for optimally solving real-world instances. In this

paper, we show—for the first time—that the randomization allows us to use linear programming (LP) instead of ILP while preserving optimality, which drops the worst-case complexity from NP to P. Specifically, we introduce the first linear programs for incrementally approaching the optimal value function starting from upper- and lower-bound functions. We further extends state-of-the-art algorithm for Dec-POMDPs to use randomized policies. Finally, empirical results demonstrate significant improvements in time needed to find an ε-optimal solution on all tested benchmarks.

7.5.3.5. Reinforcement learning approach for active perception using multiple robots

Participants: Jilles S. Dibangoye, Jacques Saradaryan, Laëtitia Matignon, Trad Ahmed Yahia [Master Internship], Lorcan Charonnat [Internship INSA], Yuting Zhang [Internship INSA], Yifan Xiong [Internship INSA].

We consider cooperative, decentralized stochastic control problems represented as a decentralized partially observable Markov decision process. A critical issue that limits the applicability of that setting to realistic domains is how agents can learn to act optimally by interacting with the environment and with one another, given only an incomplete knowledge about the model. Reinforcement learning has previously been applied to decentralized decision making with a focus on distributed methods, which often results in suboptimal solutions. On the contrary, we build upon the idea that plans that are to be executed in a decentralized fashion can nonetheless be formulated in a centralized manner using a generative model of the environment. Following this line of thought, we propose the first (centralized) reinforcement learning algorithm for computing the optimal Q-value functions for cooperative, decentralized stochastic control problems. Experiments show our approach can learn to act optimally in many domains from the literature. We currently investigate robotic applications of this approach, including unknown scene reconstruction by a fleet of mobile robots.

EXMO Project-Team

6. New Results

6.1. Ontology matching and alignments

6.1.1. Evaluation

Participant: Jérôme Euzenat [Correspondent].

Since 2004, we run the Ontology Alignment Evaluation Initiative (OAEI) which organises evaluation campaigns for assessing the degree of achievement of actual ontology matching algorithms [3].

This year, we used again our generator for generating a new version of benchmarks. The Alignment API was used for manipulating alignments and evaluating results [8].

The participating systems and evaluation results were presented in the 11th Ontology Matching workshop [14], [15], held Kobe (JP). More information on OAEI can be found at http://oaei.ontologymatching.org/.

6.1.2. Algebras of alignment relations

Participants: Manuel Atencia Arcas, Jérôme Euzenat [Correspondent], Armen Inants.

Qualitative calculi are central in qualitative binary constraint satisfaction problems. All these qualitative calculi share an implicit assumption that the universe is homogeneous, i.e., consists of objects of the same kind. However, objects of different kinds may also entertain relations. Many applications discriminate between different kinds of objects. For example, some spatial models discriminate between regions, lines and points, and different relations are used for each kind of objects. In ontology matching, qualitative calculi were shown useful for expressing alignments between only one kind of entities, such as concepts or individuals. However, relations between individuals and concepts, which impose additional constraints, are not exploited.

We introduced modularity in qualitative calculi and provided a methodology for modeling qualitative calculi with heterogeneous universes [5]. It is based on a special class of partition schemes which we call modular. For a qualitative calculus generated by a modular partition scheme, we can define a structure that associates each relation symbol with an abstract domain and codomain from a Boolean lattice of sorts. A module of such a qualitative calculus is a sub-calculus restricted to a given sort, which is obtained through an operation called relativisation to a sort. Of a greater practical interest is the opposite operation, which allows for combining several qualitative calculi into a single calculus. We defined an operation called combination modulo glue, which combines two or more qualitative calculi over different universes, provided some glue relations between these universes. The framework is general enough to support most known qualitative spatio-temporal calculi.

In 2012, we introduced a semantics supporting confidences in correspondences as weights. However, it introduced a discontinuity between weighted and non-weighted interpretations. Moreover, it does not provide a calculus for reasoning with weighted ontology alignments. We introduced a calculus for such alignments [11] provided by an infinite relation-type algebra, the elements of which are weighted taxonomic relations. In addition, it approximates the non-weighted case in a continuous manner.

This work has been part of the PhD of Armen Inants [5] partially funded by the LINDICLE project (§7.1.1).

6.2. Data interlinking

The web of data uses semantic web technologies to publish data on the web in such a way that they can be interpreted and connected together. It is thus important to be able to establish links between these data [7], both for the web of data and for the semantic web that it contributes to feed. We consider this problem from different perspectives.

6.2.1. Interlinking cross-lingual RDF data sets

Participants: Tatiana Lesnikova, Jérôme David [Correspondent], Jérôme Euzenat.

RDF data sets are being published with labels that may be expressed in different languages. Even systems based on graph structure, ultimately rely on anchors based on language fragments. In this context, data interlinking requires specific approaches in order to tackle cross-lingualism. We proposed a general framework for interlinking RDF data in different languages and implemented two approaches: one approach is based on machine translation, the other one takes advantage of multilingual references, such as BabelNet.

This year, we evaluated machine translation for interlinking concepts, i.e., generic entities named with a common noun or term, as opposed to individual entities. In previous work, the evaluated method has been applied on named entities. We conducted two experiments involving different thesauri in different languages. The first experiment involved concepts from the TheSoz multilingual thesaurus in three languages: English, French and German. The second experiment involved concepts from the EuroVoc and AGROVOC thesauri in English and Chinese respectively. We demonstrated that machine translation can be beneficial for cross-lingual thesauri interlining independently of a dataset structure [12].

This work has been part of the PhD of Tatiana Lesnikova [6] developed in the LINDICLE project (§7.1.1).

6.2.2. Uncertainty-sensitive reasoning for inferring sameAs facts in Linked Data

Participants: Manuel Atencia Arcas [Correspondent], Jérôme David.

A major challenge in data interlinking is to design tools that effectively deal with incomplete and noisy data, and exploit uncertain knowledge. We modelled data interlinking as a reasoning problem with uncertainty. For that purpose, we introduced a probabilistic framework for modelling and reasoning over uncertain RDF facts and rules that is based on the semantics of probabilistic Datalog. We have designed an algorithm, ProbFR, based on this framework. Experiments on real-world datasets have shown the usefulness and effectiveness of our approach for data linkage and disambiguation [9].

This work was carried out in collaboration with Mustafa Al-Bakri and Marie-Christine Rousset (LIG).

6.2.3. Tableau extensions for reasoning with link keys

Participants: Manuel Atencia Arcas [Correspondent], Jérôme Euzenat, Maroua Gmati.

Link keys allow for generating links across datasets expressed in different ontologies (see §3.3). But they can also be thought of as axioms in a description logic. As such, they can contribute to infer ABox axioms, such as links, or terminological axioms and other link keys. Yet, no reasoning support existed for link keys. We extended the tableau method designed for ALC to take link keys into account [10]. We showed how this extension enables combining link keys with classical terminological reasoning with and without ABox and TBox and generate non trivial link keys.

IMAGINE Project-Team

6. New Results

6.1. User-centered Models for Shapes and Shape Assemblies

- **Scientist in charge**: Stefanie Hahmann.
- **Other permanent researchers**: Marie-Paule Cani, Jean-Claude Léon, Damien Rohmer.

Our goal, is to develop responsive shape models, i.e. 3D models that respond in the expected way under any user action, by maintaining specific application-dependent constraints (such as a volumetric objects keeping their volume when bent, or cloth-like surfaces remaining developable during deformation, etc). We are extending this approach to composite objects made of distributions and/or combination of sub-shapes of various dimensions.

6.1.1. Shape analysis

Figure 3. Left: Illustration of comparative study of 3D medial axis quality in [21]. Right: Hierarchies for similar shapes (dancers) in different poses to show that the proposed hierarchy is stable under articulation [22]. Coarser levels of the hierarchy are consistent even if finer levels are added in the presence of finer details. Also, note that the hierarchy is retained even with occlusion: The pink level of the left arm of the first dancer is occluded, but the blue level begins as it should.

Within the post-doc of Thomas Delame a comparative study between 3D skeletonization methods has been achieved. This work has been summarized as a Eurographics State of the Art [15]. Moreover, a comparative study of the quality between 3D medial axis was assessed and published in Vision, Modeling and Visualization [21].

We developed a multilevel analysis method of 2D shapes and used it to find similarities between the different parts of a shape [22]. Such an analysis is important for many applications such as shape comparison, editing, and compression. Our robust and stable method decomposes a shape into parts, determines a parts hierarchy, and measures similarity between parts based on a salience measure on the medial axis, the Weighted Extended Distance Function, providing a multi-resolution partition of the shape that is stable across scale and articulation. Comparison with our extensive user study on the MPEG-7 database, see below, demonstrates that our geometric results are consistent with user perception. This work has been accomplished within a collaboration between S. Hahmann, Kathryn Leonard (CSUCI), and Geraldine Morin and Axel Carlier (IRIT, ENSEEIHT). K. Leonard was visiting the Imagine team during several month in 2016 as an invited professor funded by the ERC Expressive grant.

We also conducted a large user-study and made the results available throughout an open access data base: The 2D Shape Structure database [9] is a public, user-generated dataset of 2D shape decompositions into a hierarchy of shape parts with geometric relationships retained. It is the outcome of a large-scale user study obtained by crowdsourcing, involving over 1200 shapes in 70 shape classes, and 2861 participants. A total of 41953 annotations has been collected with at least 24 annotations per shape. For each shape, user decompositions into main shape, one or more levels of parts, and a level of details are available. This database reinforces a philosophy that understanding shape structure as a whole, rather than in the separated categories of parts decomposition, parts hierarchy, and analysis of relationships between parts, is crucial for full shape understanding. We provide initial statistical explorations of the data to determine representative ("mean") shape annotations and to determine the number of modes in the annotations. The primary goal of this work is to make this rich and complex database openly available (through the website http://2dshapesstructure.github.io), providing the shape community with a ground truth of human perception of holistic shape structure. This paper has received the « Reproducibility Award » (http://www.reproducibilitystamp.com).

6.1.2. Surface design

Figure 4. Left: Illustration of results of [14].

Recent surface acquisition technologies based on micro-sensors produce three-space tangential curve data which can be transformed into a network of space curves with surface normals. In the thesis of Tibor Stanko, which is funded by the CEA-LETI, we dispose such a mobile device equipped with several micro-sensors. The goal of the thesis is to develop surface acquisitions methods using this equipped mobile device. As a first step, we address the theoretical and algorithmic problem of surfacing an arbitrary closed 3D curve network with given surface normals. Thanks to the normal vector input, the patch finding problem can be solved unambiguously and an initial piecewise smooth triangle mesh is computed. The input normals are propagated throughout the mesh. Together with the initial mesh, the propagated normals are used to compute mean curvature vectors. We then compute the final mesh as the solution of a new variational optimization method based on the mean curvature vectors. The intuition behind this original approach is to guide the standard Laplacian-based variational methods by the curvature information extracted from the input normals. The normal input increases shape fidelity and allows to achieve globally smooth and visually pleasing shapes. This work has been presented at Eurographics 2016 as a short paper [25] and GTMG 2016 [26] and has been published as a journal paper [14].

Morse-Smale complexes have been proposed to visualize topological features of scalar fields defined on manifold domains. Herein, three main problems have been addressed in the past: (a) efficient computation of the initial combinatorial structure connecting the critical points; (b) simplification of these combinatorial structures; (c) reconstruction of a scalar field in accordance to the simplified Morse-Smale complex. The PhD thesis of Leo Allemand-Giorgis faces the third problem by proposing a novel approach for computing a scalar field coherent with a given simplified MS complex that privileges the use of piecewise polynomial functions [31]. Based on techniques borrowed from shape preserving design in Computer Aided Geometric Design, our method constructs the surface cell by cell using piecewise polynomial curves and surfaces.

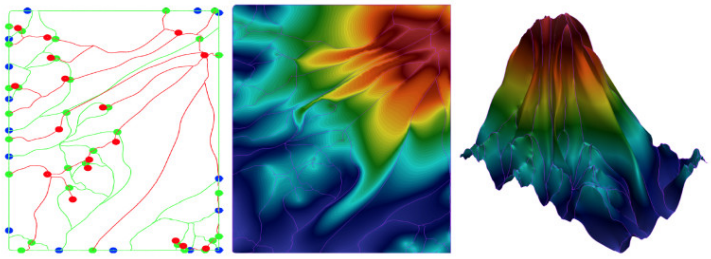


Figure 5. The terrain data set of Mt Rainier: Surface reconstruction (c) with contour lines (b) from the MS complex with 69 critical points(a) [31].

6.2. Motion & Sound Synthesis

- **Scientist in charge**: François Faure.
- **Other permanent researchers**: Marie-Paule Cani, Damien Rohmer, Rémi Ronfard.

Animating objects in real-time is mandatory to enable user interaction during motion design. Physically-based models, an excellent paradigm for generating motions that a human user would expect, tend to lack efficiency for complex shapes due to their use of low-level geometry (such as fine meshes). Our goal is therefore two-folds: first, develop efficient physically-based models and collisions processing methods for arbitrary passive objects, by decoupling deformations from the possibly complex, geometric representation; second, study the combination of animation models with geometric responsive shapes, enabling the animation of complex constrained shapes in real-time. The last goal is to start developing coarse to fine animation models for virtual creatures, towards easier authoring of character animation for our work on narrative design.

6.2.1. Physically-based models

We proposed a survey on the exhisting adaptative physically based models in Computer Graphics in collaboration with IST Austria, University of Minnesota, and NANO-D Inria team. Models where classified according to the strategy they use for adaptation, from time-stepping and freezing techniques to geometric adaptivity in the form of structured grids, meshes, and particles. Applications range from fluids, through deformable bodies, to articulated solids. The survey has been published as a Eurographics state of the art [13].

In collaboration with the *Reproduction et Développement des Plantes* Lab (ENS Lyon), we proposed a realistic three-dimensional mechanical model of the indentation of a flower bud using the SOFA library, in order to provide a framework for the analysis of force-displacement curves obtained experimentally [12].

6.2.2. Simulating paper material with sound

We developed within the PhD from Camille Schreck a dedicated approach to model a real time deforming virtual sheet of paper. First we developed a geometrical model interleaving physically based elastic deformation with a dedicated geometrical correction and remeshing. The key idea consists in modeling the surface using a set of generalized cones able to model developable ruled surfaces instead of the more traditional set of triangles. This surface can handle length preservation with respect to the 2D pattern, and permanent non smooth crumpling appearance. This geometrical model published in ACM Transactions on Graphics in Dec. 2015 [5] has been presented at ACM SIGGRAPH this summer and is currently under investigation to be part of Inria Showroom. This model has then been extended to real time sound synthesis of crumpled paper within the collaboration with Doug James (Stanford University). This method was the first to handle real-time shape dependent sound synthesis. During the interactive deformation, sudden curvature changes and friction are detected. These sound generating events are then associated to a geometrical region where the sound resonates and defined efficiently using previous geometrical model. Finally, the sound is synthesized using a

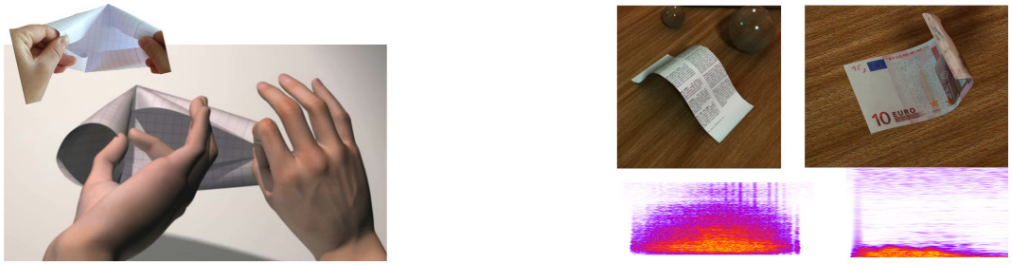


Figure 6. Left: Geometrical deformation using our geometrical model from [5]. Right: Various paper deformation and type leading to different synthesized sounds [24].

pre-recorded sound data base of crumple and friction events sorted with respect to the resonator region size. This work has been published at Symposium on Computer Animation [24] and received the best paper award.

6.2.3. Human motion

Figure 7. Live tracking and visualization of a plausible anatomical skeleton following the pose the subject [17].

Armelle Bauer defended her PhD in November, co-advised with TIMC (Jocelyne Troccaz as principal advisor), on Augmented Reality for the interactive visualization of human anatomy. This is one of the main achievements of the Living Book of Anatomy project, funded by Labex Persyval. This work was partly published at the Motion in Games conference (MIG 2016) [17]. It served as a basis for the follow-up ANR project Anatomy2020 involving Anatoscope, TIMC and LIG laboratories, and Univ Lyon 2.

6.3. Knowledge-based Models for Narrative Design

- **Scientist in charge**: Rémi Ronfard.
- **Other permanent researchers**: Marie-Paule Cani, Frédéric Devernay, François Faure, Jean-Claude Léon, Olivier Palombi.

