THE CHALLENGES OF THE NEXT DECADE IN NUMERICAL COSMOLOGY : BIG DATA AND EXTREME-SCALE COMPUTING .

Jean-Michel Alimi
Laboratoire Univers et Théories, CNRS, Observatoire de Paris, Univ. Paris Diderot ;
5 Place Jules Janssen, 92190 Meudon, France, http://www.deus-consortium.org

Abstract : With the DEUS experiment, we were able to perform a N-body gravitational simulation with 0.5 trillion particles evolving in the entire volume of the Observable Universe with 2.5 trillion computing points. In this paper, we present what we think to be the critical points regarding the possibility to perform a new simulation with 100 times more particles. This simulation for the next decade is scientifically motivated both from a theoretical point of view, to study the properties of non linear clustering of dark matter and the nature of Dark Energy; and from an observational point of view, to support future large surveys of our Universe provided by the next generation of satellite like Euclid (ESA/NASA), to be launched in 2020.

95 % of the cosmos is in the form of two invisible components: ~ 25 % is in Cold Dark Matter (CDM) particles, which are primarily responsible for the formation of the visible structures in the universe [Peebles1980]; ~ 70 % is in a unknown exotic form, dubbed « dark energy » (DE), which is responsible for the present phase of cosmic accelerated expansion [Copeland et al. 2006]. Dark Energy may well be of different origin. Several alternative scenarios have been advanced, but to date a coherent physical understanding of DE is still missing. In the lack of a complete theoretical guidance, cosmologists are also turning to observations and several observational big-science projects such as BOSS [BOSS], DES [DES], LSST [LSST] and the EUCLID mission [EUCLID] which will map with great accuracy the distribution of matter in the universe. These projects may eventually shed new light on the problem, as the clustering of matter may be key to infer the properties of DE. In order to study both the properties of the clustering of matter and the imprint of Dark Energy on the cosmic structure formation, one has to follow the gravitational collapse of Dark Matter throughout the history of the universe and across several orders of magnitude length scales for different cosmological models with different nature of Dark Energy. During the past 10 years several groups have pushed to the limits both size and resolution of cosmological N-body simulations. The Millenium Simulation in 2005 has run a 2.2 billion light-years simulation box with 10 billion particles [Springel et al. 2005]. Since then, the performance of cosmological simulations has rapidly increased. The Millenium-XXL simulation has evolved more recently 303 billion particles in a 13 billion light-years box [Angulo et al. 2012], while the Horizon Run 3 has followed the evolution of 374 billion particles in a 49 billion light-years box [Kim et al. 2011]. DEUS Simulation [DEUS, Alimi et al 2012] have performed the first-ever numerical N-body simulations of the full observable universe (~95 billion light-years box) for three different cosmological models. These simulations have evolved 550 billion particles on an Adaptive Mesh Refinement grid with more than two and half trillion computing points along the entire evolutionary history of the universe and across 6 orders of magnitudes length scales, from the size of the Milky Way (mass of one particle) to that of the whole observable Universe. Such runs provide unique information on the formation and evolution of the largest structure in the universe and an exceptional support to future observational programs dedicated to the mapping of the distribution of matter and galaxies in the universe. Each simulation has run on 4752 (of 5040) thin nodes of GENCI's supercomputer CURIE, using more than 300 TB of memory for 5 million hours of computing time. About 50 PBytes of rough data were generated throughout each run. Using an advanced and innovative reduction workflow the amount of useful stored data has been reduced almost one the fly for each to 500 TBytes. Overall the realization of such large simulations required the development of a global application which integrated all aspects of the physical computational problem: initial conditions, dynamical computing, data validation, reduction and storage. Hence, it required optimizing not only the efficiency of the numerical dynamical solver, but also the memory usage, communications and I/O at the same time. Previous cosmological "grand-challenge" simulations could limit the use of efficient parallel HPC schemes to the dynamical computational solver only. Instead the DEUS experiment and similar projects clearly show the need for the use of MPI instructions in all parts of the application, scalable to more than 10,000 to 100 000 processes and optimized to match the specificities of supercomputing machine in which these have been run.

The numerical challenge in cosmology for the dark Universe for the next decade probably needs in the first place a better physical description for the two dark components than ones performed until now. We need for example N-body numerical simulation for a wide class of modified gravity and with dynamical dark energy theories. But these physical aspects of the problem of the cosmic structure formation need to be better understood to perform in such theoretical frameworks very high performance numerical simulation. In the following we focus on a numerical simulation where the gravitational physics is supposed to be well defined and locally well understood but where we want to follow the formation of all DM halos with a mass corresponding to one galaxy (typically) 10^{12} (10^{11}) solar masses evolving on a large fraction of the observable universe. That implies we need to gain a factor 100 in the number of particles compared with today's numerical simulation like

DEUS simulation where only all halos (where forms galaxy clusters) with a mass larger than 10^{14} solar masses are described with at least 100 particles. How such a simulation is possible ?

