Software infrastructure for the computers of tomorrow

New software infrastructure is required to fully exploit the potential of the emerging generation of high performance computers. We spoke to **Ignacio Pagonabarraga**, **Sara Bonella**, **Jony Castagna** and **Donal MacKernan** of the E-CAM project about their work in developing new software modules targeted at the needs of both academic and industrial end-users.

The ongoing development of high performance computers (HPC) is opening up new possibilities in research, helping scientists gain deeper insights into fundamental questions across all fields, from molecular dynamics, to electronic structure, to multiscale models. The E-CAM project has been established to support the ongoing development of Europe's HPC infrastructure, as its technical manager Professor Ignacio Pagonabarraga explains. "E-CAM aims to create an infrastructure of software development, and to train users in the new HPC landscape. We liaise with industrialists, with the goal of enlarging the community," he says. The focus of the project is on modelling materials and biological processes, work which holds wider importance to society and industry. There are three complementary themes to this work. "We develop, test, maintain and disseminate software modules targeted at end-user needs," outlines Professor Pagonabarraga. "Second, we train academic and industrial researchers to exploit these capabilities. Third, we provide multi-disciplinary, applied consultancy to industrial end-users."

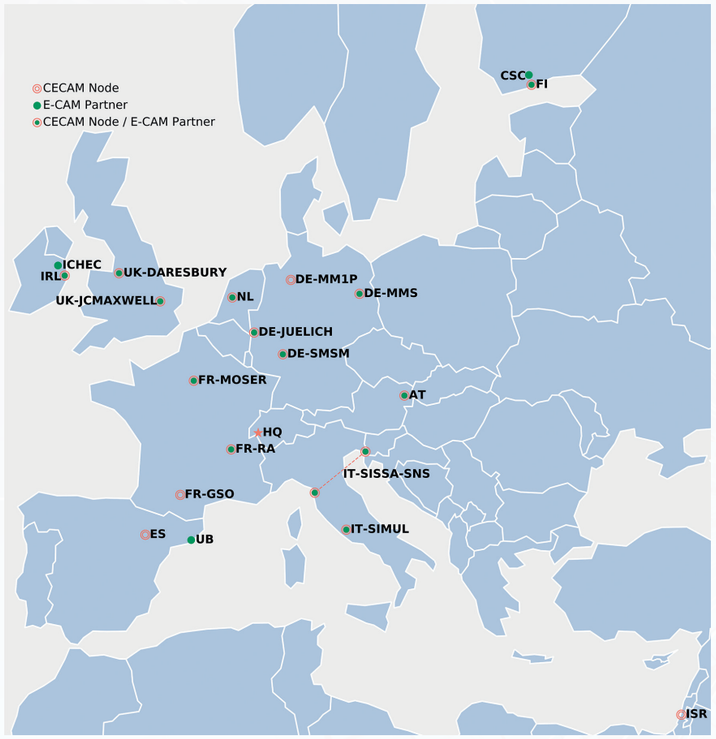


Figure 1: Map of the CECAM nodes and of E-CAM partners, showing how the E-CAM consortium is based around the CECAM's distributed Network of nodes. E-CAM is coordinated at the École Polytechnique Fédérale de Lausanne (Switzerland), which is also where the Headquarters (HQ) of CECAM is located.

CECAM

This research builds on the network of CECAM, an organisation which aims to facilitate and catalyse international cooperation in computational science. The computational science research community was relatively small when CECAM was established in 1969, but the field has developed rapidly over the last 50 years, and CECAM's overall mission has widened. "The community has grown, and the areas of expertise covered by CECAM have also increased accordingly," says Professor Pagonabarraga. A number of workshops, schools and other events are organised by CECAM, activities which Professor Pagonabarraga says are central to its wider agenda. "We aim to bring people together to collaborate and work on different topics, for example questions around molecular simulations and modelling, and the development of new algorithms," he continues. "CECAM today is organised as a network of 17 nodes distributed across Europe, and with Headquarters in Lausanne (Figure 1). These nodes also help to bring research initiatives closer to local communities."

