8th. World Congress on Computational Mechanics (WCCM8)
5th European Congress on Computational Methods in Applied Sciences and Engineeering (ECCOMAS 2008)
June 30 –July 5, 2008
Venice, Italy

A 21 billion degrees of freedom, 2.5 terabyte simulation of seismic wave propagation in the inner core of the Earth on MareNostrum

*** Dimitri Komatitsch[1,2], Jesús Labarta[3] and David Michéa[1]**

[1] CNRS, Laboratoire de Modélisation et Imagerie en Géosciences de Pau
Université de Pau and INRIA Magique3D, France
dimitri.komatitsch@univ-pau.fr, web.univ-pau.fr/~dkomati1

[2] Institut universitaire de France, 103, boulevard Saint-Michel, 75005 Paris, France

[3] Barcelona Supercomputing Center
Technical University of Catalonia,
c/ Jordi Girona, 31 - 08034 Barcelona, Spain

Key Words: *MPI, load balancing, cache misses, mesh partitioning, Spectral-Element Method, Seismic Wave Propagation*

ABSTRACT

Modeling of seismic wave propagation resulting from large earthquakes in the three-dimensional (3D) Earth is of considerable interest in seismology because analyzing seismic wave propagation in the Earth is one of the few ways of studying the structure of the Earth's interior, based upon seismic tomography. The field of numerical modeling of seismic wave propagation in 3D geological media has significantly evolved in the last few years due to the introduction of the spectral-element method (SEM), which is a high-degree version of the finite-element method that is very accurate for linear hyperbolic problems such as wave propagation, having very little intrinsic numerical dispersion. In addition, the mass matrix is exactly diagonal by construction, which makes it much easier to implement on parallel machines because no linear system needs to be inverted. Here we implement the SEM on MareNostrum, the world's number 9 supercomputer and fastest supercomputer in Europe (as of the June 2007 Top500 list of supercomputers, `www.top500.org`), which is located in Barcelona, Catalonia, Spain. We show that on 2166 of its IBM PowerPC 970 processors we can simulate seismic waveforms accurately up to a maximum frequency of 0.5 Hertz based upon message passing with MPI.

The SEM combines the flexibility of the finite-element method with the accuracy of the pseudospectral method. It uses a mesh of hexahedral finite elements on which the wave field is interpolated by high-degree Lagrange polynomials at Gauss-Lobatto-Legendre (GLL) integration points [1, 2, 3].

Here we are interested in differential effects on very high frequency (0.5 Hertz) seismic phases when they propagate inside the solid inner core of the Earth, therefore to significantly reduce the computational cost we suppress the crust of the Earth and replace it with an extended upper mantle, and convert the whole mantle from elastic to acoustic, thus reducing the problem in that part of the model from a vectorial unknown to a scalar unknown, i.e. reducing memory usage and CPU cost by a factor of

roughly three in 3D. In the acoustic mantle and crust we solve the acoustic wave equation in terms of a fluid potential [3]. We keep a (much more expensive to solve) elastic anisotropic medium in the inner core only. In that small part of the mesh we also model seismic attenuation (i.e., loss of energy by viscoelasticity), which has a significant impact on the cost of the simulation because memory requirements increase by a factor of roughly 2 and CPU time by a factor of roughly 1.5 [3].

The total number of spectral elements in this mesh is 323 million, which corresponds to a total of approximately 21 billion global grid points (the 'equivalent' of a $2770 \times 2770 \times 2770$ grid), since each spectral element contains $5 \times 5 \times 5 = 125$ grid points but with points on its faces shared by neighboring elements. This in turn also corresponds to approximately 21 billion degrees of freedom because a scalar unknown is used in most of the mesh (everywhere except in the inner core of the Earth, as mentioned above). Such simulations use a total of approximately 2.5 terabytes of memory. The mesh files, created once and for all by our in-house parallel mesh generator, are stored on the local disk of each blade of MareNostrum to avoid overloading the parallel GPFS file system with very large files that are only written once and then read back once at the beginning of the simulation.

