Activity Report 2015

Section Scientific Foundations

CAGIRE Team

3. Research Program

3.1. Computational fluid mechanics: modeling or not before discretizing ?

A typical continuous solution of the Navier Stokes equations at sufficiently high values of the Reynolds number is governed by a spectrum of time and space scales fluctuations closely connected with the turbulent nature of the flow. The term deterministic chaos employed by Frisch in its enlightening book [33] is certainly conveying most adequately the difficulty in analyzing and simulating this kind of flows. The broadness of the turbulence spectrum is directly controlled by the Reynolds number defined as the ratio between the inertial forces and the viscous forces. This number is not only useful to determine the transition from a laminar to a turbulent flow regime, it also indicates the range of scales of fluctuations that are present in the flow under consideration.

Typically, for the velocity field and far from solid walls, the ratio between the largest scale (the integral length scale) to the smallest one (Kolmogorov scale) scales as $Re^{3/4}$ per dimension. In addition, for internal flows, the viscous effects near the solid walls yield a scaling proportional to Re per dimension. The smallest scales play a crucial role in the dynamics of the largest ones which implies that an accurate framework for the computation of turbulent flows must take into account all these scales.

Thus, the usual practice to deal with turbulent flows is to choose between an a priori modeling (in most situations) or not (low Re number and rather simple configurations) before proceeding to the discretization step followed by the simulation runs themselves. If a modeling phase is on the agenda, then one has to choose again among the above mentioned variety of approaches.

As it is illustrated in Fig. 1 , this can be achieved either by directly solving the Navier-Stokes equations (DNS) or by first applying a statistical averaging (RANS) or a spatial filtering operator to the Navier-Stokes equations (LES). The new terms brought about by the filtering operator have to be modeled. From a computational point of view, the RANS approach is the least demanding, which explains why historically it has been the workhorse in both the academic and the industrial sectors. It has permitted quite a substantial progress in the understanding of various phenomena such as turbulent combustion or heat transfer. Its inherent inability to provide a time-dependent information has led to promote in the last decade the recourse to either LES or DNS to supplement if not replace RANS. By simulating the large scale structures while modeling the smallest ones supposed to be more isotropic, LES proved to be quite a step through that permits to fully take advantage of the increasing power of computers to study complex flow configurations. At the same time, DNS was progressively applied to geometries of increasing complexity (channel flows with values of Re_τ multiplied by 10 during the last 15 years, jets, turbulent premixed flames, among many others), and proved to be a formidable tool that permits **(i)** to improve our knowledge on turbulent flows and **(ii)** to test (i.e., validate or invalidate) and improve the modeling hypotheses inherently associated to the RANS and LES approaches. From a numerical point of view, if the steady nature of the RANS equations allows to perform iterative convergence on finer and finer meshes, the high computational cost of LES or DNS makes necessary the use of highly accurate numerical schemes in order to optimize the use of computational resources.

To the noticeable exception of the hybrid RANS-LES modeling, which is not yet accepted as a reliable tool for industrial design, as mentioned in the preamble of the Go4hybrid European program [0], once chosen, a single turbulence model will (try to) do the job for modeling the whole flow. Thus, depending on its intrinsic strengths and weaknesses, the accuracy will be a rather volatile quantity strongly dependent on the flow configuration. The turbulence modeling and industrial design communities waver between the desire to continue to rely on the RANS approach, which is unrivaled in terms of computational cost, but is still not able to accurately represent all the complex phenomena; and the temptation to switch to LES, which outperforms RANS in many situations but is prohibitively expensive in high-Reynolds number wall-bounded flows. In order to account for

---

[0]http://www.transport-research.info/web/projects/project_details.cfm?id=46810

the deficiencies of both approaches and to combine them for significantly improving the overall quality of the modeling, the hybrid RANS-LES approach has emerged during the last decade as a viable, intermediate way, and we are definitely inscribing our project in this innovative field of research, with an original approach though, connected with a time filtered hybrid RANS-LES and a systematic and progressive validation process against experimental data produced by the team.
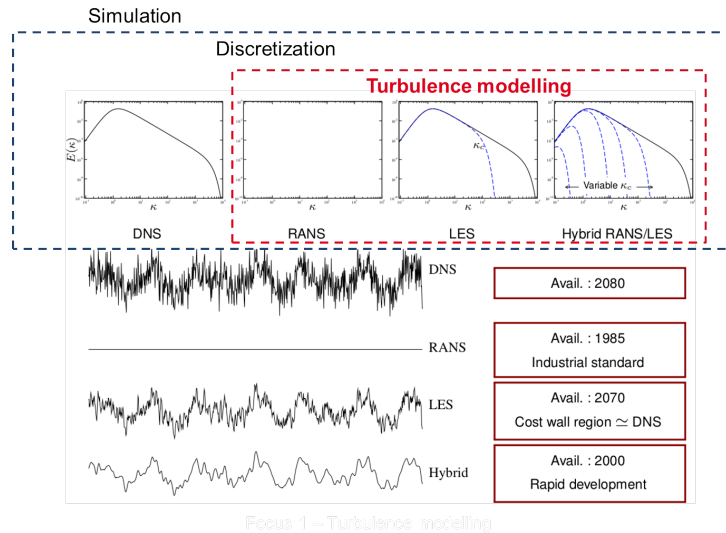

Figure 1. A schematic view of the different nested steps for turbulent flows simulation: from DNS to hybrid RANS-LES. The approximate dates at which the different approaches are or will be routinely used in the industry are indicated in the boxes on the right (extrapolations based on the present rate of increase in computer performances).

3.2. Computational fluid mechanics: high order discretization on unstructured meshes and efficient methods of solution

All the methods considered in the project are mesh-based methods: the computational domain is divided into cells, that have an elementary shape: triangles and quadrangles in two dimensions, and tetrahedra, hexahedra, pyramids, and prisms in three dimensions. If the cells are only regular hexahedra, the mesh is said to be structured. Otherwise, it is said to be unstructured. If the mesh is composed of more than one sort of elementary shape, the mesh is said to be hybrid.

The AeroSol library developed in the team is based on discontinuous Galerkin methods. These methods were introduced by Reed and Hill [38] and first studied by Lesaint and Raviart [36]. The extension to the Euler system with explicit time integration was mainly led by Shu, Cockburn and their collaborators. The steps of time integration and slope limiting were similar to high order ENO schemes, whereas specific constraints given by the finite element nature of the scheme were progressively solved, for scalar conservation laws [29], [28], one dimensional systems [27], multidimensional scalar conservation laws [26], and multidimensional systems [30]. For the same system, we can also cite the work of [32], [35], which is slightly different: the stabilization is made by adding a nonlinear term, and the time integration is implicit.

Contrary to continuous Galerkin methods, the discretization of diffusive operators is not straightforward. This is due to the discontinuous approximation space, which does not fit well with the space function in which

the diffusive system is well posed. A first stabilization was proposed by Arnold [20]. The first application of discontinuous Galerkin methods to Navier-Stokes equations was proposed in [24] by mean of a mixed formulation. Actually, this first attempt led to a non compact computation stencil, and was later proved to be not stable. A compactness improvement was made in [25], which was later analyzed, and proved to be stable in a more unified framework [21]. The combination with the $k - \omega$ RANS model was made in [23]. As far as Navier Stokes equations are concerned, we can also cite the work of [34], in which the stabilization is closer to the one of [21], the work of [37] on local time stepping, or the first use of discontinuous Galerkin methods for direct numerical simulation of a turbulent channel flow done in [31]. Discontinuous Galerkin methods are so popular because

- They can be developed for any order of approximation.
- The computational stencil of one given cell is limited to the cells with which it has a common face. This stencil does not depend on the order of approximation. This is a pro, compared for example with high order finite volumes, which require as more and more neighbors as the order increases.
- They can be developed for any kind of mesh, structured, unstructured, but also for aggregated grids [22]. This is a pro compared not only with finite differences schemes, which can be developed only on structured meshes, but also compared with continuous finite elements methods, for which the definition of the approximation basis is not clear on aggregated elements.
- p-adaptivity is easier than with continuous finite elements, because neighboring elements having a different order are only weakly coupled.
- Upwinding is as natural as for finite volumes methods, which is a benefit for hyperbolic problems.
- As the formulation is weak, boundary conditions are naturally weakly formulated. This is a benefit compared with strong formulations, for example point centered formulation when a point is at the intersection of two kinds of boundary conditions.

For concluding this section, there already exist numerical schemes based on the discontinuous Galerkin method which proved to be efficient for computing compressible viscous flows. Nevertheless, there remain many things to be improved, which include: efficient shock capturing methods for supersonic flows, high order discretization of curved boundaries, low Mach number behavior of these schemes and combination with second-moment RANS models. Another drawback of the discontinuous Galerkin methods is that they can be computationally costly, due to the accurate representation of the solution calling for a particular care of implementation for being efficient. We believe that this cost can be balanced by the strong memory locality of the method, which is an asset for porting on emerging many-core architectures.

3.3. Flow analysis and CFD assessment: experimental aspects

With the considerable and constant development of computer performance, many people were thinking at the turn of the 21st century that in the short term, CFD would replace experiments considered as too costly and not flexible enough. Simply flipping through scientific journals such as Journal of Fluid Mechanics, Combustion of Flame, Physics of Fluids or Journal of Computational Physics or through websites such that of Ercoftac [0] is sufficient to convince oneself that the recourse to experiments to provide either a quantitative description of complex phenomena or reference values for the assessment of the predictive capabilities of the physical modeling and of the related simulations is still necessary. The major change that can be noted though concerns the content of the interaction between experiments and CFD (understood in the broad sense). Indeed, LES or DNS assessment calls for the experimental determination of time and space turbulent scales as well as time resolved measurements and determination of single or multi-point statistical properties of the velocity field. Thus, the team methodology incorporates from the very beginning an experimental component that is in strong interaction with the physical modeling and the simulation activities. The capability of producing in-situ experimental data is another originality of our project. By carefully controlling the flow configuration and the type of data we are measuring, we are in situation of assessing in depth the quality of our simulations results over the complete spectrum of possible approaches ranging from DNS, LES, RANS and Hybrid RANS-LES

---

[0]http://www.ercoftac.org

Figure 2. View of the team's test facility MAVERIC.

models that the team is developing. The flow configuration we have chosen is that of a jet in cross-flow since it features large scale coherent structures, flow separation, turbulence and wall-flow interaction. Thus, this test facility called MAVERIC (Fig. 2) is extensively used in the framework of the present project to investigate a 1-hole cylindrical inclined jet interacting with a turbulent crossflow. PIV [0] and LDV [0] are the workhorses as far as metrology is concerned.

---

[0]Particle image velocimetry
[0]Laser Doppler velocimetry

CARDAMOM Team

3. Research Program

3.1. Variational discrete asymptotic modelling

In many of the applications we consider, intermediate fidelity models are or can be derived using an asymptotic expansion for the relevant scale resolving PDEs, and eventually considering some averaged for of the resulting continuous equations. The resulting systems of PDEs are often very complex and their characterization, e.g. in terms of stability, unclear, or poor, or too complex to allow to obtain discrete analogy of the continuous properties. This makes the numerical approximation of these PDE systems a real challenge. Moreover, most of these models are often based on asymptotic expansions involving small geometrical scales. This is true for many applications considered here involving flows in/of thin layers (free surface waves, liquid films on wings generating ice layers, oxide flows in material cracks, etc). This asymptotic expansion is nothing else than a discretization (some sort of Taylor expansion) in terms of the small parameter. The actual discretization of the PDE system is another expansion in space involving as a small parameter the mesh size. What is the interaction between these two expansions ? Could we use the spatial discretization (truncation error) as means of filtering undesired small scales instead of having to explicitly derive PDEs for the large scales ? We will investigate in depth the relations between asymptotics and discretization by :

- comparing the asymptotic limits of discretized forms of the relevant scale resolving equations with the discretization of the analogous continuous asymptotic PDEs. Can we discretize a well understood system of PDEs instead of a less understood and more complex one ? ;

- study the asymptotic behaviour of error terms generated by coarse one-dimensional discretization in the direction of the "small scale". What is the influence of the number of cells along the vertical direction, and of their clustering ? ;

- derive equivalent continuous equations (modified equations) for anisotropic discretizations in which the direction is direction of the "small scale" is approximated with a small number of cells. What is the relation with known asymptotic PDE systems ?

Our objective is to gain sufficient control of the interaction between discretization and asymptotics to be able to replace the coupling of several complex PDE systems by adaptive strongly anisotrotropic finite element approximations of relevant and well understood PDEs. Here the anisotropy is intended in the sense of having a specific direction in which a much poorer (and possibly variable with the flow conditions) polynomial approximation (expansion) is used. The final goal is, profiting from the availability of faster and cheaper computational platforms, to be able to automatically control numerical *and* physical accuracy of the model with the same techniques. This activity will be used to improve our modelling in coastal engineering as well as for de-anti icing systems, wave energy converters, composite materials (cf. next sections).

In parallel to these developments, we will make an effort in to gain a better understanding of continuous asymptotic PDE models. We will in particular work on improving, and possibly, simplifying their numerical approximation. An effort will be done in trying to embed in these more complex nonlinear PDE models discrete analogs of operator identities necessary for stability (see e.g. the recent work of [115], [118] and references therein).

3.2. High order discretizations on moving adaptive meshes

We will work on both the improvement of high order mesh generation and adaptation techniques, and the construction of more efficient, adaptive high order discretisation methods.

Concerning curved mesh generation, we will focus on two points. First propose a robust and automatic method to generate curved simplicial meshes for realistic geometries. The untangling algorithm we plan to develop is a hybrid technique that gathers a local mesh optimization applied on the surface of the domain and a linear elasticity analogy applied in its volume. Second we plan to extend the method proposed in [67] to hybrid meshes (prism/tetra).

For time dependent adaptation we will try to exploit as much as possible the use of $r-$adaptation techniques based on the solution of some PDE system for the mesh. We will work on enhancing the initial results of [71], [73] by developing more robust nonlinear variants allowing to embed rapidly moving objects. For this the use of non-linear mesh PDEs (cf e.g. [127], [134], [83]), combined with Bezier type approximations for the mesh displacements to accommodate high order curved meshes [67], and with improved algorithms to discretize accurately and fast the elliptic equations involved. For this we will explore different type of relaxation methods, including those proposed in [116], [121], [120] allowing to re-use high order discretizations techniques already used for the flow variables. All these modelling approaches for the mesh movement are based on some minimization argument, and do not allow easily to take into account explicitly properties such as e.g. the positivity of nodal volumes. An effort will be made to try to embed these properties, as well as to improve the control on the local mesh sizes obtained. Developments made in numerical methods for Lagrangian hydrodynamics and compressible materials may be a possible path for these objectives (see e.g. [93], [140], [139] and references therein). We will stretch the use of these techniques as much as we can, and couple them with remeshing algorithms based on local modifications plus conservative, high order, and monotone ALE (or other) remaps (cf. [68], [104], [141], [91] and references therein).

The development of high order schemes for the discretization of the PDE will be a major part of our activity. We will work from the start in an Arbitrary Lagrangian Eulerian setting, so that mesh movement will be easily accommodated, and investigate the following main points:

- the ALE formulation is well adapted both to handle moving meshes, and to provide conservative, high order, and monotone remaps between different meshes. We want to address the issue of cost-accuracy of adaptive mesh computations by exploring different degrees of coupling between the flow and the mesh PDEs. Initial experience has indicated that a clever coupling may lead to a considerable CPU time reduction for a given resolution [73], [71]. This balance is certainly dependent on the nature of the PDEs, on the accuracy level sought, on the cost of the scheme, and on the time stepping technique. All these elements will be taken into account to try to provide the most efficient formulation ;

- the conservation of volume, and the subsequent preservation of constant mass-momentum-energy states on deforming domains is one of the most primordial elements of Arbitrary Lagrangian-Eulerian formulations. For complex PDEs as the ones considered here, of especially for some applications, there may be a competition between the conservation of e.g. mass, an the conservation of other constant states, as important as mass. This is typically the case for free surface flows, in which mass preservation is in competitions with the preservation of constant free surface levels [72]. Similar problems may arise in other applications. Possible solutions to this competition may come from super-approximation (use of higher order polynomials) of some of the data allowing to reduce (e.g. bathymetry) the error in the preservation of one of the competing quantities. This is similar to what is done in super-parametric approximations of the boundaries of an object immersed in the flow, except that in our case the data may enter the PDE explicitly and not only through the boundary conditions. Several efficient solutions for this issue will be investigated to obtain fully conservative moving mesh approaches:

- an issue related to the previous one is the accurate treatment of wall boundaries. It is known that even for standard lower order (second) methods, a higher order, curved, approximation of the boundaries may be beneficial. This, however, may become difficult when considering moving objects, as in the case e.g. of the study of the impact of ice debris in the flow. To alleviate this issue, we plan to follow on with our initial work on the combined use of immersed boundaries techniques with high order, anisotropic (curved) mesh adaptation. In particular, we will develop combined approaches involving high order hybrid meshes on fixed boundaries with the use of penalization techniques and immersed

boundaries for moving objects. We plan to study the accuracy obtainable across discontinuous functions with $r-$adaptive techniques, and otherwise use whenever necessary anisotropic meshes to be able to provide a simplified high order description of the wall boundary (cf. [114]). The use of penalization will also provide a natural setting to compute immediate approximations of the forces on the immersed body [119], [122]. An effort will be also made on improving the accuracy of these techniques using e.g. higher order approaches, either based on generalizations of classical splitting methods [105], or on some iterative Defect Correction method (see e.g. [85]) ;

- the proper treatment of different physics may be addressed by using mixed/hybrid schemes in which different variables/equations are approximated using a different polynomial expansion. A typical example is our work on the discretization of highly non-linear wave models [100] in which we have shown how to use a standard continuous Galerkin method for the elliptic equation/variable representative of the dispersive effects, while the underlying hyperbolic system is evolved using a (discontinuous) third order finite volume method. This technique will be generalized to other classes of discontinuous methods, and similar ideas will be used in other context to provide a flexible approximation. Such mathods have clear advantages in multiphase flows but not only. A typical example where such mixed methods are beneficial are flows involving different species and tracer equations, which are typically better treated with a discontinuous approximation. Another example is the use of this mixed approximation to describe the topography with a high order continuous polynomial even in discontinuous method. This allows to greatly simplify the numerical treatment of the bathymetric source terms ;

- the enhancement of stabilized methods based on some continuous finite element approximation will remain a main topic. We will further pursue the study on the construction of simplified stabilization operators which do not involve any contributions to the mass matrix. We will in particular generalize our initial results [129], [70], [130] to higher order spatial approximations using cubature points, or Bezier polynomials, or also hierarchical approximations. This will also be combined with time dependent variants of the reconstruction techniques initially proposed by D. Caraeni [84], allowing to have a more flexible approach similar to the so-called P^nP^m method [96], [133]. How to localize these enhancements, and to efficiently perform local reconstructions/enrichment, as well as $p-$adaptation, and handling hanging nodes will also be a main line of work. A clever combination of hierarchical enrichment of the polynomials, with a constrained approximation will be investigated. All these developments will be combined with the shock capturing/positivity preserving construction we developed in the past. Other discontinuity resolving techniques will be investigated as well, such as face limiting techniques as those partially studied in [102] ;

- time stepping is an important issue, especially in presence of local mesh adaptation. The techniques we use will force us to investigate local and multilevel techniques. We will study the possibility constructing semi-implicit methods combining extrapolation techniques with space-time variational approaches. Other techniques will be considered, as multi-stage type methods obtained using Defect-Correction, Multi-step Runge-Kutta methods [81], as well as spatial partitioning techniques [111]. A major challenge will be to be able to guarantee sufficient locality to the time integration method to allow to efficiently treat highly refined meshes, especially for viscous reactive flows. Another challenge will be to embed these methods in the stabilized methods we will develop.

3.3. Coupled approximation/adaptation in parameter and physical space

As already remarked, classical methods for uncertainty quantification are affected by the so-called Curse-of-Dimensionality. Adaptive approaches proposed so far, are limited in terms of efficiency, or of accuracy. Our aim here is to develop methods and algorithms permitting a very high-fidelity simulation in the physical and in the stochastic space at the same time. We will focus on both non-intrusive and intrusive approaches.

Simple non-intrusive techniques to reduce the overall cost of simulations under uncertainty will be based on adaptive quadrature in stochastic space with mesh adaptation in physical space using error monitors related to the variance of to the sensitivities obtained e.g. by an ANOVA decomposition. For steady state problems,

remeshing using metric techniques is enough. For time dependent problems both mesh deformation and re-meshing techniques will be used. This approach may be easily used in multiple space dimensions to minimize the overall cost of model evaluations by using high order moments of the properly chosen output functional for the adaptation (as in optimization). Also, for high order curved meshes, the use of high order moments and sensitivities issued from the UQ method or optimization provides a viable solution to the lack of error estimators for high order schemes.

Despite the coupling between stochastic and physical space, this approach can be made massively parallel by means of extrapolation/interpolation techniques for the high order moments, in time and on a reference mesh, guaranteeing the complete independence of deterministic simulations. This approach has the additional advantage of being feasible for several different application codes due to its non-intrusive character.

To improve on the accuracy of the above methods, intrusive approaches will also be studied. To propagate uncertainties in stochastic differential equations, we will use Harten's multiresolution framework, following [66]. This framework allows a reduction of the dimensionality of the discrete space of function representation, defined in a proper stochastic space. This reduction allows a reduction of the number of explicit evaluations required to represent the function, and thus a gain in efficiency. Moreover, multiresolution analysis offers a natural tool to investigate the local regularity of a function and can be employed to build an efficient refinement strategy, and also provides a procedure to refine/coarsen the stochastic space for unsteady problems. This strategy should allow to capture and follow all types of flow structures, and, as proposed in [66], allows to formulate a non-linear scheme in terms of compression capabilities, which should allow to handle non-smooth problems. The potential of the method also relies on its moderate intrusive behaviour, compared to e.g. spectral Galerkin projection, where a theoretical manipulation of the original system is needed.

Several activities are planned to generalize our initial work, and to apply it to complex flows in multiple (space) dimensions and with many uncertain parameters.

The first is the improvement of the efficiency. This may be achieved by means of anisotropic mesh refinement, and by experimenting with a strong parallelization of the method. Concerning the first point, we will investigate several anisotropic refinement criteria existing in literature (also in the UQ framework), starting with those already used in the team to adapt the physical grid. Concerning the implementation, the scheme formulated in [66] is conceived to be highly parallel due to the external cycle on the number of dimensions in the space of uncertain parameters. In principle, a number of parallel threads equal to the number of spatial cells could be employed. The scheme should be developed and tested for treating unsteady and discontinuous probability density function, and correlated random variables. Both the compression capabilities and the accuracy of the scheme (in the stochastic space) should be enhanced with a high-order multidimensional conservative and non-oscillatory polynomial reconstruction (ENO/WENO).

Another main objective is related to the use of multiresolution in both physical and stochastic space. This requires a careful handling of data and an updated definition of the wavelet. Until now, only a weak coupling has been performed, since the number of points in the stochastic space varies according to the physical space, but the number of points in the physical space remains unchanged. Several works exist on the multiresolution approach for image compression, but this could be the first time i in which this kind of approach would be applied at the same time in the two spaces with an unsteady procedure for refinement (and coarsening). The experimental code developed using these technologies will have to fully exploit the processing capabilities of modern massively parallel architectures, since there is a unique mesh to handle in the coupled physical/stochastic space.

3.4. Robust multi-fidelity modelling for optimization and certification

Due to the computational cost, it is of prominent importance to consider multi-fidelity approaches gathering high-fidelity and low-fidelity computations. Note that low-fidelity solutions can be given by both the use of surrogate models in the stochastic space, and/or eventually some simplified choices of physical models of some element of the system. Procedures which deal with optimization considering uncertainties for complex problems may require the evaluation of costly objective and constraint functions hundreds or even thousands of times. The associated costs are usually prohibitive. For these reason, the robustness of the optimal

solution should be assessed, thus requiring the formulation of efficient methods for coupling optimization and stochastic spaces. Different approaches will be explored. Work will be developed along three axes:

1. a robust strategy using the statistics evaluation will be applied separately, *i.e.* using only low or high-fidelity evaluations. Some classical optimization algorithms will be used in this case. Influence of high-order statistics and model reduction in the robust design optimization will be explored, also by further developing some low-cost methods for robust design optimization working on the so-called Simplex^2 method [89] ;

2. a multi-fidelity strategy by using in an efficient way low fidelity and high-fidelity estimators both in physical and stochastic space will be conceived, by using a Bayesian framework for taking into account model discrepancy and a PC expansion model for building a surrogate model ;

3. develop advanced methods for robust optimization. In particular, the Simplex^2 method will be modified for introducing a hierarchical refinement with the aim to reduce the number of stochastic samples according to a given design in an adaptive way.

This work is related to the activities foreseen in the EU contract MIDWEST, in the ANR LabCom project VIPER (currently under evaluation), in a joint project with DGA and VKI, in two projects under way with AIRBUS and SAFRAN-HERAKLES.

CARMEN Team

3. Research Program

3.1. Complex models for the propagation of cardiac action potentials

The contraction of the heart is coordinated by a complex electrical activation process which relies on about a million ion channels, pumps, and exchangers of various kinds in the membrane of each cardiac cell. Their interaction results in an activation wave that rapidly propagates through the tissue. The spatio-temporal pattern of this propagation is related both to the function of the cellular membrane and to the structural organisation of the cells into tissues. Cardiac arrythmias originate from malfunctions in this process. The field of cardiac electrophysiology studies the multiscale organisation of the cardiac activation process from the subcellular scale up to the scale of the body. It relates the molecular processes in the cell membranes to the propagation process and to measurable signals in the heart and to the electrocardiogram, an electrical signal on the torso.

Several improvements of current models of the propagation of the action potential will be developped, based on previous work and on the data available at the LIRYC:

- Enrichment of the current monodomain and bidomain models by accounting for structural hetero-geneities of the tissue at an intermediate scale. Here we focus on multiscale analysis techniques applied to the various high-resolution structural data available at the LIRYC.

- Coupling of the tissues from the different cardiac compartments and conduction systems. Here, we want to develop model that couples 1D, 2D and 3D phenomena described by reaction-diffusion PDEs.

These models are essential to improve our in-depth understanding of cardiac electrical dysfunction. To this aim, we use high-performance computing techniques in order to numerically explore the complexity of these models and check that they are reliable experimental tools.

3.2. Simplified models and inverse problems

The medical and clinical exploration of the electrical signals is based on accurate reconstruction of the typical patterns of propagation of the action potential. The correct detection of these complex patterns by non-invasive electrical imaging techniques has to be developed. Both problems involve solving inverse problems that cannot be addressed with the more compex models. We want both to develop simple and fast models of the propagation of cardiac action potentials and improve the solutions to the inverse problems found in cardiac electrical imaging techniques.

The cardiac inverse problem consists in finding the cardiac activation maps or, more generally, the whole cardiac electrical activity, from high density body surface electrocardiograms. It is a new and a powerful diagnosis technique, which success would be considered as a breakthrough in the cardiac diagnosis. Although widely studied during the last years, it remains a challenge for the scientific community. In many cases the quality of reconstructed electrical potential is not sufficiently accurate. The methods used consist in solving the Laplace equation on the volume delimited by the body surface and the epicardial surface. We want to

- Study in depth the dependance of this inverse problem inhomogeneities in the torso, conductivity values, the geometry, electrode placements...

- Improve the solution to the inverse problem be using new regularization strategies and the theory of optimal control, both in the quasistatic and in the dynamic contexts.

Of course we will use our models as a basis to regularize these inverse problems. We will conside the follwong strategies:

- using complete propagation models in the inverse problem, like the bidomain equations; for instance in order to localize electrical sources;
- construct families of reduced-order models, using e.g. statistical learning techniques, which would accurately represent some families of well-identified pathologies; and
- construct simple models of the propagation of the activation front, based on eikonal or level-sets equations, but which would incorporate the representation of complex activation patterns.

Additionaly, we will need to develop numerical techniques dedicated to our simplified eikonal/level-set equations.

3.3. Numerical techniques

We want the numerical simulations of the previous direct or inverse models to be efficient and reliable with respect to the needs of the medical community. They should qualify and guarantee the accuracy and robustness of the numerical techniques and the efficiency of the resolution algorithms.

Based on previous work on solving the monodomain and bidomain equations [4], [5] and [6] and [1], we will focus on

- High-order numerical techniques with respect to the variables with physiological meaning, like velocity, AP duration and restitution properties.
- Efficient, dedicated preconditionning techniques coupled with parallel computing.

Existing simulation tools used in our team rely, among others, on mixtures of explicit and implicit integration methods for ODEs, hybrid MPI-OpenMP parallellization, algebraic multigrid preconditioning, and a BiCGStab algorithm with adaptations to retain numerical accuracy while handling large underdetermined systems.