Our long term goal is to develop high-level models helping users to express and convey their own narrative content (from fiction stories to more practical educational or demonstrative scenarios). Before being able to specify the narration, a first step is to define models able to express some a priori knowledge on the background scene and on the object(s) or character(s) of interest. Our first goal is to develop 3D ontologies able to express such knowledge. The second goal is to define a representation for narration, to be used in future storyboarding frameworks and virtual direction tools. Our last goal is to develop high-level models

for virtual cinematography such as rule-based cameras able to automatically follow the ongoing action and semi-automatic editing tools enabling to easily convey the narration via a movie.

6.3.1. Virtual cameras

Filming live action requires a coincidence of many factors: actors of course, but also lighting, sound capture, set design, and finally the camera (position, frame, and motion). Some of these, such as sound and lighting, can be more or less reworked in post-production, but the camera parameters are usually considered to be fixed at shooting time. We developed two kinds of image-based rendering technique, which allows to change in post-production either the camera frame (in terms of pan, tilt, and zoom), or the camera position.

To be able to change the camera frame after shooting, we developed techniques to construct a video panorama from a set of cameras placed roughly at the same position. Video panorama exhibits a specific problem, which is not present in photo panorama: because the projection centers of the cameras can not physically be at the same location, there is residual parallax between the video sequences, which produce artifacts when the videos are stitched together. Sandra Nabil has worked during her PhD on producing video panoramas without visible artifacts, which can be used to freely pick the camera frame in terms of pan, tilt and zoom during the post-production phase.

Modifying the camera position itself is an even greater challenge, since it either requires a perfect 3D reconstruction of the scene or a dense sampling of the 4D space of optical rays at each time (called the 4D lightfield). During the PhD of Gregoire Nieto, we developed image-based rendering (IBR) techniques which are designed to work in cases where the 3D reconstruction cannot be obtained with a high precision, and the number of cameras used to capture the scene is low, resulting in a sparse sampling of the 4D lightfield.

6.3.2. Virtual actors

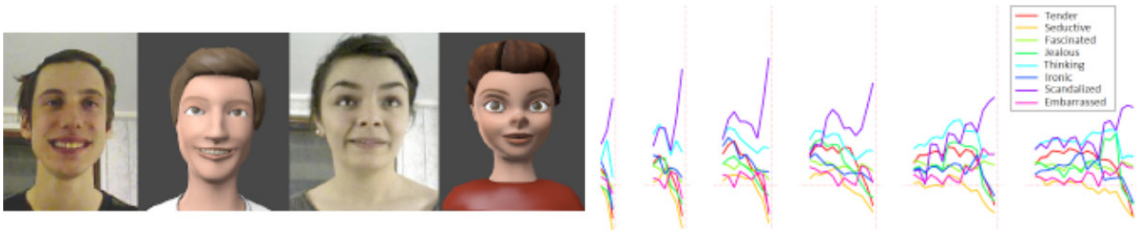


Figure 8. Left: Examples of video and animation frames for a dramatic attitude (seductive) played by two semi-professional actors. Right: Prosodic contours for 8 dramatic attitudes, showing evidence that "scandalized" and "thinking" strongly stand out from other attitudes.

Following up on Adela Barbelescu's PhD thesis, we tested the capability of audiovisual parameters (voice frequency, rhythm, head motion and facial expressions) to discriminate among different dramatic attitudes in both real actors (video) and virtual actors (3D animation). Using Linear Discriminant Analysis classifiers, we showed that sentence-level features present a higher discriminating rate among the attitudes and are less dependent on the speaker than frame and sylable features. We also performed perceptual evaluation tests, showing that voice frequency is correlated to the perceptual results for all attitudes, while other features, such as head motion, contribute differently, depending both on the attitude and the speaker. Those new results were presented at the Interspeech conference [16].

6.4. Creating and Interacting with Virtual Prototypes

- **Scientist in charge**: Jean-Claude Léon.

- **Other permanent researchers**: Marie-Paule Cani, Frédéric Devernay, Olivier Palombi, Damien Rohmer, Rémi Ronfard.

The challenge is to develop more effective ways to put the user in the loop during content authoring. We generally rely on sketching techniques for quickly drafting new content, and on sculpting methods (in the sense of gesture-driven, continuous distortion) for further 3D content refinement and editing. The objective is to extend these expressive modeling techniques to general content, from complex shapes and assemblies to animated content. As a complement, we are exploring the use of various 2D or 3D input devices to ease interactive 3D content creation.

6.4.1. Sculpting Virtual Worlds

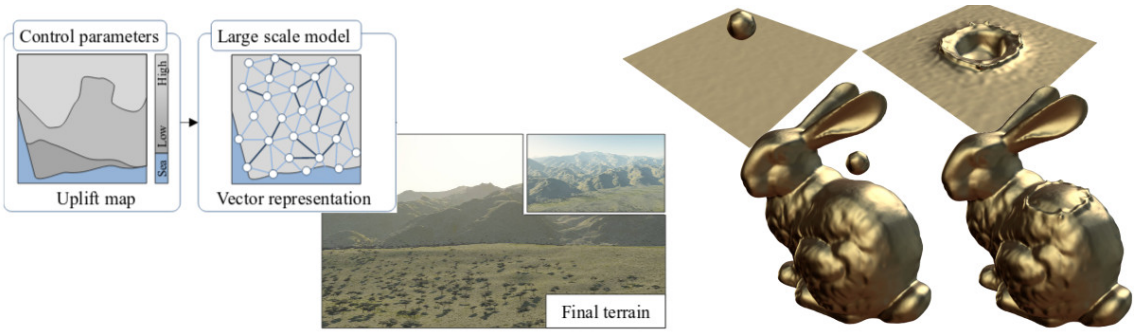


Figure 9. Left: Generating a large-scale terrain following fluvial erosion principle from a coarse set of control parameters [10]. Right: Copy of an animated drop of fluid and its effect on the underlying surface into another animation [23].

Extending expressive modeling paradigms to full virtual worlds, with complex terrains, streams and oceans, and vegetation is a challenging goal. To achieve this, we need to combine procedural methods that accurately simulate physical, geological and biological phenomena shaping the world, which high level user control. This year, our work in the area was three-folds:

Firstly, in collaboration with Jean Braun, professor in geo-morphology and other colleagues, we designed the fist efficient simulation method able to take into account large-scale fluvial erosion to shape mountains. This method was published at Eurographics [10]. We also designed an interactive sculpting system with multi-touch finger interaction, able to shape mountain ranges based on tectonic forces. This method, combined in real-time with our erosion simulation process, was submitted for a journal publication.

Secondly, we extended the "Worldbrush" system proposed in 2015 (Emilien et al, Siggraph 2015) in order to consistently populate virtual worlds with learned statistical distributions of trees and plants. The main contributor to this project was James Gain, our visiting professor. After clustering the input terrain into a number of characteristic environmental conditions, we computed sand-box (small-scale) simulations of ecosystems (plant growth) for each of these conditions, and then used learned statistical models (an extension of worldbrush) to populate the full terrain with consistent sets of species. This work was submitted for publication.

Third, we extended interactive sculpting paradigms to the sculpting of liquid simulation results, such as editing waves on a virtual ocean [23]. Liquid simulations are both compute intensive and very hard to control, since they are typically edited by re-launching the simulations with slightly different initial conditions until the user is satisfied. In contrast, our method enables users to directly edit liquid animation results (coming in the

form of animated meshes) in order to directly output new animations. More precisely, the method offer semi-automatic clustering methods enabling users to select features such as droplets and waves, edit them in space and time and them paste them back into the current liquid animation or to another one.

6.4.2. Sketch based design

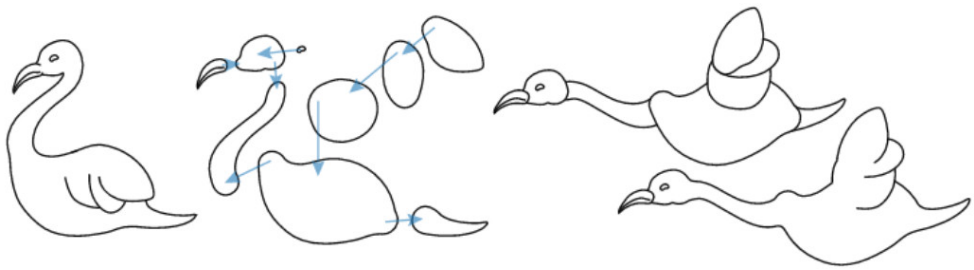


Figure 10. Progressive extraction of sub-parts of an input sketch in depth with automatic contour completion [37].

Using 2D sketches is one of the easiest way for creating 3D contents. While prior knowledge on the object being sketch can be used to retrieve the missing information, and thus consistently inferring 3D, interpreting more general sketches and generating 3D shapes from them in indeed a challenging long-term goals. This year, our work in the area was two-folds:

Firstly, we participated to a course on Sketch-based Modeling, presented at both Eurographics 2016 and Siggraph Asia 2016 [18]. The parts we worked on was sketch-based modeling from prior knowledge, with the examples of our works on animals, garments (developable surfaces) and trees.

Secondly, we advanced towards the interpretation of general sketches representing smooth, organic shapes. The key features of our methods are a new approach for aesthetic contour completion, and an interactive algorithm for progressively interpreting internal silhouettes (suggestive contours) in order to progressively extract sub-parts of the shape from the drawing. These parts are ordered in depth. Our first results were presented as a poster at the Siggraph 2016 conference [37], and then extended an submitted for publication. We are now extending them towards the inference of 3D, organic shapes from a 2D sketch.

MAVERICK Project-Team

7. New Results

7.1. Computer-aided image manipulation

7.1.1. Automatic lighting design from photographic rules

Participants: Jérémy Wambecke, Romain Vergne, Georges-Pierre Bonneau, Joëlle Thollot.

Figure 2. Our lighting setup produces realistic images for any kind of opaque surfaces, where shapes of objects are always properly conveyed.

Lighting design is crucial in 3D scenes modeling for its ability to provide cues to understand the objects shape. However a lot of time, skills, trials and errors are required to obtain a desired result. Existing automatic lighting methods for conveying the shape of 3D objects are based either on costly optimizations or on non-realistic shading effects. Also they do not take the material information into account. In this work, we propose a new method that automatically suggests a lighting setup to reveal the shape of a 3D model, taking into account its material and its geometric properties (see Figure 2). Our method is independent from the rendering algorithm. It is based on lighting rules extracted from photography books, applied through a fast and simple geometric analysis. We illustrate our algorithm on objects having different shapes and materials, and we show by both visual and metric evaluation that it is comparable to optimization methods in terms of lighting setups quality. Thanks to its genericity our algorithm could be integrated in any rendering pipeline to suggest appropriate lighting. It has been published in WICED'2016 [8].

7.1.2. Automatic Texture Guided Color Transfer and Colorization

Participants: Benoit Arbelot, Romain Vergne, Thomas Hurtut, Joëlle Thollot.

This work targets two related color manipulation problems: *Color transfer* for modifying an image colors and *colorization* for adding colors to a greyscale image. Automatic methods for these two applications propose to modify the input image using a reference that contains the desired colors. Previous approaches usually do not target both applications and suffer from two main limitations: possible misleading associations between input and reference regions and poor spatial coherence around image structures. In this work, we propose a unified framework that uses the textural content of the images to guide the color transfer and colorization (see Figure 3). Our method introduces an edge-aware texture descriptor based on region covariance, allowing for local color transformations. We show that our approach is able to produce results comparable or better than state-of-the-art methods in both applications. It has been published in Expressive'2016 [4] and an extended version has been submitted to C&G.

7.1.3. Flow-Guided Warping for Image-Based Shape Manipulation

Participants: Romain Vergne, Pascal Barla, Georges-Pierre Bonneau, Roland W. Fleming.

Color transfer Colorization

Figure 3. Our framework allows for automatic local color transfer (left) and colorization (right) based on textural properties.

(a) Input image - ©Expertissim (b) Shape sharpening (c) Shape rounding

Figure 4. Our warping technique takes as input (a) a single image (Jules Bennes, after Barye: "walking lion") and modifies its perceived surface shape, either making it sharper in (b) or rounder in (c).

We present an interactive method that manipulates perceived object shape from a single input color image thanks to a warping technique implemented on the GPU. The key idea is to give the illusion of shape sharpening or rounding by exaggerating orientation patterns in the image that are strongly correlated to surface curvature. We build on a growing literature in both human and computer vision showing the importance of orientation patterns in the communication of shape, which we complement with mathematical relationships and a statistical image analysis revealing that structure tensors are indeed strongly correlated to surface shape features. We then rely on these correlations to introduce a flow-guided image warping algorithm, which in effect exaggerates orientation patterns involved in shape perception. We evaluate our technique by 1) comparing it to ground truth shape deformations, and 2) performing two perceptual experiments to assess its effects. Our algorithm produces convincing shape manipulation results on synthetic images and photographs, for various materials and lighting environments (see Figure 4). This work has been published in ACM TOG 2016 [3].

7.1.4. Local Shape Editing at the Compositing Stage

Participants: Carlos Jorge Zubiaga Peña, Gael Guennebaud, Romain Vergne, Pascal Barla.

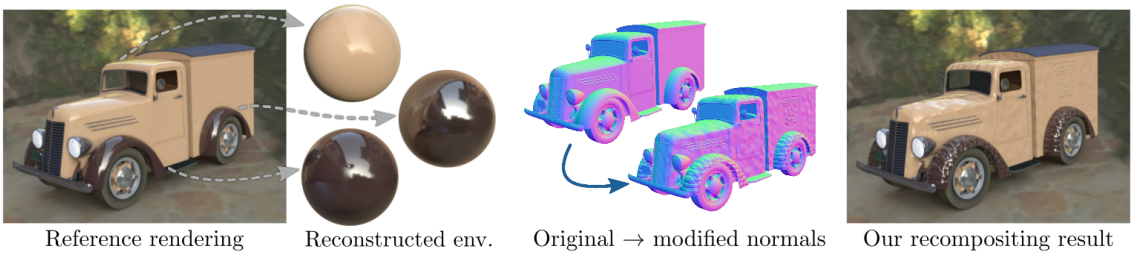


Figure 5. Our method permits to modify surface shape by making use of the shading and auxiliary buffers output by modern renderers. We first reconstruct shading environments for each object/material combination of the Truck scene, relying on normal and shading buffers. When normals are then modified by the compositing artist, the color image is recomposited in real-time, enabling interactive exploration. Our method reproduces inter-reflections between objects, as seen when comparing the reconstructed environments for rear and front mudguards.

Modern compositing software permit to linearly recombine different 3D rendered outputs (e.g., diffuse and reflection shading) in post-process, providing for simple but interactive appearance manipulations. Renderers also routinely provide auxiliary buffers (e.g., normals, positions) that may be used to add local light sources or depth-of-field effects at the compositing stage. These methods are attractive both in product design and movie production, as they allow designers and technical directors to test different ideas without having to re-render an entire 3D scene. We extend this approach to the editing of local shape: users modify the rendered normal buffer, and our system automatically modifies diffuse and reflection buffers to provide a plausible result (see Figure 5). Our method is based on the reconstruction of a pair of diffuse and reflection prefiltered environment maps for each distinct object/material appearing in the image. We seamlessly combine the reconstructed buffers in a recompositing pipeline that works in real-time on the GPU using arbitrarily modified normals. This work has been published in EGSR (EI & I) 2016 [13].

7.1.5. Map Style Formalization: Rendering Techniques Extension for Cartography

Participants: Hugo Loi, Benoit Arbelot, Romain Vergne, Joëlle Thollot.