In principle, a simulation with a number of particles 100 times larger than in DEUS simulations could be performed with a supercomputer with 8 million cores (the thin partition of Curie Computer has 80 000 cores) and up to 30 PB of Memory (exascale supercomputer) (the thin partition of Curie Computer has 320 TB of memory). However, with such a large system with such number of cores, a new deep difficulty appears concerning the fault-tolerance of our application. For such a large system the probability of failure during the run becomes then non negligible. For the moment, it is still very complex to solve this kind of problems for a calculation like the ones we are carrying (gravitational N-body problem with usual algorithm for dynamical solver and usual algorithm for analysis program). In the following, we should then assume that this problem is solved!

In the figure (1) extracted from DEUS results [Alimi et al 2012], we can see that up to 4752 MPI tasks (4096^3 particles simulation), PM-AMR code (N-body/Poisson solver) reached an efficiency of around 90% compared to the reference simulation at 74 MPI processes (1024^3 particles). This exceeds the theoretical efficiency of a standard « Particle-Mesh » solver (without refinement) which solves the Poisson equation by FFT. This was due to the solver of the original version of our PM-AMR code which is based on a multi-grid method [Guillet 11]. In the case of 38016 MPI tasks, the efficiency slightly drops. However, this decrease depends on the computational phase of the system dynamics. At the beginning, when there is not refinement (the clustering of the matter is low), the efficiency is about 60 %, then it falls to 55 % when the first level of refinement is triggered. Nevertheless, the decline remains limited to about 15 % of the total computation time. After this phase, the efficiency increases all along the simulation run when the total number of cells becomes very large. It remains above 65 % for more than half of the run, reaching up to 75 % efficiency, which is nearly the ideal efficiency of a PM-FFT solver. From this figure, we can estimate the efficiency for 100 times more tasks to ~50 %. But with 100 times more tasks for simulating 100 times more particles, the increase of the resolution will lead also to greatly increase the clustering, thus the refinement, and finally to reduce the efficiency again by a factor 2 or 3. Finally we can estimate for a simulation with 100 times more particles than in DEUS run, i.e. with 50 trillions particles, the efficiency should be reduced by a factor 5. As the simulation with 0.5 trillion particles (DEUS) took about three days on Curie Supercomputer, we estimate to 15 days a simulation with 50 trillion particles on an pre-exascale supercomputer as previously configured. Such a usage time remains reasonable.

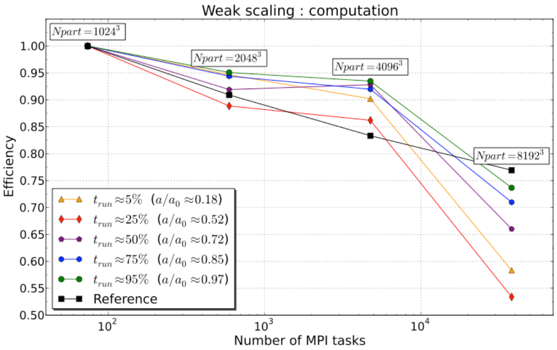

Figure 1 : Efficiency of N-body/Poisson solver as a function of the number of MPI tasks in a weak-scaling configuration. The reference corresponds to 74 MPI tasks. The efficiency is shown at the beginning of the run (yellow), at 1/4th (red), half (purple), 3/4th (blue) and at the end of the run (green). The efficiency is first of the order of 60 %, it falls to about 55 % during a short time when the first refinements are triggered and finally it increases to 75%. Multigrid acceleration allows us to reach higher efficiencies comparatively to the efficiency of an ideal PM-FFT code in black.

Memory usage is the second challenge for such an ambitious/ challenging simulation. When the number of processes is multiplied by a factor 100, it is critical to reduce as much as possible the memory usage due to MPI communication buffers and user defined communication arrays which could scale as the number of MPI tasks. Local allows limited use of memory per process, in the opposite any collective based on global point-to-point communication, when a process communicates with all other ones as well as barriers is obviously to reduce. Implementing a memory scalable N-body code is a numerical challenge especially when one deals with long range forces such as gravity which involves global communication and frequent load balancing. Given the large number of tasks, one could not avoid some memory overhead due to communications. As a consequence, a minimum of 8 GB per process seems reasonable; this is two times the memory of each cores of the Curie Thin partition. The challenge of memory usage may ultimately require a new Memory technology which will provide enough capacity and bandwidth to support such volume of data.