3.4. Cardiac Electrophysiology at the Microscopic Scale

Numerical models of cardiac physiology are based on the approximation of a perfect muscle using homogenisation method. However, due to the age and due to some cardiomyopathies, the cellular structure of the tissue changes. These modifications give rise to life-threatening arrhythmias. For our research on this subject and with cardiologists of the IHU LIRYC Bordeaux, we aim to design and implement models that describe the strong heterogeneity of the tissue at the cellular level and numerically explore the mechanisms of these diseases.

The problem is that literature on this type of model is still very poor and existing models are bidimensionels or limited to idealised geometries. We propose an approach in opposition with the usual homogenisation way. We want to describe the muscle as a system of three-dimensional cells, whose dynamics is given by the modeling of ion fluxes across cell membranes in equilibrium with the electrostatic potentials in the intracellular and extracellular environments.

The goals are to design, analyse, and explore numerically a model of cardiac electrophysiology at a level of discretisation of about $1\mu m$ (that means 10 to 100 times smaller than the size of cardiomyocytes), develop model and its numerical discretisation, define realistic geometries or actual cells.

Issues are scale simulations for thousands of cores to take into account thousands or tens of thousands cells. For this, a hybrid parallelism approach OpenMP and MPI will be considered.

CQFD Project-Team

3. Research Program

3.1. Introduction

The scientific objectives of the team are to provide mathematical tools for modeling and optimization of complex systems. These systems require mathematical representations which are in essence dynamic, multi-model and stochastic. This increasing complexity poses genuine scientific challenges in the domain of modeling and optimization. More precisely, our research activities are focused on stochastic optimization and (parametric, semi-parametric, multidimensional) statistics which are complementary and interlinked topics. It is essential to develop simultaneously statistical methods for the estimation and control methods for the optimization of the models.

3.2. Main research topics

- Stochastic modeling: Markov chain, Piecewise Deterministic Markov Processes (PDMP), Markov Decision Processes (MDP).
  The mathematical representation of complex systems is a preliminary step to our final goal corresponding to the optimization of its performance. For example, in order to optimize the predictive maintenance of a system, it is necessary to choose the adequate model for its representation. The step of modeling is crucial before any estimation or computation of quantities related to its optimization. For this we have to represent all the different regimes of the system and the behavior of the physical variables under each of these regimes. Moreover, we must also select the dynamic variables which have a potential effect on the physical variable and the quantities of interest. The team CQFD works on the theory of Piecewise Deterministic Markov Processes (PDMP's) and on Markov Decision Processes (MDP's). These two classes of systems form general families of controlled stochastic processes suitable for the modeling of sequential decision-making problems in the continuous-time (PDMPs) and discrete-time (MDP's) context. They appear in many fields such as engineering, computer science, economics, operations research and constitute powerful class of processes for the modeling of complex system.

- Estimation methods: estimation for PDMP; estimation in non- and semi parametric regression modeling.
  To the best of our knowledge, there does not exist any general theory for the problems of estimating parameters of PDMPs although there already exist a large number of tools for sub-classes of PDMPs such as point processes and marked point processes. However, to fill the gap between these specific models and the general class of PDMPs, new theoretical and mathematical developments will be on the agenda of the whole team. In the framework of non-parametric regression or quantile regression, we focus on kernel estimators or kernel local linear estimators for complete data or censored data. New strategies for estimating semi-parametric models via recursive estimation procedures have also received an increasing interest recently. The advantage of the recursive estimation approach is to take into account the successive arrivals of the information and to refine, step after step, the implemented estimation algorithms. These recursive methods do require restarting calculation of parameter estimation from scratch when new data are added to the base. The idea is to use only the previous estimations and the new data to refresh the estimation. The gain in time could be very interesting and there are many applications of such approaches.

- Dimension reduction: dimension-reduction via SIR and related methods, dimension-reduction via multidimensional and classification methods.

  Most of the dimension reduction approaches seek for lower dimensional subspaces minimizing the loss of some statistical information. This can be achieved in modeling framework or in exploratory data analysis context.

  In modeling framework we focus our attention on semi-parametric models in order to conjugate the advantages of parametric and nonparametric modeling. On the one hand, the parametric part of the model allows a suitable interpretation for the user. On the other hand, the functional part of the model offers a lot of flexibility. In this project, we are especially interested in the semi-parametric regression model $Y = f(X'\theta) + \varepsilon$, the unknown parameter θ belongs to $\mathbb{R}^p$ for a single index model, or is such that $\theta = [\theta_1, \cdots, \theta_d]$ (where each θ_k belongs to $\mathbb{R}^p$ and $d \leq p$ for a multiple indices model), the noise ε is a random error with unknown distribution, and the link function f is an unknown real valued function. Another way to see this model is the following: the variables X and Y are independent given $X'\theta$. In our semi-parametric framework, the main objectives are to estimate the parametric part θ as well as the nonparametric part which can be the link function f, the conditional distribution function of Y given X or the conditional quantile q_α. In order to estimate the dimension reduction parameter θ we focus on the Sliced Inverse Regression (SIR) method which has been introduced by Li [44] and Duan and Li [42]

  Methods of dimension reduction are also important tools in the field of data analysis, data mining and machine learning.They provide a way to understand and visualize the structure of complex data sets.Traditional methods among others are principal component analysis for quantitative variables or multiple component analysis for qualitative variables. New techniques have also been proposed to address these challenging tasks involving many irrelevant and redundant variables and often comparably few observation units. In this context, we focus on the problem of synthetic variables construction, whose goals include increasing the predictor performance and building more compact variables subsets. Clustering of variables is used for feature construction. The idea is to replace a group of "similar" variables by a cluster centroid, which becomes a feature. The most popular algorithms include K-means and hierarchical clustering. For a review, see, e.g., the textbook of Duda [43]

- Stochastic optimal control: optimal stopping, impulse control, continuous control, linear programming.

  The first objective is to focus on the development of computational methods.

  – In the continuous-time context, stochastic control theory has from the numerical point of view, been mainly concerned with Stochastic Differential Equations (SDEs in short). From the practical and theoretical point of view, the numerical developments for this class of processes are extensive and largely complete. It capitalizes on the connection between SDEs and second order partial differential equations (PDEs in short) and the fact that the properties of the latter equations are very well understood. It is, however, hard to deny that the development of computational methods for the control of PDMPs has received little attention. One of the main reasons is that the role played by the familiar PDEs in the diffusion models is here played by certain systems of integro-differential equations for which there is not (and cannot be) a unified theory such as for PDEs as emphasized by M.H.A. Davis in his book. To the best knowledge of the team, there is only one attempt to tackle this difficult problem by O.L.V. Costa and M.H.A. Davis. The originality of our project consists in studying this unexplored area. It is very important to stress the fact that these numerical developments will give rise to a lot of theoretical issues such as type of approximations, convergence results, rates of convergence,....

  – Theory for MDP's has reached a rather high degree of maturity, although the classical tools such as value iteration, policy iteration and linear programming, and their various extensions, are not applicable in practice. We believe that the theoretical progress of MDP's must be in parallel with the corresponding numerical developments. Therefore, solving

MDP's numerically is an awkward and important problem both from the theoretical and practical point of view. In order to meet this challenge, the fields of neural networks, neuro-dynamic programming and approximate dynamic programming became recently an active area of research. Such methods found their roots in heuristic approaches, but theoretical results for convergence results are mainly obtained in the context of finite MDP's. Hence, an ambitious challenge is to investigate such numerical problems but for models with general state and action spaces. Our motivation is to develop theoretically consistent computational approaches for approximating optimal value functions and finding optimal policies.

–   An effort has been devoted to the development of efficient computational methods in the setting of communication networks. These are complex dynamical systems composed of several interacting nodes that exhibit important congestion phenomena as their level of interaction grows. The dynamics of such systems are affected by the randomness of their underlying events (e.g., arrivals of http requests to a web-server) and are described stochastically in terms of queueing network models. These are mathematical tools that allow one to predict the performance achievable by the system, to optimize the network configuration, to perform capacity-planning studies, etc. These objectives are usually difficult to achieve without a mathematical model because Internet systems are huge in size. However, because of the exponential growth of their state spaces, an exact analysis of queueing network models is generally difficult to obtain. Given this complexity, we have developed analyses in some limiting regime of practical interest (e.g., systems size grows to infinity). This approach is helpful to obtain a simpler mathematical description of the system under investigation, which leads to the direct definition of efficient, though approximate, computational methods and also allows to investigate other aspects such as Nash equilibria.

The second objective of the team is to study some theoretical aspects related to MDPs such as convex analytical methods and singular perturbation. Analysis of various problems arising in MDPs leads to a large variety of interesting mathematical problems.

<p style="text-align:center; color:red">**FLOWERS Project-Team**</p>

3. Research Program

3.1. Research Program

Research in artificial intelligence, machine learning and pattern recognition has produced a tremendous amount of results and concepts in the last decades. A blooming number of learning paradigms - supervised, unsupervised, reinforcement, active, associative, symbolic, connectionist, situated, hybrid, distributed learning... - nourished the elaboration of highly sophisticated algorithms for tasks such as visual object recognition, speech recognition, robot walking, grasping or navigation, the prediction of stock prices, the evaluation of risk for insurances, adaptive data routing on the internet, etc... Yet, we are still very far from being able to build machines capable of adapting to the physical and social environment with the flexibility, robustness, and versatility of a one-year-old human child.

Indeed, one striking characteristic of human children is the nearly open-ended diversity of the skills they learn. They not only can improve existing skills, but also continuously learn new ones. If evolution certainly provided them with specific pre-wiring for certain activities such as feeding or visual object tracking, evidence shows that there are also numerous skills that they learn smoothly but could not be "anticipated" by biological evolution, for example learning to drive a tricycle, using an electronic piano toy or using a video game joystick. On the contrary, existing learning machines, and robots in particular, are typically only able to learn a single pre-specified task or a single kind of skill. Once this task is learnt, for example walking with two legs, learning is over. If one wants the robot to learn a second task, for example grasping objects in its visual field, then an engineer needs to re-program manually its learning structures: traditional approaches to task-specific machine/robot learning typically include engineer choices of the relevant sensorimotor channels, specific design of the reward function, choices about when learning begins and ends, and what learning algorithms and associated parameters shall be optimized.

As can be seen, this requires a lot of important choices from the engineer, and one could hardly use the term "autonomous" learning. On the contrary, human children do not learn following anything looking like that process, at least during their very first years. Babies develop and explore the world by themselves, focusing their interest on various activities driven both by internal motives and social guidance from adults who only have a folk understanding of their brains. Adults provide learning opportunities and scaffolding, but eventually young babies always decide for themselves what activity to practice or not. Specific tasks are rarely imposed to them. Yet, they steadily discover and learn how to use their body as well as its relationships with the physical and social environment. Also, the spectrum of skills that they learn continuously expands in an organized manner: they undergo a developmental trajectory in which simple skills are learnt first, and skills of progressively increasing complexity are subsequently learnt.

A link can be made to educational systems where research in several domains have tried to study how to provide a good learning experience to learners. This includes the experiences that allow better learning, and in which sequence they must be experienced. This problem is complementary to that of the learner that tries to learn efficiently, and the teacher here has to use as efficienlty the limited time and motivational resources of the learner. Several results from psychology [76] and neuroscience [10] have argued that the human brain feels intrinsic pleasure in practicing activities of optimal difficulty or challenge. A teacher must exploit such activities to create positive psychological states of flow [83].

A grand challenge is thus to be able to build robotic machines that possess this capability to discover, adapt and develop continuously new know-how and new knowledge in unknown and changing environments, like human children. In 1950, Turing wrote that the child's brain would show us the way to intelligence: "Instead of trying to produce a program to simulate the adult mind, why not rather try to produce one which simulates the child's" [139]. Maybe, in opposition to work in the field of Artificial Intelligence who has focused on mechanisms trying to match the capabilities of "intelligent" human adults such as chess playing or natural language

dialogue [95], it is time to take the advice of Turing seriously. This is what a new field, called developmental (or epigenetic) robotics, is trying to achieve [106] [142]. The approach of developmental robotics consists in importing and implementing concepts and mechanisms from developmental psychology [111], cognitive linguistics [82], and developmental cognitive neuroscience [99] where there has been a considerable amount of research and theories to understand and explain how children learn and develop. A number of general principles are underlying this research agenda: embodiment [78] [123], grounding [93], situatedness [69], self-organization [137] [126], enaction [141], and incremental learning [80].

Among the many issues and challenges of developmental robotics, two of them are of paramount importance: exploration mechanisms and mechanisms for abstracting and making sense of initially unknown sensorimotor channels. Indeed, the typical space of sensorimotor skills that can be encountered and learnt by a developmental robot, as those encountered by human infants, is immensely vast and inhomogeneous. With a sufficiently rich environment and multimodal set of sensors and effectors, the space of possible sensorimotor activities is simply too large to be explored exhaustively in any robot's life time: it is impossible to learn all possible skills and represent all conceivable sensory percepts. Moreover, some skills are very basic to learn, some other very complicated, and many of them require the mastery of others in order to be learnt. For example, learning to manipulate a piano toy requires first to know how to move one's hand to reach the piano and how to touch specific parts of the toy with the fingers. And knowing how to move the hand might require to know how to track it visually.

Exploring such a space of skills randomly is bound to fail or result at best on very inefficient learning [16]. Thus, exploration needs to be organized and guided. The approach of epigenetic robotics is to take inspiration from the mechanisms that allow human infants to be progressively guided, i.e. to develop. There are two broad classes of guiding mechanisms which control exploration:

1. **internal guiding mechanisms,** and in particular intrinsic motivation, responsible of spontaneous exploration and curiosity in humans, which is one of the central mechanisms investigated in FLOWERS, and technically amounts to achieve online active self-regulation of the growth of complexity in learning situations;

2. **social learning and guidance,** a learning mechanisms that exploits the knowledge of other agents in the environment and/or that is guided by those same agents. These mechanisms exist in many different forms like emotional reinforcement, stimulus enhancement, social motivation, guidance, feedback or imitation, some of which being also investigated in FLOWERS;

3.1.1. Internal guiding mechanisms

In infant development, one observes a progressive increase of the complexity of activities with an associated progressive increase of capabilities [111], children do not learn everything at one time: for example, they first learn to roll over, then to crawl and sit, and only when these skills are operational, they begin to learn how to stand. The perceptual system also gradually develops, increasing children perceptual capabilities other time while they engage in activities like throwing or manipulating objects. This make it possible to learn to identify objects in more and more complex situations and to learn more and more of their physical characteristics.

Development is therefore progressive and incremental, and this might be a crucial feature explaining the efficiency with which children explore and learn so fast. Taking inspiration from these observations, some roboticists and researchers in machine learning have argued that learning a given task could be made much easier for a robot if it followed a developmental sequence and "started simple" [72] [87]. However, in these experiments, the developmental sequence was crafted by hand: roboticists manually build simpler versions of a complex task and put the robot successively in versions of the task of increasing complexity. And when they wanted the robot to learn a new task, they had to design a novel reward function.

Thus, there is a need for mechanisms that allow the autonomous control and generation of the developmental trajectory. Psychologists have proposed that intrinsic motivations play a crucial role. Intrinsic motivations are mechanisms that push humans to explore activities or situations that have intermediate/optimal levels of novelty, cognitive dissonance, or challenge [76] [83] [86]. The role and structure of intrinsic motivation in humans have been made more precise thanks to recent discoveries in neuroscience showing the implication

of dopaminergic circuits and in exploration behaviors and curiosity [85] [96] [134]. Based on this, a number of researchers have began in the past few years to build computational implementation of intrinsic motivation [16] [121] [132] [75] [97] [108] [133]. While initial models were developed for simple simulated worlds, a current challenge is to manage to build intrinsic motivation systems that can efficiently drive exploratory behaviour in high-dimensional unprepared real world robotic sensorimotor spaces [121][16] [122] [131]. Specific and complex problems are posed by real sensorimotor spaces, in particular due to the fact that they are both high-dimensional as well as (usually) deeply inhomogeneous. As an example for the latter issue, some regions of real sensorimotor spaces are often unlearnable due to inherent stochasticity or difficulty, in which case heuristics based on the incentive to explore zones of maximal unpredictability or uncertainty, which are often used in the field of active learning [81] [94] typically lead to catastrophic results. The issue of high dimensionality does not only concern motor spaces, but also sensory spaces, leading to the problem of correctly identifying, among typically thousands of quantities, those latent variables that have links to behavioral choices. In FLOWERS, we aim at developing intrinsically motivated exploration mechanisms that scale in those spaces, by studying suitable abstraction processes in conjunction with exploration strategies.

3.1.2. *Socially Guided and Interactive Learning*

Social guidance is as important as intrinsic motivation in the cognitive development of human babies [111]. There is a vast literature on learning by demonstration in robots where the actions of humans in the environment are recognized and transferred to robots [71]. Most such approaches are completely passive: the human executes actions and the robot learns from the acquired data. Recently, the notion of interactive learning has been introduced in [138], [77], motivated by the various mechanisms that allow humans to socially guide a robot [128]. In an interactive context the steps of self-exploration and social guidances are not separated and a robot learns by self exploration and by receiving extra feedback from the social context [138], [101] [109].

Social guidance is also particularly important for learning to segment and categorize the perceptual space. Indeed, parents interact a lot with infants, for example teaching them to recognize and name objects or characteristics of these objects. Their role is particularly important in directing the infant attention towards objects of interest that will make it possible to simplify at first the perceptual space by pointing out a segment of the environment that can be isolated, named and acted upon. These interactions will then be complemented by the children own experiments on the objects chosen according to intrinsic motivation in order to improve the knowledge of the object, its physical properties and the actions that could be performed with it.

In FLOWERS, we are aiming at including intrinsic motivation system in the self-exploration part thus combining efficient self-learning with social guidance [116], [117]. We also work on developing perceptual capabilities by gradually segmenting the perceptual space and identifying objects and their characteristics through interaction with the user [107] and robots experiments [98]. Another challenge is to allow for more flexible interaction protocols with the user in terms of what type of feedback is provided and how it is provided [104].

Exploration mechanisms are combined with research in the following directions:

3.1.3. *Cumulative learning, reinforcement learning and optimization of autonomous skill learning*

FLOWERS develops machine learning algorithms that can allow embodied machines to acquire cumulatively sensorimotor skills. In particular, we develop optimization and reinforcement learning systems which allow robots to discover and learn dictionaries of motor primitives, and then combine them to form higher-level sensorimotor skills.

3.1.4. *Autonomous perceptual and representation learning*

In order to harness the complexity of perceptual and motor spaces, as well as to pave the way to higher-level cognitive skills, developmental learning requires abstraction mechanisms that can infer structural information out of sets of sensorimotor channels whose semantics is unknown, discovering for example the topology of the body or the sensorimotor contingencies (proprioceptive, visual and acoustic). This process is meant to

be open- ended, progressing in continuous operation from initially simple representations towards abstract concepts and categories similar to those used by humans. Our work focuses on the study of various techniques for:

- autonomous multimodal dimensionality reduction and concept discovery;
- incremental discovery and learning of objects using vision and active exploration, as well as of auditory speech invariants;
- learning of dictionaries of motion primitives with combinatorial structures, in combination with linguistic description;
- active learning of visual descriptors useful for action (e.g. grasping);

3.1.5. *Embodiment and maturational constraints*

FLOWERS studies how adequate morphologies and materials (i.e. morphological computation), associated to relevant dynamical motor primitives, can importantly simplify the acquisition of apparently very complex skills such as full-body dynamic walking in biped. FLOWERS also studies maturational constraints, which are mechanisms that allow for the progressive and controlled release of new degrees of freedoms in the sensorimotor space of robots.

3.1.6. *Discovering and abstracting the structure of sets of uninterpreted sensors and motors*

FLOWERS studies mechanisms that allow a robot to infer structural information out of sets of sensorimotor channels whose semantics is unknown, for example the topology of the body and the sensorimotor contingencies (proprioceptive, visual and acoustic). This process is meant to be open-ended, progressing in continuous operation from initially simple representations to abstract concepts and categories similar to those used by humans.

<p style="text-align:center">**GEOSTAT Project-Team**</p>

3. Research Program

3.1. Multiscale description in terms of multiplicative cascade

GEOSTAT is studying complex signals under the point of view of *nonlinear* methods developed *statistical physics* i.e. the methodologies developed to study complex systems, with a strong emphasis on multiresolution analysis. Linear methods in signal processing refer to the standard point of view under which operators are expressed by simple convolutions with impulse responses. Linear methods in signal processing are widely used, from least-square deconvolution methods in adaptive optics to source-filter models in speech processing. Because of the absence of localization of the Fourier transform, linear methods are not successful to unlock the multiscale structures and cascading properties of variables which are of primary importance as stated by the physics of the phenomena. This is the reason why new approaches, such as DFA (Detrented Fluctuation Analysis), Time-frequency analysis, variations on curvelets [49] etc. have appeared during the last decades. Recent advances in dimensionality reduction, and notably in Compressive Sensing, go beyond the Nyquist rate in sampling theory using nonlinear reconstruction, but data reduction occur at random places, independently of geometric localization of information content, which can be very useful for acquisition purposes, but of lower impact in signal analysis. One important result obtained in GEOSTAT is the effective use of multiresolution analysis associated to optimal inference along the scales of a complex system. The multiresolution analysis is performed on dimensionless quantities given by the *singularity exponents* which encode properly the geometrical structures associated to multiscale organization. This is applied successfully in the derivation of high resolution ocean dynamics, or the high resolution mapping of gaseous exchanges between the ocean and the atmosphere; the latter is of primary importance for a quantitative evaluation of global warming. Understanding the dynamics of complex systems is recognized as a new discipline, which makes use of theoretical and methodological foundations coming from nonlinear physics, the study of dynamical systems and many aspects of computer science. One of the challenges is related to the question of *emergence* in complex systems: large-scale effects measurable macroscopically from a system made of huge numbers of interactive agents [40], [37], [54], [44]. Some quantities related to nonlinearity, such as Lyapunov exponents, Kolmogorov-Sinai entropy etc. can be computed at least in the phase space [38]. Consequently, knowledge from acquisitions of complex systems (which include *complex signals*) could be obtained from information about the phase space. A result from F. Takens [50] about strange attractors in turbulence has motivated the determination of discrete dynamical systems associated to time series [42], and consequently the theoretical determination of nonlinear characteristics associated to complex acquisitions. Emergence phenomena can also be traced inside complex signals themselves, by trying to localize information content geometrically. Fundamentally, in the nonlinear analysis of complex signals there are broadly two approaches: characterization by attractors (embedding and bifurcation) and time-frequency, multiscale/multiresolution approaches. Time-frequency analysis [39] and multiscale/multiresolution are the subjects of intense research and are profoundly reshaping the analysis of complex signals by nonlinear approaches [36], [41]. In real situations, the phase space associated to the acquisition of a complex phenomenon is unknown. It is however possible to relate, inside the signal's domain, local predictability to local reconstruction and deduce from that singularity exponents (SEs) [11] [7]. The SEs are defined at any point in the signal's domain, they relate, but are different, to other kinds of exponents used in the nonlinear analysis of complex signals. We are working on their relation with:

- properties in universality classses,
- the geometric localization of multiscale properties in complex signals,
- cascading characteristics of physical variables,
- optimal wavelets and inference in multiresolution analysis.

The alternative approach taken in GEOSTAT is microscopical, or geometrical: the multiscale structures which have their "fingerprint" in complex signals are being isolated in a single realization of the complex system, i.e. using the data of the signal itself, as opposed to the consideration of grand ensembles or a wide set of realizations. This is much harder than the ergodic approaches, but it is possible because a reconstruction formula such as the one derived in [51] is local and reconstruction in the signal's domain is related to predictability. This approach is analogous to the consideration of "microcanonical ensembles" in statistical mechanics.

Nonlinear signal processing is making use of various quantities related to predictability (for instance first Lyapunov exponent is related, from Osedelec's theorem, to the limiting behaviour of the response to perturbation in phase space [38]). In GEOSTAT our aim is to relate such classical quantities (among others) to the behaviour of SEs, which are defined by a limiting behaviour

$$\mu\left(\mathcal{B}_r(\mathbf{x})\right) \; = \; \alpha(\mathbf{x})\, r^{d+h(\mathbf{x})} \; + \; o\left(r^{d+h(\mathbf{x})}\right) \; (r \to 0) \tag{1}$$

(d: dimension of the signal's domain, μ: multiscale measure, typically whose density is the gradient's norm, $\mathcal{B}_r(\mathbf{x})$: ball of radius r centered at $\mathbf{x}$). For precise computation, SEs can be smoothly interpolated by projecting wavelets:

$$\mathcal{T}_\Psi \mu(\mathbf{x}, r) = \int_{\mathbb{R}^d} \mathrm{d}\mu(\mathbf{x}') \frac{1}{r^d} \Psi\left(\frac{\mathbf{x} - \mathbf{x}'}{r}\right) \tag{2}$$

(Ψ: mother wavelet, admissible or not), but the best numerical method in computing singularity exponents lies in the definition of a measure related to predictability [5]:

$$h(\mathbf{x}) = \frac{\log \mathcal{T}_\Psi \mu(\mathbf{x}, r_0) / \langle \mathcal{T}_\Psi \mu(\cdot, r_0) \rangle}{\log r_0} + o(\frac{1}{\log r_0}) \tag{3}$$

with: r_0 is a scale choosen to diminish the amplitude of the correction term, and $\langle \mathcal{T}_\Psi \mu(\cdot, r_0) \rangle$ is the average value of the wavelet projection (mother wavelet Ψ) over the whole signal. Singularity exponents computed with this formula generalize the elementary "gradient's norm" in a very statistically coherent way across the scales.

SEs are related to the framework of reconstructible systems, and consequently to predictability. They unlock the geometric localization of a multiscale structure in a complex signal:

$$\mathcal{F}_h = \{\mathbf{x} \in \Omega \; | \; h(\mathbf{x}) = h\}, \tag{4}$$

(Ω: signal's domain). This multiscale structure is a fundamental feature of a complex system, it is related to the cascade description of turbulence. We make use of this kind of description when analyzing turbulent signals: intermittency is observed within the inertial range and is related to the fact that, in the case of FDT, symmetry is restored only in a statistical sense, a fact that has consequences on the quality of any nonlinear signal representation by frames or dictionaries.

The example of FDT as a standard "template" for developing general methods that apply to a vast class of complex systems and signals is of fundamental interest because, in FDT, the existence of a multiscale hierarchy (i.e. the collection of sets $\mathcal{F}_h$ of equation 4) which is of multifractal nature and geometrically localized can be derived from physical considerations. This geometric hierarchy of sets is responsible for the shape of the computed singularity spectra, which in turn is related to the statistical organization of information content in a signal. It explains scale invariance, a characteristic feature of complex signals. The analogy from statistical physics comes from the fact that singularity exponents are direct generalizations of *critical exponents* which explain the macroscopic properties of a system around critical points, and the quantitative characterization of *universality classes*, which allow the definition of methods and algorithms that apply to general complex signals and systems, and not only turbulent signals: signals which belong to a same universality class share common statistical organization. In GEOSTAT, the approach to singularity exponents is done within a microcanonical setting, which can interestingly be compared with other approaches such that wavelet leaders, WTMM or DFA. During the past decades, classical approaches (here called "canonical" because they use the analogy taken from the consideration of "canonical ensembles" in statistical mechanics) permitted the development of a well-established analogy taken from thermodynamics in the analysis of complex signals: if $\mathcal{F}$ is the free energy, $\mathcal{T}$ the temperature measured in energy units, $\mathcal{U}$ the internal energy per volume unit $\mathcal{S}$ the entropy and $\widehat{\beta} = 1/\mathcal{T}$, then the scaling exponents associated to moments of intensive variables $p \to \tau_p$ corresponds to $\widehat{\beta}\mathcal{F}$, $\mathcal{U}(\widehat{\beta})$ corresponds to the singularity exponents values, and $\mathcal{S}(\mathcal{U})$ to the singularity spectrum.

The singularity exponents belong to a universality class, independently of microscopic properties in the phase space of various complex systems, and beyond the particular case of turbulent data (where the existence of a multiscale hierarchy, of multifractal nature, can be inferred directly from physical considerations). They describe common multiscale statistical organizations in different complex systems [48], and this is why GEOSTAT is working on nonlinear signal processing tools that are applied to very different types of signals.