The wider aim of the E-CAM project is to develop the sophisticated software infrastructure required to fully exploit the capabilities of the emerging generation of supercomputing technologies. These machines are far more powerful than their predecessors, and their architecture is also different. "Therefore there is a need to develop this software infrastructure," says Professor Pagonabarraga. Researchers are investigating how to model and simulate materials and biological processes on scales that have not previously been achievable. "There is no single, monolithic code that can be used with all the different types of materials and biological processes that we're interested in. There are different types of codes depending on the type of system, or the material properties," explains Professor Pagonabarraga. "We are developing what we call modular software, that can be adapted to a particular type of code or family of codes. We have also created libraries that can interface with these codes to enable the effective exploitation of HPC resources."

This modular approach allows researchers to identify the specific pieces of software or workflows that may be of interest with respect to different codes. Another important activity in the project is the co-design of scientific codes to run on HPC machines. GPU accelerated and scalable applications can help reduce the time-to-market of novel products (Figure 2).

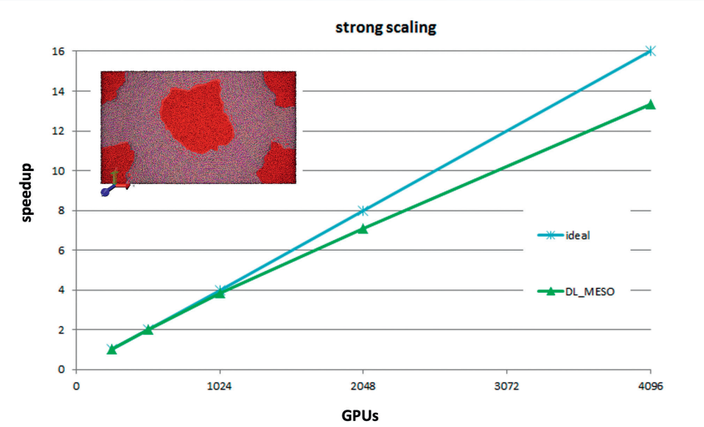


Figure 2: Strong scaling efficiency of DL_MESO (E-CAM Meso and Multi-scale Modelling code) versus the number of GPUs for a simulation of a complex mixed phase system consisting of 1.8 billion atoms. J. Castagna et al., Comput. Phys. Commun. 251, 107159 (2020)

There are several workpackages within the project, each covering different challenges around simulation and modelling. "They involve different types of codes and systems descriptions at different scales," says Professor Pagonabarraga. For example, one of the workpackages concerns the development of coarse-grained models and multi-scale modelling. "The general goal of coarse-graining is to identify - out of the whole detail of a system composed of atoms and molecules - how much detail is required and which atomic or molecular details we can effectively average out," continues Professor Pagonabarraga. "There is no single answer to this question. There are different ways to approach coarse-graining, depending on the scientific or technological problem."

of communicating with each other," he points out. This topic is being addressed, for example, in a pilot project together with Michelin where the aim is to develop new codes to study polymer systems. "If you want to understand how polymers are connected to each other in a tyre, and the mechanical properties of a tyre, then you need to use a coarse-grain approach," says Professor Pagonabarraga. "In another pilot project, an algorithm called GC-AdResS is being developed for both coarse-grained and more detailed descriptions of a material."

