HPC: 2015 prospects

I. OVERVIEW

High performance computing (HPC) and technology watch has always been a main activity at CERFACS. If the global evolution trend of massively parallel architectures observed on the top500 list continues (http://www.top500.org), the first exascale computer will exist by the year 2020. The preparatory years towards this revolution are therefore extremely important. Even though the exact composition and architecture of such a machine are not fixed yet, it is possible to anticipate algorithmic hurdles to overcome in the way to exaflop/s capable applications.

To take advantage of the next generation of machines the problems to be tackled in the next year at CERFACS include peak performance utilization improvement, hybrid-programming strategies for extreme computing, resilient algorithm design, data management and high performance I/O methodologies.

II. OBJECTIVES

The objective of this transverse activity is improve the knowledge base of CERFACS on how to take maximum advantage of hardware and software improvements to use leadership class systems more efficiently and ease their usage for non-experts. This is done on real codes, mostly used by CERFACS partners and not only demo tools.

Currently CERFACS's applications (especially CFD and Climate modeling) scale up to tens of thousand of cores. Top tier systems contain thousands to millions of cores and exascale systems are expected to have higher numbers. 2015 will see the evolution of the PRACE and GENCI computational landscape with renewed systems at LRZ and CINES for example. We need to modify our codes to take advantage of the extreme parallelism and new hardware present in these machines. The increased complexity of these systems enhances the importance of taking into account the resilience of each component and its impact on the simulation. Alternative methods to simple checkpoint restart will be investigated.

Besides resilience, architecture updates are a great challenge to existing codes. Indeed, the new architectures (in 2015 Intel's Haswell and IBM's Power 8 for example) possess new features that increase vector and data management capabilities. Efficiently using these components is the key to improved computation efficiency. Numerical and code refactoring methods started in 2014 will be continued next year to devise the best practices for these new chips that will equip the machines of the next three years.

III. MAIN STEPS

Increased scalability

The current trend of massively many core systems will continue in 2015. The next generation of processors will make 12 core chips a standard (compared to a more frequent 8 core today) and go to increased hyper-threading (or SMT) capabilities. For optimum performance we will not be able to rely solely on message passing protocols. We started in 2014 to implement hierarchical hybrid parallelism with collaborations with CINES (on the jaguar code) or with Juelich (DEEP project) and Intel (on AVBP). These first approaches used MPI-OPENMP, MPI-CUDA and MPI-OMPSS techniques.

In 2015 these studies will be continued increasing the parallelism of these applications and later reworking the algorithms to allow for asynchronous task execution a necessary step for heterogeneous architectures. In some cases, the OASIS coupler will be used in order to modularize and parallelize model tasks like sea-ice, icebergs and wave-related algorithms in NEMO for example.

Since hardware choices for the future exascale systems are still at a very early stage we will remain vigilant for further alternatives on efficient parallel programing paradigms. We will keep a close eye on partitioned global address space programming models (PGAS) and runtime systems like StarPU developed at INRIA. Optimization of data movement will be the key to performance in heterogeneous systems and we will keep watch of upcoming tools and methods on the field.

Using millions of cores will increase the probability of hardware/middleware failures and the code redesign will include a brainstorming on code resiliency at the algorithm level as well as the parallelism level. We will follow closely the evolution of the MPI3 fault tolerant standard and its first implementations.

Increase computational efficiency

Next generation hardware roadmaps focus on massively many core and co-processing hardware. This evolution is addressed with the previous paragraph strategies, however there is also a planned increase on the vector capabilities of these architectures. For example the current generation of INTEL processors is capable of quadri-vector operations, which are mostly unused today by most codes. The next generations will double these vector operations by next year with new FMA instructions introduced on Haswell and there are plans for more complex functionalities in the future. In order to take full advantage of these architectures we will need to improve vectorization in our algorithms. This work started in late 2014 on AVBP via collaboration with the team of Dr Jalby (Exascale Lab) and Intel (IPCC program) will continue, our aim being to devise comprehensive strategies to refactor existing codes for increase performance. The main target applications will be the CFD code AVBP, but later on the methodologies will be tested on others solvers as well. For ARPEGE and NEMO, the existing vectorization from the historical vector architecture implementation will be refactored.

These strategies are architecture dependent and will take into account the availability of new systems with diversified components. The main targets will be Intel Haswell, Intel Xeon Phi and IBM Power 8.