For example we give some insight about the collaboration with LEGOS Dynbio team [0] about high-resolution ocean dynamics from microcanonical formulations in nonlinear complex signal analysis. Indeed, synoptic determination of ocean circulation using data acquired from space, with a coherent depiction of its turbulent characteristics remains a fundamental challenge in oceanography. This determination has the potential of revealing all aspects of the ocean dynamic variability on a wide range of spatio-temporal scales and will enhance our understanding of ocean-atmosphere exchanges at super resolution, as required in the present context of climate change. We show that the determination of a multiresolution analysis associated to the multiplicative cascade of a typical physical variable like the Sea Surface Temperature permits an *optimal inference* of oceanic motion field across the scales, resulting in a new method for deriving super resolution oceanic motion from lower resolution altimetry data; the resulting oceanic motion field is validated at super resolution with the use of Lagrangian buoy data available from the Global Drifter Program [0]. In FDT, singularity exponents range in a bounded interval: $]h_\infty, h_{\max}[$ with $h_\infty < 0$ being the most singular exponent. Points $\mathbf{r}$ for which $h(\mathbf{r}) < 0$ localize the strongest transitions in the turbulent fluid, where an intensive physical variable like sea surface temperature behaves like $1/\mathbf{r}^{|h(\mathbf{r})|}$. The links between the geometrically localized singularity exponents, the scaling exponents of structure functions, the multiplicative cascade and the multiscale hierarchy $\mathcal{F}_h$ is the following:

$$\begin{cases} \mathcal{F}_h = \{\mathbf{r} \ \mid \ h(\mathbf{r}) = h\} \\ D(h) = \dim \mathcal{F}_h \\ \tau_p = \inf_h \{ph + 3 - D(h)\} \\ D(h) = \inf_p \{ph + 3 - \tau_p\} \end{cases} \tag{5}$$

Let $\mathfrak{S}(\mathbf{x})$ be the bidimensionnal signal recording, for each sample point $\mathbf{x}$ representing a pixel on the surface of the ocean of given resolution, the sea surface temperature (sst). To this signal we associate a measure μ

---

[0]http://www.legos.obs-mip.fr/recherches/equipes/dynbio.
[0]http://www.aoml.noaa.gov/phod/dac/index.php.

whose density w.r.t Lebesgue measure is the signal's gradient norm, and from which the singularity exponents are computed [6]. It is fundamental to notice here that, contrary to other types of exponents computed in Oceanography, such as Finite Size Lyapunov exponents, singularity exponents are computed at instantaneous time, and do not need time series.

Having computed the singularity exponents at each point of a SST signal, a microcanonical version of the multiplicative cascade associated to the scaling properties of the sst become available. The idea of the existence of a geometrically localized multiplicative cascade goes back to [47]. The multiplicative cascade, written pointwise, introduces random variables $\eta_{l'/l}(\mathbf{x})$ for $0 < l' < l$ such that

$$\mathcal{T}_\psi \mu(\mathbf{x}, l') \; = \; \eta_{l'/l}(\mathbf{x}) \mathcal{T}_\psi \mu(\mathbf{x}, l) \tag{6}$$

in which the equality is valid pointwise and not only in distribution. Any mother wavelet ψ such that the process $\eta_{l'/l}(\mathbf{x})$ is independant of $\mathcal{T}_\psi \mu(\mathbf{x}, l')$ is called an optimal wavelet: it optimizes inference of physical variables across the scales and consequently describes the multiplicative cascade at each point $\mathbf{x}$ in the signal domain. The injection variables $\eta_{l'/l}(\mathbf{x})$ are indefinitely divisible: $\eta_k(\mathbf{x})\eta_{k'}(\mathbf{x}) \dot{=} \eta_{kk'}(\mathbf{x})$. It is possible to optimize cross-scale inference of physical variables by considering a *multiresolution analysis* associated to a discrete covering of the "space-frequency" domain. Denoting as usual $(V_j)_{j\in\mathbb{Z}}$ and $(W_j)_{j\in\mathbb{Z}}$ the discrete sequence of approximation and detail spaces associated to a given scaling function, and denoting by $\psi \in L^2(\mathbb{R}^2)$ a wavelet which generates an Hilbertian basis on each detail space W_j, it is known that the detail spaces encode borders and transition information, which is ideally described in the case of turbulent signals by the singularity exponents $\mathbf{h}(\mathbf{x})$. Consequently, a novel idea for super-resolution consists in computing a multiresolution analysis on the signal of singularity exponents $\mathbf{h}(\mathbf{x})$, and to consider that the detail information coming from spaces W_j is given the signal $\mathbf{h}(\mathbf{x})$. The associated orthogonal projection $\pi_j : L^2(\mathbb{R}^2) \to W_j$ defined by $\pi_j(\mathbf{h}) = \sum_{n\in\mathbb{Z}} \langle\, \mathbf{h} \mid \psi_{j,n} \,\rangle \psi_{j,n}$ is then used in the reconstruction formula for retrieving a physical variable at higher resolution from its low resolution counterpart. If $\mathfrak{S}(\mathbf{x})$ is such a variable, we use a reconstruction formula: $A_{j-1}\mathfrak{S} = A_j\mathfrak{S} + \pi_j\mathbf{h}$ with $A_j : L^2(\mathbb{R}^2) \to V_j$ is the orthogonal projection on the space V_j (approximation operator) and π_j is the orthogonal projection on the detail spaces W_j associated to the signal of singularity exponents $\mathbf{h}(\mathbf{x})$. Validation is performed using Lagrangian buoy data with very good results [10]. We have realized a demonstration movie showing the turbulent ocean dynamics at an SST resolution of 4 km computed from the SST microcanonical cascade and the low-resolution GEKCO product for the year 2006 over the southwestern part of the Indian Ocean. We replace the missing data in the SST MODIS product (clouds and satellite swath) by the corresponding data available from the Operational SST and Sea Ice Analysis (OSTIA) provided by the Group for High-Resolution SST Project [11], which, however, is of lower quality. Two images per day are generated for the whole year of 2006. The resulting images show the norm of the vector field in the background rendered using the line integral convolution algorithm. In the foreground, we show the resulting vector field in a linear gray-scale color map. See link to movie (size: 800 Mo).

3.2. Excitable systems and heartbeat signal analysis

In collaboration with IHU LIRYC, we are developing novel approaches to heartbeat signal analysis for understanding chronic atrial fibrillation. The noisy aspect of data recorded by electrodes, on the inner surface of human atria during episodes of atrial fibrillation, exhibit intriguing features for excitable media. Instead of phase chaos as typically expected, it shares many common traits of non-equilibrium fluctuations in disordered systems or strong turbulence. To assess those peculiar observations we investigate a *synaptic plasticity* that affects conduction properties. Electrical synapses comprise many different kinds of connexins, which may be affected by diverse factors, so we use a generic approach. Slight detuning of their linear response leads to an instability of the modulating agents, here an excess charge. Acting on slow time scales of repolarisation, it is understood as *collective modes* propagating through and retroacting on each synapse: the medium is *desynchronised*. It is not a syncytium. We propose to associate transient states with a phenomenon called

electrical remodelling, which has not received any accepted description thus far. Moreover, from the properties of the model it is possible to start exploring phase space. Transitions between different regimes could help decipher stages in the evolution of the disease from acute to chronic, one main goal of cardiovascular research.

Theoretically, a myocardium is an excitable tissue acting under normal circumstances as a functional syncytium of myocardial cells. Models of excitability for the heart are reaction-diffusion systems describing the propagation of electric pulses called action potentials similarly to models for axons. Reaction results from ionic exchange cycles between the cytoplasm of excitable cells and their extra-cellular medium, when initiated by a stimulus above some threshold. Pulses are robust topological structures.

Considering the stable fixed point as a phase resetting state, even under periodic forcing, no return map can be drawn, therefore no chaos. Chaos may arise in temporal sequences. As propagation becomes further involved in the arrhythmia, spatio-temporal chaos amounts to the breakup of spirals from direct fore front and back front collisions within the pulse train. This is the paradigm for cardiac fibrillation.

The distributions of amplitudes all collapse on a scaling function G. We map exponents on data patients provided by IHU LIRYC showing non-universal properties. Singular exponents are observed with consistent Hausdorff dimension of sets $D(h)$. Negative contribution is high, suggesting an underlying multiplicative process.

Excess charge in cells like of *Ca* may perturb the dynamics of synapses. We consider a physiologically plausible linear response of synapses to the electro-chemical potential. This response is unknown as of today. The new dynamics may interact with excitability. It has the specific form of a Rayleigh instability. Cycles become retarded or advanced. Hopf bifurcation and chaos are allowed creating EADs (Early After Depolarization). Regarding propagation, pulses are pinned and released on a chaotic background. Cycle modulations create defects via facilitation through the third dimension. Defects proliferate creating a glassy phase, which back-scatter fronts in 1D and roughens them in 2D. Further effective inhibitor diffusion splits them. Electrical remodelling is the time by which the background has pervaded the tissue.

There are features of Self Organized Criticality (SOC) in large regions of phase space. Pulses have a phase and propagate on a random medium. For instance one paradigm we investigate would be:

$$\partial_t \theta + \sin\left(\theta + \widetilde{\phi}\right) = \Omega + \partial_{xx}\theta \tag{7}$$

(θ: phase of activation front, Ω: tachycardia frequency, $\widetilde{\phi}$: phase perturbation). Randomness reactualises non-linearly, which tells that the noise is quenched and reset. For instance in 1 + 1D, spatio-temporal maps look very much like optimal directed paths along diagonals. In 1 + 2D, we are guessing that pulses do propagate in the (q)KPZ universality class, just as the remodelling front does. We will need refined numerical computations to confirm this. This is consistent with a multiplicative process. Physiologically, one interesting bonus is the interpretation of non-reentrant Tachycardia as dislocation patterns slowly evolving.

3.3. Speech analysis

Our research in speech processing focus on the development of novel nonlinear analysis methods for the characterization and classification of pathological and affective speech. For the latter, classical linear methods do not generally capture the nonlinearity, aperiodicity, turbulence and noise that can be present in pathological voices. We thus aim to design and extract new features that allow better characterization/classification of such voices, while being easy to interpret by clinicians. For the former, recent research have shown that the voice source signal information allow significant improvement of speech emotion detection systems. Our goal is to develop novel nonlinear techniques to extract relevant voice source features and to design efficient machine learning algorithms for robust emotion classification.

<div align="center">

HIEPACS Project-Team

</div>

3. Research Program

3.1. Introduction

The methodological component of HIEPACS concerns the expertise for the design as well as the efficient and scalable implementation of highly parallel numerical algorithms to perform frontier simulations. In order to address these computational challenges a hierarchical organization of the research is considered. In this bottom-up approach, we first consider in Section 3.2 generic topics concerning high performance computational science. The activities described in this section are transversal to the overall project and their outcome will support all the other research activities at various levels in order to ensure the parallel scalability of the algorithms. The aim of this activity is not to study general purpose solution but rather to address these problems in close relation with specialists of the field in order to adapt and tune advanced approaches in our algorithmic designs. The next activity, described in Section 3.3 , is related to the study of parallel linear algebra techniques that currently appear as promising approaches to tackle huge problems on extreme scale platforms. We highlight the linear problems (linear systems or eigenproblems) because they are in many large scale applications the main computational intensive numerical kernels and often the main performance bottleneck. These parallel numerical techniques, which are involved in the IPL C2S@EXA, will be the basis of both academic and industrial collaborations, some are described in Section 4.1 , but will also be closely related to some functionalities developed in the parallel fast multipole activity described in Section 3.4 . Finally, as the accuracy of the physical models increases, there is a real need to go for parallel efficient algorithm implementation for multiphysics and multiscale modeling in particular in the context of code coupling. The challenges associated with this activity will be addressed in the framework of the activity described in Section 3.5 .

Currently, we have one major application (see Section 4.1) that is in material physics. We will collaborate to all steps of the design of the parallel simulation tool. More precisely, our applied mathematics skill will contribute to the modelling, our advanced numerical schemes will help in the design and efficient software implementation for very large parallel simulations. We also participate to a few co-design actions in close collaboration with some applicative groups. The objective of this activity is to instantiate our expertise in fields where they are critical for designing scalable simulation tools. We refer to Section 4.2 for a detailed description of these activities.

3.2. High-performance computing on next generation architectures

Participants: Emmanuel Agullo, Olivier Coulaud, Luc Giraud, Mathieu Faverge, Abdou Guermouche, Matías Hastaran, Xavier Lacoste, Guillaume Latu, Stojce Nakov, Florent Pruvost, Pierre Ramet, Jean Roman, Emrullah-Fatih Yetkin, Mawussi Zounon.

The research directions proposed in HIEPACS are strongly influenced by both the applications we are studying and the architectures that we target (i.e., massively parallel heterogeneous many-core architectures, ...). Our main goal is to study the methodology needed to efficiently exploit the new generation of high-performance computers with all the constraints that it induces. To achieve this high-performance with complex applications we have to study both algorithmic problems and the impact of the architectures on the algorithm design.

From the application point of view, the project will be interested in multiresolution, multiscale and hierarchical approaches which lead to multi-level parallelism schemes. This hierarchical parallelism approach is necessary to achieve good performance and high-scalability on modern massively parallel platforms. In this context, more specific algorithmic problems are very important to obtain high performance. Indeed, the kind of applications we are interested in are often based on data redistribution for example (e.g., code coupling applications). This well-known issue becomes very challenging with the increase of both the number of computational nodes and the amount of data. Thus, we have both to study new algorithms and to adapt the

existing ones. In addition, some issues like task scheduling have to be restudied in this new context. It is important to note that the work developed in this area will be applied for example in the context of code coupling (see Section 3.5).

Considering the complexity of modern architectures like massively parallel architectures or new generation heterogeneous multicore architectures, task scheduling becomes a challenging problem which is central to obtain a high efficiency. Of course, this work requires the use/design of scheduling algorithms and models specifically to tackle our target problems. This has to be done in collaboration with our colleagues from the scheduling community like for example O. Beaumont (Inria REALOPT Project-Team). It is important to note that this topic is strongly linked to the underlying programming model. Indeed, considering multicore architectures, it has appeared, in the last five years, that the best programming model is an approach mixing multi-threading within computational nodes and message passing between them. In the last five years, a lot of work has been developed in the high-performance computing community to understand what is critic to efficiently exploit massively multicore platforms that will appear in the near future. It appeared that the key for the performance is firstly the granularity of the computations. Indeed, in such platforms the granularity of the parallelism must be small so that we can feed all the computing units with a sufficient amount of work. It is thus very crucial for us to design new high performance tools for scientific computing in this new context. This will be developed in the context of our solvers, for example, to adapt to this new parallel scheme. Secondly, the larger the number of cores inside a node, the more complex the memory hierarchy. This remark impacts the behaviour of the algorithms within the node. Indeed, on this kind of platforms, NUMA effects will be more and more problematic. Thus, it is very important to study and design data-aware algorithms which take into account the affinity between computational threads and the data they access. This is particularly important in the context of our high-performance tools. Note that this work has to be based on an intelligent cooperative underlying run-time (like the tools developed by the Inria STORM Project-Team) which allows a fine management of data distribution within a node.

Another very important issue concerns high-performance computing using "heterogeneous" resources within a computational node. Indeed, with the deployment of the GPU and the use of more specific co-processors, it is important for our algorithms to efficiently exploit these new type of architectures. To adapt our algorithms and tools to these accelerators, we need to identify what can be done on the GPU for example and what cannot. Note that recent results in the field have shown the interest of using both regular cores and GPU to perform computations. Note also that in opposition to the case of the parallelism granularity needed by regular multicore architectures, GPU requires coarser grain parallelism. Thus, making both GPU and regular cores work all together will lead to two types of tasks in terms of granularity. This represents a challenging problem especially in terms of scheduling. From this perspective, we investigate new approaches for composing parallel applications within a runtime system for heterogeneous platforms.

In that framework, the SOLHAR project aims at studying and designing algorithms and parallel programming models for implementing direct methods for the solution of sparse linear systems on emerging computers equipped with accelerators. Several attempts have been made to accomplish the porting of these methods on such architectures; the proposed approaches are mostly based on a simple offloading of some computational tasks (the coarsest grained ones) to the accelerators and rely on fine hand-tuning of the code and accurate performance modeling to achieve efficiency. SOLHAR proposes an innovative approach which relies on the efficiency and portability of runtime systems, such as the StarPU tool developed in the STORM team. Although the SOLHAR project will focus on heterogeneous computers equipped with GPUs due to their wide availability and affordable cost, the research accomplished on algorithms, methods and programming models will be readily applicable to other accelerator devices. Our final goal would be to have high performance solvers and tools which can efficiently run on all these types of complex architectures by exploiting all the resources of the platform (even if they are heterogeneous).

In order to achieve an advanced knowledge concerning the design of efficient computational kernels to be used on our high performance algorithms and codes, we will develop research activities first on regular frameworks before extending them to more irregular and complex situations. In particular, we will work first on optimized dense linear algebra kernels and we will use them in our more complicated direct and hybrid

solvers for sparse linear algebra and in our fast multipole algorithms for interaction computations. In this context, we will participate to the development of those kernels in collaboration with groups specialized in dense linear algebra. In particular, we intend develop a strong collaboration with the group of Jack Dongarra at the University of Tennessee and collaborating research groups. The objectives will be to develop dense linear algebra algorithms and libraries for multicore architectures in the context the PLASMA project and for GPU and hybrid multicore/GPU architectures in the context of the MAGMA project. The framework that hosts all these research activities is the associate team MORSE.

A more prospective objective is to study the resiliency in the context of large-scale scientific applications for massively parallel architectures. Indeed, with the increase of the number of computational cores per node, the probability of a hardware crash on a core or of a memory corruption is dramatically increased. This represents a crucial problem that needs to be addressed. However, we will only study it at the algorithmic/application level even if it needed lower-level mechanisms (at OS level or even hardware level). Of course, this work can be performed at lower levels (at operating system) level for example but we do believe that handling faults at the application level provides more knowledge about what has to be done (at application level we know what is critical and what is not). The approach that we will follow will be based on the use of a combination of fault-tolerant implementations of the run-time environments we use (like for example ULFM) and an adaptation of our algorithms to try to manage this kind of faults. This topic represents a very long range objective which needs to be addressed to guaranty the robustness of our solvers and applications. In that respect, we are involved in a ANR-Blanc project entitles RESCUE jointly with two other Inria EPI, namely ROMA and GRAND-LARGE as well as in the EXA2CT FP7 project. The main objective of the RESCUE project is to develop new algorithmic techniques and software tools to solve the exascale resilience problem. Solving this problem implies a departure from current approaches, and calls for yet-to-be- discovered algorithms, protocols and software tools.

Finally, it is important to note that the main goal of HIEPACS is to design tools and algorithms that will be used within complex simulation frameworks on next-generation parallel machines. Thus, we intend with our partners to use the proposed approach in complex scientific codes and to validate them within very large scale simulations as well as designing parallel solution in co-design collaborations.

3.3. High performance solvers for large linear algebra problems

Participants: Emmanuel Agullo, Astrid Casadei, Olivier Coulaud, Mathieu Faverge, Romain Garnier, Luc Giraud, Abdou Guermouche, Yuval Harness, Matthieu Kuhn, Xavier Lacoste, Stojce Nakov, Julien Pedron, Gregoire Pichon, Louis Poirel, Florent Pruvost, Pierre Ramet, Jean Roman.

Starting with the developments of basic linear algebra kernels tuned for various classes of computers, a significant knowledge on the basic concepts for implementations on high-performance scientific computers has been accumulated. Further knowledge has been acquired through the design of more sophisticated linear algebra algorithms fully exploiting those basic intensive computational kernels. In that context, we still look at the development of new computing platforms and their associated programming tools. This enables us to identify the possible bottlenecks of new computer architectures (memory path, various level of caches, inter processor or node network) and to propose ways to overcome them in algorithmic design. With the goal of designing efficient scalable linear algebra solvers for large scale applications, various tracks will be followed in order to investigate different complementary approaches. Sparse direct solvers have been for years the methods of choice for solving linear systems of equations, it is nowadays admitted that classical approaches are not scalable neither from a computational complexity nor from a memory view point for large problems such as those arising from the discretization of large 3D PDE problems. We will continue to work on sparse direct solvers on the one hand to make sure they fully benefit from most advanced computing platforms and on the other hand to attempt to reduce their memory and computational costs for some classes of problems where data sparse ideas can be considered. Furthermore, sparse direct solvers are a key building boxes for the design of some of our parallel algorithms such as the hybrid solvers described in the sequel of this section. Our activities in that context will mainly address preconditioned Krylov subspace methods; both components, preconditioner and Krylov solvers, will be investigated. In this framework, and possibly in relation with the

research activity on fast multipole, we intend to study how emerging $\mathcal{H}$-matrix arithmetic can benefit to our solver research efforts.

3.3.1. *Parallel sparse direct solver*

Solving large sparse systems $Ax = b$ of linear equations is a crucial and time-consuming step, arising in many scientific and engineering applications. Consequently, many parallel techniques for sparse matrix factorization have been studied and implemented.

Sparse direct solvers are mandatory when the linear system is very ill-conditioned; such a situation is often encountered in structural mechanics codes, for example. Therefore, to obtain an industrial software tool that must be robust and versatile, high-performance sparse direct solvers are mandatory, and parallelism is then necessary for reasons of memory capability and acceptable solution time. Moreover, in order to solve efficiently 3D problems with more than 50 million unknowns, which is now a reachable challenge with new multicore supercomputers, we must achieve good scalability in time and control memory overhead. Solving a sparse linear system by a direct method is generally a highly irregular problem that induces some challenging algorithmic problems and requires a sophisticated implementation scheme in order to fully exploit the capabilities of modern supercomputers.

New supercomputers incorporate many microprocessors which are composed of one or many computational cores. These new architectures induce strongly hierarchical topologies. These are called NUMA architectures. In the context of distributed NUMA architectures, in collaboration with the Inria STORM team, we study optimization strategies to improve the scheduling of communications, threads and I/O. We have developed dynamic scheduling designed for NUMA architectures in the `PaStiX` solver. The data structures of the solver, as well as the patterns of communication have been modified to meet the needs of these architectures and dynamic scheduling. We are also interested in the dynamic adaptation of the computation grain to use efficiently multi-core architectures and shared memory. Experiments on several numerical test cases have been performed to prove the efficiency of the approach on different architectures.

In collaboration with the ICL team from the University of Tennessee, and the STORM team from Inria, we are evaluating the way to replace the embedded scheduling driver of the `PaStiX` solver by one of the generic frameworks, `PaRSEC` or `StarPU`, to execute the task graph corresponding to a sparse factorization. The aim is to design algorithms and parallel programming models for implementing direct methods for the solution of sparse linear systems on emerging computer equipped with GPU accelerators. More generally, this work will be performed in the context of the associate team MORSE and the ANR SOLHAR project which aims at designing high performance sparse direct solvers for modern heterogeneous systems. This ANR project involves several groups working either on the sparse linear solver aspects (HIEPACS and ROMA from Inria and APO from IRIT), on runtime systems (STORM from Inria) or scheduling algorithms (REALOPT and ROMA from Inria). The results of these efforts will be validated in the applications provided by the industrial project members, namely CEA-CESTA and Airbus Group Innovations.

On the numerical side, we are studying how the data sparsness that might exist in some dense blocks appearing during the factorization can be exploited using different compression techniques based on $\mathcal{H}$-matrix (and variants) arithmetics. This research activity will be conducted in the framework of the FASTLA associate team and will naturally irrigate the hybrid solvers described below as well as closely interact with the sparse direct solver actions as well as the other research efforts where similar data sparseness might be exploited.

3.3.2. *Hybrid direct/iterative solvers based on algebraic domain decomposition techniques*

One route to the parallel scalable solution of large sparse linear systems in parallel scientific computing is the use of hybrid methods that hierarchically combine direct and iterative methods. These techniques inherit the advantages of each approach, namely the limited amount of memory and natural parallelization for the iterative component and the numerical robustness of the direct part. The general underlying ideas are not new since they have been intensively used to design domain decomposition techniques; those approaches cover a fairly large range of computing techniques for the numerical solution of partial differential equations (PDEs) in time and space. Generally speaking, it refers to the splitting of the computational domain into sub-domains with or

without overlap. The splitting strategy is generally governed by various constraints/objectives but the main one is to express parallelism. The numerical properties of the PDEs to be solved are usually intensively exploited at the continuous or discrete levels to design the numerical algorithms so that the resulting specialized technique will only work for the class of linear systems associated with the targeted PDE.

In that context, we intend to continue our effort on the design of algebraic non-overlapping domain decomposition techniques that rely on the solution of a Schur complement system defined on the interface introduced by the partitioning of the adjacency graph of the sparse matrix associated with the linear system. Although it is better conditioned than the original system the Schur complement needs to be precondition to be amenable to a solution using a Krylov subspace method. Different hierarchical preconditioners will be considered, possibly multilevel, to improve the numerical behaviour of the current approaches implemented in our software libraries `HIPS` and `MaPHyS`. This activity will be developed in the context of the ANR DEDALES project. In addition to this numerical studies, advanced parallel implementation will be developed that will involve close collaborations between the hybrid and sparse direct activities.

3.3.3. *Linear Krylov solvers*

Preconditioning is the main focus of the two activities described above. They aim at speeding up the convergence of a Krylov subspace method that is the complementary component involved in the solvers of interest for us. In that framework, we believe that various aspects deserve to be investigated; we will consider the following ones:

- preconditioned block Krylov solvers for multiple right-hand sides. In many large scientific and industrial applications, one has to solve a sequence of linear systems with several right-hand sides given simultaneously or in sequence (radar cross section calculation in electromagnetism, various source locations in seismic, parametric studies in general, ...). For "simultaneous" right-hand sides, the solvers of choice have been for years based on matrix factorizations as the factorization is performed once and simple and cheap block forward/backward substitutions are then performed. In order to effectively propose alternative to such solvers, we need to have efficient preconditioned Krylov subspace solvers. In that framework, block Krylov approaches, where the Krylov spaces associated with each right-hand side are shared to enlarge the search space will be considered. They are not only attractive because of this numerical feature (larger search space), but also from an implementation point of view. Their block-structures exhibit nice features with respect to data locality and re-usability that comply with the memory constraint of multicore architectures. We will continue the numerical study and design of the block GMRES variant that combines inexact breakdown detection, deflation at restart and subspace recycling. Beyond new numerical investigations, a software implementation to be included in our linear solver library will be developed in the context of the DGA HIBOX project.

- Extension or modification of Krylov subspace algorithms for multicore architectures: finally to match as much as possible to the computer architecture evolution and get as much as possible performance out of the computer, a particular attention will be paid to adapt, extend or develop numerical schemes that comply with the efficiency constraints associated with the available computers. Nowadays, multicore architectures seem to become widely used, where memory latency and bandwidth are the main bottlenecks; investigations on communication avoiding techniques will be undertaken in the framework of preconditioned Krylov subspace solvers as a general guideline for all the items mentioned above. This research activity will benefit from the FP7 EXA2CT project led by HIEPACS on behalf of the IPL C2S@EXA that involves two other Inria projects namely ALPINES and SAGE.

3.3.4. *Eigensolvers*

Many eigensolvers also rely on Krylov subspace techniques. Naturally some links exist between the Krylov subspace linear solvers and the Krylov subspace eigensolvers. We plan to study the computation of eigenvalue problems with respect to the following two different axes:

- Exploiting the link between Krylov subspace methods for linear system solution and eigensolvers, we intend to develop advanced iterative linear methods based on Krylov subspace methods that use some spectral information to build part of a subspace to be recycled, either though space augmentation or through preconditioner update. This spectral information may correspond to a certain part of the spectrum of the original large matrix or to some approximations of the eigenvalues obtained by solving a reduced eigenproblem. This technique will also be investigated in the framework of block Krylov subspace methods.

- In the context of the calculation of the ground state of an atomistic system, eigenvalue computation is a critical step; more accurate and more efficient parallel and scalable eigensolvers are required.

3.3.5. $\mathcal{H}$-*matrix arithmetics to reduce the memory and computation complexity of sparse direct solvers*

In the context of FastLA Associate Team, during the last 2 years, we are collaborating with Eric Darve, professor in the Institute for Computational and Mathematical Engineering and the Mechanical Engineering Department at Stanford, on the design of a new efficient sparse direct solvers.

Sparse direct solvers such as `PaStiX` are currently limited by their memory requirements and computational cost. They are competitive for small matrices but are often less efficient than iterative methods for large matrices in terms of memory. We are currently accelerating the dense algebra components of direct solvers using hierarchical matrices algebra. In the first step, we are targeting an $O(N^{4/3})$ solver. Preliminary benchmarks indicate that a speed up of 2x to 10x is possible (on the largest test cases).

Moreover, in a second step, we are looking into a new class of direct solvers with cost $O(N)$. These solvers have the potential to significantly change the capabilities of linear solvers to solve complex partial differential equations on large scale parallel machines. In particular, many existing iterative methods and preconditioners fail at large scale (size of matrix N and size of parallel cluster). Instead, this new generation of solvers has been shown to scale like $O(N)$ even for very large matrices. On parallel computers, the amount of communication is essentially optimal (through near-optimal low-rank compression methodologies), leading to very scalable methods.

Such improvements will allow these solvers to outperform current state-of-the-art methods such as algebraic multigrid methods. `PaStiX` is a key platform to support these developments since it has been highly optimized for multicore and heterogeneous platforms. This project, investigated in the PhD of Gregoire Pichon (granted by DGA), will leverage several codes written at Stanford to solve linear systems.

3.4. High performance Fast Multipole Method for N-body problems

Participants: Emmanuel Agullo, Bérenger Bramas, Arnaud Etcheverry, Olivier Coulaud, Cyrille Piacibello, Guillaume Sylvand.