Cartographic design requires controllable methods and tools to produce maps that are adapted to users' needs and preferences. The formalized rules and constraints for cartographic representation come mainly from the conceptual framework of graphic semiology. Most current Geographical Information Systems (GIS) rely on

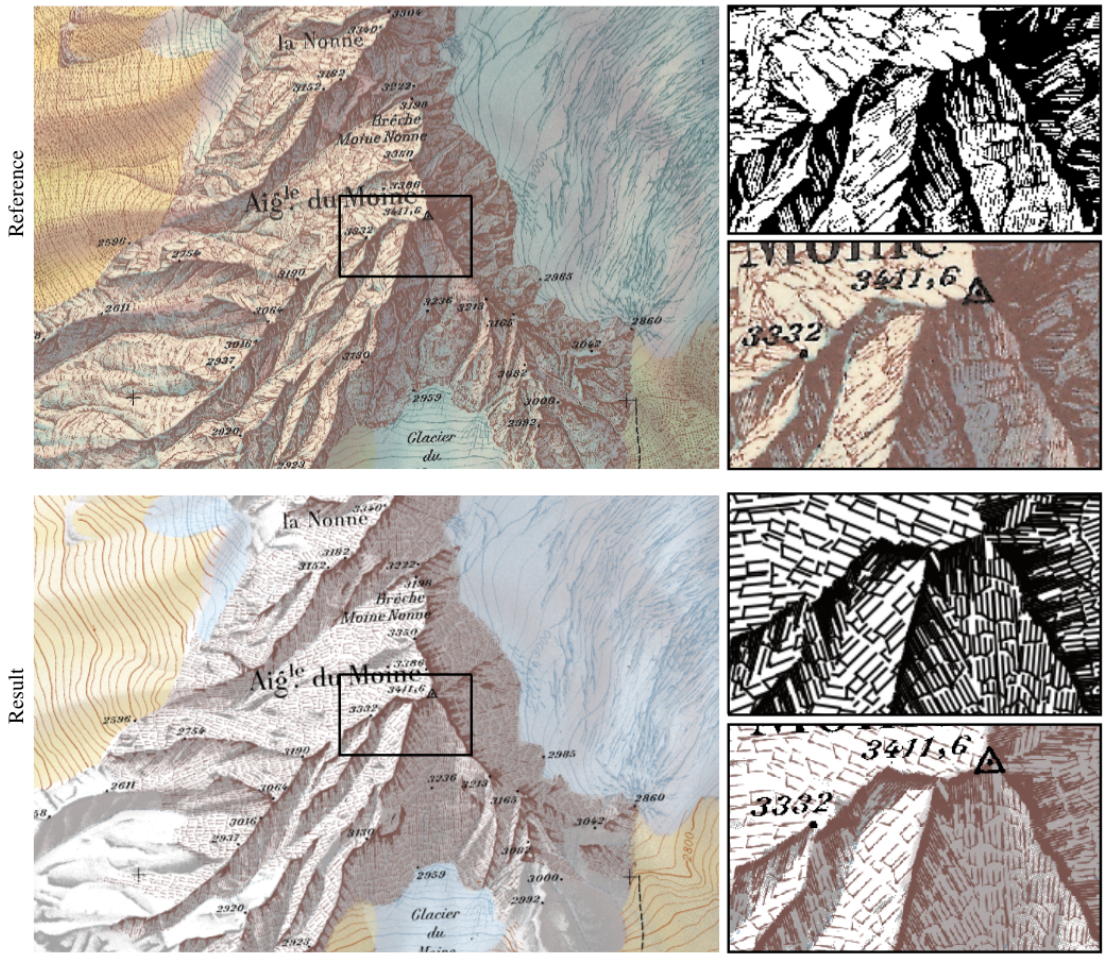


Figure 6. Reference and Resulting Mountain map "Aiguille du Moine", 1:10k scale: extracts of reference (first line) and resulting rocky areas (second line): on the right, zooms on, first the hatching primitives, second the stylized same ones. For a fair comparison, we provide resulting map at a resolution similar to the reference map.

the Styled Layer Descriptor and Semiology Encoding (SLD/SE) specifications which provide an XML schema describing the styling rules to be applied on geographic data to draw a map. Although this formalism is relevant for most usages in cartography, it fails to describe complex cartographic and artistic styles. In order to overcome these limitations, we propose an extension of the existing SLD/SE specifications to manage extended map stylizations, by the means of controllable expressive methods. Inspired by artistic and cartographic sources (Cassini maps, mountain maps, artistic movements, etc.), we propose to integrate into our system three main expressive methods: linear stylization, patch-based region filling and vector texture generation. We demonstrate how our pipeline allows to personalize map rendering with expressive methods in several examples. This work is the result of the MAPSTYLE ANR and has been published at Expressive 20016 [5].

7.2. Illumination Simulation and Materials

7.2.1. A Physically-Based Reflectance Model Combining Reflection and Diffraction

Participant: Nicolas Holzschuch.

Reflectance properties express how objects in a virtual scene interact with light; they control the appearance of the object: whether it looks shiny or not, whether it has a metallic or plastic appearance. Having a good reflectance model is essential for the production of photo-realistic pictures. Measured reflectance functions provide high realism at the expense of memory cost. Parametric models are compact, but finding the right parameters to approximate measured reflectance can be difficult. Most parametric models use a model of the surface micro-geometry to predict the reflectance at the macroscopic level. We have shown that this micro-geometry causes two different physical phenomena: reflection and diffraction. Their relative importance is connected to the surface roughness. Taking both phenomena into account, we developped a new reflectance model that is compact, based on physical properties and provides a good approximation of measured reflectance (See Figure 7).

7.2.2. A Robust and Flexible Real-Time Sparkle Effect

Participant: Beibei Wang.

We present a fast and practical procedural sparkle effect for snow and other sparkly surfaces which we integrated into a recent video game. Following from previous work, we generate the sparkle glints by intersecting a jittered 3D grid of sparkle seed points with the rendered surface. By their very nature, the sparkle effect consists of high frequencies which must be dealt with carefully to ensure an anti-aliased and noise free result (See Figure 8). We identify a number of sources of aliasing and provide effective techniques to construct a signal that has an appropriate frequency content ready for sampling at pixels at both foreground and background ranges of the scene. This enables artists to push down the sparkle size to the order of 1 pixel and achieve a solid result free from noisy flickering or other aliasing problems, with only a few intuitive tweakable inputs to manage [9].

7.2.3. Capturing Spatially Varying Anisotropic Reflectance Parameters using Fourier Analysis

Participants: Nicolas Holzschuch, Alban Fichet.

Reflectance parameters condition the appearance of objects in photorealistic rendering. Practical acquisition of reflectance parameters is still a difficult problem. Even more so for spatially varying or anisotropic materials, which increase the number of samples required. We present an algorithm for acquisition of spatially varying anisotropic materials, sampling only a small number of directions. Our algorithm uses Fourier analysis to extract the material parameters from a sub-sampled signal. We are able to extract diffuse and specular reflectance, direction of anisotropy, surface normal and reflectance parameters from as little as 20 sample directions (See Figure 9). Our system makes no assumption about the stationarity or regularity of the materials, and can recover anisotropic effects at the pixel level. This work has been published at Graphics Interface 2016 [6].

7.2.4. Estimating Local Beckmann Roughness for Complex BSDFs

Participant: Nicolas Holzschuch.

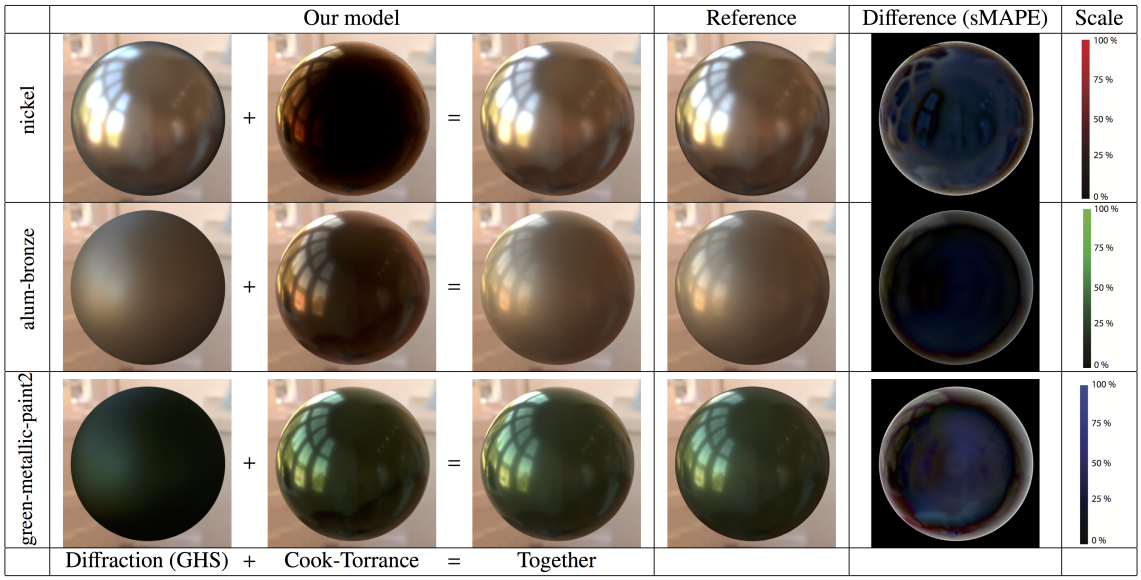


Figure 7. Surface micro-geometry contributes to its visible aspect (material reflectance). Two physical phenomena are acting together: reflection on micro-facets and diffraction. Our reflectance model combines them, with the proper energy repartition between them. The importance of diffraction depends on the roughness of the material. Even when it it relatively small, as for `green-metallic-paint2`*, it has a significant impact of the aspect of the material. Our model explains even a very difficult material like* `alum-bronze` *(middle row) as a single material.*

Figure 8. Two scenes rendered with our sparkle effect

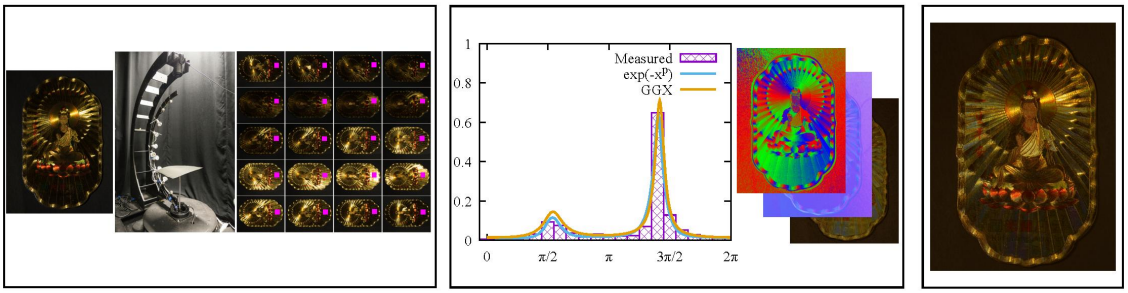


Figure 9. Our acquisition pipeline: first, we place a material sample on our acquisition platform, and acquire photographs with varying incoming light direction. In a second step, we extract anisotropic direction, shading normal, albedo and reflectance parameters from these photographs and store them in texture maps. We later use these texture maps to render new views of the material.

Many light transport related techniques require an analysis of the blur width of light scattering at a path vertex, for instance a Beckmann roughness. Such use cases are for instance analysis of expected variance (and potential biased countermeasures in production rendering), radiance caching or directionally dependent virtual point light sources, or determination of step sizes in the path space Metropolis light transport framework: recent advanced mutation strategies for Metropolis Light Transport, such as Manifold Exploration and Half Vector Space Light Transport employ local curvature of the BSDFs (such as an average Beckmann roughness) at all interactions along the path in order to determine an optimal mutation step size. A single average Beckmann roughness, however, can be a bad fit for complex measured materials and, moreover, such curvature is completely undefined for layered materials as it depends on the active scattering layer. We propose a robust estimation of local curvature for BSDFs of any complexity by using local Beckmann approximations, taking into account additional factors such as both incident and outgoing direction (See Figure 10). This work has been published as a Siggraph 2016 Talk [18].

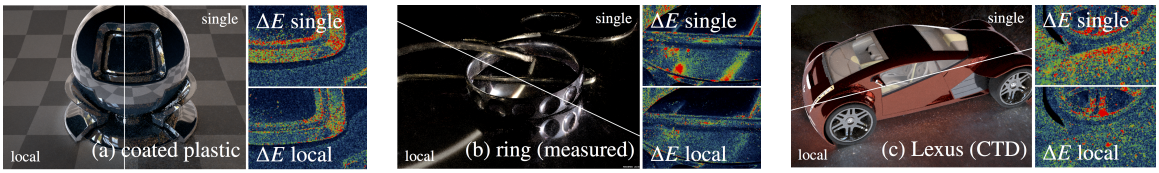


Figure 10. Indirect lighting (exposure in b and c increased for printouts) on three test scenes rendered with different materials: (a) multilayer coated plastic material, (b) measured materials on a ring, (c) CTD material on a car. The insets show difference to reference in CIE'76 ΔE. Top: single Gaussian, bottom: our local Gaussian approximation. We can render both analytic (a, c) and measured materials (b) more robustly because the local Gaussian approximation facilitates more even exploration of path space.

7.2.5. MIC based PBGI

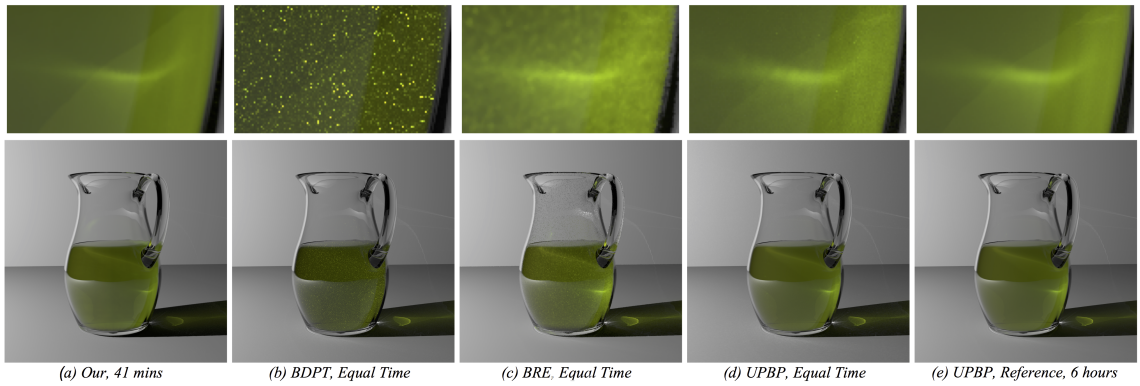
Participant: Beibei Wang.

Point-Based Global Illumination (PBGI) is a popular rendering method in special effects and motion picture productions. The tree-cut computation is in gen eral the most time consuming part of this algorithm, but it can be formulated for efficient parallel execution, in particular regarding wide-SIMD hardware. In this context, we propose several vectorization schemes, namely single, packet and hybrid, to maximize the utilization of modern CPU architectures. Whil e for the single scheme, 16 nodes from the hierarchy are processed for a single receiver in parallel, the packet scheme handles one node for 16 receivers. These two schemes work well for scenes having smooth geometry and diffuse material. When the scene contains high frequency bumps maps and glossy reflection s, we use a hybrid vectorization method. We conduct experiments on an Intel Many Integrated Core architecture and report preliminary results on several sce nes, showing that up to a 3x speedup can be achieved when compared with non-vectorized execution [19].

7.2.6. Point-Based Light Transport for Participating Media with Refractive Boundaries

Participants: Beibei Wang, Jean-Dominique Gascuel, Nicolas Holzschuch.

Illumination effects in translucent materials are a combination of several physical phenomena: absorption and scattering inside the material, refraction at its surface. Because refraction can focus light deep inside the material, where it will be scattered, practical illumination simulation inside translucent materials is difficult. In this paper, we present an a Point-Based Global Illumination method for light transport on translucent materials with refractive boundaries. We start by placing volume light samples inside the translucent material and organising them into a spatial hierarchy. At rendering, we gather light from these samples for each camera ray. We compute separately the samples contributions to single, double and multiple scattering, and add them (See Figure 11). Our approach provides high-quality results, comparable to the state of the art, with significant speed-ups (from 9× to 60× depending on scene c omplexity) and a much smaller memory footprint [10], [12].



(a) Our, 41 mins *(b) BDPT, Equal Time* *(c) BRE, Equal Time* *(d) UPBP, Equal Time* *(e) UPBP, Reference, 6 hours*

Figure 11. Our algorithm (a), compared with Bi-Directional Path Tracing (BDPT) (b), Photon Mapping with Beam-Radiance Estimate (BRE) (c) and Unified Points, Beams and Paths (UPBP) (d) (e). Our algorithm is up to 60 times faster than UPBP, with similar quality. Material: olive oil, $\alpha = 0.0042, 0.4535, 0.0995$*;* $\ell = 9.7087, 11.6279, 2.7397$*. For this material with low albedo* α *and large mean-free-path* ℓ*, low-order scattering effects dominate.*

7.3. Complex Scenes

In order to render both efficiently and accurately ultra-detailed large scenes, this approach consists in developing representations and algorithms able to account compactly for the quantitative visual appearance of a regions of space projecting on screen at the size of a pixel.

7.3.1. Appearance pre-filtering

Participants: Guillaume Loubet, Fabrice Neyret.

We address the problem of constructing appearance-preserving level of details (LoDs) of complex 3D models such as trees and propose a hybrid method that combines the strength of mesh and volume representations. Our main idea is to separate macroscopic (i.e. larger than the target spatial resolution) and microscopic (sub-resolution) surfaces at each scale and to treat them differently, because meshes are very efficient at representing macroscopic surfaces while sub-resolution geometry benefit from volumetric approximations. We introduce a new algorithm based on mesh analysis that detects the macroscopic surfaces of a 3D model at a given resolution. We simplify these surfaces with edge collapses and provide a method for pre-filtering their BRDFs parameters. To approximate microscopic details, we use a heterogeneous microflake participating medium and provide a new artifact-free voxelization algorithm that preserves local occlusion. Thanks to our macroscopic surface analysis, our algorithm is fully automatic and can generate seamless LoDs at arbitrarily coarse resolutions for a wide range of 3D models. We validated our method on highly complex geometry and show that appearance is consistent across scales while memory usage and loading times are drasticall y reduced (see Figure 12). This work has been submitted to EG2017.