The different scientific workpackages in the project provide the flexibility to address a problem at different scales. This is an important attribute in terms of collaborating with industrial partners, as Dr Sara Bonella explains. "You might find that answering a specific question posed by an industrial partner requires the use of **multi-scale modelling**. E-CAM covers modelling and simulations at different scales, from the **atomic level** right up to the **coarse-grain models**, providing the flexibility to **address a problem from different points of view**

Modelling and simulations

A further part of that workpackage involves multi-scale modelling, where researchers are developing software that can reproduce the detail of a system at different levels. For example it may be that one part of a particular system can be dealt with in a fairly approximate manner, while with another it may be important to capture the chemical specificity, which Professor Pagonabarraga says represents a significant technical challenge. "This is quite a complex task, as it requires different types of models and in some cases paradigms capable

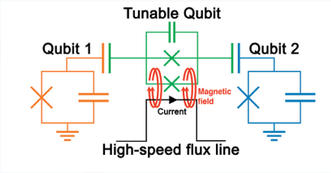


Figure 3: E-CAM developed a new method and dedicated software for designing control pulses to manipulate qubit systems based on the local control theory (LCT). The system is schematically represented above: two fixed frequency superconducting transmon qubits (Qubit 1 and Qubit 2) are coupled to a Tunable Qubit (TQ) whose frequency is controlled by an external magnetic field. Changing the frequency, the TQ behaves as a targeted quantum logic gate, effectively enabling an operation on the qubit states. M. Mališ et al., Phys. Rev. A 99, 052316 (2019)

find that answering a specific question posed by an industrial partner requires the use of multi-scale modelling. E-CAM covers modelling and simulations at different scales, from the atomic level right up to the coarse-grain models, providing the flexibility to address a problem from different points of view", she outlines. As the leader of the workpackage on quantum dynamics, Dr Bonella is working on the development of superconducting qubits for quantum computing, which again involves close collaboration with industry (Figure 3). "We're looking at how laser pulses can be used to effectively tune the response of a group of these qubits. By controlling this response, we hope to create different logical gates," she continues. "This is a very ambitious objective, and we are looking at it in collaboration with a theory and simulation group at IBM."

This is a case in which the goal of the simulations is to provide indicators on experimental setups. The hope is that the set of indicators that have been gathered, in particular about what the laser should do to guide the behaviour of each qubit, can inform future experiments. The ability to model a system at multiple scales holds great relevance to the development of quantum computing. "A qubit is a quantum dynamical object, and its behaviour can be profoundly affected through something called decoherence. This is essentially the effect of the environment on the qubit, and it tends to destroy its quantum properties," says Dr Bonella. "If these quantum properties are destroyed, then we lose the advantage to the hardware that is provided by the qubit."

A multi-scale description of a quantum system, in which the qubit is described at a quantum dynamical level, can help researchers build a fuller picture in this respect. This quantum dynamical description of the qubit can then be matched with a classical description of the environment in which it is embedded, which Dr Bonella says will help scientists understand how the environment affects its performance. "It's another example of the importance of multi-scale modelling," she outlines. The ability to control and tune the properties of a qubit is central to the wider goal of developing quantum computers, which could have significantly greater computational power than conventional machines. "There is an added richness that comes from the fact that you can have these linear combinations of states and entangled states and coherent transfer of information. These things exponentially multiply the computational capacity," explains Dr Bonella. "But the development of these machines is still at a relatively early stage."

Machine learning

The project's overall agenda also includes research into machine learning and artificial intelligence, areas which have developed rapidly over the last decade as a new way of exploring science. One topic being addressed in this part of the project is building reliable models for interactions using neural networks based approaches. "We aim to understand how to build neural networks which are correct, and which make reliable predictions on potential energy surfaces," says Dr Jony Castagna. A neural network is based on a layered structure, where each layer is made of neurons, referred to in the project as perceptrons, which can be thought of as a mathematical model of a neuron. "We use graphic processing units (GPUs), which are able to calculate necessary matrix-matrix multiplications very rapidly," explains Dr Castagna. "So they really enhance the power dramatically. They effectively train the neural network, and then you can use it to make predictions."

Researchers are also considering the potential applications of these machine learning techniques. There are pilot projects within E-CAM looking to exploit machine learning for important problems in specific fields. "One of them is the modelling of metal ions in proteins. These are basically metal co-factors, and they play very important roles in certain areas of biology. In a pilot project we are using machine learning to effectively generate more accurate coarse-grained models capturing biochemically important details (Figure 4)," outlines Dr Donal Mackernan.