In most scientific computing applications considered nowadays as computational challenges (like biological and material systems, astrophysics or electromagnetism), the introduction of hierarchical methods based on an octree structure has dramatically reduced the amount of computation needed to simulate those systems for a given accuracy. For instance, in the N-body problem arising from these application fields, we must compute all pairwise interactions among N objects (particles, lines, ...) at every timestep. Among these methods, the Fast Multipole Method (FMM) developed for gravitational potentials in astrophysics and for electrostatic (coulombic) potentials in molecular simulations solves this N-body problem for any given precision with $O(N)$ runtime complexity against $O(N^2)$ for the direct computation.

The potential field is decomposed in a near field part, directly computed, and a far field part approximated thanks to multipole and local expansions. We introduced a matrix formulation of the FMM that exploits the cache hierarchy on a processor through the Basic Linear Algebra Subprograms (BLAS). Moreover, we developed a parallel adaptive version of the FMM algorithm for heterogeneous particle distributions, which is very efficient on parallel clusters of SMP nodes. Finally on such computers, we developed the first hybrid MPI-thread algorithm, which enables to reach better parallel efficiency and better memory scalability. We plan to work on the following points in HIEPACS.

3.4.1. Improvement of calculation efficiency

Nowadays, the high performance computing community is examining alternative architectures that address the limitations of modern cache-based designs. GPU (Graphics Processing Units) and the Cell processor have thus already been used in astrophysics and in molecular dynamics. The Fast Mutipole Method has also been implemented on GPU. We intend to examine the potential of using these forthcoming processors as a building block for high-end parallel computing in N-body calculations. More precisely, we want to take advantage of our specific underlying BLAS routines to obtain an efficient and easily portable FMM for these new architectures. Algorithmic issues such as dynamic load balancing among heterogeneous cores will also have to be solved in order to gather all the available computation power. This research action will be conduced on close connection with the activity described in Section 3.2 .

3.4.2. Non uniform distributions

In many applications arising from material physics or astrophysics, the distribution of the data is highly non uniform and the data can grow between two time steps. As mentioned previously, we have proposed a hybrid MPI-thread algorithm to exploit the data locality within each node. We plan to further improve the load balancing for highly non uniform particle distributions with small computation grain thanks to dynamic load balancing at the thread level and thanks to a load balancing correction over several simulation time steps at the process level.

3.4.3. Fast multipole method for dislocation operators

The engine that we develop will be extended to new potentials arising from material physics such as those used in dislocation simulations. The interaction between dislocations is long ranged ($O(1/r)$) and anisotropic, leading to severe computational challenges for large-scale simulations. Several approaches based on the FMM or based on spatial decomposition in boxes are proposed to speed-up the computation. In dislocation codes, the calculation of the interaction forces between dislocations is still the most CPU time consuming. This computation has to be improved to obtain faster and more accurate simulations. Moreover, in such simulations, the number of dislocations grows while the phenomenon occurs and these dislocations are not uniformly distributed in the domain. This means that strategies to dynamically balance the computational load are crucial to achieve high performance.

3.4.4. Fast multipole method for boundary element methods

The boundary element method (BEM) is a well known solution of boundary value problems appearing in various fields of physics. With this approach, we only have to solve an integral equation on the boundary. This implies an interaction that decreases in space, but results in the solution of a dense linear system with $O(N^3)$ complexity. The FMM calculation that performs the matrix-vector product enables the use of Krylov subspace methods. Based on the parallel data distribution of the underlying octree implemented to perform the FMM, parallel preconditioners can be designed that exploit the local interaction matrices computed at the finest level of the octree. This research action will be conduced on close connection with the activity described in Section 3.3 . Following our earlier experience, we plan to first consider approximate inverse preconditionners that can efficiently exploit these data structures.

3.5. Efficient algorithmic for load balancing and code coupling in complex simulations

Participants: Astrid Casadei, Olivier Coulaud, Aurélien Esnard, Maria Predari, Pierre Ramet, Jean Roman.

Many important physical phenomena in material physics and climatology are inherently complex applications. They often use multi-physics or multi-scale approaches, that couple different models and codes. The key idea is to reuse available legacy codes through a coupling framework instead of merging them into a standalone application. There is typically one model per different scale or physics; and each model is implemented by a parallel code. For instance, to model a crack propagation, one uses a molecular dynamic code to represent the atomistic scale and an elasticity code using a finite element method to represent the continuum scale. Indeed,

fully microscopic simulations of most domains of interest are not computationally feasible. Combining such different scales or physics is still a challenge to reach high performance and scalability. If the model aspects are often well studied, there are several open algorithmic problems, that we plan to investigate in the HIEPACS project-team.

3.5.1. Efficient schemes for multiscale simulations

As mentioned previously, many important physical phenomena, such as material deformation and failure (see Section 4.1), are inherently multiscale processes that cannot always be modeled via continuum model. Fully microscopic simulations of most domains of interest are not computationally feasible. Therefore, researchers must look at multiscale methods that couple micro models and macro models. Combining different scales such as quantum-atomistic or atomistic, mesoscale and continuum, are still a challenge to obtain efficient and accurate schemes that efficiently and effectively exchange information between the different scales. We are currently involved in two national research projects, that focus on multiscale schemes. More precisely, the models that we start to study are the quantum to atomic coupling (QM/MM coupling) in the ANR NOSSI and the atomic to dislocation coupling in the ANR OPTIDIS.

3.5.2. Dynamic load balancing for massively parallel coupled codes

In this context of code coupling, one crucial issue is undoubtedly the load balancing of the whole coupled simulation that remains an open question. The goal here is to find the best data distribution for the whole coupled simulation and not only for each standalone code, as it is most usually done. Indeed, the naive balancing of each code on its own can lead to an important imbalance and to a communication bottleneck during the coupling phase, that can drastically decrease the overall performance. Therefore, one argues that it is required to model the coupling itself in order to ensure a good scalability, especially when running on massively parallel architectures (tens of thousands of processors/cores). In other words, one must develop new algorithms and software implementation to perform a *coupling-aware* partitioning of the whole application.

Another related problem is the problem of resource allocation. This is particularly important for the global coupling efficiency and scalabilty, because each code involved in the coupling can be more or less computationally intensive, and there is a good trade-off to find between resources assigned to each code to avoid that one of them waits for the other(s). And what happens if the load of one code dynamically changes relatively to the other? In such a case, it could be convenient to dynamically adapt the number of resources used at runtime.

For instance, the conjugate heat transfer simulation in complex geometries (as developed by the CFD team of CERFACS) requires to couple a fluid/convection solver (AVBP) with a solid/conduction solver (AVTP). The AVBP code is much more CPU consuming than the AVTP code. As a consequence, there is an important computational imbalance between the two solvers. The use of new algorithms to correctly load balance coupled simulations with enhanced graph partitioning techniques appears as a promising way to reach better performances of coupled application on massively parallel computers.

3.5.3. Graph partitioning for hybrid solvers

Graph handling and partitioning play a central role in the activity described here but also in other numerical techniques detailed in Section 3.3 .

The Nested Dissection is now a well-known heuristic for sparse matrix ordering to both reduce the fill-in during numerical factorization and to maximize the number of independent computation tasks. By using the block data structure induced by the partition of separators of the original graph, very efficient parallel block solvers have been designed and implemented according to supernodal or multifrontal approaches. Considering hybrid methods mixing both direct and iterative solvers such as `HIPS` or `MaPHyS`, obtaining a domain decomposition leading to a good balancing of both the size of domain interiors and the size of interfaces is a key point for load balancing and efficiency in a parallel context. We intend to revisit some well-known graph partitioning techniques in the light of the hybrid solvers and design new algorithms to be tested in the `Scotch` package.

LFANT Project-Team

3. Research Program

3.1. Number fields, class groups and other invariants

Participants: Bill Allombert, Karim Belabas, Cyril Bouvier, Julio Brau Avila, Jean-Paul Cerri, Iuliana Ciocanea Teodorescu, Henri Cohen, Jean-Marc Couveignes, Andreas Enge, Fredrik Johansson, Pınar Kılıçer, Pierre Lezowski.

Modern number theory has been introduced in the second half of the 19th century by Dedekind, Kummer, Kronecker, Weber and others, motivated by Fermat's conjecture: There is no non-trivial solution in integers to the equation $x^n + y^n = z^n$ for $n \geqslant 3$. For recent textbooks, see [5]. Kummer's idea for solving Fermat's problem was to rewrite the equation as $(x + y)(x + \zeta y)(x + \zeta^2 y) \cdots (x + \zeta^{n-1} y) = z^n$ for a primitive n-th root of unity ζ, which seems to imply that each factor on the left hand side is an n-th power, from which a contradiction can be derived.

The solution requires to augment the integers by *algebraic numbers*, that are roots of polynomials in $\mathbb{Z}[X]$. For instance, ζ is a root of $X^n - 1$, $\sqrt[3]{2}$ is a root of $X^3 - 2$ and $\frac{\sqrt{3}}{5}$ is a root of $25X^2 - 3$. A *number field* consists of the rationals to which have been added finitely many algebraic numbers together with their sums, differences, products and quotients. It turns out that actually one generator suffices, and any number field K is isomorphic to $\mathbb{Q}[X]/(f(X))$, where $f(X)$ is the minimal polynomial of the generator. Of special interest are *algebraic integers*, "numbers without denominators", that are roots of a monic polynomial. For instance, ζ and $\sqrt[3]{2}$ are integers, while $\frac{\sqrt{3}}{5}$ is not. The *ring of integers* of K is denoted by $\mathcal{O}_K$; it plays the same role in K as $\mathbb{Z}$ in $\mathbb{Q}$.

Unfortunately, elements in $\mathcal{O}_K$ may factor in different ways, which invalidates Kummer's argumentation. Unique factorisation may be recovered by switching to *ideals*, subsets of $\mathcal{O}_K$ that are closed under addition and under multiplication by elements of $\mathcal{O}_K$. In $\mathbb{Z}$, for instance, any ideal is *principal*, that is, generated by one element, so that ideals and numbers are essentially the same. In particular, the unique factorisation of ideals then implies the unique factorisation of numbers. In general, this is not the case, and the *class group* Cl_K of ideals of $\mathcal{O}_K$ modulo principal ideals and its *class number* $h_K = |\mathrm{Cl}_K|$ measure how far $\mathcal{O}_K$ is from behaving like $\mathbb{Z}$.

Using ideals introduces the additional difficulty of having to deal with *units*, the invertible elements of $\mathcal{O}_K$: Even when $h_K = 1$, a factorisation of ideals does not immediately yield a factorisation of numbers, since ideal generators are only defined up to units. For instance, the ideal factorisation $(6) = (2) \cdot (3)$ corresponds to the two factorisations $6 = 2 \cdot 3$ and $6 = (-2) \cdot (-3)$. While in $\mathbb{Z}$, the only units are 1 and -1, the unit structure in general is that of a finitely generated $\mathbb{Z}$-module, whose generators are the *fundamental units*. The *regulator* R_K measures the "size" of the fundamental units as the volume of an associated lattice.

One of the main concerns of algorithmic algebraic number theory is to explicitly compute these invariants (Cl_K and h_K, fundamental units and R_K), as well as to provide the data allowing to efficiently compute with numbers and ideals of $\mathcal{O}_K$; see [30] for a recent account.

The *analytic class number formula* links the invariants h_K and R_K (unfortunately, only their product) to the ζ-function of K, $\zeta_K(s) := \prod_{\mathfrak{p} \text{ prime ideal of } \mathcal{O}_K} \left(1 - \mathrm{N}\,\mathfrak{p}^{-s}\right)^{-1}$, which is meaningful when $\Re(s) > 1$, but which may be extended to arbitrary complex $s \neq 1$. Introducing characters on the class group yields a generalisation of ζ- to L-functions. The *generalised Riemann hypothesis (GRH)*, which remains unproved even over the rationals, states that any such L-function does not vanish in the right half-plane $\Re(s) > 1/2$. The validity of the GRH has a dramatic impact on the performance of number theoretic algorithms. For instance, under GRH, the class group admits a system of generators of polynomial size; without GRH, only exponential bounds are known. Consequently, an algorithm to compute Cl_K via generators and relations (currently the only viable practical approach) either has to assume that GRH is true or immediately becomes exponential.

When $h_K = 1$ the number field K may be norm-Euclidean, endowing $\mathcal{O}_K$ with a Euclidean division algorithm. This question leads to the notions of the Euclidean minimum and spectrum of K, and another task in algorithmic number theory is to compute explicitly this minimum and the upper part of this spectrum, yielding for instance generalised Euclidean gcd algorithms.

3.2. Function fields, algebraic curves and cryptology

Participants: Karim Belabas, Julio Brau Avila, Guilhem Castagnos, Jean-Marc Couveignes, Andreas Enge, Hamish Ivey-Law, Enea Milio, Damien Robert, Emmanouil Tzortzakis.

Algebraic curves over finite fields are used to build the currently most competitive public key cryptosystems. Such a curve is given by a bivariate equation $\mathcal{C}(X, Y) = 0$ with coefficients in a finite field $\mathbb{F}_q$. The main classes of curves that are interesting from a cryptographic perspective are *elliptic curves* of equation $\mathcal{C} = Y^2 - (X^3 + aX + b)$ and *hyperelliptic curves* of equation $\mathcal{C} = Y^2 - (X^{2g+1} + \cdots)$ with $g \geqslant 2$.

The cryptosystem is implemented in an associated finite abelian group, the *Jacobian* $\mathrm{Jac}_{\mathcal{C}}$. Using the language of function fields exhibits a close analogy to the number fields discussed in the previous section. Let $\mathbb{F}_q(X)$ (the analogue of $\mathbb{Q}$) be the *rational function field* with subring $\mathbb{F}_q[X]$ (which is principal just as $\mathbb{Z}$). The *function field* of $\mathcal{C}$ is $K_{\mathcal{C}} = \mathbb{F}_q(X)[Y]/(\mathcal{C})$; it contains the *coordinate ring* $\mathcal{O}_{\mathcal{C}} = \mathbb{F}_q[X, Y]/(\mathcal{C})$. Definitions and properties carry over from the number field case $K/\mathbb{Q}$ to the function field extension $K_{\mathcal{C}}/\mathbb{F}_q(X)$. The Jacobian $\mathrm{Jac}_{\mathcal{C}}$ is the divisor class group of $K_{\mathcal{C}}$, which is an extension of (and for the curves used in cryptography usually equals) the ideal class group of $\mathcal{O}_{\mathcal{C}}$.

The size of the Jacobian group, the main security parameter of the cryptosystem, is given by an L-function. The GRH for function fields, which has been proved by Weil, yields the Hasse–Weil bound $(\sqrt{q} - 1)^{2g} \leqslant |\mathrm{Jac}_{\mathcal{C}}| \leqslant (\sqrt{q} + 1)^{2g}$, or $|\mathrm{Jac}_{\mathcal{C}}| \approx q^g$, where the *genus* g is an invariant of the curve that correlates with the degree of its equation. For instance, the genus of an elliptic curve is 1, that of a hyperelliptic one is $\frac{\deg_X \mathcal{C} - 1}{2}$. An important algorithmic question is to compute the exact cardinality of the Jacobian.

The security of the cryptosystem requires more precisely that the *discrete logarithm problem* (DLP) be difficult in the underlying group; that is, given elements D_1 and $D_2 = xD_1$ of $\mathrm{Jac}_{\mathcal{C}}$, it must be difficult to determine x. Computing x corresponds in fact to computing $\mathrm{Jac}_{\mathcal{C}}$ explicitly with an isomorphism to an abstract product of finite cyclic groups; in this sense, the DLP amounts to computing the class group in the function field setting.

For any integer n, the *Weil pairing* e_n on $\mathcal{C}$ is a function that takes as input two elements of order n of $\mathrm{Jac}_{\mathcal{C}}$ and maps them into the multiplicative group of a finite field extension $\mathbb{F}_{q^k}$ with $k = k(n)$ depending on n. It is bilinear in both its arguments, which allows to transport the DLP from a curve into a finite field, where it is potentially easier to solve. The *Tate-Lichtenbaum pairing*, that is more difficult to define, but more efficient to implement, has similar properties. From a constructive point of view, the last few years have seen a wealth of cryptosystems with attractive novel properties relying on pairings.

For a random curve, the parameter k usually becomes so big that the result of a pairing cannot even be output any more. One of the major algorithmic problems related to pairings is thus the construction of curves with a given, smallish k.

3.3. Complex multiplication

Participants: Karim Belabas, Henri Cohen, Jean-Marc Couveignes, Andreas Enge, Hamish Ivey-Law, Fredrik Johansson, Chloë Martindale, Enea Milio, Damien Robert.

Complex multiplication provides a link between number fields and algebraic curves; for a concise introduction in the elliptic curve case, see [32], for more background material, [31]. In fact, for most curves $\mathcal{C}$ over a finite field, the endomorphism ring of $\mathrm{Jac}_{\mathcal{C}}$, which determines its L-function and thus its cardinality, is an order in a special kind of number field K, called *CM field*. The CM field of an elliptic curve is an imaginary-quadratic field $\mathbb{Q}(\sqrt{D})$ with $D < 0$, that of a hyperelliptic curve of genus g is an imaginary-quadratic extension of a totally real number field of degree g. Deuring's lifting theorem ensures that $\mathcal{C}$ is the reduction modulo some prime of a curve with the same endomorphism ring, but defined over the *Hilbert class field* H_K of K.

Algebraically, H_K is defined as the maximal unramified abelian extension of K; the Galois group of H_K/K is then precisely the class group Cl_K. A number field extension H/K is called *Galois* if $H \simeq K[X]/(f)$ and H contains all complex roots of f. For instance, $\mathbb{Q}(\sqrt{2})$ is Galois since it contains not only $\sqrt{2}$, but also the second root $-\sqrt{2}$ of $X^2 - 2$, whereas $\mathbb{Q}(\sqrt[3]{2})$ is not Galois, since it does not contain the root $e^{2\pi i/3}\sqrt[3]{2}$ of $X^3 - 2$. The *Galois group* $\mathrm{Gal}_{H/K}$ is the group of automorphisms of H that fix K; it permutes the roots of f. Finally, an *abelian* extension is a Galois extension with abelian Galois group.

Analytically, in the elliptic case H_K may be obtained by adjoining to K the *singular value* $j(\tau)$ for a complex valued, so-called *modular* function j in some $\tau \in \mathcal{O}_K$; the correspondence between $\mathrm{Gal}_{H/K}$ and Cl_K allows to obtain the different roots of the minimal polynomial f of $j(\tau)$ and finally f itself. A similar, more involved construction can be used for hyperelliptic curves. This direct application of complex multiplication yields algebraic curves whose L-functions are known beforehand; in particular, it is the only possible way of obtaining ordinary curves for pairing-based cryptosystems.

The same theory can be used to develop algorithms that, given an arbitrary curve over a finite field, compute its L-function.

A generalisation is provided by *ray class fields*; these are still abelian, but allow for some well-controlled ramification. The tools for explicitly constructing such class fields are similar to those used for Hilbert class fields.

<div align="center">

MAGIQUE-3D Project-Team

</div>

3. Research Program

3.1. Introduction

Probing the invisible is a quest that is shared by a wide variety of scientists such as archaeologists, geologists, astrophysicists, physicists, etc... Magique-3D is involved in Geophysical imaging which aims at understanding the internal structure of the Earth from the propagation of waves. Both qualitative and quantitative information are required and two geophysical techniques can be used: **seismic reflection** and **seismic inversion**. Seismic reflection provides a qualitative description of the subsurface from reflected seismic waves by indicating the position of the reflectors while seismic inversion transforms seismic reflection data into a quantitative description of the subsurface. Both techniques are inverse problems based upon the numerical solution of wave equations. Oil and Gas explorations have been pioneering application domains for seismic reflection and inversion and even if numerical seismic imaging is computationally intensive, oil companies promote the use of numerical simulations to provide synthetic maps of the subsurface. This is due to the tremendous progresses of scientific computing which have pushed the limits of existing numerical methods and it is now conceivable to tackle realistic 3D problems. However, mathematical wave modeling has to be well-adapted to the region of interest and the numerical schemes which are employed to solve wave equations have to be both accurate and scalable enough to take full advantage of parallel computing. Today, geophysical imaging tackles more and more realistic problems and we can contribute to this task by improving the modeling and by deriving advanced numerical methods for solving wave problems.

Magique-3D proposes to organize its research around three main axes:

1. Mathematical modeling of multi-physics involving wave equations;
2. Supercomputing for Helmholtz problems;
3. Construction of high-order hybrid schemes.

These three research fields will be developed with the main objective of solving inverse problems dedicated to geophysical imaging.

3.2. Mathematical modeling of multi-physics involving wave equations

Wave propagation modeling is of great interest for many applications like oil and gas exploration, non destructive testing, medical imaging, etc. It involves equations which can be solved in time or frequency domain and their numerical approximation is not easy to handle, in particular when dealing with real-world problems. In both cases, the propagation domain is either infinite or its dimensions are much greater than the characteristic wavelength of the phenomenon of interest. But since wave problems are hyperbolic, the physical phenomenon can be accurately described by computing solutions in a bounded domain including the sources which have generated the waves. Until now, we have mainly worked on imaging techniques based on acoustic or elastic waves and we have developed advanced finite element software packages which are used by Total for oil exploration. Nevertheless, research on modeling must go on because there are simulations which can still not be performed because their computational cost is much too high. This is particularly true for complex tectonics involving coupled wave equations. We then propose to address the issue of coupling wave equations problems by working on the mathematical construction of reduced systems. By this way, we hope to improve simulations of elastoacoustic and electroseismic phenomena and then, to perform numerical imaging of strongly heterogeneous media. Even in the simplest situation where the wavelengths are similar (elasto-acoustic coupling), the dimension of the discrete coupled problem is huge and it is a genuine issue in the prospect of solving 3D inverse problems.

The accurate numerical simulation of full wave problems in heterogeneous media is computationally intensive since it needs numerical schemes based on grids. The size of the cells depends on the propagation velocity of waves. When coupling wave problems, conversion phenomena may occur and waves with very different propagation velocity coexist. The size of the cells is then defined from the smallest velocity and in most of the real-world cases, the computational cost is crippling. Regarding existing computing capabilities, we propose to derive intermediate models which require less computational burden and provide accurate solutions for a wide-ranging class of problems including Elasto-acoustics and Electroseismology.

When it comes to mathematical analysis, we have identified two tasks which could help us simulate realistic 3D multi-physics wave problems and which are in the scope of our savoir-faire. They are construction of approximate and multiscale models which are different tasks. The construction of approximate problems aims at deriving systems of equations which discrete formulation involves middle-sized matrices and in general, they are based on high frequency hypothesis. Multiscale models are based on a rigorous analysis involving a small parameter which does not depend on the propagation velocity necessarily.

Recently, we have conducted research on the construction of approximate models for offshore imaging. Elastic and acoustic wave equations are coupled and we investigate the idea of eliminating the computations inside water by introducing equivalent interface conditions on the sea bottom. We apply an On-Surface-Radiation-Condition (OSRC) which is obtained from the approximation of the acoustic Dirichlet-to-Neumann (DtN) operator [90], [68]. To the best of our knowledge, OSRC method has never been used for solving reduced coupling wave problems and preliminary promising results are available at [71]. We would like to investigate this technique further because we could form a battery of problems which can be solved quickly. This would provide a set of solutions which we could use as initial guess for solving inverse problems. But we are concerned with the performance of the OSRC method when wave conversions with different wavelengths occur. Anyway, the approximation of the DtN operator is not obvious when the medium is strongly heterogeneous and multiscale analysis might be more adapted. For instance, according to existing results in Acoustics and Electromagnetism for the modeling of wire antennas [80], multiscale analysis should turn out to be very efficient when the propagation medium includes well logs, fractures and faults which are very thin structures when compared to the wavelength of seismic waves. Moreover, multiscale analysis should perform well when the medium is strongly oscillating like porous media. It could thus provide an alternative to homogenization techniques which can be applied only when the medium is periodic. We thus propose to develop reduced multi-scale models by performing rigorous mathematical procedure based on regular and singular multiscale analysis. Our approach distinguishes itself from others because it focuses on the numerical representation of small structures by time-dependent problems. This could give rise to the development of new finite element methods which would combine DG approximations with XFEM (Extended Finite Element Method) which has been created for the finite element treatment of thin structures like cracks.

But Earth imaging must be more than using elasto-acoustic wave propagation. Electromagnetic waves can also be used and in collaboration with Prof. D. Pardo (Iker Basque Foundation and University of Bilbao), we conduct researches on passive imaging to probe boreholes. Passive imaging is a recent technique of imaging which uses natural electromagnetic fields as sources. These fields are generated by hydromagnetic waves propagating in the magnetosphere which transform into electromagnetic waves when they reach the ionosphere. This is a mid-frequency imaging technique which applies also to mineral and geothermal exploration, to predict seismic hazard or for groundwater monitoring. We aim at developing software package for resistivity inversion, knowing that current numerical methods are not able to manage 3D inversion. We have obtained results based on a Petrov-Galerkin approximation [65], but they are limited to 2D cases. We have thus proposed to reduce the 3D problem by using 1D semi-analytic approximation of Maxwell equations [95]. This work has just started in the framework of a PhD thesis and we hope that it will give us the possibility of imaging 3D problems.

Magique-3D would like to expand its know-how by considering electro-seismic problems which are in the scope of coupling electromagnetic waves with seismic waves. Electroseismic waves are involved in porous media imaging which is a tricky task because it is based on the coupling of waves with very different wavelengths described by Biot equations and Maxwell equations. Biot equations govern waves in saturated porous media and they represent a complex physical phenomenon involving a slow wave which is very difficult to simulate numerically. In [88], interesting results have been obtained for the simulation of piezoelectric

sensors. They are based on a quasi-static approximation of the Maxwell model coupled with Elastodynamics. Now, we are concerned with the capability of using this model for Geophysical Imaging and we believe that the derivation and/or the analysis of suitable modelings is necessary. Collaborations with Geophysicists are thus mandatory in the prospect of using both experimental and numerical approaches. We would like to collaborate with Prof. C. Bordes and Prof. D. Brito (Laboratory of Complex Fluids and their Reservoirs, CNRS and University of Pau) who have efficient experimental devices for the propagation of electromagnetic waves inside saturated porous media [70]. This collaboration should be easy to organize since Magique-3D has a long-term experience in collaborating with geophysicists. We then believe that we will not need a lot of time to get joint results since we can use our advanced software packages Hou10ni and Montjoie and our colleagues have already obtained data. Electroseismology is a very challenging research domain for us and we would like to enforce our collaborations with IsTerre (Institute of Earth Science, University of Grenoble) and for that topic with Prof. S. Garambois who is an expert in Electroseismology [97], [98], [85], [86]. A joint research program could gather Geophysicists from the University of Pau and from IsTerre and Magique-3D. In particular, it would be interesting to compare simulations performed with Hou10ni, Montjoie, with the code developed by Prof. S. Garambois and to use experimental simulations for validation.

3.3. Supercomputing for Helmholtz problems

Probing invisible with harmonic equations is a need for many scientists and it is also a topic offering a wealth of interesting problems for mathematicians. It is well-known that Helmholtz equations discretization is very sensitive to the frequency scale which can be wide-ranging for some applications. For example, depth imaging is searching for deeper layers which may contain hydrocarbons and frequencies must be of a few tens of Hertz with a very low resolution. If it is to detect hidden objects, the depth of the explored region does not exceed a few tens of meters and frequencies close to the kiloHertz are used. High performing numerical methods should thus be stable for a widest as possible frequency range. In particular, these methods should minimize phenomena of numerical pollution that generate errors which increase faster with frequency than with the inverse of space discretization step. As a consequence, there is a need of mesh refinement, in particular at high frequency.

During the period 2010-2014, the team has worked extensively on high order discontinuous Galerkin (DG) methods. Like standard Finite Element Methods, they are elaborated with polynomial basis functions and they are very popular because they are defined locally for each element. It is thus easy to use basis polynomial functions with different degrees and this shows the perfect flexibility of the approximation in case of heterogeneous media including homogeneous parts. Indeed, low degree basis functions can be used in heterogeneous regions where a fine grid is necessary while high degree polynomials can be used for coarse elements covering homogeneous parts. In particular, Magique-3D has developed Hou10ni that solves harmonic wave equations with DG methods and curved elements. We found that both the effects of pollution and dispersion, which are very significant when a conventional finite element method is used, are limited [72]. However, bad conditioning is persisting and reliability of the method is not guaranteed when the coefficients vary considerably. In addition, the number of unknowns of the linear system is too big to hope to solve a realistic 3D problem. So it is important to develop approximation methods that require fewer degrees of freedom. Magique-3D wishes to invest heavily in the development of new approximation methods for harmonic wave equations. It is a difficult subject for which we want to develop different tasks, in collaboration with academic researchers with whom we are already working or have established contacts. Research directions that we would like to follow are the following.

First, we will continue our long-term collaboration with Prof. Rabia Djellouli. We want to continue to work on hybrid finite element methods that rely on basis functions composed of plane waves and polynomials. These methods have demonstrated good resistance to the phenomenon of numerical pollution [66], [67], but their capability of solving industrial problems has not been illustrated. This is certainly due to the absence of guideline for choosing the plane waves. We are thus currently working on the implementation of a methodology that makes the choice of plane waves automatic for a given simulation (fixed propagation domain, data source, etc.). This is up-front investigation and there is certainly a lot of remaining work before

being applied to geophysical imaging. But it gives the team the opportunity to test new ideas while remaining in contact with potential users of the methods.