Our SEM solver is based upon a pure MPI implementation. We first performed a ParaVer analysis of the code on 96 processors (Figure 1). The figure shows that in version 3.6 of SPECFEM3D, the number of L2 cache misses was very different between mesh slices, thus inducing severe load imbalance. In version 4.0, L2 cache misses have been drastically reduced and very well balanced. The number of instructions executed is also very well balanced. As a result, useful duration of the calculations is well balanced as well. In total, we gain a huge factor of 3.3 in terms of wall-clock time. This shows that the IBM PowerPC 970 is very sensitive to cache misses because the same run performed on an Intel Itanium and also on an AMD Opteron cluster shows a factor of 'only' 1.55 to 1.60.

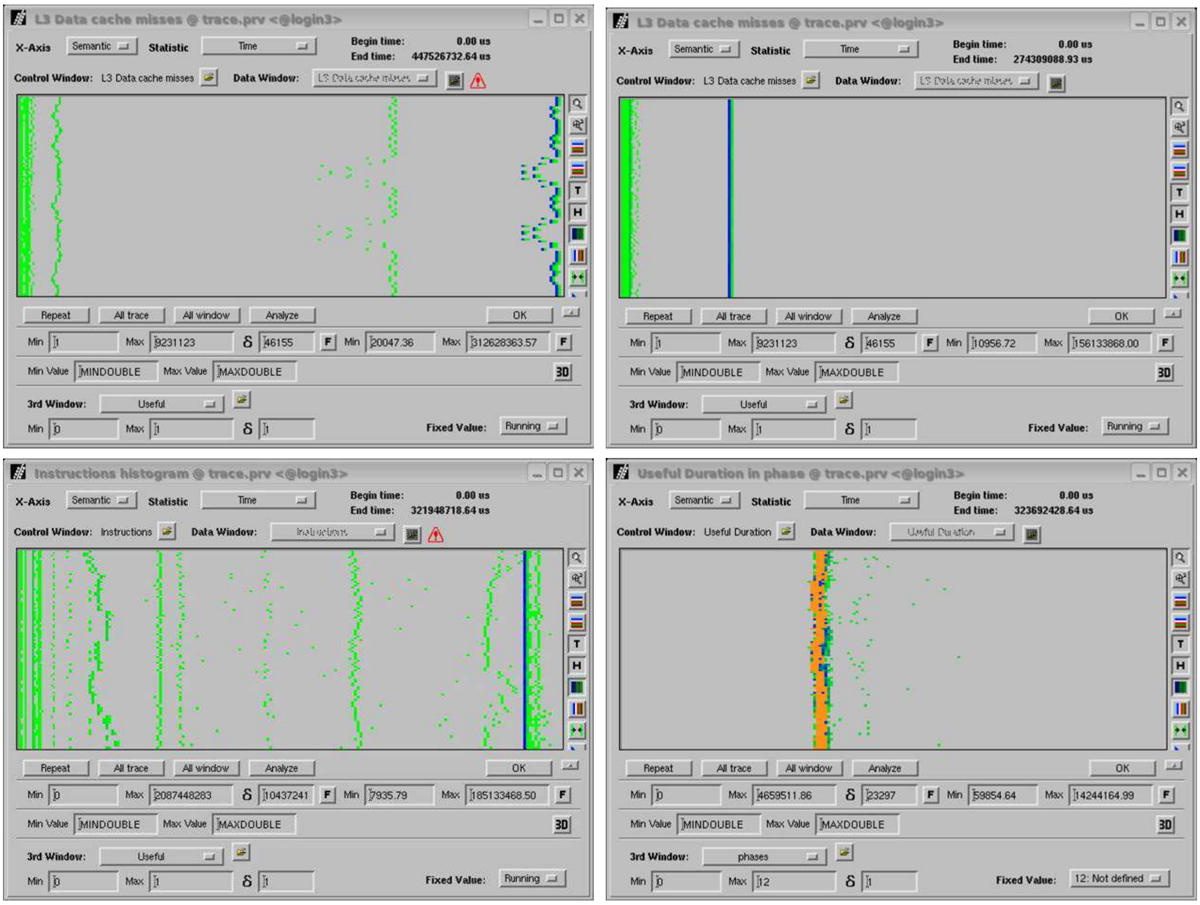
MareNostrum in Barcelona, Catalonia, Spain, has 2560 two-biprocessor blades, for a total of 10240 processor cores. Each blade has 8 gigabytes of memory, for a total of 20480 gigabytes of memory. Measured peak and sustained performance in the June 2007 Top500 list of supercomputers (`www.top500.org`) are $R_{\mathrm{peak}} = 94$ teraflops and R_{max} = 62 teraflops for the LINPACK benchmark. For the final runs, we computed 50600 time steps of the explicit time integration scheme of the SEM algorithm in 60 hours of wall-clock time on 2166 processors (being the only user running on the corresponding dedicated blades). Total memory used was 2.5 terabytes. The code performed well and performance levels obtained were very satisfactory. The geophysical analysis of the seismograms is now under way.