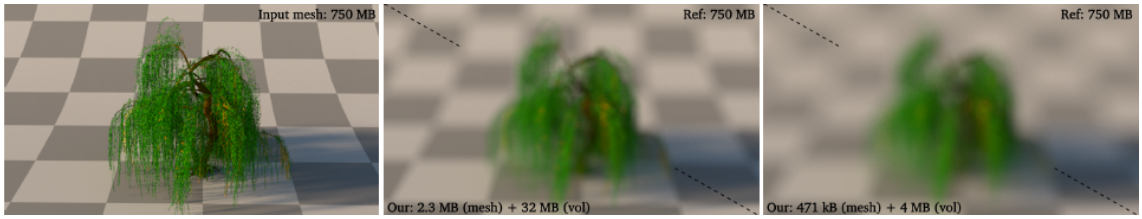


Figure 12. A weeping willow 3D model pre-filtered with our method. Our LoDs use meshes for representing macroscopic surfaces and a volumetric representation to approximate sub-resolution geometry. This approach allows for accurate preservation of the appearance of complex geometry acro ss scales while memory usage is drastic reduced. These images have been rendered with 256spp and a thin lense camera model in Mitsuba

7.4. Texture Synthesis

7.4.1. Understanding and controlling contrast oscillations in stochastic texture algorithms using Spectrum of Variance

Participants: Fabrice Neyret, Eric Heitz.

We identify and analyze a major issue pertaining to all power-spectrum based texture synthesis algorithms from Fourier synthesis to procedural noise algori thms like Perlin or Gabor noise, namely, the oscillation of contrast (see Figure 13). One of our key contributions is to introduce a simple yet powerf ul descriptor of signals, the Spectrum of Variance (not to be confused with the PSD), which, to our surprise, has never been leveraged before. In this new framework, several issues get easy to understand measure and control, with new handles, as we illustrate. We finally show that fixing oscillation of contra st opens many doors to a more controllable authoring of stochastic texturing. We explore some of the new reachable possibilities such as constrained noise content and bridges towards very different families of look such as cellular patterns, points-like distributions or reaction-diffusion [17].

7.5. Visualization and Geometric Design

7.5.1. Surfacing Curve Networks with Normal Control

Participant: Georges-Pierre Bonneau.

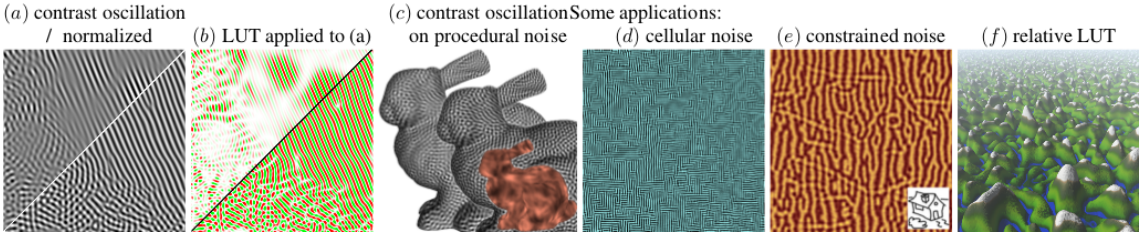


Figure 13. Power-spectrum based texturing algorithms (e.g., Gabor, Fourier synthesis) suffer from unexpected low frequency contrast variations (a,b,c top) even when the spectrum has no low frequency (the contrast field is display in red in (c)). This prevents precise authoring with non- linear transform, like color LUT (b top). Our renormalization method allows to control the stationarity (a,b,c bottom). It also opens many doors for noise authoring such as the generation of reaction-diffusion-like strips and spots (b bottom), cellular-like patterns (d), content constraints (e), or the parame trization of height maps relative to local extrema (f).

Members of Maverick involved: Georges-Pierre Bonneau

This is a joint work with team-project IMAGINE (Tibor Stanko and Stefanie Hahmann) at Inria-Grenoble and CEA-Leti (Nathalie Saguin). Recent surface acquisition technologies based on microsensors produce three-space tangential curve data which can be transformed into a network of space curves with surface normals. This work addresses the problem of surfacing an arbitrary closed 3D curve network with given surface normals. Thanks to the normal vector input, the patch finding problem can be solved unambiguously and an initial piecewise smooth triangle mesh is computed. The input normals are propagated throughout the mesh. Together with the initial mesh, the propagated normals are used to compute mean curvature vectors. We compute the final mesh as the solution of a new variational optimization method based on the mean curvature vectors. The intuition behind this original approach is to guide the standard Laplacian-based variational methods by the curvature information extracted from the input normals. The normal input increases shape fidelity and allows to achieve globally smooth and visually pleasing shapes [2], [7]. This is a joint work with team-project IMAGINE (Tibor Stanko and Stefanie Hahmann) at Inria-Grenoble and CEA-Leti (Nathalie Saguin).

7.5.2. Piecewise polynomial Reconstruction of Scalar Fields from Simplified Morse-Smale Complexes

Participants: Léo Allemand-Giorgis, Georges-Pierre Bonneau.

Morse-Smale (MS) complexes have been proposed to visualize topological features of scalar fields defined on manifold domains. Herein, three main problems have been addressed in the past: (a) efficient computation of the initial combinatorial structure connecting the critical points; (b) simplification of these combinatorial structures; (c) reconstruction of a scalar field in accordance to the simplified Morse-Smale complex. The present work faces the third problem by proposing a novel approach for computing a scalar field coherent with a given simplified MS complex that privileges the use of piecewise polynomial functions. Based on techniques borrowed from shape preserving design in Computer Aided Geometric Design, our method constructs the surface cell by cell using piecewise polynomial curves and surfaces. The benefit and limitations of using polynomials for reconstruction surfaces from topological data are studied in this work [14].

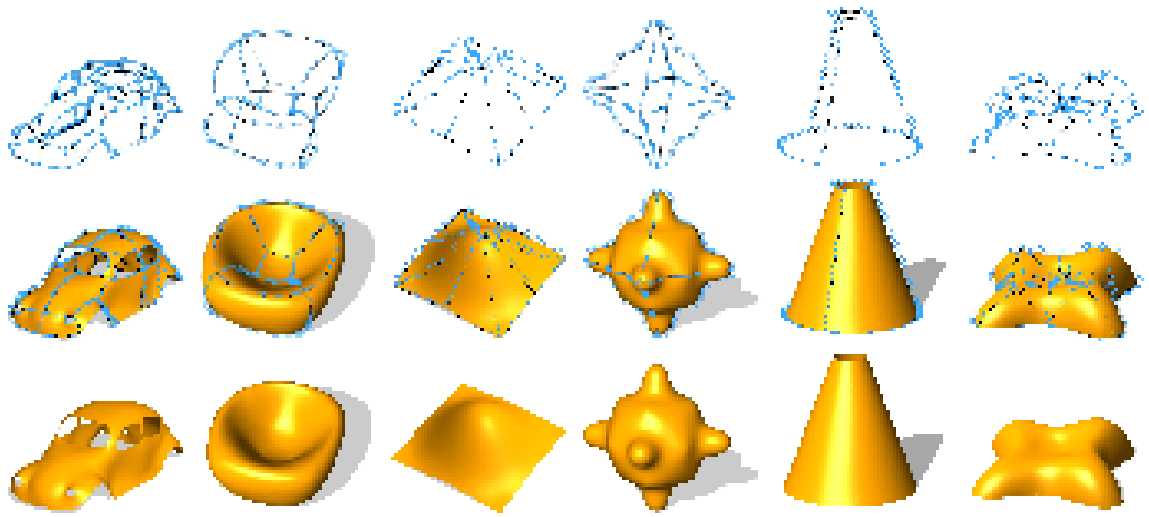


Figure 14. In [2] and [7] we address the problem of surfacing an arbitrary closed 3D curve network with given surface normals (top row). Our interpolating surfaces are visualized with (middle row) and without (bottom row) input curves.

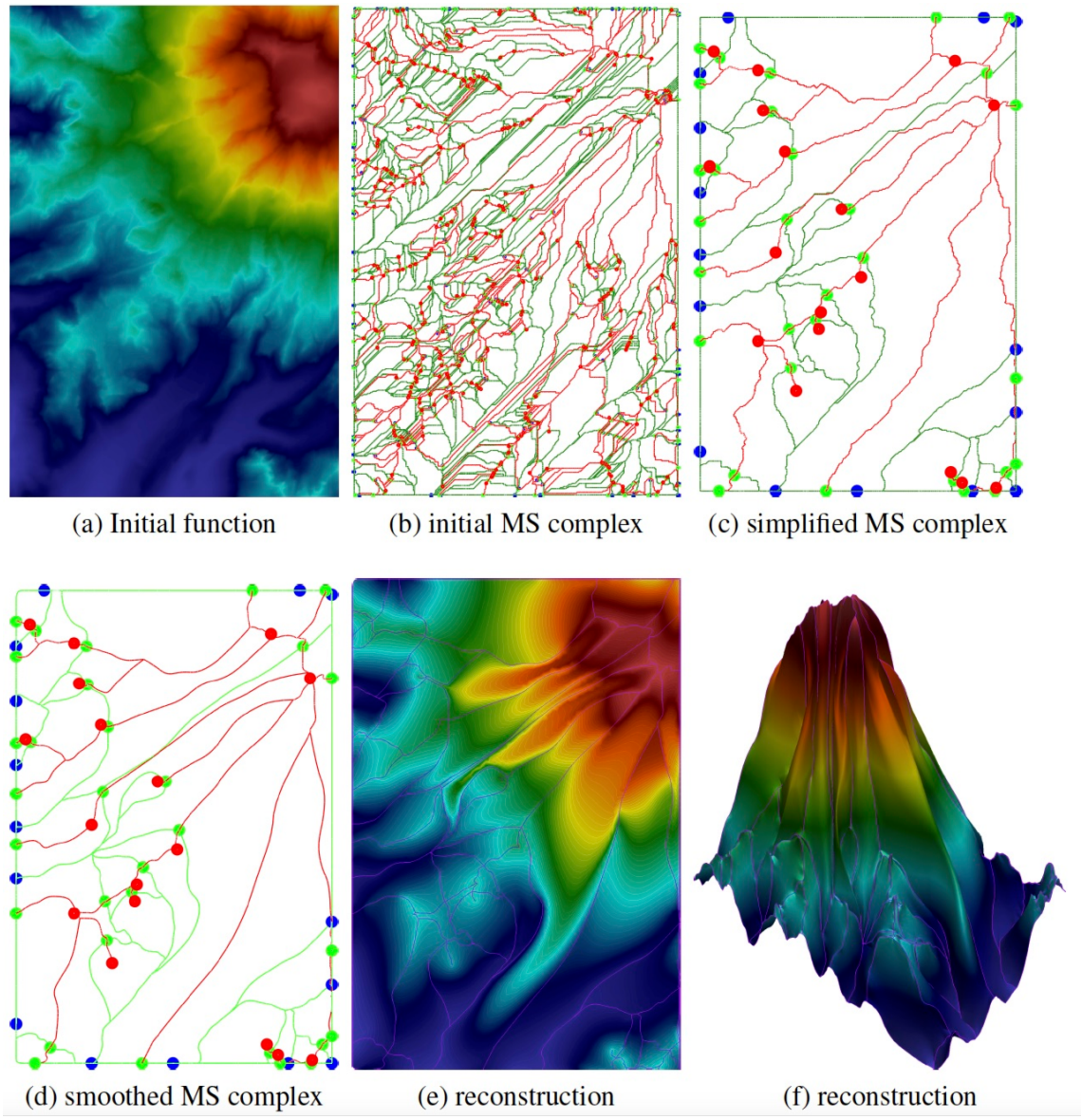


Figure 15. The terrain data set of Mt Rainier (a) has 1931 critical points (b). The simplified Morse-Smale complex with 69 critical points is reconstructed using our methods. The final function approximates the original one, with a topology that is simplified in a controlled-manner.

MORPHEO Project-Team

7. New Results

7.1. Cotemporal Multi-View Video Segmentation

We address the problem of multi-view video segmentation of dynamic scenes in general and outdoor environments with possibly moving cameras. Multi-view methods for dynamic scenes usually rely on geometric calibration to impose spatial shape constraints between viewpoints. In this paper, we show that the calibration constraint can be relaxed while still getting competitive segmentation results using multi-view constraints. We introduce new multi-view cotemporality constraints through motion correlation cues, in addition to common appearance features used by co-segmentation methods to identify co-instances of objects. We also take advantage of learning based segmentation strategies by casting the problem as the selection of monocular proposals that satisfy multi-view constraints. This yields a fully automated method that can segment subjects of interest without any particular pre-processing stage, as depicted in Figure 2 . Results on several challenging outdoor datasets demonstrate the feasibility and robustness of our approach.

This work has been presented at the International Conference on 3D Vision (3DV) 2016 [9].

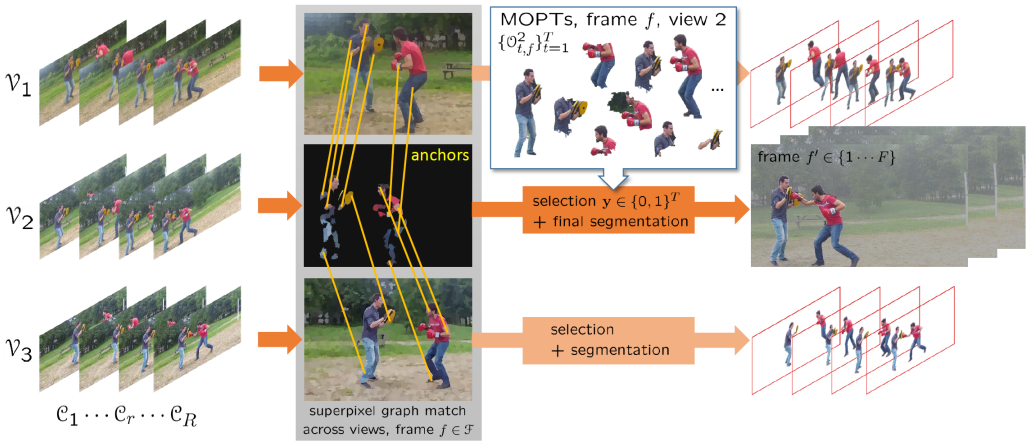


Figure 2. Overview of multiview segmentation pipeline

7.2. Volumetric Shape Reconstruction from Implicit Forms

In this work we evaluate volumetric shape reconstruction methods that consider as input implicit forms in 3D. Many visual applications build implicit representations of shapes that are converted into explicit shape representations using geometric tools such as the Marching Cubes algorithm. This is the case with image based reconstructions that produce point clouds from which implicit functions are computed, with for instance a Poisson reconstruction approach. While the Marching Cubes method is a versatile solution with proven efficiency, alternative solutions exist with different and complementary properties that are of interest for shape modeling. In this paper, we propose a novel strategy that builds on Centroidal Voronoi Tessellations (CVTs). These tessellations provide volumetric and surface representations with strong regularities in addition to provably more accurate approximations of the implicit forms considered. In order to compare the existing

strategies, we present an extensive evaluation that analyzes various properties of the main strategies for implicit to explicit volumetric conversions: Marching cubes, Delaunay refinement and CVTs, including accuracy and shape quality of the resulting shape mesh.

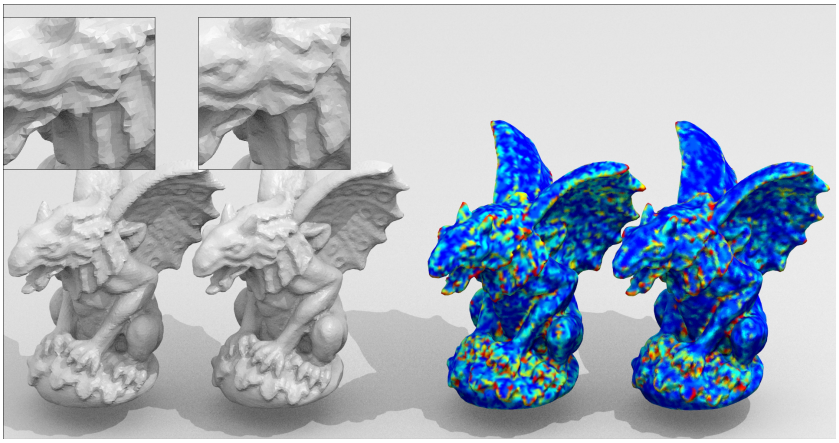


Figure 3. Poisson volumetric reconstructions from a Gargoyle point cloud with Marching Cubes (left) and CVT (right) [16]. Distances to the implicit form are color encoded on the right, from low (blue) to high (red).

This work has been presented at the ECCV 2016 conference [16].

7.3. Bayesian 3D imaging from X-rays and video

A new method for estimating 3D dense attenuation of moving samples such as body parts from multiple video and a single planar X-ray device has been devised [12]. Most dense modeling methods consider samples observed with a moving X-ray device and cannot easily handle moving samples. We proposed a novel method that uses a surface motion capture system associated to a single low-cost/low-dose planar X-ray imaging device for dense in-depth attenuation information. Our key contribution is to rely on Bayesian inference to solve for a dense attenuation volume given planar radioscopic images of a moving sample. The approach enables multiple sources of noise to be considered and takes advantage of limited prior information to solve an otherwise ill-posed problem. Results show that the proposed strategy is able to reconstruct dense volumetric attenuation models from a very limited number of radiographic views over time on simulated and in-vivo data, as illustrated in Figure 4 .