Finally the last but the most critical question remains the I/O and the post-processing workflow. Because in cosmology, we do not know a priori which physical observable we should measure to be sure to solve our scientific issue/problem like understanding the non linear clustering properties of dark matter or the nature of dark energy, we need to keep as much as possible all information on the distribution of cosmic matter and its evolution. In the DEUS run we finally back up data consisting of (i) 16 « snapshots » corresponding to the positions, velocities and identifiers of all particles followed during the computation, (ii) 16 « snapshots » corresponding to the positions, velocities and identifiers of all particles inside halos, (iii) 474 « samples » for the backup of a 512th of the simulation box at all computational coarse time-steps. We stored not only the particles and their properties, but also AMR cells describing the gravity field, (iv) 5 light-cones built during the dynamical computation stored at all time-steps containing the particles and the AMR grid in spherical shells around 5 observers at different space-time points. Such data represent 500 TB for each run. With a similar back up policy data for a simulation with 100 more particles, we should save at least 50 PB for each run. In fact as the clustering (and consequently the number of refinement) improves with the number of particles, over hundred PB of data should be saved. Over hundred PB for each numerical simulation seems to be prohibitive. As a matter of fact, in the DEUS experiment, the time for the I/O was comparable to the computation time (respectively about 1/3 and 2/3 of the total time of the run). To get such a performance a dynamic system of I/O delegation based on tokens has been implemented in all the parts of our application [Alimi et al 2012]. This token system, using MPI blocking instructions, and parallel I/O allowed to saturate the bandwidth allocated for our simulations: finally up to 594 simultaneous writings were allowed in the case of snapshots, whereas in the case of the samples all tasks could write at the same time. The large variation in the size of shell outputs required the use of an adaptive token system. This has been set up to the extent that at each time-step the ratio of the volume of the overall box to the shell volume defines the number of concomitant writings. A part of the first level private LUSTRE parallel file system has been dedicated to DEUS experiment: 1.7 PB with a ~60GB/s bandwidth, it was used at almost full speed: more than 40 GB/s writing during numerous periods of about half an hour and the same reading speed [Alimi et al 2012]. With similar performances for the I/O bandwidth but with a factor 100 in the number of particles, the time for the I/O should increase by a factor 100. The distribution between the time I/O and computing time is then completely reversed. It was 1/3 +2/3 = 1, it becomes 1/3 * 100 + 2/3 * 5 ~ 33. This time, I / O will represent nearly 90% of the run time. Such a situation seems totally inadequate. Efforts for increasing file system bandwidth and scalability as well as robustness must be followed otherwise this will represent the major bottleneck of future systems. Efforts for moving a first stage of the filesystem fitted into next generation of memory or flash devices at the compute nodes level must be considered. Our current and future massive simulations will need to assess new data analytics methodologies like Map/Reduce or NoSQL on top of existing filesystems (Lustre, GPFS) or through new intermediate layers able to handle large objects (like the Inria Blobseer approach).

In addition with such a quantity of data there is the difficulty of data integrity and the absolute necessity to further optimize the post-processing workflow and data stating/archiving issues. At the European level initiatives like EUDAT aims to develop a pan European collaborative data infrastructure linked to big HPC infrastructure like PRACE.

Finally, we need to consider data retention as if we need to generate as much data, it does not seem possible to keep such amount because of the storage volume, the time of writing and reading, the difficulties for post-processing the data « on the fly » during the computation, and finally because of the data integrity. This is why two options are available to us. Either we have such capability computing resources (CPU and memory) required to perform the calculation on demand whenever a specific scientific subject is covered and that the matter requires to calculate a specific physical observable. Either we have the capacity computing resources (CPU and memory) to perform not a simulation with 100 times more particles, but a simulation of comparable size as we do today, maybe with a factor of 2 or 3 on the number of particles, but we would realize such a simulation thousands of times with a different set of initial conditions. Accurate measurement on large spatial scales (the size of the box computing) of a given physical observable then will result from the statistical average of the measures obtained in each simulation. This last option however requires an optimization of the new post-processing programs to allow combining the results of analyzes of 1000 or more simulations. This will represent not Exascale computing but much more Extreme Computing.

Acknowledgements : The author wishes to thank S. Requena (CTO Genci), Y. Rasera (LUTh) and V. Bouillot (LUTh) for the discussions that led writing of this white paper

References :

[Alimi et al 2012], Alimi, J.-M. et al. ; IEEEComputer Soc. Press, CA, USA, SC2012, Article No 73.
[Angulo et al. 2012], Angulo, R. E., Springel, V., White, S. D. M., et al. 2012, arXiv:1203.3216
[Copeland et al. 2006], Copeland E. J., Sami M., Tsujikawa S., 2006, Int. J. Mod. Phys. D, 15, 1753
[BOSS], http://cosmology.lbl.gov/BOSS
[DES], http://www.darkenergysurvey.org
[Kim et al. 2011], Kim J. et al 2011, Journal of The Korean Astronomical Society, 44.6.217.
[LSST], http://www.lsst.org
[EUCLID] EUCLID mission, http://sci.esa.int/euclid
[Peebles80], Peebles P. J. E., 1993, Principles of Physical Cosmology. Princeton University Press, Princeton, NJ
[Springel et al. 2005], Springel, V., White, S. D. M., Jenkins, A., et al. 2005, nat, 435, 629