Then we want to work with Prof. A. Bendali on developing methods of local integral equations which allow calculation of numerical fluxes on the edges of elements. One could then use these fluxes in a DG method for reconstructing the solution throughout the volume of calculation. This research is motivated by recent results which illustrate the difficulties of the existing methods which are not always able to approximate the propagating modes (plane waves) and the evanescent modes (polynomials) that may coexist, especially when one considers realistic applications. Integral equations are direct tools for computing fluxes and they are known for providing very good accuracy. They thus should help to improve the quality of approximation of DG methods which are fully flux-dependent. In addition, local integral equations would limit calculations at the interfaces, which would have the effect of limiting the number of unknowns generally high, especially for DG methods. Again, it is a matter of long-term research which success requires a significant amount of mathematical analysis, and also the development of non-trivial code.

To limit the effects of pollution and dispersion is not the only challenge that the team wants to tackle. Our experience alongside Total has made us aware of the difficulties in constructing meshes that are essential to achieve our simulations. There are several teams at Inria working on mesh generation and we are in contact with them, especially with Gamma3 (Paris-Rocquencourt Research Center). These teams develop meshes increasingly sophisticated to take account of the constraints imposed by realistic industrial benchmarks. But in our opinion, issues which are caused by the construction of meshes are not the only downside. Indeed, we have in mind to solve inverse problems and in this case it is necessary to mesh the domain at each iteration of Newton-type solver. It is therefore interesting to work on methods that either do not use mesh or rely on meshes which are very easy to construct. Regarding meshless methods, we have begun a collaboration with Prof. Djellouli which allowed us to propose a new approach called Mesh-based Frontier Free Formulation (MF3). The principle of this method is the use of fundamental solutions of Helmholtz equations as basic functions. One can then reduce the volumic variational formulation to a surfacic variational formulation which is close to an integral equation, but which does not require the calculation of singularities. The results are very promising and we hope to continue our study in the context of the application to geophysical imaging. An important step to validate this method will be particularly its extension to 3D because the results we have achieved so far are for 2D problems.

Keeping in mind the idea of limiting the difficulties of mesh, we want to study the method of virtual elements. This method attracts us because it relies on meshes that can be made of arbitrarily-shaped polygon and meshes should thus be fairly straightforward. Existing works on the subject have been mainly developed by the University of Pavia, in collaboration with Los Alamos National Laboratory [69], [76], [75], [73], [77]. None of them mentions the feasibility of the method for industrial applications and to our knowledge, there are no results on the method of virtual elements applied to the wave equations. First, we aim at applying the method described in [74] to the scalar Helmholtz equation and explore opportunities to use discontinuous elements within this framework. Then hp-adaptivity could be kept, which is particularly interesting for wave propagation in heterogeneous media.

DG methods are known to require a lot of unknowns that can exceed the limits accepted by the most advanced computers. This is particularly true for harmonic wave equations that require a large number of discretization points, even in the case of a conventional finite element method. We therefore wish to pursue a research activity that we have just started in collaboration with the project-team Nachos (Sophia-Antipolis Méditerranée Research Center). In order to reduce the number of degrees of freedom, we are interested in "hybrid mixed" Discontinuous Galerkin methods that provides a two-step procedure for solving the Helmholtz equations [89], [94], [92]. First, Lagrange multipliers are introduced to represent the flux of the numerical solution through the interface (edge or face) between two elements. The Lagrange multipliers are solution to a linear system which is constructed locally element by element. The number of degrees of freedom is then strongly reduced since for a standard DG method, there is a need of considering unknowns including volumetric values inside the element. And obviously, the gain is even more important when the order of the element is high. Next, the solution is reconstructed from the values of the multipliers and the cost of this step is negligible since it only requires inverting small-sized matrices. We have obtained promising results in the framework of the PhD

thesis of Marie Bonnasse Gahot and we want to apply it to the simulation of complex phenomena such as the 3D viscoelastic wave propagation.

Obviously, the success of all these works depends on our ability to consider realistic applications such as wave propagation in the Earth. And in these cases, it is quite possible that even if we manage to develop accurate less expensive numerical methods, the solution of inverse problems will still be computationally intensive. It is thus absolutely necessary that we conduct our research by taking advantage of the latest advances in high-performance computing. We have already initiated discussions with the project team HIEPACS (Bordeaux Sud-Ouest research Center) to test the performance of the latest features of Mumps http://mumps.enseeiht.fr/, such as Low Rank Approximation or adaptation to hybrid CPU / GPU architectures and to Intel Xeon Phi, on realistic test cases. We are also in contact with the team Algorithm at Cerfacs (Toulouse) for the development of local integral equations solvers. These collaborations are essential for us and we believe that they will be decisive for the simulation of three-dimensional elastodynamic problems. However, our scientific contribution will be limited in this area because we are not experts in HPC.

3.4. Hybrid time discretizations of high-order

Most of the meshes we consider are composed of cells greatly varying in size. This can be due to the physical characteristics (propagation speed, topography, ...) which may require to refine the mesh locally, very unstructured meshes can also be the result of dysfunction of the mesher. For practical reasons which are essentially guided by the aim of reducing the number of matrix inversions, explicit schemes are generally privileged. However, they work under a stability condition, the so-called Courant Friedrichs Lewy (CFL) condition which forces the time step being proportional to the size of the smallest cell. Then, it is necessary to perform a huge number of iterations in time and in most of the cases because of a very few number of small cells. This implies to apply a very small time step on grids mainly composed of coarse cells and thus, there is a risk of creating numerical dispersion that should not exist. However, this drawback can be avoided by using low degree polynomial basis in space in the small meshes and high degree polynomials in the coarse meshes. By this way, it is possible to relax the CFL condition and in the same time, the dispersion effects are limited. Unfortunately, the cell-size variations are so important that this strategy is not sufficient. One solution could be to apply implicit and unconditionally stable schemes, which would obviously free us from the CFL constraint. Unfortunately, these schemes require inverting a linear system at each iteration and thus needs huge computational burden that can be prohibitive in 3D. Moreover, numerical dispersion may be increased. Then, as second solution is the use of local time stepping strategies for matching the time step to the different sizes of the mesh. There are several attempts [81], [78], [96], [91], [84] and Magique 3D has proposed a new time stepping method which allows us to adapt both the time step and the order of time approximation to the size of the cells. Nevertheless, despite a very good performance assessment in academic configurations, we have observed to our detriment that its implementation inside industrial codes is not obvious and in practice, improvements of the computational costs are disappointing, especially in a HPC framework. Indeed, the local time stepping algorithm may strongly affect the scalability of the code. Moreover, the complexity of the algorithm is increased when dealing with lossy media [87].

Recently, Dolean *et al* [83] have considered a novel approach consisting in applying hybrid schemes combining second order implicit schemes in the thin cells and second order explicit discretization in the coarse mesh. Their numerical results indicate that this method could be a good alternative but the numerical dispersion is still present. It would then be interesting to implement this idea with high-order time schemes to reduce the numerical dispersion. The recent arrival in the team of J. Chabassier should help us to address this problem since she has the expertise in constructing high-order implicit time scheme based on energy preserving Newmark schemes [79]. We propose that our work be organized around the two following tasks. The first one is the extension of these schemes to the case of lossy media because applying existing schemes when there is attenuation is not straightforward. This is a key issue because there is artificial attenuation when absorbing boundary conditions are introduced and if not, there are cases with natural attenuation like in visco-elastic media. The second one is the coupling of high-order implicit schemes with high-order explicit schemes. These two tasks can be first completed independently, but the ultimate goal is obviously to couple the schemes for lossy media. We will consider two strategies for the coupling. The first one will be based on the method

proposed by Dolean *et al*, the second one will consist in using Lagrange multiplier on the interface between the coarse and fine grids and write a novel coupling condition that ensures the high order consistency of the global scheme. Besides these theoretical aspects, we will have to implement the method in industrial codes and our discretization methodology is very suitable for parallel computing since it involves Lagrange multipliers. We propose to organize this task as follows. There is first the crucial issue of a systematic distribution of the cells in the coarse/explicit and in the fine/implicit part. Based on our experience on local time stepping, we claim that it is necessary to define a criterion which discriminates thin cells from coarse ones. Indeed, we intend to develop codes which will be used by practitioners, in particular engineers working in the production department of Total. It implies that the code will be used by people who are not necessarily experts in scientific computing. Considering real-world problems means that the mesh will most probably be composed of a more or less high number of subsets arbitrarily distributed and containing thin or coarse cells. Moreover, in the prospect of solving inverse problems, it is difficult to assess which cells are thin or not in a mesh which varies at each iteration.

Another important issue is the load balancing that we can not avoid with parallel computing. In particular, we will have to choose one of these two alternatives: dedicate one part of processors to the implicit computations and the other one to explicit calculus or distribute the resolution with both schemes on all processors. A collaboration with experts in HPC is then mandatory since we are not expert in parallel computing. We will thus continue to collaborate with the team-projects Hiepacs and Runtime with whom we have a long-term experience of collaborations. The load-balancing leads then to the issue of mesh partitioning. Main mesh partitioners are very efficient for the coupling of different discretizations in space but to the best of our knowledge, the case of non-uniform time discretization has never been addressed. The study of meshes being out of the scopes of Magique-3D, we will collaborate with experts on mesh partitioning. We get already on to François Pellegrini who is the principal investigator of Scotch (http://www.labri.fr/perso/pelegrin/scotch) and permanent member of the team project Bacchus (Inria Bordeaux Sud Ouest Research Center).

In the future, we aim at enlarging the application range of implicit schemes. The idea will be to use the degrees of freedom offered by the implicit discretization in order to tackle specific difficulties that may appear in some systems. For instance, in systems involving several waves (as P and S waves in porous elastic media, or coupled wave problems as previously mentioned) the implicit parameter could be adapted to each wave and optimized in order to reduce the computational cost. More generally, we aim at reducing numeric bottlenecks by adapting the implicit discretization to specific cases.

<p style="text-align:center; color:red;">**MANAO Project-Team**</p>

3. Research Program
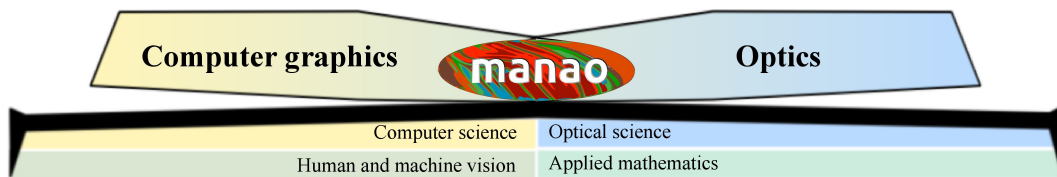
3.1. Related Scientific Domains

Figure 3. Related scientific domains of the MANAO project.

The *MANAO* project aims at studying, acquiring, modeling, and rendering the interactions between the three components that are light, shape, and matter from the viewpoint of an observer. As detailed more lengthily in the next section, such a work will be done using the following approach: first, we will tend to consider that these three components do not have strict frontiers when considering their impacts on the final observers; then, we will not only work in **computer graphics**, but also at the intersection of computer graphics and **optics**, exploring the mutual benefits that the two domains may provide. It is thus intrinsically a **transdisciplinary** project (as illustrated in Figure 3) and we expect results in both domains.

Thus, the proposed team-project aims at establishing a close collaboration between computer graphics (e.g., 3D modeling, geometry processing, shading techniques, vector graphics, and GPU programming) and optics (e.g., design of optical instruments, and theories of light propagation). The following examples illustrate the strengths of such a partnership. First, in addition to simpler radiative transfer equations [44] commonly used in computer graphics, research in the later will be based on state-of-the-art understanding of light propagation and scattering in real environments. Furthermore, research will rely on appropriate instrumentation expertise for the measurement [56], [57] and display [55] of the different phenomena. Reciprocally, optics researches may benefit from the expertise of computer graphics scientists on efficient processing to investigate interactive simulation, visualization, and design. Furthermore, new systems may be developed by unifying optical and digital processing capabilities. Currently, the scientific background of most of the team members is related to computer graphics and computer vision. A large part of their work have been focused on simulating and analyzing optical phenomena as well as in acquiring and visualizing them. Combined with the close collaboration with the optics laboratory LP2N (http://www.lp2n.fr) and with the students issued from the "Institut d'Optique" (http://www.institutoptique.fr), this background ensures that we can expect the following results from the project: the construction of a common vocabulary for tightening the collaboration between the two scientific domains and creating new research topics. By creating this context, we expect to attract (and even train) more trans-disciplinary researchers.

At the boundaries of the *MANAO* project lie issues in **human and machine vision**. We have to deal with the former whenever a human observer is taken into account. On one side, computational models of human vision are likely to guide the design of our algorithms. On the other side, the study of interactions between light, shape, and matter may shed some light on the understanding of visual perception. The same kind of connections are expected with machine vision. On the one hand, traditional computational methods for acquisition (such as photogrammetry) are going to be part of our toolbox. On the other hand, new display technologies (such as the ones used for augmented reality) are likely to benefit from our integrated approach

and systems. In the *MANAO* project we are mostly users of results from human vision. When required, some experimentation might be done in collaboration with experts from this domain, like with the European PRISM project (cf. Section TODO). For machine vision, provided the tight collaboration between optical and digital systems, research will be carried out inside the *MANAO* project.

Analysis and modeling rely on **tools from applied mathematics** such as differential and projective geometry, multi-scale models, frequency analysis [46] or differential analysis [80], linear and non-linear approximation techniques, stochastic and deterministic integrations, and linear algebra. We not only rely on classical tools, but also investigate and adapt recent techniques (e.g., improvements in approximation techniques), focusing on their ability to run on modern hardware: the development of our own tools (such as Eigen, see Section 6.3) is essential to control their performances and their abilities to be integrated into real-time solutions or into new instruments.

3.2. Research axes

The *MANAO* project is organized around four research axes that cover the large range of expertise of its members and associated members. We briefly introduce these four axes in this section. More details and their inter-influences that are illustrated in the Figure 2 will be given in the following sections.

Axis 1 is the theoretical foundation of the project. Its main goal is to increase the understanding of light, shape, and matter interactions by combining expertise from different domains: optics and human/machine vision for the analysis and computer graphics for the simulation aspect. The goal of our analyses is to identify the different layers/phenomena that compose the observed signal. In a second step, the development of physical simulations and numerical models of these identified phenomena is a way to validate the pertinence of the proposed decompositions.

In Axis 2, the final observers are mainly physical captors. Our goal is thus the development of new acquisition and display technologies that combine optical and digital processes in order to reach fast transfers between real and digital worlds, in order to increase the convergence of these two worlds.

Axes 3 and 4 focus on two aspects of computer graphics: rendering, visualization and illustration in Axis 3, and editing and modeling (content creation) in Axis 4. In these two axes, the final observers are mainly human users, either generic users or expert ones (e.g., archaeologist [84], computer graphics artists).

3.3. Axis 1: Analysis and Simulation

Challenge: Definition and understanding of phenomena resulting from interactions between light, shape, and matter as seen from an observer point of view.

Results: Theoretical tools and numerical models for analyzing and simulating the observed optical phenomena.

To reach the goals of the *MANAO* project, we need to **increase our understanding** of how light, shape, and matter act together in synergy and how the resulting signal is finally observed. For this purpose, we need to identify the different phenomena that may be captured by the targeted observers. This is the main objective of this research axis, and it is achieved by using three approaches: the simulation of interactions between light, shape, and matter, their analysis and the development of new numerical models. This resulting improved knowledge is a foundation for the researches done in the three other axes, and the simulation tools together with the numerical models serve the development of the joint optical/digital systems in Axis 2 and their validation.

One of the main and earliest goals in computer graphics is to faithfully reproduce the real world, focusing mainly on light transport. Compared to researchers in physics, researchers in computer graphics rely on a subset of physical laws (mostly radiative transfer and geometric optics), and their main concern is to efficiently use the limited available computational resources while developing as fast as possible algorithms. For this purpose, a large set of theoretical as well as computational tools has been introduced to take a **maximum benefit of hardware** specificities. These tools are often dedicated to specific phenomena (e.g., direct or indirect lighting, color bleeding, shadows, caustics). An efficiency-driven approach needs such a classification

of light paths [52] in order to develop tailored strategies [96]. For instance, starting from simple direct lighting, more complex phenomena have been progressively introduced: first diffuse indirect illumination [50], [88], then more generic inter-reflections [59], [44] and volumetric scattering [85], [41]. Thanks to this search for efficiency and this classification, researchers in computer graphics have developed a now recognized expertise in fast-simulation of light propagation. Based on finite elements (radiosity techniques) or on unbiased Monte Carlo integration schemes (ray-tracing, particle-tracing, ...), the resulting algorithms and their combination are now sufficiently accurate to be used-back in physical simulations. The *MANAO* project will continue the search for **efficient and accurate simulation** techniques, but extending it from computer graphics to optics. Thanks to the close collaboration with scientific researchers from optics, new phenomena beyond radiative transfer and geometric optics will be explored.

Search for algorithmic efficiency and accuracy has to be done in parallel with **numerical models**. The goal of visual fidelity (generalized to accuracy from an observer point of view in the project) combined with the goal of efficiency leads to the development of alternative representations. For instance, common classical finite-element techniques compute only basis coefficients for each discretization element: the required discretization density would be too large and to computationally expensive to obtain detailed spatial variations and thus visual fidelity. Examples includes texture for decorrelating surface details from surface geometry and high-order wavelets for a multi-scale representation of lighting [40]. The numerical complexity explodes when considering directional properties of light transport such as radiance intensity (Watt per square meter and per steradian - $W.m^{-2}.sr^{-1}$), reducing the possibility to simulate or accurately represent some optical phenomena. For instance, Haar wavelets have been extended to the spherical domain [87] but are difficult to extend to non-piecewise-constant data [90]. More recently, researches prefer the use of Spherical Radial Basis Functions [93] or Spherical Harmonics [79]. For more complex data, such as reflective properties (e.g., BRDF [73], [60] - 4D), ray-space (e.g., Light-Field [69] - 4D), spatially varying reflective properties (6D - [83]), new models, and representations are still investigated such as rational functions [76] or dedicated models [28] and parameterizations [86], [91]. For each (newly) defined phenomena, we thus explore the space of possible numerical representations to determine the **most suited one for a given application**, like we have done for BRDF [76].
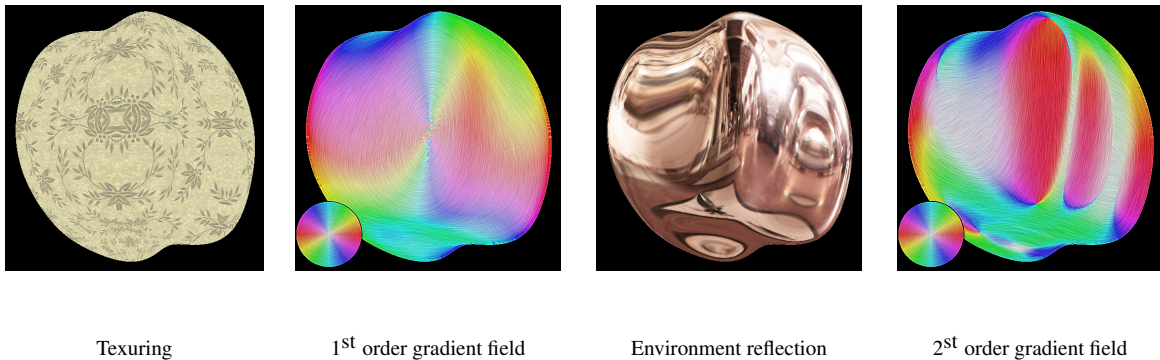

| Texuring | 1st order gradient field | Environment reflection | 2st order gradient field |

Figure 4. First-oder analysis [97] have shown that shading variations are caused by depth variations (first-order gradient field) and by normal variations (second-order fields). These fields are visualized using hue and saturation to indicate direction and magnitude of the flow respectively.

Before being able to simulate or to represent the different **observed phenomena**, we need to define and describe them. To understand the difference between an observed phenomenon and the classical light, shape, and matter decomposition, we can take the example of a highlight. Its observed shape (by a human user or a sensor) is the resulting process of the interaction of these three components, and can be simulated this way. However, this does not provide any intuitive understanding of their relative influence on the final shape: an artist will directly describe the resulting shape, and not each of the three properties. We thus want to

decompose the observed signal into models for each scale that can be easily understandable, representable, and manipulable. For this purpose, we will rely on the **analysis** of the resulting interaction of light, shape, and matter as observed by a human or a physical sensor. We first consider this analysis from an **optical point of view**, trying to identify the different phenomena and their scale according to their mathematical properties (e.g., differential [80] and frequency analysis [46]). Such an approach has leaded us to exhibit the influence of surfaces flows (depth and normal gradients) into lighting pattern deformation (see Figure 4). For a **human observer**, this correspond to one recent trend in computer graphics that takes into account the human visual systems [47] both to evaluate the results and to guide the simulations.

3.4. Axis 2: From Acquisition to Display

Challenge: Convergence of optical and digital systems to blend real and virtual worlds.

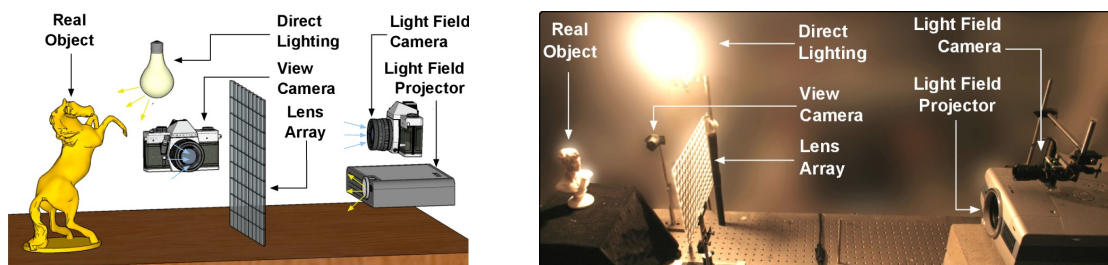**Results:** Instruments to acquire real world, to display virtual world, and to make both of them interact.

Figure 5. Light-Field transfer: global illumination between real and synthetic objects [39]

In this axis, we investigate *unified acquisition and display systems*, that is systems which combine optical instruments with digital processing. From digital to real, we investigate new display approaches [69], [55]. We consider projecting systems and surfaces [36], for personal use, virtual reality and augmented reality [31]. From the real world to the digital world, we favor direct measurements of parameters for models and representations, using (new) optical systems unless digitization is required [49], [48]. These resulting systems have to acquire the different phenomena described in Axis 1 and to display them, in an efficient manner [53], [29], [54], [57]. By efficient, we mean that we want to shorten the path between the real world and the virtual world by increasing the data bandwidth between the real (analog) and the virtual (digital) worlds, and by reducing the latency for real-time interactions (we have to prevent unnecessary conversions, and to reduce processing time). To reach this goal, the systems have to be designed as a whole, not by a simple concatenation of optical systems and digital processes, nor by considering each component independently [58].

To increase data bandwidth, one solution is to **parallelize more and more the physical systems**. One possible solution is to multiply the number of simultaneous acquisitions (e.g., simultaneous images from multiple viewpoints [57], [78]). Similarly, increasing the number of viewpoints is a way toward the creation of full 3D displays [69]. However, full acquisition or display of 3D real environments theoretically requires a continuous field of viewpoints, leading to huge data size. Despite the current belief that the increase of computational power will fill the missing gap, when it comes to visual or physical realism, if you double the processing power, people may want four times more accuracy, thus increasing data size as well. To reach the best performances, a trade-off has to be found between the amount of data required to represent accurately the reality and the amount of required processing. This trade-off may be achieved using **compressive sensing**. Compressive sensing is a new trend issued from the applied mathematics community that provides tools to accurately reconstruct a signal from a small set of measurements assuming that it is sparse in a transform domain (e.g., [77], [100]).

We prefer to achieve this goal by avoiding as much as possible the classical approach where acquisition is followed by a fitting step: this requires in general a large amount of measurements and the fitting itself may consume consequently too much memory and preprocessing time. By **preventing unnecessary conversion** through fitting techniques, such an approach increase the speed and reduce the data transfer for acquisition but also for display. One of the best recent examples is the work of Cossairt et al. [39]. The whole system is designed around a unique representation of the energy-field issued from (or leaving) a 3D object, either virtual or real: the Light-Field. A Light-Field encodes the light emitted in any direction from any position on an object. It is acquired thanks to a lens-array that leads to the capture of, and projection from, multiple simultaneous viewpoints. A unique representation is used for all the steps of this system. Lens-arrays, parallax barriers, and coded-aperture [65] are one of the key technologies to develop such acquisition (e.g., Light-Field camera [0] [58] and acquisition of light-sources [49]), projection systems (e.g., auto-stereoscopic displays). Such an approach is versatile and may be applied to improve classical optical instruments [63]. More generally, by designing unified optical and digital systems [74], it is possible to leverage the requirement of processing power, the memory footprint, and the cost of optical instruments.

Those are only some examples of what we investigate. We also consider the following approaches to develop new unified systems. First, similar to (and based on) the analysis goal of Axis 1, we have to take into account as much as possible the characteristics of the measurement setup. For instance, when fitting cannot be avoided, integrating them may improve both the processing efficiency and accuracy [76]. Second, we have to integrate signals from multiple sensors (such as GPS, accelerometer, ...) to prevent some computation (e.g., [66]). Finally, the experience of the group in surface modeling help the design of optical surfaces [61] for light sources or head-mounted displays.

3.5. Axis 3: Rendering, Visualization and Illustration

Challenge: How to offer the most legible signal to the final observer in real-time?

Results: High-level shading primitives, expressive rendering techniques for object depiction, real-time realistic rendering algorithms

Realistic   Rendering   Visualization   and Illustration

(a) Global illumination [75]   (b) Shadows [30]   (c) Shape enhancement [10]   (d) Shape depiction [27]
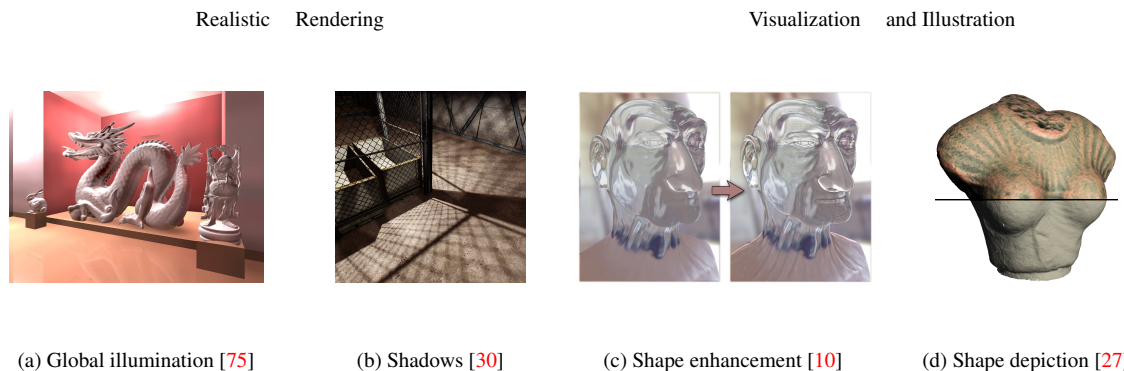
Figure 6. In the MANAO project, we are investigating rendering techniques from realistic solutions (e.g., inter-reflections (a) and shadows (b)) to more expressive ones (shape enhancement (c) with realistic style and shape depiction (d) with stylized style) for visualization.

The main goal of this axis is to offer to the final observer, in this case mostly a human user, the most legible signal in real-time. Thanks to the analysis and to the decomposition in different phenomena resulting from interactions between light, shape, and matter (Axis 1), and their perception, we can use them to convey essential information in the most pertinent way. Here, the word *pertinent* can take various forms depending on the application.

---

[0]Lytro, http://www.lytro.com/

In the context of scientific illustration and visualization, we are primarily interested in tools to convey shape or material characteristics of objects in animated 3D scenes. **Expressive rendering** techniques (see Figure 6 c,d) provide means for users to depict such features with their own style. To introduce our approach, we detail it from a shape-depiction point of view, domain where we have acquired a recognized expertise. Prior work in this area mostly focused on stylization primitives to achieve line-based rendering [98], [62] or stylized shading [34],[10] with various levels of abstraction. A clear representation of important 3D **object features** remains a major challenge for better shape depiction, stylization and abstraction purposes. Most existing representations provide only local properties (e.g., curvature), and thus lack characterization of broader shape features. To overcome this limitation, we are developing higher level descriptions of shape [26] with increased robustness to sparsity, noise, and outliers. This is achieved in close collaboration with Axis 1 by the use of higher-order local fitting methods, multi-scale analysis, and global regularization techniques. In order not to neglect the observer and the material characteristics of the objects, we couple this approach with an analysis of the appearance model. To our knowledge, this is an approach which has not been considered yet. This research direction is at the heart of the *MANAO* project, and has a strong connection with the analysis we plan to conduct in Axis 1. Material characteristics are always considered at the light ray level, but an understanding of **higher-level primitives** (like the shape of highlights and their motion) would help us to produce more legible renderings and permit novel stylizations; for instance, there is no method that is today able to create stylized renderings that follow the motion of highlights or shadows. We also believe such tools also play a fundamental role for geometry processing purposes (such as shape matching, reassembly, simplification), as well as for editing purposes as discussed in Axis 4.