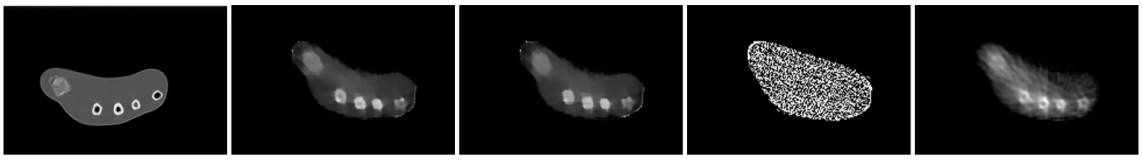


Figure 4. Results of the proposed method on a forearm phantom (2 selected slices). Left-to-right: ground-truth CT scan, proposed method, without optical flow, without TVL_1prior, ART. Without optical flow, artefacts are visible, for example in the bone cavities. The ART method produces much noisier results.

7.4. Robust Multilinear Model Learning Framework for 3D Faces

Statistical models are widely used to represent the variations of 3D human faces. Multilinear models in particular are common as they decouple shape changes due to identity and expression. Existing methods to learn a multilinear face model degrade if not every person is captured in every expression, if face scans are noisy or partially occluded, if expressions are erroneously labeled, or if the vertex correspondence is inaccurate. These limitations impose requirements on the training data that disqualify large amounts of available 3D face data from being usable to learn a multilinear model. To overcome this, we have developed an effective framework to robustly learn a multilinear model from 3D face databases with missing data, corrupt data, wrong semantic correspondence, and inaccurate vertex correspondence. To achieve this robustness to erroneous training data, our framework jointly learns a multilinear model and fixes the data. This framework is significantly more efficient than prior methods based on linear statistical models. This work was presented at CVPR 2016 [7].

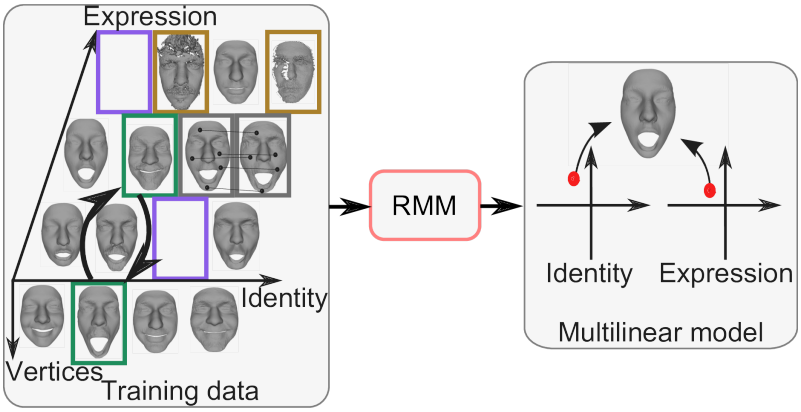


Figure 5. Overview of our robust multilinear model (RMM) learning framework that is robust to missing data (purple), corrupt data (brown), wrong semantic correspondence (green), and inaccurate vertex correspondence (gray).

7.5. Segmentation of Tree Seedling Point Clouds into Elementary Units

We propose a new semi-automatic method to cluster TLS data into meaningful sets of points to extract plant components. The approach is designed for small plants with distinguishable branches and leaves, such as tree seedlings. It first creates a graph by connecting each point to its most relevant neighbours, then embeds the graph into a spectral space, and finally segments the embedding into clusters of points. The process can then be iterated on each cluster separately. The main idea underlying the approach is that the spectral embedding of the graph aligns the points along the shape's principal directions. A quantitative evaluation of the segmentation accuracy, as well as of leaf area estimates, is provided on a poplar seedling mock-up. It shows that the segmentation is robust with false positive and false negative rates around 1%. Qualitative results on four contrasting plant species with three different scan resolution levels each are also shown in the paper, which has been published in the International Journal of Remote Sensing [2].

7.6. Estimation of Human Body Shape in Motion with Wide Clothing

Estimating 3D human body shape in motion from a sequence of unstructured oriented 3D point clouds is important for many applications. We propose the first automatic method to solve this problem that works in

the presence of loose clothing. The problem is formulated as an optimization problem that solves for identity and posture parameters in a shape space capturing likely body shape variations. The automation is achieved by leveraging a recent robust pose detection method Stitched Puppet. To account for clothing, we take advantage of motion cues by encouraging the estimated body shape to be inside the observations. The method is evaluated on a new benchmark containing different subjects, motions, and clothing styles that allows to quantitatively measure the accuracy of body shape estimates. Furthermore, we compare our results to existing methods that require manual input and demonstrate that results of similar visual quality can be obtained.

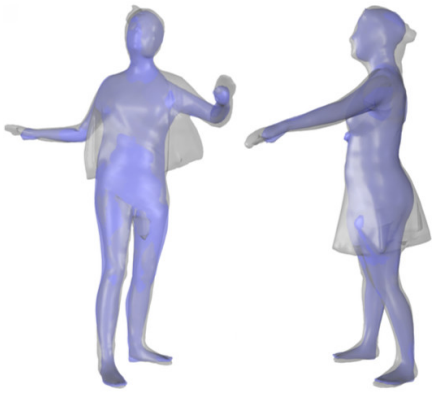


Figure 6. Two frames of an input point cloud sequence (in gray) with the estimated body shape shown in blue [14].

This work has been presented at the ECCV 2016 conference [14].

7.7. Computing Temporal Alignments of Human Motion Sequences in Wide Clothing using Geodesic Patches

In this work, we address the problem of temporal alignment of surfaces for subjects dressed in wide clothing, as acquired by calibrated multi-camera systems. Most existing methods solve the alignment by fitting a single surface template to each instant's 3D observations, relying on a dense point-to-point correspondence scheme, e.g. by matching individual surface points based on local geometric features or proximity. The wide clothing situation yields more geometric and topological difficulties in observed sequences, such as apparent merging of surface components, misreconstructions, and partial surface observation, resulting in overly sparse, erroneous point-to-point correspondences, and thus alignment failures. To resolve these issues, we propose an alignment framework where point-to-point correspondences are obtained by growing isometric patches from a set of reliably obtained body landmarks. This correspondence decreases the reliance on local geometric features subject to instability, instead emphasizing the surface neighborhood coherence of matches, while improving density given sufficient landmark coverage. We validate and verify the resulting improved alignment performance in our experiments.
This work has been presented at the International Conference on 3D Vision (3DV) 2016 [13].

7.8. A 3D+t Laplace Operator for Temporal Mesh Sequences

The Laplace operator plays a fundamental role in geometry processing. Several discrete versions have been proposed for 3D meshes and point clouds, among others. We have defined a discrete Laplace operator for temporally coherent mesh sequences, which allows to process mesh animations in a simple yet efficient way. This operator is a discretization of the Laplace-Beltrami operator using Discrete Exterior Calculus on CW complexes embedded in a four-dimensional space. A parameter is introduced to tune the influence of the

motion with respect to the geometry. This enables straightforward generalization of existing Laplacian static mesh processing works to mesh sequences. An application to spacetime editing has been provided as example.

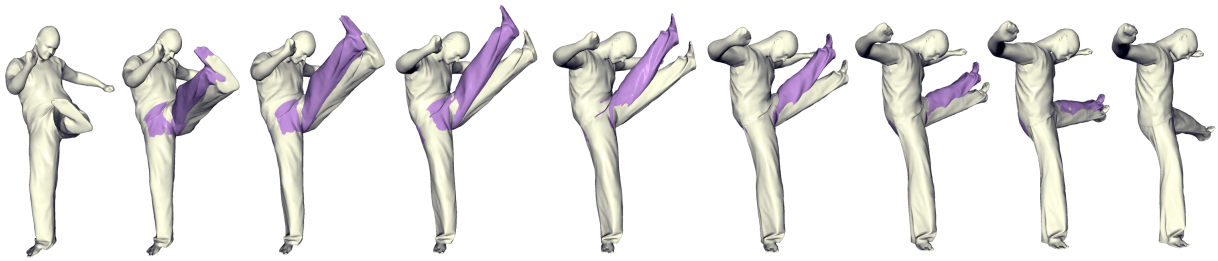


Figure 7. Spacetime editing of a temporal mesh sequence using the proposed 3D+t Laplace operator [1].

This work has been published in Computer & Graphics [1] and presented at the Shape Modeling International (SMI) 2016 conference.

7.9. Volumetric 3D Tracking by Detection

In this collaboration with TU Munich, we investigated a new solutions for 3D tracking by detection based on fully volumetric representations. On one hand, 3D tracking by detection has shown robust use in the context of interaction (Kinect) and surface tracking. On the other hand, volumetric representations have recently been proven efficient both for building 3D features and for addressing the 3D tracking problem. We leveraged these benefits by unifying both families of approaches into a single, fully volumetric tracking-by-detection framework. We used a centroidal Voronoi tessellation (CVT) representation to compactly tessellate shapes with optimal discretization, construct a feature space, and perform the tracking according to the correspondences provided by trained random forests (see figure 8). Our results show improved tracking and training computational efficiency and improved memory performance. This in turn enables the use of larger training databases than state of the art approaches, which we leveraged by proposing a cross-tracking subject training scheme to benefit from all subject sequences for all tracking situations, thus yielding better detection and less overfitting. The approach has been presented at CVPR 2016 [10].

7.10. Eigen Appearance Maps of Dynamic Shapes

In this work, we considered the problem of building efficient appearance rep- resentations of shapes observed from multiple viewpoints and in several movements. Multi-view systems now allow the acquisition of spatio-temporal models of such moving objects. While efficient geometric representations for these models have been widely studied, appearance information, as provided by the observed images, is mainly considered on a per frame basis, and no global strategy yet addresses the case where several temporal sequences of a shape are available. We proposed a per subject representation that builds on PCA to identify the underlying manifold structure of the appearance information relative to a shape. The resulting eigen representation encodes shape appearance variabilities due to viewpoint and motion, with Eigen textures, and due to local inaccuracies in the geometric model, with Eigen warps. In addition to providing compact representations, such decompositions also allow for appearance interpolation and appearance completion. We evaluated their performances over different characters and with respect to their ability to reproduce compelling appearances in a compact way. This work was presented at ECCV 2016.

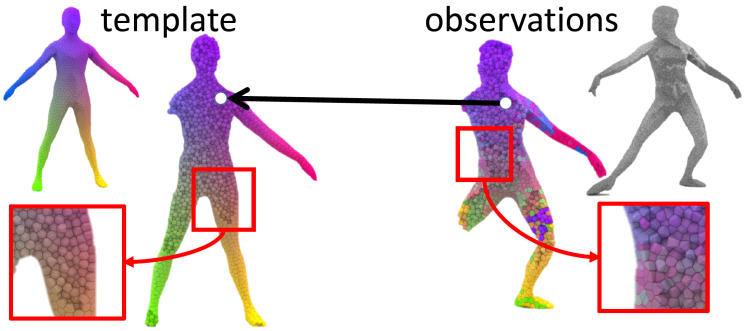


Figure 8. 3D shapes are represented using centroidal Voronoi tessellations. The volumetric cells of the observations are matched to cells of the template.

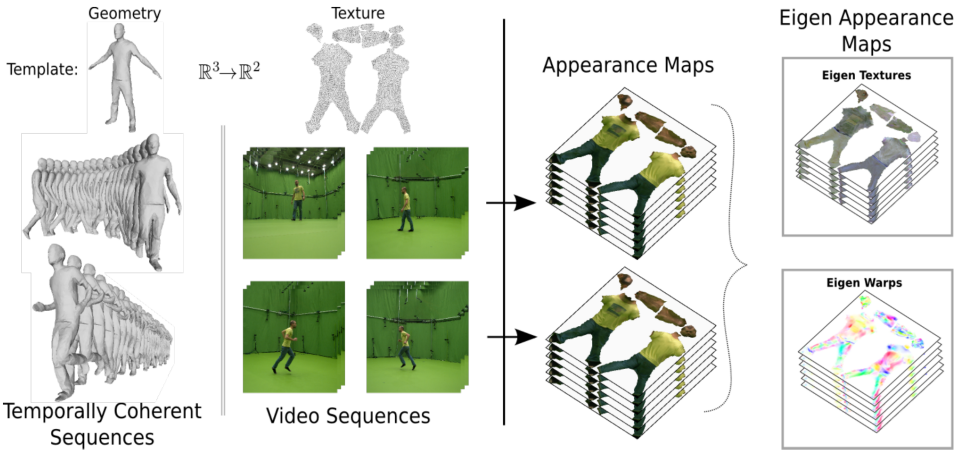


Figure 9. Given time consistent shape models and their appearance maps, our method exploits the manifold structure of these appearance information through PCA decomposition to generate the Eigen appearance maps relative to a shape.

7.11. Visual Contrast Sensitivity and Discrimination for 3D Meshes

In this work, we first introduce an algorithm for estimating the visual contrast on a 3D mesh. We then perform a series of psychophysical experiments to study the effects of contrast sensitivity and contrast discrimination of the human visual system for the task of differentiating between two contrasts on a 3D mesh. The results of these experiments allow us to propose a perceptual model that is able to predict whether a change in local contrast on 3D mesh, induced by a local geometric distortion, is visible or not. Finally, we illustrate the utility of the proposed perceptual model in a number of applications: we compute the Just Noticeable Distortion (JND) profile for smooth-shaded 3D meshes and use the model to guide mesh processing algorithms.
This work has been published in Computer Graphics Forum [] and has received the best paper award at the Pacific Graphics 2016 conference.

PERCEPTION Project-Team

6. New Results

6.1. Audio-Source Localization

In previous years we have developed several *supervised* sound-source localization algorithms. The general principle of these algorithms was based on the learning of a mapping (regression) between binaural feature vectors and source locations [5], [7]. While fixed-length wide-spectrum sounds (white noise) are used for training to reliably estimate the model parameters, we show that the testing (localization) can be extended to variable-length sparse-spectrum sounds (such as speech), thus enabling a wide range of realistic applications. Indeed, we demonstrate that the method can be used for audio-visual fusion, namely to map speech signals onto images and hence to spatially align the audio and visual modalities, thus enabling to discriminate between speaking and non-speaking faces. We released a novel corpus of real-room recordings that allow quantitative evaluation of the co-localization method in the presence of one or two sound sources. Experiments demonstrate increased accuracy and speed relative to several state-of-the-art methods. During the period 2015-2016 we extended this method to an arbitrary number of microphones based on the *relative transfer function – RTF* (between any channel and a reference channel). Then we extended this work and developed a novel transfer function that contains the direct path between the source and the microphone array, namely the *direct-path relative transfer function* [29], [36].

Websites:
https://team.inria.fr/perception/research/acoustic-learning/
https://team.inria.fr/perception/research/binaural-ssl/
https://team.inria.fr/perception/research/ssl-rtf/

6.2. Audio-Source Separation

We address the problem of separating audio sources from time-varying convolutive mixtures. We proposed an unsupervised probabilistic framework based on the local complex-Gaussian model combined with non-negative matrix factorization [33], [28]. The time-varying mixing filters are modeled by a continuous temporal stochastic process. This model extends the case of static filters which corresponds to static audio sources. While static filters can be learnt in advance, e.g. [5], time-varying filters cannot and therefore the problem is more complex. We present a variational expectation-maximization (VEM) algorithm that employs a Kalman smoother to estimate the time-varying mixing matrix, and that jointly estimates the source parameters. The sound sources are then separated by Wiener filters constructed with the estimators provided by the VEM algorithm. Extensive experiments on simulated data show that the proposed method outperforms a block-wise version of a state-of-the-art baseline method. This work is part of the PhD topic of Dionyssos Kounades Bastian and is conducted in collaboration with Sharon Gannot (Bar Ilan University) and Xavier Alameda Pineda (University of Trento). Our journal paper [28] is an extended version of a paper presented at IEEE WASPAA in 2015 which received the best student paper award.

Website:
https://team.inria.fr/perception/research/vemove/
https://team.inria.fr/perception/research/nmfig/

6.3. Single-Channel Audio Processing

While most of our audio scene analysis work involves microphone arrays, it is important to develop single-channel (one microphone) signal processing methods as well. In particular, it is important to detect speech signal (or voice) in the presence of various types of noise (stationary or non-stationary). In this context, we developed the following methods [39], [37]:

- Statistical likelihood ratio test is a widely used voice activity detection (VAD) method, in which the likelihood ratio of the current temporal frame is compared with a threshold. A fixed threshold is always used, but this is not suitable for various types of noise. In this work, an adaptive threshold is proposed as a function of the local statistics of the likelihood ratio. This threshold represents the upper bound of the likelihood ratio for the non-speech frames, whereas it remains generally lower than the likelihood ratio for the speech frames. As a result, a high non-speech hit rate can be achieved, while maintaining speech hit rate as large as possible.
- Estimating the noise power spectral density (PSD) is essential for single channel speech enhancement algorithms. We propose a noise PSD estimation approach based on regional statistics which consist of four features representing the statistics of the past and present periodograms in a short-time period. We show that these features are efficient in characterizing the statistical difference between noise PSD and noisy-speech PSD. We therefore propose to use these features for estimating the speech presence probability (SPP). The noise PSD is recursively estimated by averaging past spectral power values with a time-varying smoothing parameter controlled by the SPP. The proposed method exhibits good tracking capability for non-stationary noise, even for abruptly increasing noise level.