Highly sophisticated algorithms are required to investigate changes in protein structure and function, a topic central to Dr Mackernan's work in the project focused on the development of diagnostic protein sensors for diseases, such as influenza and Covid 19 . "A typical protein complex is part of a system consisting of millions of atoms. Proteins don't exist by themselves, they're usually in water, in salt, or another complex environment, so they are a formidable system to simulate," he explains.

A major factor here is the timescales over which biological processes take place. The timescales involved in the oscillations of hydrogen bonds in water are on the order of 10^{-15} seconds, which is very short, yet the timescales at which real biological mechanisms take place are often on the order of milliseconds or seconds. "No supercomputer currently available can simulate a system to the required level of detail on that kind of timescale," says Dr Mackernan. This is a problem that Dr Mackernan and his colleagues are working to overcome while developing a novel type of biosensor. "This biosensor is essentially a modular system, and is naturally coarse-grained," he explains "We can investigate how to build it and optimise it. But when you do that you're immediately confronted by the problem that the devil is in the detail - very often you need to take atomistic effects into account." A powerful set of techniques called rare event sampling methods are being applied to help develop the very detailed simulations that are required. "These methods are used to try and overcome this timescale problem, to go from say 10^{-15} of a second to milli-seconds or seconds. A lot of software has been developed within E-CAM for that purpose. It is about developing sophisticated statistical techniques to try and capture fleetingly short lived events or configurations of the system critically important to the chemical or physical process," outlines Dr Mackernan.

Another pilot project, within the general activities of creating interaction potentials via neural network, specifically focuses on using electronic structure calculations as the building blocks of the machine learning procedure. "The quantum system, specifically an electronic structure calculation, trains the neural network, so it explores and discovers the relevant parameters of a highly adjustable multi component potential. That gives rise to an effective potential, which can be used to simulate a quantum system using classical simulation," continues Dr Mackernan. The benefit of machine learning here is that it allows researchers to model the interactions at a small scale using quantum mechanics, from which point it's then possible to simulate the system on much larger scales. Whereas other methods are relatively limited in terms of simulation size, Dr Mackernan says the use of machine learning opens up the possibility of simulating a system on much larger scales, and also over longer timescales. " There's an enormous improvement, in terms of the realism of the modelling," he says. This is not a purely academic exercise, and alongside research Dr Mackernan is keen to help potential users develop their skills through extended software development workshops. "Recently we've held training events on the use of machine learning to derive potentials to model quantum systems for example," he outlines. "Another important topic here is the extraction of reaction mechanisms. So when you try to simulate a large system, you want to understand some of the key properties of that system."