References

[1] Komatitsch, D., Vilotte, J.P.: The spectral-element method: an efficient tool to simulate the seismic response of 2D and 3D geological structures. Bull. Seismol. Soc. Am. **88**(2) (1998) 368–392

[2] Komatitsch, D., Tromp, J.: Introduction to the spectral-element method for 3-D seismic wave propagation. Geophys. J. Int. **139**(3) (1999) 806–822

[3] Komatitsch, D., Tromp, J.: Spectral-element simulations of global seismic wave propagation-I. Validation. Geophys. J. Int. **149**(2) (2002) 390–412



L3 Data cache misses @ trace.prv <@login3>
X-Axis Semantic
Statistic Time
Begin time: 0.00 us
End time: 447526732.64 us
Control Window: L3 Data cache misses
Data Window: L3 Data cache misses
Min 1 Max 9231123 δ 46155 Min 20047.36 Max 312628363.57
Min Value MINDOUBLE Max Value MAXDOUBLE
3rd Window: Useful
Min 0 Max 1 δ 1
Fixed Value: Running

L3 Data cache misses @ trace.prv <@login3>
X-Axis Semantic
Statistic Time
Begin time: 0.00 us
End time: 274309088.93 us
Control Window: L3 Data cache misses
Data Window: L3 Data cache misses
Min 1 Max 9231123 δ 46155 Min 10956.72 Max 156133868.00
Min Value MINDOUBLE Max Value MAXDOUBLE
3rd Window: Useful
Min 0 Max 1 δ 1
Fixed Value: Running

Instructions histogram @ trace.prv <@login3>
X-Axis Semantic
Statistic Time
Begin time: 0.00 us
End time: 321948718.64 us
Control Window: Instructions
Data Window: Instructions
Min 0 Max 2087448283 δ 10437241 Min 7935.79 Max 185133468.50
Min Value MINDOUBLE Max Value MAXDOUBLE
3rd Window: Useful
Min 0 Max 1 δ 1
Fixed Value: Running

Useful Duration in phase @ trace.prv <@login3>
X-Axis Semantic
Statistic Time
Begin time: 0.00 us
End time: 323692428.64 us
Control Window: Useful Duration
Data Window: Useful Duration
Min 0 Max 4659511.86 δ 23297 Min 59854.64 Max 14244164.99
Min Value MINDOUBLE Max Value MAXDOUBLE
3rd Window: phases
Min 0 Max 12 δ 1
Fixed Value: 12: Not defined

Figure 1: ParaVer analysis of the code on 96 processors, from processor 1 at the top of each picture to processor 96 at the bottom. Top left: In version 3.6 of SPECFEM3D, L2 cache misses were very poorly balanced between mesh slices, thus inducing severe load imbalance. In version 4.0 (top right), L2 cache misses (represented on the same horizontal scale) have been drastically reduced and very well balanced. The number of instructions executed is also very well balanced (bottom left, blue line). As a result, useful duration of the calculations (bottom right, in orange) is well balanced as well.