Website:
https://team.inria.fr/perception/research/noise-psd/

6.4. Tracking Multiple Persons

Object tracking is an ubiquitous problem in computer vision with many applications in human-machine and human-robot interaction, augmented reality, driving assistance, surveillance, etc. Although thoroughly investigated, tracking multiple persons remains a challenging and an open problem. In this work, an online variational Bayesian model for multiple-person tracking is proposed. This yields a variational expectation-maximization (VEM) algorithm. The computational efficiency of the proposed method is made possible thanks to closed-form expressions for both the posterior distributions of the latent variables and for the estimation of the model parameters. A stochastic process that handles person birth and person death enables the tracker to handle a varying number of persons over long periods of time [24], [30].

Website:
https://team.inria.fr/perception/research/ovbt/

6.5. Audio-Visual Speaker Detection, Localization, and Diarization

Any multi-party conversation system benefits from speaker diarization, that is, the assignment of speech signals among the participants. More generally, in HRI and CHI scenarios it is important to recognize the speaker over time. We propose to address speaker detection, localization and diarization using both audio and visual data. We cast the diarization problem into a tracking formulation whereby the active speaker is detected and tracked over time. A probabilistic tracker exploits the spatial coincidence of visual and auditory observations and infers a single latent variable which represents the identity of the active speaker. Visual and auditory observations are fused using our recently developed weighted-data mixture model [25], while several options for the speaking turns dynamics are fulfilled by a multi-case transition model. The modules that translate raw audio and visual data into image observations are also described in detail. The performance of the proposed method are tested on challenging data-sets that are available from recent contributions which are used as baselines for comparison [26].

Websites:

https://team.inria.fr/perception/research/wdgmm/
https://team.inria.fr/perception/research/speakerloc/
https://team.inria.fr/perception/research/speechturndet/
https://team.inria.fr/perception/research/avdiarization/

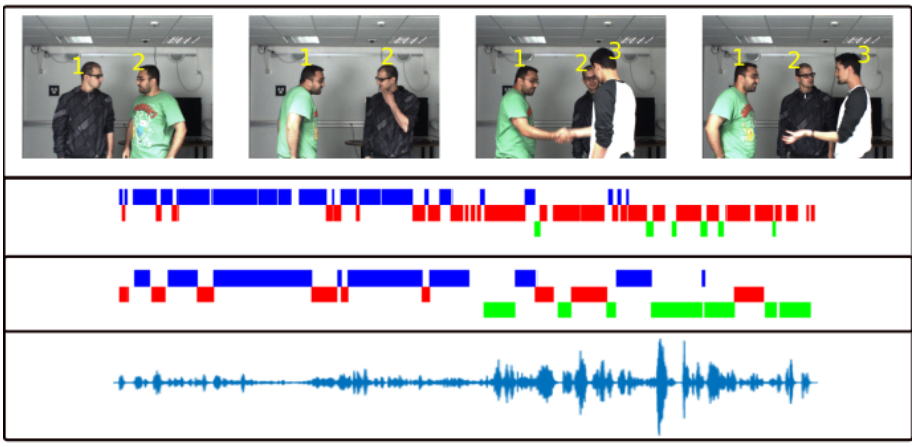


Figure 3. This figure illustrates the audiovisual tracking and diarization method that we have recently developed. First row: A number is associated with each tracked person. Second row: diarization result. Third row: the ground truth diarization. Fourth row: acoustic signal recorded by one of the two microphones.

6.6. Head Pose Estimation and Tracking

Head pose estimation is an important task, because it provides information about cognitive interactions that are likely to occur. Estimating the head pose is intimately linked to face detection. We addressed the problem of head pose estimation with three degrees of freedom (pitch, yaw, roll) from a single image and in the presence of face detection errors. Pose estimation is formulated as a high-dimensional to low-dimensional mixture of linear regression problem [6]. We propose a method that maps HOG-based descriptors, extracted from face bounding boxes, to corresponding head poses. To account for errors in the observed bounding-box position, we learn regression parameters such that a HOG descriptor is mapped onto the union of a head pose and an offset, such that the latter optimally shifts the bounding box towards the actual position of the face in the image. The performance of the proposed method is assessed on publicly available datasets. The experiments that we carried out show that a relatively small number of locally-linear regression functions is sufficient to deal with the non-linear mapping problem at hand. Comparisons with state-of-the-art methods show that our method outperforms several other techniques [42]. This work is part of the PhD of Vincent Drouard and it received the best student paper award (second place) at the IEEE ICIP'15. Currently we investigate a temporal extension of this model.

Website:
https://team.inria.fr/perception/research/head-pose/

6.7. Estimation of Eye Gaze and of Visual Focus of Attention

We address the problem of estimating the visual focus of attention (VFOA), e.g. who is looking at whom? This is of particular interest in human-robot interactive scenarios, e.g. when the task requires to identify targets of interest and to track them over time. We make the following contributions. We propose a Bayesian temporal model that links VFOA to eye-gaze direction and to head orientation. Model inference is cast into a switching Kalman filter formulation, which makes it tractable. The model parameters are estimated via training based on manual annotations. The method is tested and benchmarked using a publicly available dataset. We show that both eye-gaze and VFOA of several persons can be reliably and simultaneously estimated and tracked over time from observed head poses as well as from people and object locations [40].

Website:
https://team.inria.fr/perception/research/eye-gaze/.

6.8. High-Resolution Scene Reconstruction

We addressed the problem of range-stereo fusion for the construction of high-resolution depth maps. In particular, we combine time-of-flight (low resolution) depth [27] data with high-resolution stereo data, in a maximum a posteriori (MAP) formulation. Unlike existing schemes that build on MRF optimizers, we infer the disparity map from a series of local energy minimization problems that are solved hierarchically, by growing sparse initial disparities obtained from the depth data. The accuracy of the method is not compromised, owing to three properties of the data-term in the energy function. Firstly, it incorporates a new correlation function that is capable of providing refined correlations and disparities, via sub-pixel correction. Secondly, the correlation scores rely on an adaptive cost aggregation step, based on the depth data. Thirdly, the stereo and depth likelihoods are adaptively fused, based on the scene texture and camera geometry. These properties lead to a more selective growing process which, unlike previous seed-growing methods, avoids the tendency to propagate incorrect disparities. The proposed method gives rise to an intrinsically efficient algorithm, which runs at 3FPS on 2.0MP images on a standard desktop computer. The strong performance of the new method is established both by quantitative comparisons with state-of-the-art methods, and by qualitative comparisons using real depth-stereo data-sets [8]. This work is funded by the ANR project MIXCAM.

Website:
https://team.inria.fr/perception/research/dsfusion/

6.9. Registration of Multiple Point Sets

We have also addressed the rigid registration problem of multiple 3D point sets. While the vast majority of state-of-the-art techniques build on pairwise registration, we proposed a generative model that explains jointly registered multiple sets: back-transformed points are considered realizations of a single Gaussian mixture model (GMM) whose means play the role of the (unknown) scene points. Under this assumption, the joint registration problem is cast into a probabilistic clustering framework. We formally derive an expectation-maximization procedure that robustly estimates both the GMM parameters and the rigid transformations that map each individual cloud onto an under-construction reference set, that is, the GMM means. GMM variances carry rich information as well, thus leading to a noise- and outlier-free scene model as a by-product. A second version of the algorithm is also proposed whereby newly captured sets can be registered online. A thorough discussion and validation on challenging data-sets against several state-of-the-art methods confirm the potential of the proposed model for jointly registering real depth data [43].

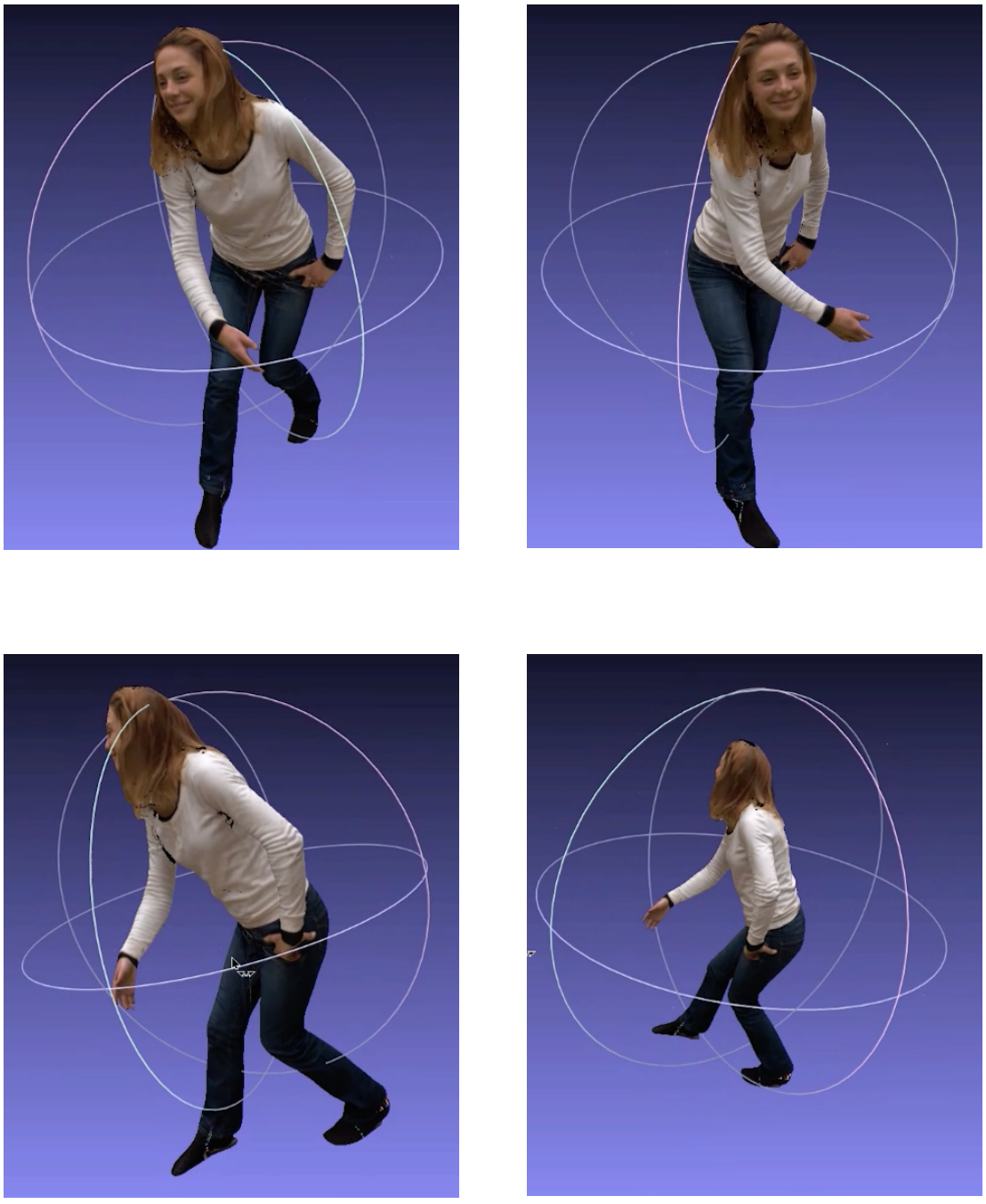


Figure 4. Four views of a 3D person reconstructed with our algorithm. In this example we used a large number of high-resolution cameras and the rendering was performed by the software of 4D View Solutions.

PERVASIVE INTERACTION Team

6. New Results

6.1. Simulating Haptic Sensations

Participants: Jingtao Chen, Sabine Coquillart, Partners: Inria GRA, LIG, GIPSA, G-SCOP

Pseudo-haptic feedback is a technique aiming to simulate haptic sensations without active haptic feedback devices. Peudo-haptic techniques have been used to simulate various haptic feedbacks such as stiffness, torques, and mass. In the framework of the Persyval project, a novel pseudo-haptic experiment has been set up. The aim of this experiment is to study the force and EMG signals during a pseudo-haptic task. A stiffness discrimination task similar to the one published in Lecuyer's PhD thesis has been chosen. The experimental set-up has been developed, as well as the software controlling the experiment. Pre-tests have been conducted. They have been followed by formal tests with subjects.

THOTH Project-Team

7. New Results

7.1. Visual recognition in images

7.1.1. Convolutional Neural Fabrics

Participants: Shreyas Saxena, Jakob Verbeek.

Despite the success of CNNs, selecting the optimal architecture for a given task remains an open problem. Instead of aiming to select a single optimal architecture, in this work [20], we propose a "fabric" that embeds an exponentially large number of architectures. See 1 for a schematic illustration of how fabrics embed different architectures. The fabric consists of a 3D trellis that connects response maps at different layers, scales, and channels with a sparse homogeneous local connectivity pattern. The only hyper-parameters of a fabric are the number of channels and layers. While individual architectures can be recovered as paths, the fabric can in addition ensemble all embedded architectures together, sharing their weights where their paths overlap. Parameters can be learned using standard methods based on back-propagation, at a cost that scales linearly in the fabric size. We present benchmark results competitive with the state of the art for image classification on MNIST and CIFAR10, and for semantic segmentation on the Part Labels dataset.

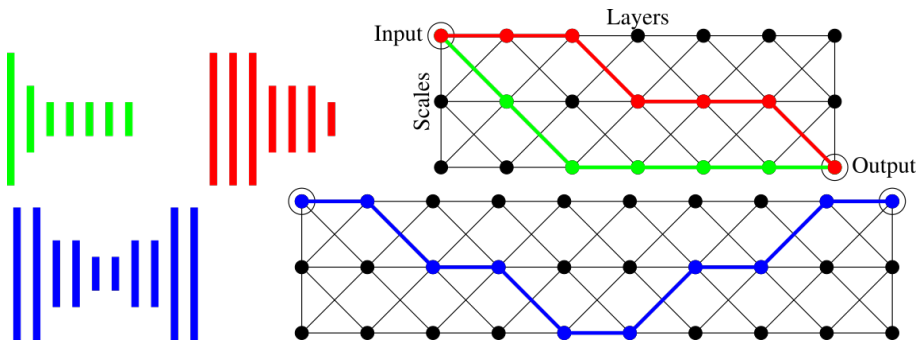


Figure 1. Fabrics embedding two seven-layer CNNs (red, green) and a ten-layer deconvolutional network (blue). Feature map size of the CNN layers are given by height. Fabric nodes receiving input and producing output are encircled. All edges are oriented to the right, down in the first layer, and towards the output in the last layer. The channel dimension of the 3D fabric is omitted for clarity.

7.1.2. Heterogeneous Face Recognition with CNNs

Participants: Shreyas Saxena, Jakob Verbeek.

Heterogeneous face recognition aims to recognize faces across different sensor modalities, see 2 for a schematic illustration. Typically, gallery images are normal visible spectrum images, and probe images are infrared images or sketches. Recently significant improvements in visible spectrum face recognition have been obtained by CNNs learned from very large training datasets. In this paper [21], we are interested in the question to what extent the features from a CNN pre-trained on visible spectrum face images can be used to perform heterogeneous face recognition. We explore different metric learning strategies to reduce the discrepancies between the different modalities. Experimental results show that we can use CNNs trained on visible spectrum images to obtain results that are on par or improve over the state-of-the-art for heterogeneous recognition with near-infrared images and sketches.

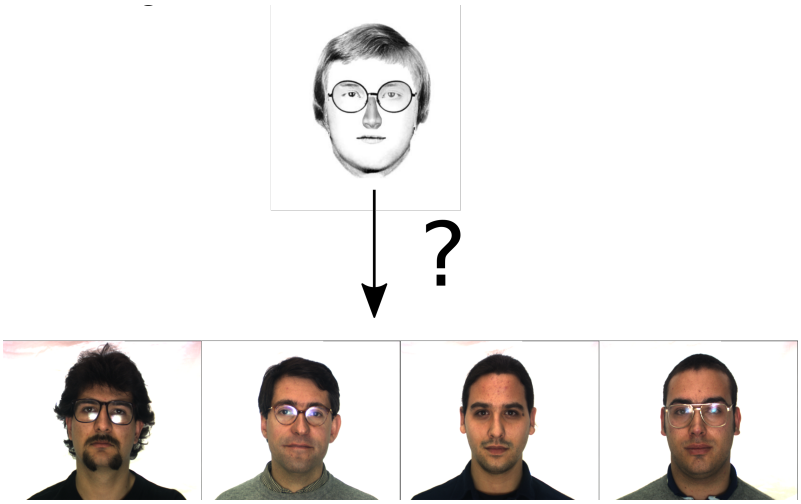


Figure 2. Schematic illustration for the task of heterogenous face recognition. The goal is to find the identity of the probe image (shown as a sketch) among one of the identities from the gallery set (shown in the bottom row). In contrast to standard face recognition, the probe and the gallery set do not share the same modality. In the illustartion, the probe image is a sketch and the galley images are normal visible spectrum images.

7.1.3. Mocap-guided Data Augmentation for 3D Pose Estimation in the Wild

Participants: Grégory Rogez, Cordelia Schmid.