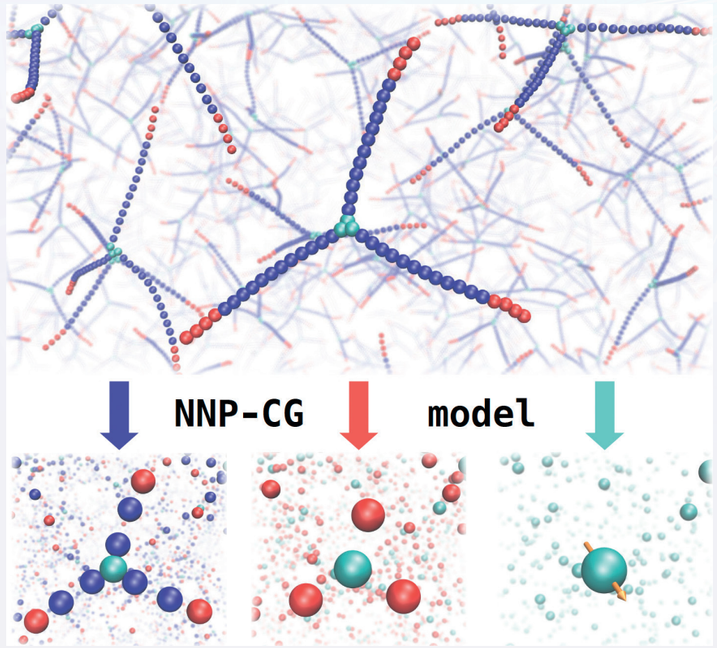


Figure 4: Neural network potentials (NNP) with their capability to reproduce arbitrary potential energy surfaces can also be used to construct coarse-grained (CG) models. The image shows different coarse-graining levels of dendrimer-like DNA molecules (top), from a description retaining more detail (left) to a single-bead representation (right). This work is part of an E-CAM pilot project.

Training

This commitment to providing training is an important aspect of E-CAM, and a recognition of the need to help the next generation of researchers develop their skills. Extended software development workshops organized in the project provide the opportunity for researchers to learn about software tools, simulations and modelling in today's rapidly evolving computing landscape (Figure 5). "Participants can work together on specific modules, on the development of software," says Professor Pagonabarraga. Another type of training environment established within E-CAM is what Professor Pagonabarraga calls a scoping workshop. "Here we bring together industrialists, software developers and academic researchers, to discuss the challenges they face," he continues. "It's very important to have people with complementary areas of expertise and to have these discussions as open and broad as possible."

A number of dedicated training events are planned for investors as well, as new simulation and modelling techniques could bring great benefits to industry, helping the commercial sector work more effectively. A couple of events are planned for the next year, including one on meso-scale simulations, as well as another in collaboration with one of the project's industrial partners Biki, an SME that develops software for drug design. "We want to work with industrial researchers, so we can train them on the use of software, and on the methodological novelties while learning from them about exciting new problems," explains Dr Bonella.

Figure 5: E-CAM Extended Software Development Workshop at the Lorentz Center in Leiden.

E-CAM

An e-infrastructure for software, training and consultancy in simulation and modelling

Project Objectives

E-CAM aims at (1) developing software to solve important simulation and modelling problems in industry and academia, with applications from drug development, to the design of new materials. (2) Tuning those codes to run on HPC, through application co-design and the provision of HPC oriented libraries and services; (3) Training scientists from industry and academia ; (4) Supporting industrial end-users in their use of simulation and modelling, via workshops and direct discussions with experts in the CECAM community.

Project Funding

The E-CAM project is funded by the European Union's Horizon 2020 research and innovation program under the grant agreement No. 676531

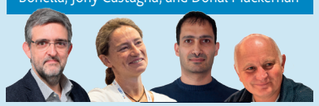
Project Partners

E-CAM is coordinated by CECAM at École Polytechnique Fédérale de Lausanne (EPFL), and is a partnership of 13 CECAM Nodes, 4 PRACE centres, 12 industrial partners and one Centre for Industrial Computing (Hartree Centre). Partners details can be found at: https://www.e-cam2020.eu/history/

Contact Details

Prof. Ignacio Pagonabarraga
E-CAM Technical Manager
CECAM Director
Ecole Polytechnique Fédérale de Lausanne (EPFL)
Batochime building BCH 3101
Avenue Forel 2
CH - 1015 Lausanne
T: +41 21 693 79 23
E: ignacio.pagonabarraga@epfl.ch
W: www.e-cam2020.eu
W: www.cecam.org

Left to right: Ignacio Pagonabarraga, Sara Bonella, Jony Castagna, and Donal Mackernan



Ignacio Pagonabarraga is the Director of CECAM, with headquarters at the École Polytechnique Fédérale de Lausanne, and Professor of Condensed Matter Physics at the University of Barcelona.

Sara Bonella is the Deputy Director of CECAM and Principal Investigator in Computational Quantum Physics, at the École Polytechnique Fédérale de Lausanne.

Jony Castagna is Computational Scientist at the Science and Technology Facilities Council, Hartree Centre.

Donal Mackernan is Principal Investigator at the University College Dublin and the Director of the CECAM Irish Node.