In the context of **real-time photo-realistic rendering** ((see Figure 6 a,b), the challenge is to compute the most plausible images with minimal effort. During the last decade, a lot of work has been devoted to design approximate but real-time rendering algorithms of complex lighting phenomena such as soft-shadows [99], motion blur [46], depth of field [89], reflexions, refractions, and inter-reflexions. For most of these effects it becomes harder to discover fundamentally new and faster methods. On the other hand, we believe that significant speedup can still be achieved through more clever use of **massively parallel architectures** of the current and upcoming hardware, and/or through more clever tuning of the current algorithms. In particular, regarding the second aspect, we remark that most of the proposed algorithms depend on several parameters which can be used to **trade the speed over the quality**. Significant speed-up could thus be achieved by identifying effects that would be masked or facilitated and thus devote appropriate computational resources to the rendering [64], [45]. Indeed, the algorithm parameters controlling the quality vs speed are numerous without a direct mapping between their values and their effect. Moreover, their ideal values vary over space and time, and to be effective such an auto-tuning mechanism has to be extremely fast such that its cost is largely compensated by its gain. We believe that our various work on the analysis of the appearance such as in Axis 1 could be beneficial for such purpose too.

Realistic and real-time rendering is closely related to Axis 2: real-time rendering is a requirement to close the loop between real world and digital world. We have to thus develop algorithms and rendering primitives that allow the integration of the acquired data into real-time techniques. We have also to take care of that these real-time techniques have to work with new display systems. For instance, stereo, and more generally multi-view displays are based on the multiplication of simultaneous images. Brute force solutions consist in independent rendering pipeline for each viewpoint. A more energy-efficient solution would take advantages of the computation parts that may be factorized. Another example is the rendering techniques based on image processing, such as our work on augmented reality [38]. Independent image processing for each viewpoint may disturb the feeling of depth by introducing inconsistent information in each images. Finally, more dedicated displays [55] would require new rendering pipelines.

3.6. Axis 4: Editing and Modeling

Challenge: Editing and modeling appearance using drawing- or sculpting-like tools through high level representations.

Results: High-level primitives and hybrid representations for appearance and shape.

During the last decade, the domain of computer graphics has exhibited tremendous improvements in image quality, both for 2D applications and 3D engines. This is mainly due to the availability of an ever increasing amount of shape details, and sophisticated appearance effects including complex lighting environments. Unfortunately, with such a growth in visual richness, even so-called *vectorial* representations (e.g., subdivision surfaces, Bézier curves, gradient meshes, etc.) become very dense and unmanageable for the end user who has to deal with a huge mass of control points, color labels, and other parameters. This is becoming a major challenge, with a necessity for novel representations. This Axis is thus complementary of Axis 3: the focus is the development of primitives that are easy to use for modeling and editing.

More specifically, we plan to investigate *vectorial representations* that would be amenable to the production of rich shapes with a minimal set of primitives and/or parameters. To this end we plan to build upon our insights on dynamic local reconstruction techniques and implicit surfaces [3] [33]. When working in 3D, an interesting approach to produce detailed shapes is by means of procedural geometry generation. For instance, many natural phenomena like waves or clouds may be modeled using a combination of procedural functions. Turning such functions into triangle meshes (main rendering primitives of GPUs) is a tedious process that appears not to be necessary with an adapted vectorial shape representation where one could directly turn procedural functions into implicit geometric primitives. Since we want to prevent unnecessary conversions in the whole pipeline (here, between modeling and rendering steps), we will also consider *hybrid representations* mixing meshes and implicit representations. Such research has thus to be conducted while considering the associated editing tools as well as performance issues. It is indeed important to keep *real-time performance* (cf. Axis 2) throughout the interaction loop, from user inputs to display, via editing and rendering operations. Finally, it would be interesting to add *semantic information* into 2D or 3D geometric representations. Semantic geometry appears to be particularly useful for many applications such as the design of more efficient manipulation and animation tools, for automatic simplification and abstraction, or even for automatic indexing and searching. This constitutes a complementary but longer term research direction.

In the *MANAO* project, we want to investigate representations beyond the classical light, shape, and matter decomposition. We thus want to directly control the appearance of objects both in 2D and 3D applications (e.g., [94]): this is a core topic of computer graphics. When working with 2D vector graphics, digital artists must carefully set up color gradients and textures: examples range from the creation of 2D logos to the photo-realistic imitation of object materials. Classic vector primitives quickly become impractical for creating illusions of complex materials and illuminations, and as a result an increasing amount of time and skill is required. This is only for still images. For animations, vector graphics are only used to create legible appearances composed of simple lines and color gradients. There is thus a need for more complex primitives that are able to accommodate complex reflection or texture patterns, while keeping the ease of use of vector graphics. For instance, instead of drawing color gradients directly, it is more advantageous to draw flow lines that represent local surface concavities and convexities. Going through such an intermediate structure then allows to deform simple material gradients and textures in a coherent way (see Figure 7), and animate them all at once. The manipulation of 3D object materials also raises important issues. Most existing material models are tailored to faithfully reproduce physical behaviors, not to be *easily controllable* by artists. Therefore artists learn to tweak model parameters to satisfy the needs of a particular shading appearance, which can quickly become cumbersome as the complexity of a 3D scene increases. We believe that an alternative approach is required, whereby material appearance of an object in a typical lighting environment is directly input (e.g., painted or drawn), and adapted to match a plausible material behavior. This way, artists will be able to create their own appearance (e.g., by using our shading primitives [94]), and replicate it to novel illumination environments and 3D models. For this purpose, we will rely on the decompositions and tools issued from Axis 1.

(a)   (b)   (c)   (d)   (e)   (f)
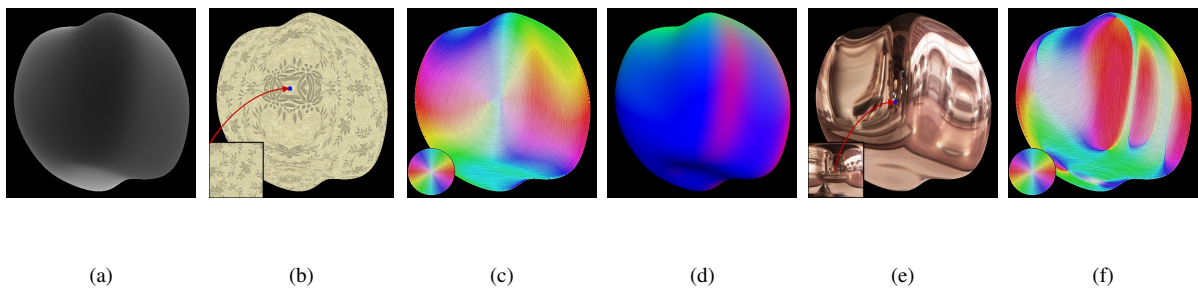
Figure 7. Based on our analysis [97] (Axis 1), we have designed a system that mimics texture (left) and shading (right) effects using image processing alone. It takes depth (a) and normal (d) images as input, and uses them to deform images (b-e) in ways that closely approximate surface flows (c-f). It provides a convincing, yet artistically controllable illusion of 3D shape conveyed through texture or shading cues.

MEMPHIS Team

3. Research Program

3.1. Hierarchical Cartesian schemes

We intend to conceive schemes that will simplify the numerical approximation of problems involving complex unsteady objects together with multi-scale physical phenomena. Rather than using extremely optimized but non-scalable algorithms, we adopt robust alternatives that bypass the difficulties linked to grid generation. Even if the mesh problem can be tackled today thanks to powerful mesh generators, it still represents a severe difficulty, in particular when highly complex unsteady geometries need to be dealt with. Industrial experience and common practice shows that mesh generation accounts for about 20% of overall analysis time, whereas creation of a simulation-specific geometry requires about 60%, and only 20% of overall time is actually devoted to analysis. The methods that we develop bypass the generation of tedious geometrical models by automatic implicit geometry representation and hierarchical Cartesian schemes.

The approach that we plan to develop combines accurate enforcement of unfitted boundary conditions with adaptive octree and overset grids. The core idea is to use an octree/overset mesh for the approximation of the solution fields, while the geometry is captured by level set functions and boundary conditions are imposed using appropriate interpolation methods. This eliminates the need for boundary conforming meshes that require time-consuming and error-prone mesh generation procedures, and opens the door for simulation of very complex geometries. In particular, it will be possible to easily import the industrial geometry and to build the associated level set function used for simulation.

Hierarchical octree grids offer several considerable advantages over classical adaptive mesh refinement for body-fitted meshes, in terms of data management, memory footprint and parallel HPC performance. Typically, when refining unstructured grids, like for example tetrahedral grids, it is necessary to store the whole data tree corresponding to successive subdivisions of the elements and eventually recompute the full connectivity graph. In the linear octree case that we develop, only the tree leaves are stored in a linear array, with a considerable memory advantage. The mapping between the tree leaves and the linear array as well as the connectivity graph is efficiently computed thanks to an appropriate space-filling curve. Concerning parallelization, linear octrees guarantee a natural load balancing thanks to the linear data structure, whereas classical non-structured meshes require sophisticated (and moreover time consuming) tools to achieve proper load distribution (SCOTCH, METIS etc.). Of course, using unfitted hierarchical meshes requires further development and analysis of methods to handle the refinement at level jumps in a consistent and conservative way, accuracy analysis for new finite-volume or finite-difference schemes, efficient reconstructions at the boundaries to recover appropriate accuracy and robustness. These subjects, that are presently virtually absent at Inria, are among the main scientific challenges of our team.

3.2. Reduced-order models

Massive parallelization and rethinking of numerical schemes will allow the solution of new problem in physics and the prediction of new phenomena thanks to simulation. However, in industrial applications fast on line responses are needed for design and control. For instance, in the design process of an aircraft, the flight conditions and manoeuvres, which provide the largest aircraft loads, are not known a priori. Therefore the aerodynamic and inertial forces are calculated at a large number of conditions to give an estimate of the maximum loads, and hence stresses, that the structure of the detailed aircraft design will experience in service. A simplistic estimate of the number of analyses required would multiply the numbers of conditions to give 10^7. Even with simplistic models of the aircraft behavior this is an unfeasible number of separate simulations. However, engineering experience is used to identify the most likely critical loads conditions, meaning that approximately 10^5 simulations are required for conventional aircraft configurations. Furthermore these analyses have to be repeated every time that there is an update in the aircraft structure...

Compared to existing approaches for ROMs, our interest will be focused on two axis. On the one hand, we start from the consideration that small, highly non-linear scales are typically concentrated in limited spatial regions of the full simulation domain. So for example, in the flow past a wing, the highly non-linear phenomena take place close to the walls at the scale of a millimeter for computational domains that are of the order of hundreds of meters. In this context our approach is characterized by a multi-scale model where the large scales are described by far field models based on ROMs and the small scales are simulated by high-fidelity models. The whole point for this approach is to optimally decouple the far field from the near field.

A second characterizing feature of our ROM approach is non-linear interpolation. We start from the consideration that dynamical models derived from the projection of the PDE model in the reduced space are neither stable to numerical integration nor robust to parameter variation when hard non-linear multi-scale phenomena are considered.

However, thanks to Proper Orthogonal Decomposition (POD) we can accurately approximate large solution databases using a small base. Recent techniques to investigate the temporal evolution of the POD modes (Koopman modes, Dynamic Mode Decomposition) allow a dynamic discrimination of the role played by each of them. This in turn can be exploited to interpolate between the modes in parameter space, thanks to ideas relying on optimal transportation that we have started developing in the FP7 project FFAST. In the following we precise these ideas on a specific example.

<p style="text-align:center;color:red;font-weight:bold;">MNEMOSYNE Project-Team</p>

3. Research Program

3.1. Integrative and Cognitive Neuroscience

The human brain is often considered as the most complex system dedicated to information processing. This multi-scale complexity, described from the metabolic to the network level, is particularly studied in integrative neuroscience, the goal of which is to explain how cognitive functions (ranging from sensorimotor coordination to executive functions) emerge from (are the result of the interaction of) distributed and adaptive computations of processing units, displayed along neural structures and information flows. Indeed, beyond the astounding complexity reported in physiological studies, integrative neuroscience aims at extracting, in simplifying models, regularities at various levels of description. From a mesoscopic point of view, most neuronal structures (and particularly some of primary importance like the cortex, cerebellum, striatum, hippocampus) can be described through a regular organization of information flows and homogenous learning rules, whatever the nature of the processed information. From a macroscopic point of view, the arrangement in space of neuronal structures within the cerebral architecture also obeys a functional logic, the sketch of which is captured in models describing the main information flows in the brain, the corresponding loops built in interaction with the external and internal (bodily and hormonal) world and the developmental steps leading to the acquisition of elementary sensorimotor skills up to the most complex executive functions.

In summary, integrative neuroscience builds, on an overwhelming quantity of data, a simplifying and interpretative grid suggesting homogenous local computations and a structured and logical plan for the development of cognitive functions. They arise from interactions and information exchange between neuronal structures and the external and internal world and also within the network of structures.

This domain is today very active and stimulating because it proposes, of course at the price of simplifications, global views of cerebral functioning and more local hypotheses on the role of subsets of neuronal structures in cognition. In the global approaches, the integration of data from experimental psychology and clinical studies leads to an overview of the brain as a set of interacting memories, each devoted to a specific kind of information processing [61]. It results also in longstanding and very ambitious studies for the design of cognitive architectures aiming at embracing the whole cognition. With the notable exception of works initiated by [57], most of these frameworks (e.g. Soar, ACT-R), though sometimes justified on biological grounds, do not go up to a *connectionist* neuronal implementation. Furthermore, because of the complexity of the resulting frameworks, they are restricted to simple symbolic interfaces with the internal and external world and to (relatively) small-sized internal structures. Our main research objective is undoubtly to build such a general purpose cognitive architecture (to model the brain *as a whole* in a systemic way), using a connectionist implementation and able to cope with a realistic environment.

3.2. Computational Neuroscience

From a general point of view, computational neuroscience can be defined as the development of methods from computer science and applied mathematics, to explore more technically and theoretically the relations between structures and functions in the brain [63], [50]. During the recent years this domain has gained an increasing interest in neuroscience and has become an essential tool for scientific developments in most fields in neuroscience, from the molecule to the system. In this view, all the objectives of our team can be described as possible progresses in computational neuroscience. Accordingly, it can be underlined that the systemic view that we promote can offer original contributions in the sense that, whereas most classical models in computational neuroscience focus on the better understanding of the structure/function relationship for isolated specific structures, we aim at exploring synergies between structures. Consequently, we target interfaces and interplay between heterogenous modes of computing, which is rarely addressed in classical computational neuroscience.

We also insist on another aspect of computational neuroscience which is, in our opinion, at the core of the involvement of computer scientists and mathematicians in the domain and on which we think we could particularly contribute. Indeed, we think that our primary abilities in numerical sciences imply that our developments are characterized above all by the effectiveness of the corresponding computations: We provide biologically inspired architectures with effective computational properties, such as robustness to noise, self-organization, on-line learning. We more generally underline the requirement that our models must also mimick biology through its most general law of homeostasis and self-adaptability in an unknown and changing environment. This means that we propose to numerically experiment such models and thus provide effective methods to falsify them.

Here, computational neuroscience means mimicking original computations made by the neuronal substratum and mastering their corresponding properties: computations are distributed and adaptive; they are performed without an homonculus or any central clock. Numerical schemes developed for distributed dynamical systems and algorithms elaborated for distributed computations are of central interest here [47], [56] and were the basis for several contributions in our group [62], [59], [64]. Ensuring such a rigor in the computations associated to our systemic and large scale approach is of central importance.

Equally important is the choice for the formalism of computation, extensively discussed in the connectionist domain. Spiking neurons are today widely recognized of central interest to study synchronization mechanisms and neuronal coupling at the microscopic level [48]; the associated formalism [53] can be possibly considered for local studies or for relating our results with this important domain in connectionism. Nevertheless, we remain mainly at the mesoscopic level of modeling, the level of the neuronal population, and consequently interested in the formalism developed for dynamic neural fields [45], that demonstrated a richness of behavior [49] adapted to the kind of phenomena we wish to manipulate at this level of description. Our group has a long experience in the study and adaptation of the properties of neural fields [59], [60] and their use for observing the emergence of typical cortical properties [52]. In the envisioned development of more complex architectures and interplay between structures, the exploration of mathematical properties such as stability and boundedness and the observation of emerging phenomena is one important objective. This objective is also associated with that of capitalizing our experience and promoting good practices in our software production (*cf.* § 6.1). In summary, we think that this systemic approach also brings to computational neuroscience new case studies where heterogenous and adaptive models with various time scales and parameters have to be considered jointly to obtain a mastered substratum of computation. This is particularly critical for large scale deployments, as we will discuss in § 6.1).

3.3. Machine Learning

The adaptive properties of the nervous system are certainly among its most fascinating characteristics, with a high impact on our cognitive functions. Accordingly, machine learning is a domain [55] that aims at giving such characteristics to artificial systems, using a mathematical framework (probabilities, statistics, data analysis, etc.). Some of its most famous algorithms are directly inspired from neuroscience, at different levels. Connectionist learning algorithms implement, in various neuronal architectures, weight update rules, generally derived from the hebbian rule, performing non supervised (e.g. Kohonen self-organizing maps), supervised (e.g. layered perceptrons) or associative (e.g. Hopfield recurrent network) learning. Other algorithms, not necessarily connectionist, perform other kinds of learning, like reinforcement learning. Machine learning is a very mature domain today and all these algorithms have been extensively studied, at both the theoretical and practical levels, with much success. They have also been related to many functions (in the living and artificial domains) like discrimination, categorisation, sensorimotor coordination, planning, etc. and several neuronal structures have been proposed as the substratum for these kinds of learning [51], [44]. Nevertheless, we believe that, as for previous models, machine learning algorithms remain isolated tools, whereas our systemic approach can bring original views on these problems.

At the cognitive level, most of the problems we face do not rely on only one kind of learning and require instead skills that have to be learned in preliminary steps. That is the reason why cognitive architectures are often referred to as systems of memory, communicating and sharing information for problem solving. Instead

of the classical view in machine learning of a flat architecture, a more complex network of modules must be considered here, as it is the case in the domain of deep learning. In addition, our systemic approach brings the question of incrementally building such a system, with a clear inspiration from developmental sciences. In this perspective, modules can generate internal signals corresponding to internal goals, predictions, error signals, able to supervise the learning of other modules (possibly endowed with a different learning rule), supposed to become autonomous after an instructing period. A typical example is that of episodic learning (in the hippocampus), storing declarative memory about a collection of past episods and supervising the training of a procedural memory in the cortex.

At the behavioral level, as mentionned above, our systemic approach underlines the fundamental links between the adaptive system and the internal and external world. The internal world includes proprioception and interoception, giving information about the body and its needs for integrity and other fundamental programs. The external world includes physical laws that have to be learned and possibly intelligent agents for more complex interactions. Both involve sensors and actuators that are the interfaces with these worlds and close the loops. Within this rich picture, machine learning generally selects one situation that defines useful sensors and actuators and a corpus with properly segmented data and time, and builds a specific architecture and its corresponding criteria to be satisfied. In our approach however, the first question to be raised is to discover what is the goal, where attention must be focused on and which previous skills must be exploited, with the help of a dynamic architecture and possibly other partners. In this domain, the behavioral and the developmental sciences, observing how and along which stages an agent learns, are of great help to bring some structure to this high dimensional problem.

At the implementation level, this analysis opens many fundamental challenges, hardly considered in machine learning : stability must be preserved despite on-line continuous learning; criteria to be satisfied often refer to behavioral and global measurements but they must be translated to control the local circuit level; in an incremental or developmental approach, how will the development of new functions preserve the integrity and stability of others? In addition, this continous re-arrangement is supposed to involve several kinds of learning, at different time scales (from msec to years in humans) and to interfer with other phenomena like variability and meta-plasticity.

In summary, our main objective in machine learning is to propose on-line learning systems, where several modes of learning have to collaborate and where the protocoles of training are realistic. We promote here a *really autonomous* learning, where the agent must select by itself internal resources (and build them if not available) to evolve at the best in an unknown world, without the help of any *deus-ex-machina* to define parameters, build corpus and define training sessions, as it is generally the case in machine learning. To that end, autonomous robotics (*cf.* § 3.4) is a perfect testbed.

3.4. Autonomous Robotics

Autonomous robots are not only convenient platforms to implement our algorithms; the choice of such platforms is also motivated by theories in cognitive science and neuroscience indicating that cognition emerges from interactions of the body in direct loops with the world (*embodiment of cognition* [46]). In addition to real robotic platforms, software implementations of autonomous robotic systems including components dedicated to their body and their environment will be also possibly exploited, considering that they are also a tool for studying conditions for a real autonomous learning.

A real autonomy can be obtained only if the robot is able to define its goal by itself, without the specification of any high level and abstract cost function or rewarding state. To ensure such a capability, we propose to endow the robot with an artificial physiology, corresponding to perceive some kind of pain and pleasure. It may consequently discriminate internal and external goals (or situations to be avoided). This will mimick circuits related to fundamental needs (e.g. hunger and thirst) and to the preservation of bodily integrity. An important objective is to show that more abstract planning capabilities can arise from these basic goals.

A real autonomy with an on-line continuous learning as described in § 3.3 will be made possible by the elaboration of protocols of learning, as it is the case, in animal conditioning, for experimental studies

where performance on a task can be obtained only after a shaping in increasingly complex tasks. Similarly, developmental sciences can teach us about the ordered elaboration of skills and their association in more complex schemes. An important challenge here is to translate these hints at the level of the cerebral architecture.

As a whole, autonomous robotics permits to assess the consistency of our models in realistic condition of use and offers to our colleagues in behavioral sciences an object of study and comparison, regarding behavioral dynamics emerging from interactions with the environment, also observable at the neuronal level.

In summary, our main contribution in autonomous robotics is to make autonomy possible, by various means corresponding to endow robots with an artificial physiology, to give instructions in a natural and incremental way and to prioritize the synergy between reactive and robust schemes over complex planning structures.

3. Research Program

3.1. Introduction

We address the problem of cancer modeling through 3 axis.

- <u>Axis 1</u>: Tumor modeling for patient-specific simulations.
- <u>Axis 2</u>: Bio-physical modeling for personalized therapies.
- <u>Axis 3</u>: Quantitative cancer modeling for biological and preclinical studies.

In the first axis, we aim at producing patient-specific simulations of the growth of a tumor or its response to treatment starting from a series of images. We hope to be able to give information to the clinicians in order to improve the decision process. It is mainly useful in the case of a relapse or for metastatic diseases.

The second axis aims at modeling the biophysical therapies like radiotherapies, but also thermo-ablations, radio-frequency ablations or electroporation that play a crucial role in the case of a relapse or for a metastatic disease, which is precisely the clinical context where the techniques of axis 1 will be applied.

The third axis, even if not directly linked to clinical perspectives, is essential since it is a way to better understand and model the biological reality of cancer growth and the (possibly complex) effects of therapeutic intervention. Modeling in this case also helps to interpret the experimental results and improve the accuracy of the models used in Axis 1. Technically speaking, some of the computing tools are similar to those of Axis 1.

3.2. Axis 1: Tumor modeling for patient-specific simulations

The gold standard treatment for most cancers is surgery. In the case where total resection of the tumor is possible, the patient often benefits from an adjuvant therapy (radiotherapy, chemotherapy, targeted therapy or a combination of them) in order to eliminate the potentially remaining cells that are not visible. In this case personalized modeling of tumor growth is useless and statistical modeling will be able to quantify the risk of relapse, the mean progression-free survival time...However if total resection is not possible or if metastases emerge from distant sites, clinicians will try to adopt a strategy in order to control the disease for as long as possible. A wide set of tools are available. Clinicians may treat the disease by physical interventions (radiofrequency ablation, cryoablation, radiotherapy, electroporation, focalized ultrasound,...) or chemical agents (chemotherapies, targeted therapies, antiangiogenic drugs, immunotherapies, hormonotherapies). One can also decide to follow the patient without any treatment (this is the case for slowly growing tumors like some metastases to the lung, some lymphomas or for some low grade gliomas). If we had a reliable patient specific model of tumor growth with or without treatment, it could have different uses for the patient follow-up.

- <u>Case without treatment</u>: the evaluation of the growth of the tumor would provide a useful indication for the time at which the tumor will reach a critical size. For example, radiofrequency ablation of pulmonary lesion is very efficient as long as the diameter of the lesion is smaller than 3 cm. Thus, the prediction can help the clinician for the planification of the intervention. For tumor with very slow growth, quantitative modeling can also help to decide at what time interval the patient has to undergo a CT-scan. CT-scans are irradiative exams and there is a challenge for decreasing their occurrence for each patient. It has also an economical impact. And if the disease evolution starts to differ from the forecast, this can mean that some events have occurred at the biological level. It can be the apparition of an agressive phenotype, cells that leave a dormancy state. This kind of events cannot be predicted, but some mismatch with respect to the prediction can be an indirect proof of their existence. It could be an indication for the clinician to start a treatment.

- **Case with treatment:** a model can help to understand and to quantify the final effect of a treatment using the early response. It can help for a redefinition of the treatment planning. Modeling can also help to anticipate the relapse by analyzing some functional aspects of the tumor. Again, a deviation with respect to reference curves can mean a lack of efficiency of the therapy or a relapse. Moreover, for a long time, the response to a treatment has been quantified by the RECIST criteria which consists in (roughly speaking) measuring the diameters of the largest tumor of the patient, as it is seen on a CT-scan. This criteria is still widely used and was quite efficient for chemotherapies and radiotherapies that induce a decrease of the size of the lesion. However, with the systematic use of targeted therapies and anti-angiogenic drugs that modify the physiology of the tumor, the size may remain unchanged even if the drug is efficient and deeply modifies the tumor behavior. One better way to estimate this effect could be to use functional imaging (Pet-scan, perfusion or diffusion MRI, ...), a model can then be used to exploit the data and to understand in what extent the therapy is efficient.

- **Optimization:** currently, we do not believe that we can optimize a particular treatment in terms of distribution of doses, number, planning with the model that we will develop in a medium term perspective. But it is an aspect that we keep in mind on a long term one.

The scientific challenge is therefore as follows: knowing the history of the patient, the nature of the primitive tumor, its histopathology, knowing the treatments that patients have undergone, knowing some biological facts on the tumor and having a sequence of images (CT-scan, MRI, PET or a mix of them), are we able to provide a numerical simulation of the extension of the tumor and of its metabolism that fits as best as possible with the data (CT-scans or functional data) and that is predictive in order to address the clinical cases described above?

Our approach relies on the elaboration of PDE models and their parametrization with the image by a coupling of gradient methods and Monte-Carlo type methods. The PDE models rely on the description of the dynamics of cell populations. The number of populations depends on the pathology. For example, for glioblastoma, one needs to use proliferative cells, invasive cells, quiescent cells as well as necrotic tissues to be able to reproduce realistic behaviors of the disease. In order to describe the relapse for hepatic metastases of gastro-intestinal stromal tumor (gist), one needs three cell populations: proliferative cells, healthy tissue and necrotic tissue. The law of proliferation is often coupled with a model for the angiogenesis. However such models of angiogenesis involve too many non measurable parameters to be used with real clinical data and therefore one has to use simplified or even simplistic versions. The law of proliferation often mimics the existence of an hypoxia threshold, it consists of an O.D.E. or a P.D.E that describes the evolution of the growth rate as a combination of sigmoidal functions of nutrients or roughly speaking oxygen concentration. Usually, several laws are available for a given pathology since at this level, there are no quantitative argument to choose a particular one. The velocity of the tumor growth differs depending on the nature of the tumor. For metastases, we will derive the velocity thanks to Darcy's law in order to express that the extension of the tumor is basically due to the increase of volume. This gives a sharp interface between the metastasis and the surrounding healthy tissues, as observed by anatomopathologists. For primitive tumors like gliomas or lung cancer, we use reaction-diffusion equations in order to describe the invasive aspects of such primitive tumors. The modeling of the drugs depends on the nature of the drug: for chemotherapies, a death term can be added into the equations of the population of cells, while antiangiogenic drugs have to be introduced in a angiogenic model. Resistance to treatment can be described either by several populations of cells or with non-constant growth or death rates. As said before, it is still currently difficult to model the changes of phenotype or mutations, we therefore propose to investigate this kind of phenomena by looking at deviations of the numerical simulations compared to the medical observations. The calibration of the model is done by using a series (at least 2) of images of the same patient and by minimizing a cost function. The cost function contains at least the difference between the volume of the tumor that is measured on the images with the computed one. It also contains elements on the geometry, on the necrosis and any information that can be obtained through the medical images. We will pay special attention to functional imaging (PET, perfusion and diffusion MRI). The inverse problem is solved using a gradient method coupled with some Monte-Carlo type algorithm. If a large number of similar cases is available, one can imagine to use statistical algorithms like random forests to use some non quantitative data like the gender, the age, the origin of the primitive tumor...for example for choosing the model for the

growth rate for a patient using this population knowledge (and then to fully adapt the model to the patient by calibrating this particular model on patient data) or for having a better initial estimation of the modeling parameters. We have obtained several preliminary results concerning lung metastases including treatments and for metastases to the liver.