I/O bandwidth / Big Data and data management

Data management and I/O strategies are a recurring bottleneck in existing applications when leaping to more advanced and complex systems with more cores. Sequential I/O or one file per process I/O strategies are still dominant but are no longer satisfactory with current systems, the road to exascale requires new strategies in this framework. These strategies must account for increase I/O traffic (Petabytes of data) but they must also account for the limitation of post-processing tools when handling these quantities of datasets. These issues are addressed using industry standards like Parallel HDF5 or NETCDF enforcing an adaptability and portability of our work. Access to new systems and the inclusion of hierarchical parallelism discussed in the first section will require an adaptation and tweaking of current I/O methodologies including the implementation of sub-process regulation I/O access (one out of N process handles I/O for a group of N processes) as well as asynchronous I/O processing with co-processing.

HPC capabilities in terms of computing power already outpace the increase in available data storage and this needs to be addressed as well. Data mining strategies to extract significant data during runtime and outputting only relevant datasets will be required to reduce data output to acceptable levels and increase software productivity. For example, a particular case where these developments are highly important is for climate ensemble simulations in which a single I/O server will be setup to performed online post-processing to generate ensemble diagnostics.

IV. COOPERATION

Preparing for the next generation of machines is not be possible without high quality collaborations and all of these challenges will not be met by CERFACS alone. International and national collaborations and lobbying will be sought-after and developed next year. On the national level we will benefit from the close relationship with INRIA through the joint CERFACS-INRIA lab on HPC. Joint projects on Hierarchical parallelism and solvers have been proposed for 2015 for example. Strong collaborations exist and will continue with the high performance computing centers in France (CINES, TGCC and IDRIS) and throughout Europe. Close ties with GENCI and its affiliated centers will be maintained, CERFACS is a candidate for two early access Challenges in the new computer OCCIGEN at CINES for late 2014.

Strong ties have been woven with JSC (Juelich Supercomputing Center) and Barcelona Supercomputing Center (BSC) via two EU programs (DEEP, COPAGT) and will be maintained with further European and bilateral collaborations. A continuation of DEEP is under study at BSC and early discussions have taken place between CERFACS and BSC. It is expected to yield a formal project proposition by the end of 2014 for a start in 2015.

Our participation to the PRACE and INCITE scientific excellence programs gives us unique opportunities to perform the best science in leadership class systems and to strengthen our collaboration with the hosting facilities. We will continue to participate and collaborate with these programs. We currently have multiple active participations on the current PRACE call, at least one new project will be submitted to the 10th PRACE Call for a start in March 2015 on climate modeling. Two INCITE proposals, one on multiscale combustion modeling (ARNL) and one on LES applied to Turbo machinery (ORNL) are currently under review and are expected (if accepted) to start in January 2015. This last proposal will be closely linked with coupling studies since it involved massively petascale coupling using OpenPALM. These projects are expected to enforce our collaboration with the Argonne and Oakridge national labs.

Our traditional IPSL collaboration for I/O handling (XIOS) and high-resolution ocean modeling still remains. A new axis will be developed with respect to computational efficiency on many core architectures (ANR HEAT) focusing on performances on new icosahedral grids for atmosphere models. This work will be performed in conjunction with the Japan Agency for Marine-Earth Science and Technology and the University of Tokyo

A fruitful collaboration with Intel via the Intel Parallel Computing Centers programs is underway and our work will be enforced next year as well thanks to a collaboration with the Intel exascale labs via the GIS SUCCES. A work group on parallel IO has already been discussed between the GIS SUCCESS partners and is expected to come together by 2015.

A collaboration with IBM on OpenPower and GPU architectures is under discussion and a second collaboration with HP to evaluate low power architectures (for example ARM) is expected to start in 2015.

References

U. Fladrich, E. Maisonnave (2014): A new set of metrics for the computational performance of IS-ENES Earth System Models, Technical Report TR/CMGC/14/XX, CERFACS, Toulouse, France

E. Maisonnave (2013): PoCO, post-processing coupled with OASIS, Technical Report TR/CMGC/13/70, CERFACS, Toulouse, France

E. Maisonnave (2014): Coupling an icosahedral grid, Working Note, WN/CMGC/14/8, CERFACS, Toulouse, France

E. Maisonnave, V. Slavnic (2013): Resilience in an ocean model, Technical Report TR/CMGC/13/110, CERFACS, Toulouse, France

E. Maisonnave, A. Caubel (2014): LUCIA, load balancing tool for OASIS coupled systems, Technical Report TR/CMGC/14/63, CERFACS, Toulouse, France

Staffelbach G. Masssively parallel computations of gas turbines and building explosions on INCITE and PRACE systems - invited conference. In *HPC Advisory Council/Thermal and Fluid Sciences Affiliates and Sponsors program, Exascale Workshop 2014*, Stanford University, February 3-5 2014 (http://insidehpc.com/2014/02/26/leadership-computing-combustion-apps-incite-prace-systems/)