In this paper [19], we address the problem of 3D human pose estimation in the wild. A significant challenge is the lack of training data, i.e., 2D images of humans annotated with 3D poses. Such data is necessary to train state-of-the-art CNN architectures. Here, we propose a solution to generate a large set of photorealistic synthetic images of humans with 3D pose annotations. We introduce an image-based synthesis engine that artificially augments a dataset of real images with 2D human pose annotations using 3D Motion Capture (MoCap) data. Given a candidate 3D pose our algorithm selects for each joint an image whose 2D pose locally matches the projected 3D pose. The selected images are then combined to generate a new synthetic image by stitching local image patches in a kinematically constrained manner. See examples in Figure 3 . The resulting images are used to train an end-to-end CNN for full-body 3D pose estimation. We cluster the training data into a large number of pose classes and tackle pose estimation as a K-way classification problem. Such an approach is viable only with large training sets such as ours. Our method outperforms the state of the art in terms of 3D pose estimation in controlled environments (Human3.6M) and shows promising results for in-the-wild images (LSP). This demonstrates that CNNs trained on artificial images generalize well to real images.

7.1.4. End-to-End Kernel Learning with Supervised Convolutional Kernel Networks

Participant: Julien Mairal.

In [16], we introduce a new image representation based on a multilayer kernel machine. Unlike traditional kernel methods where data representation is decoupled from the prediction task, we learn how to shape the kernel with supervision. We proceed by first proposing improvements of the recently-introduced convolutional kernel networks (CKNs) in the context of unsupervised learning; then, we derive backpropagation rules to take advantage of labeled training data. The resulting model is a new type of convolutional neural network, where optimizing the filters at each layer is equivalent to learning a linear subspace in a reproducing kernel Hilbert space (RKHS). We show that our method achieves reasonably competitive performance for image classification on some standard " deep learning " datasets such as CIFAR-10 and SVHN, and also for image

Figure 3. Given a candidate 3D pose, our algorithm selects for each joint an image whose annotated 2D pose locally matches the projected 3D pose. The selected images are then combined to generate a new synthetic image by stitching local image patches in a kinematically constrained manner. We show 6 examples corresponding to the same 3D pose observed from 6 different camera viewpoints.

super-resolution, demonstrating the applicability of our approach to a large variety of image-related tasks. The model is illustrated in Figure 4 .

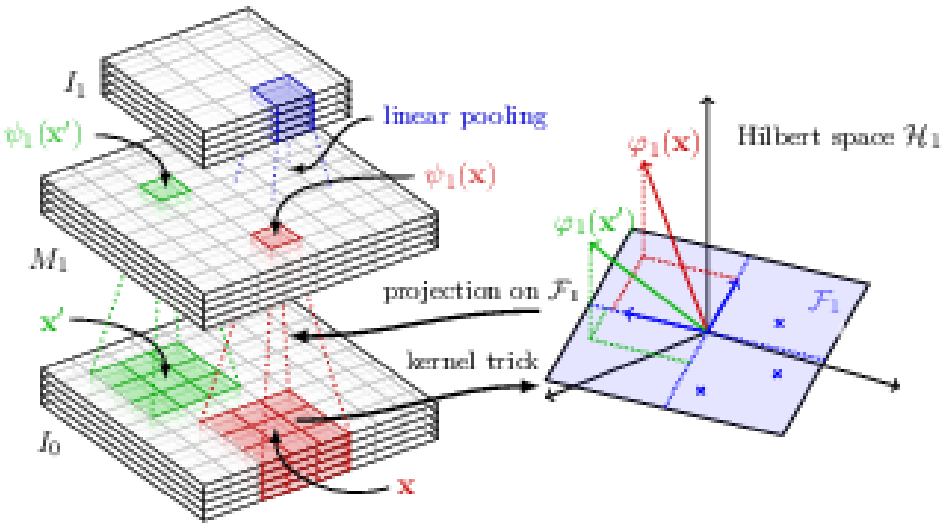


Figure 4. Our variant of convolutional kernel networks, illustrated between layers 0 and 1.

7.1.5. Semantic segmentation using Adversarial Networks

Participants: Pauline Luc, Camille Couprie [Facebook], Soumith Chintala [Facebook], Jakob Verbeek.

Adversarial training has been shown to produce state of the art results for generative image modeling. In [24], we propose an adversarial training approach to train semantic segmentation models. We train a convolutional semantic segmentation network along with an adversarial network that discriminates segmentation maps coming either from the ground truth or from the segmentation network, as shown in Figure 5 . The motivation for our approach is that it can detect and correct higher-order inconsistencies between ground truth segmentation maps and the ones produced by the segmentation net. Our experiments show that our adversarial training approach leads to improved accuracy on the Stanford Background and PASCAL VOC 2012 datasets.

7.1.6. Enhancing Energy Minimization Framework for Scene Text Recognition with Top-Down Cues

Participants: Anand Mishra [IIIT Hyderabad], Karteek Alahari, C. v. Jawahar [IIIT Hyderabad].

Recognizing scene text, i.e., text in images such as the one in Figure 6 , is a challenging problem, even more so than the recognition of scanned documents. This problem has gained significant attention from the computer vision community in recent years, and several methods based on energy minimization frameworks and deep learning approaches have been proposed. In our work presented in [8], we focus on the energy minimization framework and propose a model that exploits both bottom-up and top-down cues for recognizing cropped words extracted from street images. The bottom-up cues are derived from individual character detections from an image. We build a conditional random field model on these detections to jointly model the strength of the detections and the interactions between them. These interactions are top-down cues obtained from a lexicon-based prior, i.e., language statistics. The optimal word represented by the text image is obtained by minimizing the energy function corresponding to the random field model. We evaluate our proposed algorithm extensively on a number of cropped scene text benchmark datasets, namely Street View Text, ICDAR 2003, 2011 and 2013 datasets, and IIIT 5K-word, and show better performance than comparable methods. We perform a

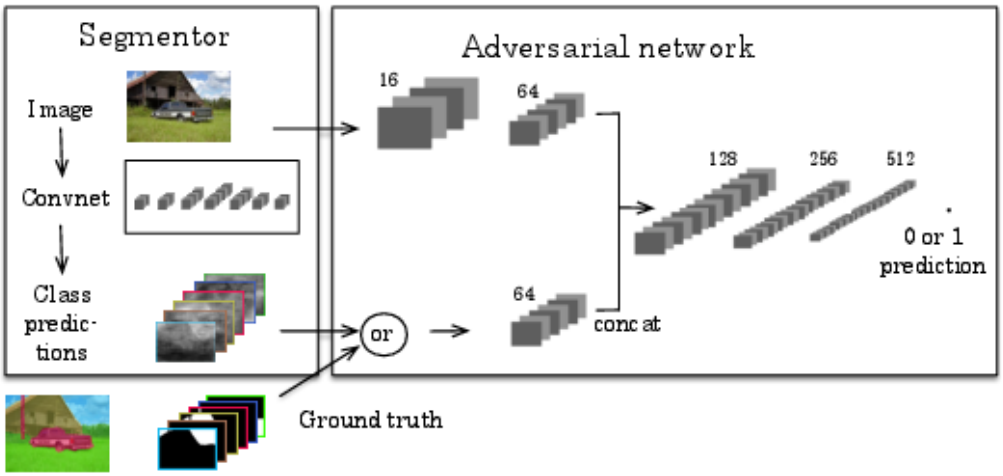


Figure 5. We use adversarial training to simultaneously learn a segmentation model (left) and a high order loss term to train it, given by the adversarial network (right). This encourages the segmentation model to output plausible segmentations, by enforcing forms of high order consistencies that are learned rather than manually designed.

rigorous analysis of all the steps in our approach and analyze the results. We also show that state-of-the-art convolutional neural network features can be integrated in our framework to further improve the recognition performance.

7.1.7. Local Convolutional Features with Unsupervised Training for Image Retrieval

Participants: Mattis Paulin, Matthijs Douze [Facebook], Zaid Harchaoui [University of Washington], Julien Mairal, Florent Perronnin [Xerox], Cordelia Schmid.

Patch-level descriptors underlie several important computer vision tasks, such as stereo-matching or content-based image retrieval. We introduce a deep convolutional architecture that yields patch-level descriptors, as an alternative to the popular SIFT descriptor for image retrieval. The proposed family of descriptors, called Patch-CKN[9], adapt the recently introduced Convolutional Kernel Network (CKN), an unsupervised framework to learn convolutional architectures. We present a comparison framework to benchmark current deep convolutional approaches along with Patch-CKN for both patch and image retrieval (see Fig. 7 for our pipeline), including our novel "RomePatches" dataset. Patch-CKN descriptors yield competitive results compared to supervised CNNs alternatives on patch and image retrieval.

7.2. Visual recognition in videos

7.2.1. Towards Weakly-Supervised Action Localization

Participants: Philippe Weinzaepfel, Xavier Martin, Cordelia Schmid.

In this paper [33], we present a novel approach for weakly-supervised action localization, i.e., that does not require per-frame spatial annotations for training. We first introduce an effective method for extracting human tubes by combining a state-of-the-art human detector with a tracking-by-detection approach. Our tube extraction leverages the large amount of annotated humans available today and outperforms the state of the art by an order of magnitude: with less than 5 tubes per video, we obtain a recall of 95% on the UCF-Sports and

Figure 6. A typical street scene image taken from Google Street View. It contains very prominent sign boards with text on the building and its windows. It also contains objects such as car, person, tree, and regions such as road, sky. Many scene understanding methods recognize these objects and regions in the image successfully, but overlook the text on the sign board, which contains rich, useful information. The goal of our work [8] is to address this gap in understanding scenes.

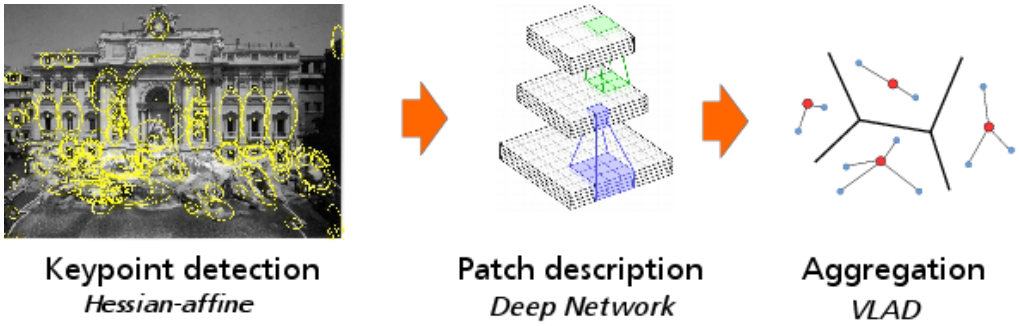


Figure 7. Image retrieval pipeline. Interest points are extracted with the Hessian-affine detector (left), encoded in descriptor space using convolutional features (middle), and aggregated into a compact representation using VLAD-pooling (right).

J-HMDB datasets. Given these human tubes, we perform weakly-supervised selection based on multi-fold Multiple Instance Learning (MIL) with improved dense trajectories and achieve excellent results. Figure 8 summarizes the approach. We obtain a mAP of 84% on UCF-Sports, 54% on J-HMDB and 45% on UCF-101, which outperforms the state of the art for weakly-supervised action localization and is close to the performance of the best fully-supervised approaches. The second contribution of this paper is a new realistic dataset for action localization, named DALY (Daily Action Localization in YouTube). It contains high quality temporal and spatial annotations for 10 actions in 31 hours of videos (3.3M frames), which is an order of magnitude larger than standard action localization datasets. On the DALY dataset, our tubes have a spatial recall of 82%, but the detection task is extremely challenging, we obtain 10.8% mAP.

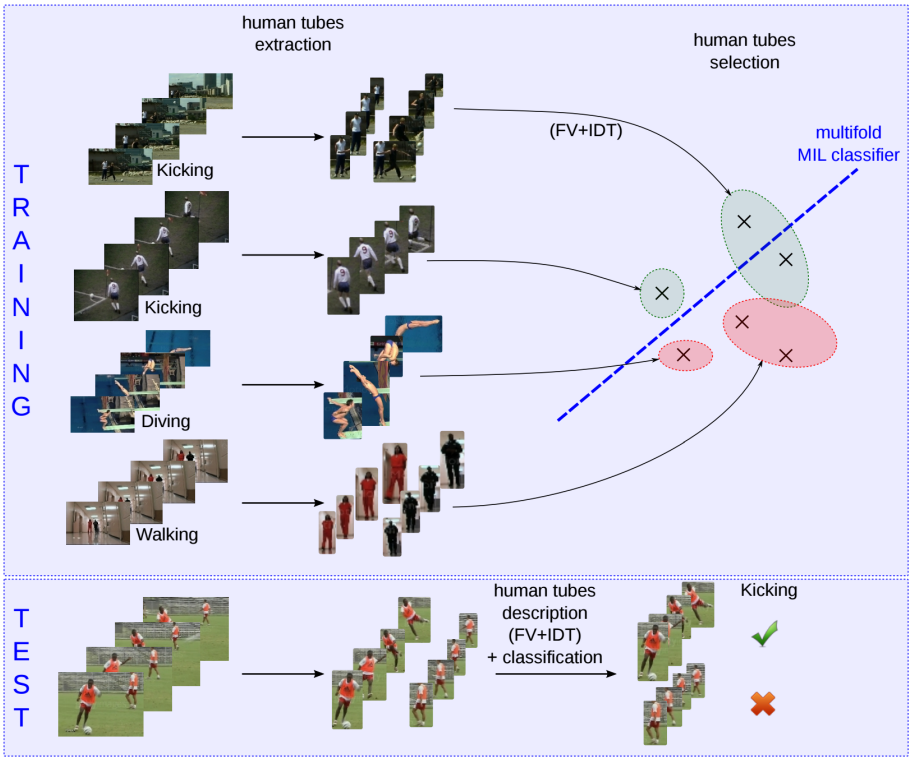


Figure 8. Overview of our approach for action localization without spatial supervision.

7.2.2. The DALY dataset

Participants: Philippe Weinzaepfel, Xavier Martin, Cordelia Schmid.

We introduce a new action localization dataset named DALY (Daily Action Localization in YouTube). DALY consists of more than 31 hours of videos (3.3M frames) from YouTube with 10 realistic daily actions, see Figure 9 , and 3.6k spatio-temporal instances. Annotations consist in the start and end time of each action instance, with high-quality spatial annotation for a sparse subset of frames. The task is to localize relatively short actions (8 seconds in average) in long untrimmed videos (3min 45 in average). Furthermore, it includes videos with multiple humans performing actions simultaneously. It overcomes the limitations of existing benchmarks that are limited to trimmed or almosttrimmed videos with specific action types, e.g. sports only, showing in most cases one human per video.

Figure 9. Overview of our approach for action localization without spatial supervision.

7.2.3. Weakly-Supervised Semantic Segmentation using Motion Cues

Participants: Pavel Tokmakov, Karteek Alahari, Cordelia Schmid.

Fully convolutional neural networks (FCNNs) trained on a large number of images with strong pixel-level annotations have become the new state of the art for the semantic segmentation task. While there have been recent attempts to learn FCNNs from image-level weak annotations, they need additional constraints, such as the size of an object, to obtain reasonable performance. To address this issue, in [23] we present motion-CNN (M-CNN), a novel FCNN framework which incorporates motion cues and is learned from video-level weak annotations. Our learning scheme to train the network uses motion segments as soft constraints, thereby handling noisy motion information, as shown in Figure 10 . When trained on weakly-annotated videos, our method outperforms the state-of-the-art EM-Adapt approach on the PASCAL VOC 2012 image segmentation benchmark. We also demonstrate that the performance of M-CNN learned with 150 weak video annotations is on par with state-of-the-art weakly-supervised methods trained with thousands of images. Finally, M-CNN substantially outperforms recent approaches in a related task of video co-localization on the YouTube-Objects dataset.

7.2.4. Multi-region two-stream R-CNN for action detection

Participants: Xiaojiang Peng, Cordelia Schmid.

This work [18] introduces a multi-region two-stream R-CNN model for action detection, see Figure 11 . It starts from frame-level action detection based on faster R-CNN and makes three contributions. The first one is the introduction of a motion region proposal network (RPN) complementary to a standard appearance RPN. The second is the stacking of optical flow over several frames, which significantly improves frame-level action detection. The third is the addition of a multi-region scheme to the faster R-CNN model, which adds complementary information on body parts. Frame-level detections are linked with the Viterbi algorithm, and action are temporally localized with the maximum subarray method. Experimental results on the UCF-Sports, J-HMDB and UCF101 action detection datasets show that the approach outperforms the state of the art with a significant margin in both frame-mAP and video-mAP.

7.2.5. Analysing domain shift factors between videos and images for object detection

Participants: Vicky Kalogeiton, Vittorio Ferrari [Univ. Edinburgh], Cordelia Schmid.