3.3. Axis 2: Bio-physical modeling for personalized therapies

In this axis, we investigate locoregional therapies such as radiotherapy, irreversible electroporation. Electroporation consists of an increase of the membrane permeability of cells due to the delivery of high voltage pulses. This phenomenon can be transient (reversible) or irreversible. This is a non-thermal phenomenon. (IRE) or electro-chemotherapy – which is a combination of reversible electroporation with a cytotoxic drug – are essential tools for the treatment of a metastatic disease. Numerical modeling of these therapies is a clear scientific challenge. Clinical applications of the modeling are the main target, which thus drives the scientific approach, even though theoretical studies in order to improve the knowledge of the biological phenomena, in particular for electroporation, should also be addressed. However, this subject is quite wide and we will focus on two particular approaches: some aspects of radiotherapies and electro-chemotherapy. This choice is motivated by some pragmatic reasons: we already have collaborations with physicians on these therapies. Other treatments could be probably tackled in the same spirit, but we do not plan to work on this subject on a medium term.

- Radiotherapy (RT) is a common therapy for cancer. Typically, using a CT scan of the patient with the structures of interest (tumor, organs at risk) delineated, the clinicians optimize the dose delivery to treat the tumor while preserving the healthy tissue. The RT is then delivered every day using low resolution scans (CBCT) to position the beams. Under treatment the patient may lose weight and the tumor shrinks. These changes may affect the propagation of the beams and subsequently change the dose that is effectively delivered. It could be harmful for the patient especially if sensitive organs are concerned. In such cases, a replanification of the RT could be done to adjust the therapeutical protocol. Unfortunately, this process takes too much time to be performed routinely. The challenges faced by clinicians are numerous, we focus on two of them:

  - *Detecting the need of replanification:* we are using the positioning scans to evaluate the movement and deformation of the various structures of interest. Thus we can detect whether or not a structure has moved out of the safe margins (fixed by clinicians) and thus if a replanification may be necessary. In a retrospective study, our work can also be used to determine RT margins when there are no standard ones. A collaboration with the RT department of Institut Bergonié is underway on the treatment of retroperitoneal sarcoma and ENT tumors (head and neck cancers). A retrospective study was performed on 11 patients with retro-peritoneal sarcoma. The results have shown that the safety margins (on the RT) that clinicians are currently using are probably not large enough. The tool used in this study is being further developed by an engineer funded by Inria (Cynthia Périer, ADT Sesar). We used well validated methods from a level-set approach and segmentation / registration methods. The originality and difficulty lie in the fact that we are dealing with real data in a clinical setup. Clinicians have currently no way to perform complex measurements with their clinical tools. This prevents them from investigating the replanification. Our work and the tools developed pave the way for easier studies on evaluation of RT plans in collaboration with Institut Bergonié. *There was no modeling involved in this work that arose during discussions with our collaborators.* The main purpose of the team is to have meaningful outcomes of our research for clinicians, sometimes it implies leaving a bit our area of expertise.

  - *Evaluating RT efficacy and finding correlation between the radiological responses and the clinical outcome:* our goal is to help doctors to identify correlation between the response to RT (as seen on images) and the longer term clinical outcome of the patient. Typically, we aim at helping them to decide when to plan the next exam after the RT. For patients whose response has been linked to worse prognosis, this exam would have to be planned

earlier. This is the subject of a starting collaboration with Institut Bergonié. The response is evaluated from image markers (*e.g.* using texture information) or with a mathematical model (another collaboration is also ongoing with LATIM team in Brest on response of colorectal tumors to RT using PET scans). The other challenges are either out of reach or not in the domain of expertise of the team. Yet our works may tackle some important issues for adaptive radiotherapy.

- Both IRE and electrochemotherapy are anticancerous treatments based on the same phenomenon: the electroporation of cell membranes. This phenomenon is known since a few decades but it is still not well understood, therefore we address the modeling two different purposes:

  1. We want to use mathematical models in order to better understand the biological behavior and the effect of the treatment. We work in tight collaboration with biologists and bioeletromagneticians to derive precise models of cell and tissue electroporation, in the continuity of the research program of the Inria team-project MC2. These studies lead to complex non-linear mathematical models involving some parameters (as less as possible). Numerical methods to compute precisely such models and the calibration of the parameters with the experimental data are then addressed. Tight collaborations with the Vectorology and Anticancerous Therapies (VAT) of IGR at Villejuif, Laboratoire Ampère of Ecole Centrale Lyon and the Karlsruhe Institute of technology will continue, and we aim at developing new collaborations with Institute of Pharmacology and Structural Biology (IPBS) of Toulouse and the Laboratory of Molecular Pathology and Experimental Oncology (LM-PEO) at CNR Rome, in order to understand differences of the electroporation of healthy cells and cancer cells in spheroids and tissues.

  2. This basic research aims at providing new understanding of electroporation, however it is necessary to address, particular questions raised by radio-oncologists that apply such treatments. One crucial question is "What pulse or what train of pulses should I apply to electroporate the tumor if the electrodes are located as given by the medical images"? Even if the real-time optimization of the placement of the electrodes for deep tumors may seem quite utopian since the clinicians face too many medical constraints that cannot be taken into account (like the position of some organs, arteries, nerves...), on can expect to produce real-time information of the validity of the placement done by the clinician. Indeed, once the placement is performed by the radiologists, medical images are usually used to visualize the localization of the electrodes. Using these medical data, a crucial goal is to provide a tool in order to compute in real-time and visualize the electric field and the electroporated region directly on theses medical images, to give the doctors a precise knowledge of the region affected by the electric field. In the long run, this research will benefit from the knowledge of the theoretical electroporation modeling, but it seems important to use the current knowledge of tissue electroporation – even quite rough –, in order to rapidly address the specific difficulty of such a goal (real-time computing of non-linear model, image segmentation and visualization). Tight collaborations with CHU Pellegrin at Bordeaux, and CHU J. Verdier at Bondy are crucial.

3.4. Axis 3: Quantitative cancer modeling for biological and preclinical studies

With the emergence and improvement of a plethora of experimental techniques, the molecular, cellular and tissue biology has operated a shift toward a more quantitative science, in particular in the domain of cancer biology. These quantitative assays generate a large amount of data that call for theoretical formalism in order to better understand and predict the complex phenomena involved. Indeed, due to the huge complexity underlying the development of a cancer disease that involves multiple scales (from the genetic, intra-cellular scale to the scale of the whole organism), and a large number of interacting physiological processes (see the so-called "hallmarks of cancer"), several questions are not fully understood. Among these, we want to focus on the most clinically relevant ones, such as the general laws governing tumor growth and the development of metastases

(secondary tumors, responsible of 90% of the deaths from a solid cancer). In this context, it is thus challenging to potentiate the diversity of the data available in experimental settings (such as *in vitro* tumor spheroids or *in vivo* mice experiments) in order to improve our understanding of the disease and its dynamics, which in turn lead to validation, refinement and better tuning of the macroscopic models used in the axes 1 and 2 for clinical applications.

In recent years, several new findings challenged the classical vision of the metastatic development biology, in particular by the discovery of organism-scale phenomena that are amenable to a dynamical description in terms of mathematical models based on differential equations. These include the angiogenesis-mediated distant inhibition of secondary tumors by a primary tumor the pre-metastatic niche or the self-seeding phenomenon Building a general, cancer type specific, comprehensive theory that would integrate these dynamical processes remains an open challenge. On the therapeutic side, recent studies demonstrated that some drugs (such as the Sunitinib), while having a positive effect on the primary tumor (reduction of the growth), could *accelerate* the growth of the metastases. Moreover, this effect was found to be scheduling-dependent. Designing better ways to use this drug in order to control these phenomena is another challenge. In the context of combination therapies, the question of the *sequence* of administration between the two drugs is also particularly relevant.

One of the recurrent technical challenge that we need to address when dealing with biological data is the presence of potentially very large inter-animal (or inter-individual) variability.

Starting from the available multi-modal data and relevant biological or therapeutic questions, our purpose is to develop adapted mathematical models (i.e., identifiable from the data) that recapitulate the existing knowledge and reduce it to its more fundamental components, with two main purposes:

1. to generate quantitative and empirically testable predictions that allow to assess biological hypotheses or

2. to investigate the therapeutic management of the disease and assist preclinical studies of anticancerous drug development.

We believe that the iterative loop between theoretical modeling and experimental studies can help to generate new knowledge and improve our predictive abilities for clinical diagnosis, prognosis, and therapeutic decision. Let us note that the first point is in direct link with the axes 1 and 2 of the team since it allows us to experimentally validate the models at the biological scale (*in vitro* and *in vivo* experiments) for further clinical applications.

More precisely, we first base ourselves on a thorough exploration of the biological literature of the biological phenomena we want to model: growth of tumor spheroids, *in vivo* tumor growth in mice, initiation and development of the metastases, effect of anti-cancerous drugs. Then we investigate, using basic statistical tools, the data we dispose, which can range from: spatial distribution of heterogeneous cell population within tumor spheroids, expression of cell makers (such as green fluorescent protein for cancer cells or specific antibodies for other cell types), bioluminescence, direct volume measurement or even intra-vital images obtained with specific imaging devices. According to the data type, we further build dedicated mathematical models that are based either on PDEs (when spatial data is available, or when time evolution of a structured density can be inferred from the data, for instance for a population of tumors) or ODEs (for scalar longitudinal data). These models are confronted to the data by two principal means:

1. when possible, experimental assays can give a direct measurement of some parameters (such as the proliferation rate or the migration speed) or

2. statistical tools to infer the parameters from observables of the model.

This last point is of particular relevance to tackle the problem of the large inter-animal variability and we use adapted statistical tools such as the mixed-effects modeling framework.

Once the models are shown able to describe the data and are properly calibrated, we use them to test or simulate biological hypotheses. Based on our simulations, we then aim at proposing to our biological collaborators new experiments to confirm or infirm newly generated hypotheses, or to test different administration protocols of the drugs. For instance, in a collaboration with the team of the professor Andreas Bikfalvi (Laboratoire

de l'Angiogénèse et du Micro-environnement des Cancers, Inserm, Bordeaux), based on confrontation of a mathematical model to multi-modal biological data (total number of cells in the primary and distant sites and MRI), we could demonstrate that the classical view of metastatic dissemination and development (one metastasis is born from one cell) was probably inaccurate, in mice grafted with metastatic kidney tumors. We then proposed that metastatic germs could merge or attract circulating cells. Experiments involving cells tagged with two different colors are currently performed in order to confirm or infirm this hypothesis.

Eventually, we use the large amount of temporal data generated in preclinical experiments for the effect of anti-cancerous drugs in order to design and validate mathematical formalisms translating the biological mechanisms of action of these drugs for application to clinical cases, in direct connection with the axis 1. We have a special focus on targeted therapies (designed to specifically attack the cancer cells while sparing the healthy tissue) such as the Sunitinib. This drug is indeed indicated as a first line treatment for metastatic renal cancer and we plan to conduct a translational study coupled between A. Bikfalvi's laboratory and medical doctors, F. Cornelis (radiologist) and A. Ravaud (head of the medical oncology department).

<p style="text-align:center; color:red; font-weight:bold;">PHOENIX Project-Team</p>

3. Research Program

3.1. Design-Driven Software Development

Raising the level of abstraction beyond programming is a very active research topic involving a range of areas, including software engineering, programming languages and formal verification. The challenge is to allow design dimensions of a software system, both functional and non-functional, to be expressed in a high-level way, instead of being encoded with a programming language. Such design dimensions can then be leveraged to verify conformance properties and to generate programming support.

Our research on this topic is to take up this challenge with an approach inspired by programming languages, introducing a full-fledged language for designing software systems and processing design descriptions both for verification and code generation purposes. Our approach is also DSL-inspired in that it defines a conceptual framework to guide software development. Lastly, to make our approach practical to software developers, we introduce a methodology and a suite of tools covering the development life-cycle.

To raise the level of abstraction beyond programming, the key approaches are model-driven engineering and architecture description languages. A number of *architecture description languages* have been proposed; they are either (1) coupled with a programming language (*e.g.,* [40]), providing some level of abstraction above programming, or (2) integrated into a programming language (*e.g.,* [33], [41]), mixing levels of abstraction. Furthermore, these approaches poorly leverage architecture descriptions to support programming, they are crudely integrated into existing development environments, or they are solely used for verification purposes. *Model-driven software development* is another actively researched area. This approach often lacks code generation and verification support. Finally, most (if not all) approaches related to our research goal are *general purpose*; their universal nature provides little, if any, guidance to design a software system. This situation is a major impediment to both reasoning about a design artifact and generating programming support.

3.2. Integrating Non-Functional Concerns into Software Design

Most existing design approaches do not address non-functional concerns. When they do, they do not provide an approach to non-functional concerns that covers the entire development life-cycle. Furthermore, they usually are general purpose, impeding the use of non-functional declarations for verification and code generation. For example, the Architecture Analysis & Design Language (AADL) is a standard dedicated to real-time embedded systems [36]. AADL provides language constructs for the specification of software systems (*e.g.,* component, port) and their deployment on execution platforms (*e.g.,* thread, process, memory). Using AADL, designers specify non-functional aspects by adding properties on language constructs (*e.g.,* the period of a thread) or using language extensions such as the Error Model Annex. [0] The software design concepts of AADL are still rather general purpose and give little guidance to the designer.

Beyond offering a conceptual framework, our language-based approach provides an ideal setting to address non-functional properties (*e.g.,* performance, reliability, security, ...). Specifically, a design language can be enriched with non-functional declarations to pursue two goals: (1) expanding further the type of conformance that can be checked between the design of a software system and its implementation, and (2) enabling additional programming support and guidance.

We are investigating this idea by extending our design language with non-functional declarations. For example, we have addressed error handling [10], access conflicts to resources [38], and quality of service constraints [37].

---

[0]The Error Model Annex is a standardized AADL extension for the description of errors [42].

Following our approach to paradigm-oriented software development, non-functional declarations are verified at design time, they generate support that guides and constrains programming, they produce a runtime system that preserves invariants.

3.3. Human-Driven Software Design

Knowledge of the human characteristics (individual, social and organizational) allow the design of complex system and artifacts for increasing their efficacy. In our approach of assistive computing, a main challenge is the integration of facets of Human Factors in order to design technology support adapted to user needs in term of ergonomic properties (acceptability, usability, utility etc) and delivered functionalities (oriented task under user abilities contraints).

We adapt this approach to improve the independent living and self-determination of users with cognitive impairments by developing a variety of orchestration scenarios of networked objects (hardware/software) to provide a pervasive support to their activities. Human factors methodologies are adopted in our approach with the direct purpose the reliability and efficiency of the performance of digital support systems in respect of objectives of health and well-being of the person (monitoring, evaluation, and rehabilitation).

Precisely, our methodologies are based on a closed iterative loop, as described in the figure below :

- Identifying the person needs in a natural situation (*i.e.,* desired but problematic activities) according to Human Factors Models of activity (*i.e.,* environmental constraints; social support networks - caregivers and family; person's abilities)

- Designing environmental support that will assist the users to bypass their cognitive impairment (according to environmental models of cognitive compensatory mechanisms); and then implement this support in terms of technological solutions (scenarios of networked objects, hardware interface, software interface, interaction style, *etc*)

- Empirically evaluating the assistive solution based on human experimentations that includes ergonomic assessments (acceptability, usability, usefulness, *etc*) as well as longitudinal evaluations of use's efficacy in terms of activities performed by the individual, of satisfaction and well-being provided to the individual but also to his/her entourage (family and caregivers).
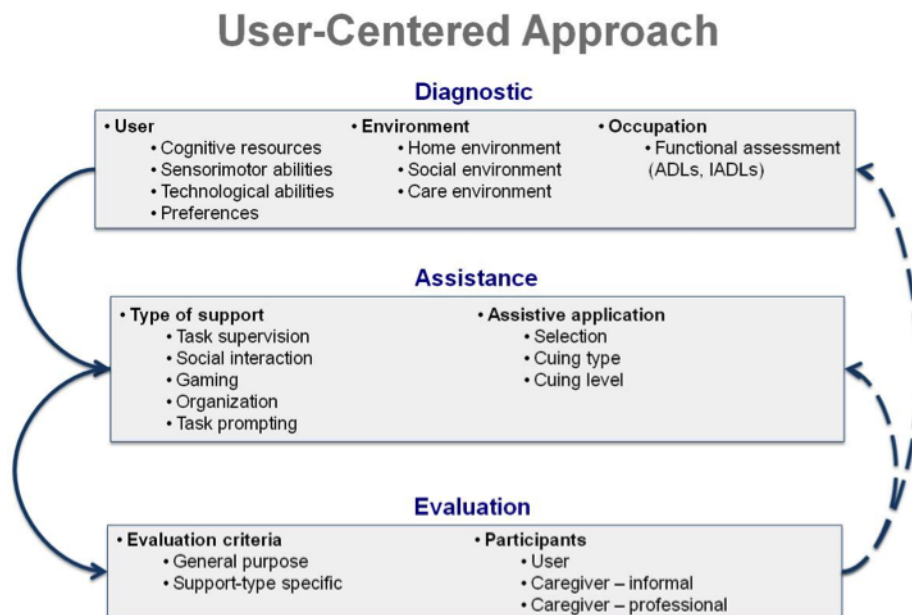
Figure 1. User-Centered Approach

PLEIADE Team

3. Research Program

3.1. Distances and pattern recognition

Diversity may be understood as a set of dissimilarities between objects. The underlying mathematical construction is the notion of distance. Knowing a set of objects, on the condition that pairwise distances can be measured, it is possible to build a Euclidan image of it as a point cloud in a space of relevant dimension. Then, diversity can be associated with the shape of the point cloud. It is still true that the reference for recognizing patterns or shapes is the human eye. One objective of our project is to narrow the gap between the story that a human eye can read, and the story that an algorithm can tell. Several directions will be explored. First, it is necessary to master dimension reduction, mainly classical algebraic tools (PCA, NGS, Isomap, eigenmaps, etc ...), and collaborate with experts in efficient methods in spectral methods. Second, a neighborhood in a point cloud naturally leads to graphs describing the neighborhood networks. There is a natural link between modular structures in distance arrays and communities on graphs. Third, points defined by DNA sequences (for example) are samples of diversity. Dimension reduction may show that they live on a given manifold. This leads to geometry (differential or Riemanian geometry). Knowing some properties of the manifold can inform us about the constraints on the space where the measured individuals live. The connection between Riemannian geometry and graphs, where weighted graphs are seen as meshes embedded in a manifold, is currently an active field of reasearch [23], [22].

To resolve these objectives computationally will require investment in research directions in computational geometry (such as convex hulls of high-dimension sets of points), on circumventing the curse of dimensionality, and on linking distance geometry with convex optimization procedures through matrix completion. None of these questions is trivial: most recent work has focused on two or three dimensions, for example for image analysis or for reconstruction of protein conformation from local distances between atoms. The methodological goal is to extend these approaches to higher dimension spaces.

3.2. Modeling by successive refinement

Describing the links between diversity in traits and diversity in function will require comprehensive models, assembled from and refining existing models. A recurring difficulty in building comprehensive models of biological systems is that accurate models for subsystems are built using different formalisms and simulation techniques, and hand-tuned models tend to be so focused in scope that it is difficult to repurpose them [15]. Our belief is that a sustainable effort in building efficient behavioral models must proceed incrementally, rather than by modeling individual processes *de novo*. *Hierarchical modeling* [11] is one way of combining specific models into networks. Effective use of hierarchical models requires both formal definition of the semantics of such composition, and efficient simulation tools for exploring the large space of complex behaviors. We have previously shown that this approach can be effective for certains kinds of systems in biotechnology [14], [16] and medicine [13]. Our challenge is to adapt incremental, hierarchical refinement to modeling organisms and communities in metagenomic and comparative genomic applications.

<p style="text-align:center"><strong style="color:red">POSET Team</p>

3. Research Program

3.1. Research Program

Our research programs is structured into three complementary research axis : models, languages and systems, allowing us to develop our multi-disciplinary approach while validating each progress in the related specific fields of computer science ranging among computer music, multi-modal system design, reactive and real-time programing, typed functional programming, formal languages, graph representation theory, applied algebra, logic in computer science, etc.

3.1.1. Models

Inverse semigroup theory has recently been shown [15], [24] to unify most string-based, tree-based or even graph-based modeling approaches. It thus provides a consistent and robust mathematical framework to model the sequential, parallel and reactive aspects of temporal media. Developing the mathematical foundations of our proposal amounts to:

- studying the combinatorial and algorithmic properties of the emerging algebra-based model of structured temporal media,
- developing formal techniques and tools for expressing and verifying properties of temporal media programs especially with a view towards capturing temporal media programing by constraint satisfaction approaches,
- deriving from the known generators of these models adequate sets of application-oriented modeling functions.

3.1.2. Languages

Functional programming is the key link between well-defined mathematical structures and their computerized realizations. Based on functional programming frameworks such as Haskell [0], we are prototyping a Domain Specific Language (DSL) [9] dedicated to the programming of interactive temporal media programming. In this research axis, we aim more specifically at

- designing a robust and modular software architecture that allows to reuse existing pieces of software as well as simply combining them together with new ones,
- defining and implementing a DSL for programming interactive multimedia systems via a simple algebra-based high-level and multi-scale control and combination layer,
- finding the right balance between generic views of temporal media when seen as abstract temporal frames and their specializations when representing concrete gestures, sound, audio, videos, animations, etc.

3.1.3. Systems

Multi-modal interactive systems gather various techniques to capture and analyze gestures, and to combine, transform and produce temporal media. Through regular experiments in collaboration with artists, we also aim at assessing, refining and extending the applicability of our proposal by:

- developing a robust and mathematically well-founded representation of systems and of their behaviors, both programmatic and visual,
- developing and evaluating the adequacy of the GUI induced by this representation when used by artists,
- relating the new models with more classical models of music formalisms and, beyond, other temporal media such as animations, videos, etc.

---

[0]See [41] for an historical presentation of the Haskell programming language.
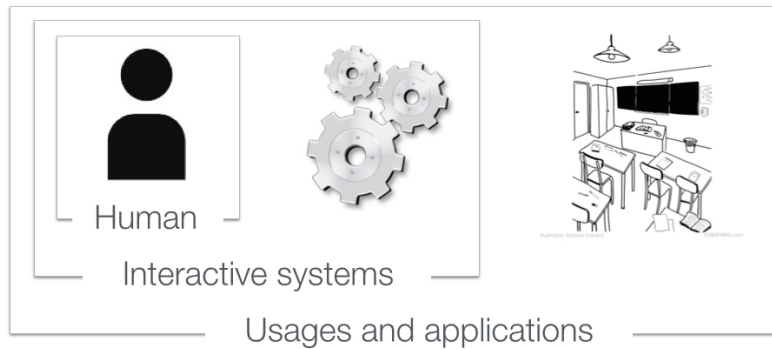
POTIOC Project-Team

3. Research Program

3.1. Introduction

The project of team potioc is oriented along three axes:

- Understanding humans interacting with the digital world
- Creating interactive systems
- Exploring new applications and usages

These axes are depicted in Figure 2 .

Figure 2. The three axes of the potioc team objectives.

Objective 1 is centered on the human sensori-motor and cognitive abilities, as well as user strategies and preferences, for completing interaction tasks. Our target contributions for this objective are a better understanding of humans interacting with interactive systems. The impact of this objective is mainly at a fundamental level.

In objective 2, our goal is to create interactive systems. This may include hardware parts where new input and output modalities are explored. This also includes software parts, that are strongly linked to the underlying hardware components. Our target contribution in objective 2 is to develop (hardware/software) interaction techniques allowing humans to perform interaction tasks.

Finally, in objective 3, we consider interaction at a higher level, taking into account factors that are linked to specific application domains and usages. Our target contribution in this area is the exploration and the emergence of new applications and usages that take benefit from the results of the project. With this objective, we target mainly a societal impact.

Of course, strong links exist between the three objectives of the project. For example, the results obtained in objective 1 guide the development of objective 2. Conversely, new systems developed in objective 2 may feed research questions of objective 1. There are similar links with objective 3.

3.2. Objective 1: Understanding humans interacting with the digital world

Our first objective is centered on the human side. Our finality is not to enhance the general knowledge about the human being as a research team in psychology would do. Instead, we focus on human skills and behaviors during interaction processes. To this end, we conduct experiments that allow us to better understand what users like, where and why they have difficulties. Thanks to these investigations, we are able to design interaction techniques and systems (described in Objective 2) that are well suited to the targeted users. We believe that this fundamental piece of work is the first step that is required for the design of usable popular interactions. We are particularly interested in 3D interaction tasks for which we design dedicated experiments. We also explore a new approach based on physiological and brain (ElectroEncephaloGraphy - EEG) signals for the evaluation of these interactions.

3.2.1. Interacting with physical and virtual environments

Interacting with digital content displayed on 2D screens has been extensively studied in HCI. On the other hand, less conventional contexts have been little studied. This is the case of 3D environments, immersive virtual environments, augmented reality, and tangible objects. With the final goal of making interaction in such contexts user-friendly, we conduct experiments to better understand user strategies and performance. This allows us to propose guidelines to help designers in the creation of tools that are accessible to non-expert users.

3.2.2. Evaluating 3DUIs with physiological signals

Recently, physiological computing has been shown to be a promising complement to Human-Computer Interfaces (HCI) in general, and to 3D User Interfaces (3DUI) in particular, in several directions. Within this research area, we are interested in using various physiological signals, and notably EEG signals, as a new tool to assess objectively the ergonomic quality of a given 3DUI, to identify where and when are the pros and cons of this interface, based on the user's mental state during interaction. For instance, estimating the user's mental workload during interaction can give insights about where and when the interface is cognitively difficult to use. This could be useful for 2D HCI in general, and even more for 3DUI. Indeed, in a 3DUI, the user perception of the 3D scene – part of which could potentially be measured in EEG - is essential. Moreover, the usual need for a mapping between the user inputs and the corresponding actions on 3D objects make 3DUI and interaction techniques more difficult to assess and to design.

3.2.3. Interacting with Brain-Computer Interfaces

Although very promising for numerous applications, BCIs mostly remain prototypes not used outside laboratories, due to their low reliability. Poor BCI performances are partly due to imperfect EEG signal processing algorithms but also to the user who may not be able to produce reliable EEG patterns. Indeed, BCI use is a skill, requiring the user to be properly trained to achieve BCI control. If he/she cannot perform the desired mental commands, no signal processing algorithm can identify them. Therefore, rather than improving EEG signal processing alone, an interesting research direction is to also guide users to learn BCI control mastery. We aim at addressing this objective. We are notably exploring theoretical models and guidelines from educational sciences to improve BCI training protocols. We also study which users' profiles (personality and cognitive profile) fail or succeed at learning BCI control. Finally, we explore new feedback types and new EEG visualization techniques in order to help users gain BCI control skills more efficiently. These new feedback and visualizations notably aim at providing BCI users with more information about their EEG patterns, in order to identify more easily relevant BCI control strategies, as well as motivating and engaging them in the learning task.

3.3. Objective 2: Creating interactive systems

Our objective here is to create interactive systems and design interaction techniques dedicated to the completion of interaction tasks. We divide our work into three main categories:
- Interaction techniques based on existing Input/Output (IO) devices.
- New IO and related techniques.
- BCI and physiological computing.

3.3.1. Interaction techniques based on existing Input/Output (IO) devices

When using desktop IOs (i.e., based on mouse/keyboards/monitors), a big challenge is to design interaction techniques that allow users to complete 3D interaction tasks. Indeed, the desktop IO space that is mainly dedicated to the completion of 2D interaction task is not well suited to 3D content and, consequently, 3D user interfaces need to be designed with a great care. In the past few years, we have been particularly interested in the problem of interaction when the 3D content is displayed on a touchscreen. Indeed, standard (2D) HCI has evolved from mouse to touch input, and numerous research projects have been conducted. On the contrary, in 3D, very little work has been proposed. We are contributing to moving desktop 3D UIs from the mouse to the touch paradigm; what we used to do with mice in front of a screen does not work well on touch devices anymore. In the future, we will continue designing new interaction techniques that are based on standard IOs (eg. pointing devices and webcams) and that target the main objectives of Potioc which are to enhance the interaction bandwidth for non expert users.