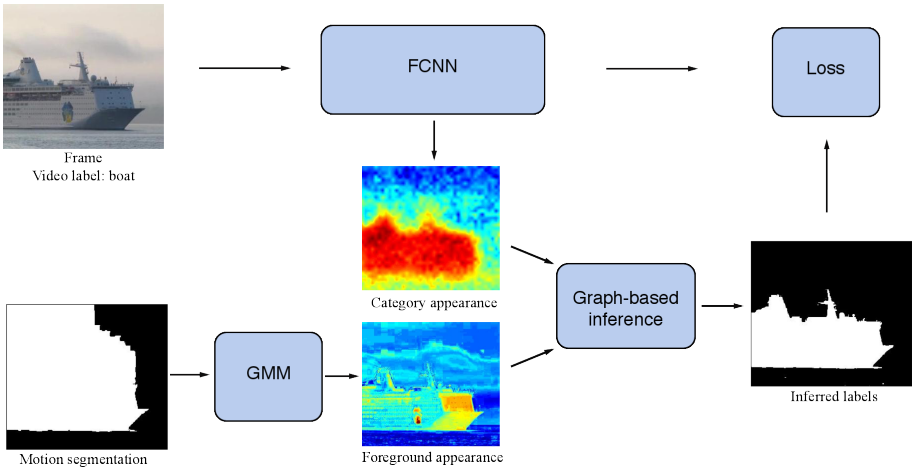


Figure 10. Overview of our M-CNN framework, where we show only one frame from a video example for clarity. The soft potentials (foreground appearance) computed from motion segmentation and the FCNN predictions (category appearance) jointly determine the latent segmentation (inferred labels) to compute the loss, and thus the network update.

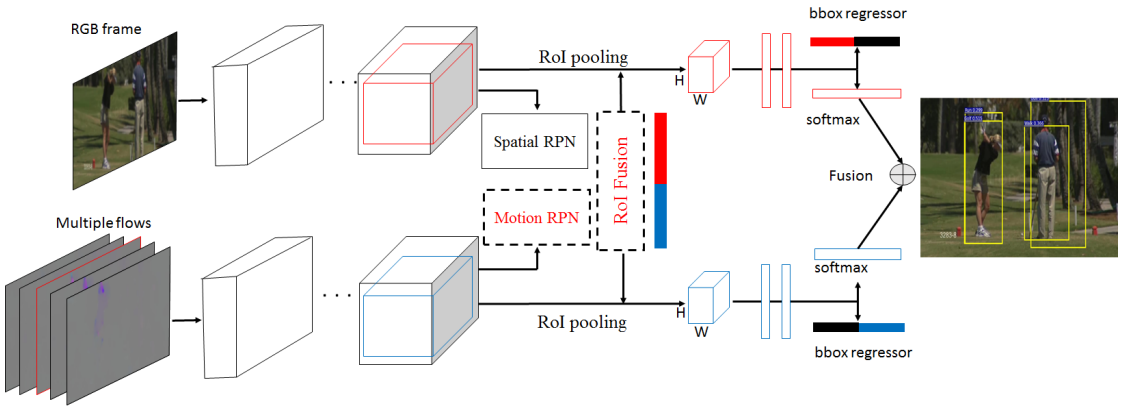


Figure 11. Two-stream faster R-CNN for spatio-temporal action detection.

Object detection is one of the most important challenges in computer vision. Object detectors are usually trained on bounding-boxes from still images. Recently, video has been used as an alternative source of data. Yet, for a given test domain (image or video), the performance of the detector depends on the domain it was trained on. In this paper [7], we examine the reasons behind this performance gap. We define and evaluate different domain shift factors (see Figure 12): spatial location accuracy, appearance diversity, image quality and aspect distribution. We examine the impact of these factors by comparing performance before and after factoring them out. The results show that all four factors affect the performance of the detectors and their combined effect explains nearly the whole performance gap.

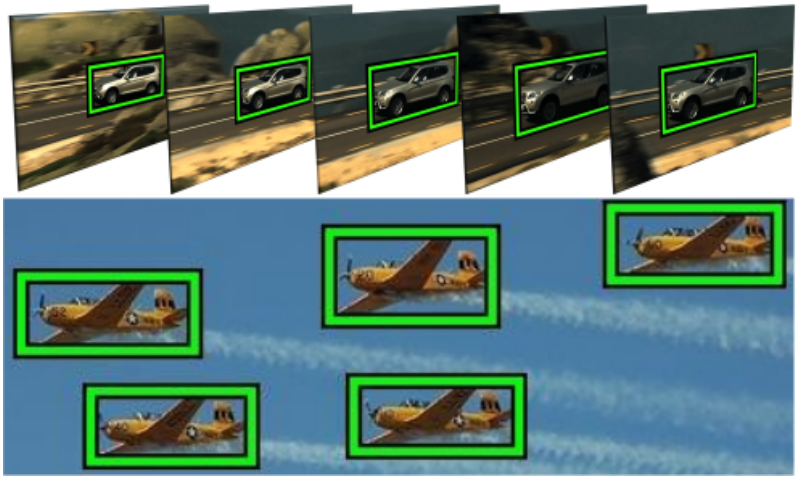


Figure 12. Example of apperance diversity domain shift factor. (top row): Frames in the same shot that contain near identical samples of an object. (bottom row): Example of near identical samples in the same image.

7.3. Large-scale statistical learning

7.3.1. Dictionary Learning for Massive Matrix Factorization

Participants: Julien Mairal, Arthur Mensch [Parietal], Gael Varoquaux [Parietal], Bertrand Thirion [Parietal].

Sparse matrix factorization is a popular tool to obtain interpretable data decompositions, which are also effective to perform data completion or denoising. Its applicability to large datasets has been addressed with online and randomized methods, that reduce the complexity in one of the matrix dimension, but not in both of them. In [25], [17], we tackle very large matrices in both dimensions. We propose a new factorization method that scales gracefully to terabyte-scale datasets. Those could not be processed by previous algorithms in a reasonable amount of time. We demonstrate the efficiency of our approach on massive functional Magnetic Resonance Imaging (fMRI) data, and on matrix completion problems for recommender systems, where we obtain significant speed-ups compared to state-of-the art coordinate descent methods. The main principle of the method is illustrated in Figure 13 .

7.3.2. Stochastic Optimization with Variance Reduction for Infinite Datasets with Finite-Sum Structure

Participants: Alberto Bietti, Julien Mairal.

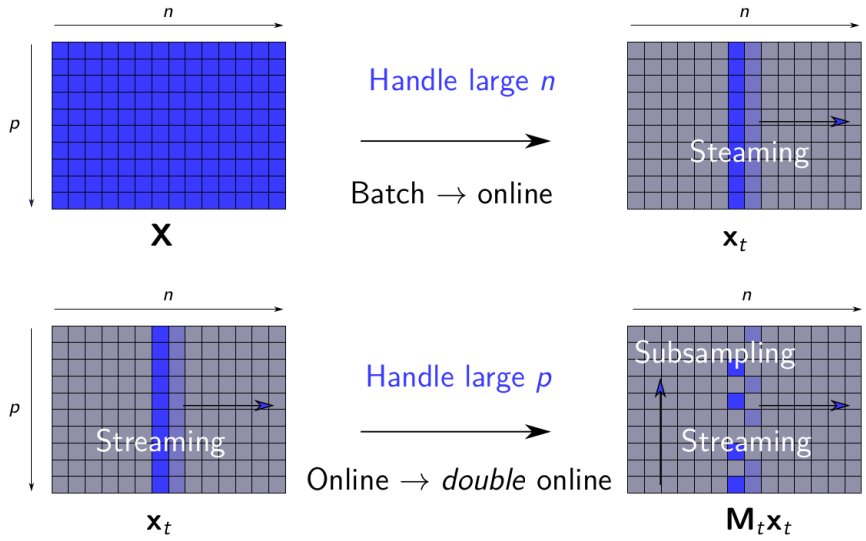


Figure 13. Illustration of the matrix factorization algorithm, which streams columns in one dimension while subsampling them.

Stochastic optimization algorithms with variance reduction have proven successful for minimizing large finite sums of functions. However, in the context of empirical risk minimization, it is often helpful to augment the training set by considering random perturbations of input examples. In this case, the objective is no longer a finite sum, and the main candidate for optimization is the stochastic gradient descent method (SGD). In this paper [26], we introduce a variance reduction approach for this setting when the objective is strongly convex. After an initial linearly convergent phase, the algorithm achieves a $O(1/t)$ convergence rate in expectation like SGD, but with a constant factor that is typically much smaller, depending on the variance of gradient estimates due to perturbations on a single example.

7.3.3. QuickeNing: A Generic Quasi-Newton Algorithm for Faster Gradient-Based Optimization

Participants: Hongzhou Lin, Julien Mairal, Zaid Harchaoui [University of Washington].

In this paper [28], we propose an approach to accelerate gradient-based optimization algorithms by giving them the ability to exploit curvature information using quasi-Newton update rules. The proposed scheme, called QuickeNing, is generic and can be applied to a large class of first-order methods such as incremental and block-coordinate algorithms; it is also compatible with composite objectives, meaning that it has the ability to provide exactly sparse solutions when the objective involves a sparsity-inducing regularization. QuickeNing relies on limited-memory BFGS rules, making it appropriate for solving high-dimensional optimization problems; with no line-search, it is also simple to use and to implement. Besides, it enjoys a worst-case linear convergence rate for strongly convex problems. We present experimental results, see Figure 14 , where QuickeNing gives significant improvements over competing methods for solving large-scale high-dimensional machine learning problems.

7.3.4. Dictionary Learning from Phaseless Measurements

Participants: Julien Mairal, Yonina Eldar [Technion], Andreas Tillmann [TU Darmstadt].

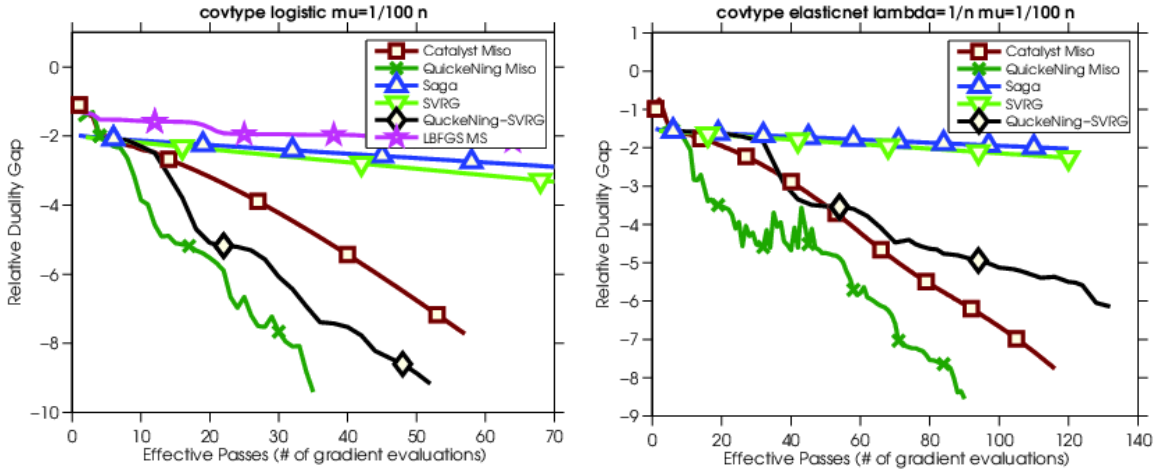


Figure 14. Relative duality gap for different number of passes performed over dataset covtype.

In [22], [12], we propose a new algorithm to learn a dictionary for reconstructing and sparsely encoding signals from measurements without phase. Specifically, we consider the task of estimating a two-dimensional image from squared-magnitude measurements of a complex-valued linear transformation of the original image. Several recent phase retrieval algorithms exploit underlying sparsity of the unknown signal in order to improve recovery performance. In this work, we consider such a sparse signal prior in the context of phase retrieval, when the sparsifying dictionary is not known in advance. Our algorithm jointly reconstructs the unknown signal—possibly corrupted by noise—and learns a dictionary such that each patch of the estimated image can be sparsely represented. Numerical experiments demonstrate that our approach can obtain significantly better reconstructions for phase retrieval problems with noise than methods that cannot exploit such "hidden" sparsity. Moreover, on the theoretical side, we provide a convergence result for our method.

TYREX Project-Team

6. New Results

6.1. Experimental evaluation of attitude estimation algorithms for smartphones

- **Context:** Pervasive applications on smartphones increasingly rely on techniques for estimating attitude. Attitude is the orientation of the smartphone with respect to Earth's local frame.

  Modern smartphones embed sensors such as accelerometer, gyroscope, and magnetometer which make it possible to leverage existing attitude estimation algorithms.
- **Contribution:** We investigated the precision of attitude estimation algorithms in the context of commodity smartphones carried by pedestrians. We considered eight typical motions (such as texting, phoning, running, etc.) with various impacts on external accelerations, as well as the presence/absence of magnetic perturbations typically found in indoor environments. We systematically analyzed, compared and evaluated eight state-of-the-art algorithms (and their variants). We precisely quantified the attitude estimation error obtained with each technique, owing to the use of a precise ground truth obtained with a motion capture system (the Inria Kinovis platform). We made our benchmark available (see Sec. 5.1 above) and payed attention to the reproducibility of results. We analyzed and discussed the obtained results and reported on lessons learned [7] [17]. We also presented a new technique which helps in improving precision by limiting the effect of magnetic perturbations with all considered algorithms.

6.2. Efficient Distributed Evaluation of SPARQL Queries

- **Context:** SPARQL is the standard query language for retrieving and manipulating data represented in the Resource Description Framework (RDF). SPARQL constitutes one key technology of the semantic web and has become very popular since it became an official W3C recommendation.

  The construction of efficient SPARQL query evaluators faces several challenges. First, RDF datasets are increasingly large, with some already containing more than a billion triples. To handle efficiently this growing amount of data, we need systems to be distributed and to scale. Furthermore, semantic data often have the characteristic of being dynamic (frequently updated). Thus being able to answer quickly after a change in the input data constitutes a very desirable property for a SPARQL evaluator.
- **Contributions:** First of all, to constitute a common basis of comparative analysis, we evaluated on the same cluster of machines various SPARQL evaluation systems from the literature [15]. These experiments led us to point several observations: (i) the solutions have very different behaviors; (ii) most of the benchmarks only use temporal metrics and forget other ones e.g. network traffic. That is why we proposed a larger set of metrics; and thanks to a new reading grid based on 5 features, we proposed new perspectives which should be considered when developing distributed SPARQL evaluators.

  Second, we developed and shared several distributed SPARQL evaluators which take into account these new considerations we introduced:

  – A SPARQL evaluator named SPARQLGX (see Sec. 5.6): an implementation of a distributed RDF datastore based on Apache Spark. SPARQLGX is designed to leverage existing Hadoop infrastructures for evaluating SPARQL queries. It relies on a translation of SPARQL queries into executable Spark code that adopts evaluation strategies according to the storage method used and statistics on data.

    In [12], [11], [8], [13], we showed that SPARQLGX makes it possible to evaluate SPARQL queries on billions of triples distributed across multiple nodes, while providing attractive

performance figures. We reported on experiments which show how SPARQLGX compares to related state-of-the-art implementations and we showed that our approach scales better than these systems in terms of supported dataset size. With its simple design, SPARQLGX represents an interesting alternative in several scenarios.

– Two SPARQL direct evaluators i.e. without a preprocessing phase: SDE (stands for Sparqlgx Direct Evaluator) lays on the same strategy than SPARQLGX but the translation process is modified in order to take the orign data files as argument. RDFHive (see Sec. 5.3) evaluates translated SPARQL queries on top of Apache Hive which is a distributed relational data warehouse based on Apache Hadoop.

6.3. An Efficient Translation from a modal μ-Calculus with Converse to Tree Automata

In [16], we presented a direct translation from a sub-logic of μ-calculus to non-deterministic binary automata of finite trees. The logic is an alternation-free modal μ-calculus, restricted to finite trees and where formulae are cycle-free. This logic is expressive enough to encode significant fragments of query languages (such as Regular XPath). The size of the generated automaton (the number of transitions) is bounded by 2^n where n is the size of a Fischer-Ladner closure of the formula. This is an improvement over previous translations in 2^{n^2}. We have implemented our translation. In practice, our prototype effectively decides static analysis problems that were beyond reach, such as the XPath containment problem with DTDs of significant size.

6.4. SPARQL Query Containment with ShEx Constraints

ShEx (Shape Expressions) is a language for expressing constraints on RDF graphs. In [14], we considered the problem of SPARQL query containment in the presence of ShEx constraints. We first investigated the complexity of the problem according to the fragments considered for SPARQL queries and for ShEx constraints. In particular, we showed that the complexity of SPARQL query containment remains the same with or without ShEx constraints. We developed two radically different approaches for solving the problem and we evaluated them. The first approach relies on the joint use of a ShEx validator and a tool for checking query containment without constraints. In a second approach, we showed how the problem can be solved by a reduction to a fragment of first-order logic with two variables. This alternative approach allows to take advantage of any of the many existing FOL theorem provers in this context. We evaluated how the two approaches compare experimentally, and reported on lessons learned. To the best of our knowledge, this is the first work addressing SPARQL query containment in the presence of ShEx constraints.

6.5. XQuery Static Type-Checking

In the context of our ongoing work on XQuery static type-checking [3], we extended our type system and improved the associated software accordingly (see Sec. 5.5 and 5.4). The type language it is based on is now a subset of RelaxNG+Schematron (instead of DTDs), which is novel in the context of static typing: Schematron is normally used to validate a document after it has been generated, whereas our system is able to ensure statically that a program will always generate a valid document.

Schematron constraints present the advantage of describing some properties in a very concise way compared to schema languages based on regular tree types, e.g. it allows writing in one line that nested anchors are forbidden in HTML, a constraint which appears in the specification but not in the formal DTD schema because of the verbosity it would involve.