3.3.2. New IO and related techniques

Beyond standard IOs, we are interested in exploring new IO modalities that may make interaction easier, more engaging and motivating. In Potioc, we design new interactive systems that exploit unconventional IO modalities such as stereoscopy, 3D spatial input, augmented reality and so on. In particular, tangible interaction and spatial augmented reality are major subjects of interest for us. Indeed, we believe that manipulating directly physical objects for interacting with the digital world has a great potential, in particular when the general public is targeted. With such approaches, the computer disappears, and the user interacts with the digital content as he or she would do with physical content, which reduces the distance to the manipulated content. As an example, we recently designed Teegi, a new system based on a unique combination of spatial augmented reality, tangible interaction and real-time neurotechnologies. With Teegi, a user can visualize and analyze his or her own brain activity in real-time, on a tangible character that can be easily manipulated, and with which it is possible to interact. Such unconventionnal user interfaces that are based on rich sensing modalities hold great promises in the field of popular interaction.

3.3.3. BCI and physiological computing

Although Brain-Computer Interfaces (BCI) have demonstrated their tremendous potential in numerous applications, they are still mostly prototypes, not used outside laboratories. This is mainly due to the following limitations:

- Performances: the poor classification accuracies of BCIs make them inconvenient to use or simply useless compared to available alternatives
- Stability and robustness: the sensibility of ElectroEncephaloGraphic (EEG) signals to noise and their inherent non-stationarity make the already poor initial performances difficult to maintain over time
- Calibration time: the need to tune current BCIs to each user's EEG signals makes their calibration times too long.

As part of our research on EEG-based BCIs, we notably aim at addressing these limitations by designing robust EEG signal processing tools with minimal calibration times, in order to design practical BCI systems, usable and useful outside laboratories. To do so we explore the design of alternative features and robust spatial filtering algorithms to make BCIs more robust to noise and non-stationarities, as well as more accurate. We also explore artificial EEG data generation and user-to-user data transfer to reduce calibration times.

3.4. Objective 3: Exploring new applications and usages

Objective 3 is centered on the applications and usages. Beyond the human sensori-motor and cognitive skills (Objective 1), and the hardware and software components (Objective 2), Objectives 3 takes into account broader criteria for the emergence of new usages and applications in various areas, and in particular in the scope of education, art, popularization of science and entertainment. Our goal here is not to develop full-fledged end-user applications. Instead, our contribution is to stimulate the evolution of current applications with new engaging interactive systems.

3.4.1. Popularization of science

Popularization of science is at the core of the motivations of the Potioc group. Focusing on this subject allows us to get inspiration for the development of new interactive approaches. In particular, we have built a strong partnership with Cap Sciences, which is a center dedicated to the popularization of science in Bordeaux that is visited by thousands of visitors every month. This was initiated with the ANR national project InSTInCT, whose goal was to study the benefits of 3D touch-based interaction in public exhibitions. This project has led to the creation of a Living Lab where several systems developed by Potioc have been tested and will be tested by the visitors. This provides us with very interesting observations that go beyond the feedback we can obtain in our controlled lab-experiments.

3.4.2. Education

Education is also a key domain for Potioc. Indeed, we are convinced that the approaches we investigate—which target motivation, curiosity, pleasure of use and high level of interactivity—may serve education purposes. To this end, we collaborate with experts in Educational Sciences and teachers for exploring new interactive systems that enhance learning processes. We are currently investigating the fields of astronomy, optics, and neurosciences. In the future, we will continue exploring new interactive approaches dedicated to education, in various fields.

3.4.3. Art

Art, which is strongly linked with emotions and user experiences, is also a target area for Potioc. We believe that the work conducted in Potioc may benefit to creation from the artist point of view, and it may open new interactive experiences from the audience point of view. As an example, we are working with colleagues who are specialists in digital music, and with musicians. We are also working with jugglers and mockup builders with the goal of enhancing interactivity and user experience.

3.4.4. Entertainment

Similarly, entertainment is a domain where our work may have an impact. We notably explored BCI-based gaming and non-medical applications of BCI, as well as mobile Augmented Reality games. Once again, we beleive that our approaches that merge the physical and the virtual world may enhance the user experience. Exploring such a domain will raise numerous scientific and technological questions.

REALOPT Project-Team

3. Research Program

3.1. Introduction

Combinatorial optimization is the field of discrete optimization problems. In many applications, the most important decisions (control variables) are binary (on/off decisions) or integer (indivisible quantities). Extra variables can represent continuous adjustments or amounts. This results in models known as *mixed integer programs* (MIP), where the relationships between variables and input parameters are expressed as linear constraints and the goal is defined as a linear objective function. MIPs are notoriously difficult to solve: good quality estimations of the optimal value (bounds) are required to prune enumeration-based global-optimization algorithms whose complexity is exponential. In the standard approach to solving an MIP is so-called *branch-and-bound algorithm* : (i) one solves the linear programming (LP) relaxation using the simplex method; (ii) if the LP solution is not integer, one adds a disjunctive constraint on a factional component (rounding it up or down) that defines two sub-problems; (iii) one applies this procedure recursively, thus defining a binary enumeration tree that can be pruned by comparing the local LP bound to the best known integer solution. Commercial MIP solvers are essentially based on branch-and-bound (such IBM-CPLEX, FICO-Xpress-mp, or GUROBI). They have made tremendous progress over the last decade (with a speedup by a factor of 60). But extending their capabilities remains a continuous challenge; given the combinatorial explosion inherent to enumerative solution techniques, they remain quickly overwhelmed beyond a certain problem size or complexity.

Progress can be expected from the development of tighter formulations. Central to our field is the characterization of polyhedra defining or approximating the solution set and combinatorial algorithms to identify "efficiently" a minimum cost solution or separate an unfeasible point. With properly chosen formulations, exact optimization tools can be competitive with other methods (such as meta-heuristics) in constructing good approximate solutions within limited computational time, and of course has the important advantage of being able to provide a performance guarantee through the relaxation bounds. Decomposition techniques are implicitly leading to better problem formulation as well, while constraint propagation are tools from artificial intelligence to further improve formulation through intensive preprocessing. A new trend is robust optimization where recent progress have been made: the aim is to produce optimized solutions that remain of good quality even if the problem data has stochastic variations. In all cases, the study of specific models and challenging industrial applications is quite relevant because developments made into a specific context can become generic tools over time and see their way into commercial software.

Our project brings together researchers with expertise in mathematical programming (polyhedral approaches, Dantzig-Wolfe decomposition, mixed integer programing, robust and stochastic programming, and dynamic programming), graph theory (characterization of graph properties, combinatorial algorithms) and constraint programming in the aim of producing better quality formulations and developing new methods to exploit these formulations. These new results are then applied to find high quality solutions for practical combinatorial problems such as routing, network design, planning, scheduling, cutting and packing problems.

3.2. Polyhedral approaches for MIP

Adding valid inequalities to the polyhedral description of an MIP allows one to improve the resulting LP bound and hence to better prune the enumeration tree. In a cutting plane procedure, one attempt to identify valid inequalities that are violated by the LP solution of the current formulation and adds them to the formulation. This can be done at each node of the branch-and-bound tree giving rise to a so-called *branch-and-cut algorithm* [61]. The goal is to reduce the resolution of an integer program to that of a linear program by deriving a linear description of the convex hull of the feasible solutions. Polyhedral theory tells us that if X is a mixed integer program: $X = P \cap \mathbb{Z}^n \times \mathbb{R}^p$ where $P = \{x \in \mathbb{R}^{n+p} : Ax \le b\}$ with matrix

$(A, b) \in \mathbb{Q}^{m \times (n+p+1)}$, then $conv(X)$ is a polyhedron that can be described in terms of linear constraints, i.e. it writes as $conv(X) = \{x \in \mathbb{R}^{n+p} : C\,x \leq d\}$ for some matrix $(C, d) \in \mathbb{Q}^{m' \times (n+p+1)}$ although the dimension m' is typically quite large. A fundamental result in this field is the equivalence of complexity between solving the combinatorial optimization problem $\min\{cx : x \in X\}$ and solving the *separation problem* over the associated polyhedron $conv(X)$: if $\widetilde{x} \notin conv(X)$, find a linear inequality $\pi\,x \geq \pi_0$ satisfied by all points in $conv(X)$ but violated by $\widetilde{x}$. Hence, for NP-hard problems, one can not hope to get a compact description of $conv(X)$ nor a polynomial time exact separation routine. Polyhedral studies focus on identifying some of the inequalities that are involved in the polyhedral description of $conv(X)$ and derive efficient *separation procedures* (cutting plane generation). Only a subset of the inequalities $C\,x \leq d$ can offer a good approximation, that combined with a branch-and-bound enumeration techniques permits to solve the problem. Using *cutting plane algorithm* at each node of the branch-and-bound tree, gives rise to the algorithm called *branch-and-cut*.

3.3. Decomposition and reformulation approaches

An hierarchical approach to tackle complex combinatorial problems consists in considering separately different substructures (subproblems). If one is able to implement relatively efficient optimization on the substructures, this can be exploited to reformulate the global problem as a selection of specific subproblem solutions that together form a global solution. If the subproblems correspond to subset of constraints in the MIP formulation, this leads to Dantzig-Wolfe decomposition [5], [8], [9], [7]. If it corresponds to isolating a subset of decision variables, this leads to Bender's decomposition. Both lead to extended formulations of the problem with either a huge number of variables or constraints. Dantzig-Wolfe approach requires specific algorithmic approaches to generate subproblem solutions and associated global decision variables dynamically in the course of the optimization. This procedure is known as *column generation*, while its combination with branch-and-bound enumeration is called *branch-and-price*. Alternatively, in Bender's approach, when dealing with exponentially many constraints in the reformulation, the *cutting plane procedures* that we defined in the previous section are well-suited tools. When optimization on a substructure is (relatively) easy, there often exists a tight reformulation of this substructure typically in an extended variable space. This gives rise powerful reformulation of the global problem, although it might be impractical given its size (typically pseudo-polynomial). It can be possible to project (part of) the extended formulation in a smaller dimensional space if not the original variable space to bring polyhedral insight (cuts derived through polyhedral studies can often be recovered through such projections).

3.4. Integration of Artificial Intelligence Techniques in Integer Programming

When one deals with combinatorial problems with a large number of integer variables, or tightly constrained problems, mixed integer programming (MIP) alone may not be able to find solutions in a reasonable amount of time. In this case, techniques from artificial intelligence can be used to improve these methods. In particular, we use primal heuristics and constraint programming.

Primal heuristics are useful to find feasible solutions in a small amount of time. We focus on heuristics that are either based on integer programming (rounding, diving, relaxation induced neighborhood search, feasibility pump), or that are used inside our exact methods (heuristics for separation or pricing subproblem, heuristic constraint propagation, ...).

Constraint Programming (CP) focuses on iteratively reducing the variable domains (sets of feasible values) by applying logical and problem-specific operators. The latter propagates on selected variables the restrictions that are implied by the other variable domains through the relations between variables that are defined by the constraints of the problem. Combined with enumeration, it gives rise to exact optimization algorithms. A CP approach is particularly effective for tightly constrained problems, feasibility problems and min-max problems Mixed Integer Programming (MIP), on the other hand, is known to be effective for loosely constrained problems and for problems with an objective function defined as the weighted sum of variables. Many problems belong to the intersection of these two classes. For such problems, it is reasonable to use algorithms that exploit complementary strengths of Constraint Programming and Mixed Integer Programming.

3.5. Robust Optimization

Decision makers are usually facing several sources of uncertainty, such as the variability in time or estimation errors. A simplistic way to handle these uncertainties is to overestimate the unknown parameters. However, this results in over-conservatism and a significant waste in resource consumption. A better approach is to account for the uncertainty directly into the decision aid model by considering mixed integer programs that involve uncertain parameters. Stochastic optimization account for the expected realization of random data and optimize an expected value representing the average situation. Robust optimization on the other hand entails protecting against the worst-case behavior of unknown data. There is an analogy to game theory where one considers an oblivious adversary choosing the realization that harms the solution the most. A full worst case protection against uncertainty is too conservative and induces very high over-cost. Instead, the realization of random data are bound to belong to a restricted feasibility set, the so-called uncertainty set. Stochastic and robust optimization rely on very large scale programs where probabilistic scenarios are enumerated. There is hope of a tractable solution for realistic size problems, provided one develops very efficient ad-hoc algorithms. The techniques for dynamically handling variables and constraints (column-and-row generation and Bender's projection tools) that are at the core of our team methodological work are specially well-suited to this context.

3.6. Polyhedral Combinatorics and Graph Theory

Many fundamental combinatorial optimization problems can be modeled as the search for a specific structure in a graph. For example, ensuring connectivity in a network amounts to building a *tree* that spans all the nodes. Inquiring about its resistance to failure amounts to searching for a minimum cardinality *cut* that partitions the graph. Selecting disjoint pairs of objects is represented by a so-called *matching*. Disjunctive choices can be modeled by edges in a so-called *conflict graph* where one searches for *stable sets* – a set of nodes that are not incident to one another. Polyhedral combinatorics is the study of combinatorial algorithms involving polyhedral considerations. Not only it leads to efficient algorithms, but also, conversely, efficient algorithms often imply polyhedral characterizations and related min-max relations. Developments of polyhedral properties of a fundamental problem will typically provide us with more interesting inequalities well suited for a branch-and-cut algorithm to more general problems. Furthermore, one can use the fundamental problems as new building bricks to decompose the more general problem at hand. For problem that let themselves easily be formulated in a graph setting, the graph theory and in particular graph decomposition theorem might help.

SISTM Project-Team

3. Research Program

3.1. Mecanistic modelling

When studying the dynamics of a given marker, say the HIV concentration in the blood (HIV viral load), one can for instance use descriptive models summarizing the dynamics over time in term of slopes of the trajectories [37]. These slopes can be compared between treatment groups or according to patients' characteristics. Another way for analyzing these data is to define a mathematical model based on the biological knowledge of what drives HIV dynamics. In this case, it is mainly the availability of target cells (the CD4+ T lymphocytes), the production and death rates of infected cells and the clearance of the viral particles that impact the dynamics. Then, a mathematical model most often based on ordinary differential equations (ODE) can be written [30]. Estimating the parameters of this model to fit observed HIV viral load gave a crucial insight in HIV pathogenesis as it revealed the very short half-life of the virions and infected cells and therefore a very high turnover of the virus, making mutations a very frequent event [29].

Having a good mechanistic model in a biomedical context such as HIV infection opens doors to various applications beyond a good understanding of the data. Global and individual predictions can be excellent because of the external validity of a model based on main biological mechanisms. Control theory may serve for defining optimal interventions or optimal designs to evaluate new interventions [22]. Finally, these models can capture explicitly the complex relationship between several processes that change over time and may therefore challenge other proposed approaches such as marginal structural models to deal with causal associations in epidemiology [21].

Therefore, we postulate that this type of model could be very useful in the context of our research that is in complex biological systems. The definition of the model needs to identify the parameter values that fit the data. In clinical research this is challenging because data are sparse, and often unbalanced, coming from populations of subjects. A substantial inter-individual variability is always present and needs to be accounted as this is the main source of information. Although many approaches have been developed to estimate the parameters of non-linear mixed models [33], [40], [25], [31], [26], [39], the difficulty associated with the complexity of ODE models and the sparsity of the data leading to identifiability issues need further research.

3.2. High dimensional data

With the availability of omics data such as genomics (DNA), transcriptomics (RNA) or proteomics (proteins), but also other types of data, such as those arising from the combination of large observational databases (e.g. in pharmacoepidemiology or environmental epidemiology), high-dimensional data have became increasingly common. Use of molecular biological technics such as Polymerase Chain Reaction (PCR) allows for amplification of DNA or RNA sequences. Nowadays, microarray and Next Generation Sequencing (NGS) techniques give the possibility to explore very large portions of the genome. Furthermore, other assays have also evolved, and traditional measures such as cytometry or imaging have became new sources of big data. Therefore, in the context of HIV research, the dimension of the datasets has much grown in term of number of variables per individual than in term of number of included patients although this latter is also growing thanks to the multi-cohort collaborations such as CASCADE or COHERE organized in the EuroCoord network [0]. As an exemple, in a recent phase 1/2 clinical trial evaluating the safety and the immunological response to a dendritic cell-based HIV vaccine, 19 infected patients were included. Bringing together data on cell count, cytokine production, gene expression and viral genome change led to a 20 Go database [36]. This is far from big databases faced in other areas but constitutes a revolution in clinical research where clinical trials of hundred of patients sized few hundred of Ko at most. Therefore, more than the storage and calculation capacities, the challenge is the comprehensive analysis of these datasets.

---

[0]see online at http://www.eurocoord.net

The objective is either to select the relevant information or to summarize it for understanding or prediction purposes. When dealing with high dimensional data, the methodological challenge arises from the fact that datasets typically contain many variables, much more than observations. Hence, multiple testing is an obvious issue that needs to be taken into account [34]. Furthermore, conventional methods, such as linear models, are inefficient and most of the time even inapplicable. Specific methods have been developed, often derived from the machine learning field, such as regularization methods [38]. The integrative analysis of large datasets is challenging. For instance, one may want to look at the correlation between two large scale matrices composed by the transcriptome in the one hand and the proteome on the other hand [27]. The comprehensive analysis of these large datasets concerning several levels from molecular pathways to clinical response of a population of patients needs specific approaches and a very close collaboration with the providers of data that is the immunologists, the virologists, the clinicians...

<p style="text-align:center">**STORM Team**</p>

3. Research Program

3.1. Parallel Computing and Architectures

Following the current trends of the evolution of HPC systems architectures, it is expected that future Exascale systems (i.e. Sustaining 10^{18} flops) will have millions of cores. Although the exact architectural details and trade-offs of such systems are still unclear, it is anticipated that an overall concurrency level of $O(10^9)$ threads/tasks will probably be required to feed all computing units while hiding memory latencies. It will obviously be a challenge for many applications to scale to that level, making the underlying system sound like "embarrassingly parallel hardware."

From the programming point of view, it becomes a matter of being able to expose extreme parallelism within applications to feed the underlying computing units. However, this increase in the number of cores also comes with architectural constraints that actual hardware evolution prefigures: computing units will feature extra-wide SIMD and SIMT units that will require aggressive code vectorization or "SIMDization", systems will become hybrid by mixing traditional CPUs and accelerators units, possibly on the same chip as the AMD APU solution, the amount of memory per computing unit is constantly decreasing, new levels of memory will appear, with explicit or implicit consistency management, etc. As a result, upcoming extreme-scale system will not only require unprecedented amount of parallelism to be efficiently exploited, but they will also require that applications generate adaptive parallelism capable to map tasks over heterogeneous computing units.

The current situation is already alarming, since European HPC end-users are forced to invest in a difficult and time-consuming process of tuning and optimizing their applications to reach most of current supercomputers' performance. It will go even worse at horizon 2020 with the emergence of new parallel architectures (tightly integrated accelerators and cores, high vectorization capabilities, etc.) featuring unprecedented degree of parallelism that only too few experts will be able to exploit efficiently. As highlighted by the ETP4HPC initiative, existing programming models and tools won't be able to cope with such a level of heterogeneity, complexity and number of computing units, which may prevent many new application opportunities and new science advances to emerge.

The same conclusion arises from a non-HPC perspective, for single node embedded parallel architectures, combining heterogeneous multicores, such as the ARM big.LITTLE processor and accelerators such as GPUs or DSPs. The need and difficulty to write programs able to run on various parallel heterogeneous architectures has led to initiatives such as HSA, focusing on making it easier to program heterogeneous computing devices. The growing complexity of hardware is a limiting factor to the emergence of new usages relying on new technology.

3.2. Parallel Computing and Architectures

In the HPC context, simulation is already considered as a third pillar of science with experiments and theory. Additional computing power means more scientific results, and the possibility to open new fields of simulation requiring more performance, such as multi-scale, multi-physics simulations. Many scientific domains able to take advantage of Exascale computers, these "Grand Challenges" cover large panels of science, from seismic, climate, molecular dynamics, theoretical and astrophysics physics... Besides, embedded applications are also able to take advantage of these performance increase. There is still an on-going trend where dedicated hardware is progressively replaced by off-the-shelf components, adding more adaptability and lowering the cost of devices. For instance, Error Correcting Codes in cell phones are still hardware chips, but with the forthcoming 5G protocol, new software and adaptative solutions relying on low power multicores are also explored. New usages are also appearing, relying on the fact that large computing capacities are becoming more affordable and widespread. This is the case for instance with Deep Neural Networks where the training phase can be done

on supercomputers and then used in embedded mobile systems. The same consideration applies for big data problems, of internet of things, where small sensors provide large amount of data that need to be processed in short amount of time. Even though the computing capacities required for such applications are in general a different scale from HPC infrastructures, there is still a need in the future for high performance computing applications.

However, the outcome of new scientific results and the development of new usages for mobile, embedded systems will be hindered by the complexity and high level of expertise required to tap the performance offered by future parallel heterogeneous architectures.

3.3. Towards More Abstraction

As emphasized by initiatives such as the European Exascale Software Initiative (EESI), the European Technology Platform for High Performance Computing (ETP4HPC), or the International Exascale Software Initiative (IESP), the HPC community needs new programming APIs and languages for expressing heterogeneous massive parallelism in a way that provides an abstraction of the system architecture and promotes high performance and efficiency. The same conclusion holds for mobile, embedded applications that require performance on heterogeneous systems.

This crucial challenge given by the evolution of parallel architectures therefore comes from this need to make high performance accessible to the largest number of developers, abstracting away architectural details providing some kind of performance portability. Disruptive uses of the new technology and groundbreaking new scientific results will not come from code optimization or task scheduling, but they require the design of new algorithms that require the technology to be tamed in order to reach unprecedented levels of performance.

Runtime systems and numerical libraries are part of the answer, since they may be seen as building blocks optimized by experts and used as-is by application developers. The first purpose of runtime systems is indeed to provide *abstraction*. Runtime systems offer a uniform programming interface for a specific subset of hardware (e.g., OpenGL or DirectX are well-established examples of runtime systems dedicated to hardware-accelerated graphics) or low-level software entities (e.g., POSIX-thread implementations). They are designed as thin user-level software layers that complement the basic, general purpose functions provided by the operating system calls. Applications then target these uniform programming interfaces in a portable manner. Low-level, hardware dependent details are hidden inside runtime systems. The adaptation of runtime systems is commonly handled through drivers. The abstraction provided by runtime systems thus enables portability. Abstraction alone is however not enough to provide portability of performance, as it does nothing to leverage low-level-specific features to get increased performance. Consequently, the second role of runtime systems is to *optimize* abstract application requests by dynamically mapping them onto low-level requests and resources as efficiently as possible. This mapping process makes use of scheduling algorithms and heuristics to decide the best actions to take for a given metric and the application state at a given point in its execution time. This allows applications to readily benefit from available underlying low-level capabilities to their full extent without breaking their portability. Thus, optimization together with abstraction allows runtime systems to offer portability of performance. Numerical libraries provide sets of highly optimized kernels for a given field (dense or sparse linear algebra, FFT, etc.) either in an autonomous fashion or using an underlying runtime system.

Application domains cannot resort to libraries for all codes however, computation patterns such as stencils are a representative example of such difficulty. The compiler technology plays here a central role, in managing high level semantics, either through templates, domain specific languages or annotations. Compiler optimizations, and the same applies for runtime optimizations, are limited by the level of semantics they manage. Providing part of the algorithmic knowledge of an application, for instance knowing that it computes a 5-point stencil and then performs a dot product, would lead to more opportunities to adapt parallelism, memory structures, and is a way to leverage the evolving hardware.

Compilers and runtime play a crucial role in the future of high performance applications, by defining the input language for users, and optimizing/transforming it into high performance code. The objective of STORM is to propose better interactions between compiler and runtime and more semantics for both approaches. We recall in the following section the expertise of the team.

<p style="text-align:center">**TADAAM Team**</p>

3. Research Program

3.1. Need for System-Scale Optimization

Firstly, in order for applications to make the best possible use of the available resources, it is impossible to expose all the low-level details of the hardware to the program, as it would make impossible to achieve portability. Hence, the standard approach is to add intermediate layers (programming models, libraries, compilers, runtime systems, etc.) to the software stack so as to bridge the gap between the application and the hardware. With this approach, optimizing the application requires to express its parallelism (within the imposed programming model), organize the code, schedule and load-balance the computations, etc. In other words, in this approach, the way the code is written and the way it is executed and interpreted by the lower layers drives the optimization. In any case, this approach is centered on how computations are performed. Such an approach is therefore no longer sufficient, as the way an application is executing does depend less and less on the organization of computation and more and more on the way its data is managed.

Secondly, modern large-scale parallel platforms comprise tens to hundreds of thousand nodes [0]. However, very few applications use the whole machine. In general, an application runs only on a subset of the nodes [0]. Therefore, most of the time, an application shares the network, the storage and other resources with other applications running concurrently during its execution. Depending on the allocated resources, it is not uncommon that the execution of one application interferes with the execution of a neighboring one.

Lastly, even if an application is running alone, each element of the software stack often performs its own optimization independently. For instance, when considering an hybrid MPI/OpenMP application, one may realize that threads are concurrently used within the OpenMP runtime system, within the MPI library for communication progression, and possibly within the computation library (BLAS) and even within the application itself (pthreads). However, none of these different classes of threads are aware of the existence of the others. Consequently, the way they are executed, scheduled, prioritized does not depend on their relative roles, their locations in the software stack nor on the state of the application.

The above remarks show that in order to go beyond the state-of-the-art, it is necessary to design a new set of mechanisms allowing cross-layer and system-wide optimizations so as to optimize the way data is allocated, accessed and transferred by the application.

3.2. Scientific Challenges and Research Issues

In TADAAM, we will tackle the problem of efficiently executing an application, at system-scale, on an HPC machine. We assume that the application is already optimized (efficient data layout, use of effective libraries, usage of state-of-the-art compilation techniques, etc.). Nevertheless, even a statically optimized application will not be able to be executed at scale without considering the following dynamic constraints: machine topology, allocated resources, data movement and contention, other running applications, access to storage, etc. Thanks to the proposed layer, we will provide a simple and efficient way for already existing applications, as well as new ones, to express their needs in terms of resource usage, locality and topology, using a high-level semantic.

It is important to note that we target the optimization of each application independently but also several applications at the same time and at system-scale, taking into account their resource requirement, their network usage or their storage access. Furthermore, dealing with code-coupling application is an intermediate use-case that will also be considered.

---

[0]More than 22,500 XE6 compute node for the BlueWaters system; 5040 B510 Bullx Nodes for the Curie machine; more than 49,000 BGQ nodes for the MIRA machine.

[0]In 2014, the median case was 2048 nodes for the BlueWaters system and, for the first year of the Curie machine, the median case was 256 nodes

Several issues have to be considered. The first one consists in providing relevant **abstractions and models to describe the topology** of the available resources **and the application behavior**.

Therefore, the first question we want to answer is: **"How to build scalable models and efficient abstractions enabling to understand the impact of data movement, topology and locality on performance?"** These models must be sufficiently precise to grasp the reality, tractable enough to enable efficient solutions and algorithms, and simple enough to remain usable by non-hardware experts. We will work on (1) better describing the memory hierarchy, considering new memory technologies; (2) providing an integrated view of the nodes, the network and the storage; (3) exhibiting qualitative knowledge; (4) providing ways to express the multi-scale properties of the machine. Concerning abstractions, we will work on providing general concepts to be integrated at the application or programming model layers. The goal is to offer means, for the application, to express its high-level requirements in terms of data access, locality and communication, by providing abstractions on the notion of hierarchy, mesh, affinity, traffic metrics, etc.

In addition to the abstractions and the aforementioned models we need to **define a clean and expressive API in a scalable way**, in order for applications to express their needs (memory usage, affinity, network, storage access, model refinement, etc.).

Therefore, the second question we need to answer is: "**how to build a system-scale, stateful, shared layer that can gather applications needs expressed with a high-level semantic?**". This work will require not only to define a clean API where applications will express their needs, but also to define how such a layer will be shared across applications and will scale on future systems. The API will provide a simple yet effective way to express different needs such as: memory usage of a given portion of the code; start of a compute intensive part; phase where the network is accessed intensively; topology-aware affinity management; usage of storage (in read and/or write mode); change of the data layout after mesh refinement, etc. From an engineering point of view, the layer will have a hierarchical design matching the hardware hierarchy, so as to achieve scalability.

Once this has been done, the service layer, will have all the information about the environment characteristics and application requirements. We therefore need to design a set of **mechanisms to optimize applications execution**: communication, mapping, thread scheduling, data partitioning/mapping/movement, etc.

Hence, the last scientific question we will address is: "**How to design fast and efficient algorithms, mechanisms and tools to enable execution of applications at system-scale, in full a HPC ecosystem, taking into account topology and locality?**" A first set of research is related to thread and process placement according to the topology and the affinity. Another large field of study is related to data placement, allocation and partitioning: optimizing the way data is accessed and processed especially for mesh-based applications. The issues of transferring data across the network will also be tackled, thanks to the global knowledge we have on the application behavior and the data layout. Concerning the interaction with other applications, several directions will be tackled. Among these directions we will deal with matching process placement with resource allocation given by the batch scheduler or with the storage management: switching from a best-effort application centric strategy to global optimization scheme.