

Summary of AWM simulations for Blue Gene Watson Consortium Days

Project Title:

“Benchmarking Very-Large Scale Earthquake Wave Propagation Simulation with 32 Billion Mesh Points and Very-High 100m Resolution on BGW”

Investigator(s):

Dr. Yifeng Cui, Computational Scientist, San Diego Supercomputer Center

Dr. Kim Bak Olsen, Associate Professor, San Diego State University

Dr. Steven Day, Professor, San Diego State University

Dr. Bernard Minster, Professor, Scripps Institution of Oceanography, University of California, San Diego

Dr. Thomas Jordan, Professor and Director of Southern California Earthquake Center, University of Southern California

1. Goals

The Southern California Earthquake Center Community Modeling Environment (SCEC/CME) collaboration is currently engaged in an active research program in earthquake system science that seeks to develop and apply the cyberinfrastructure needed for better understanding of earthquake processes. The SCEC/CME collaboration has developed a computational framework that supports large scale simulations of earthquake processes on high performance computers. In order to advance the SCEC earthquake modeling efforts and solve larger size problems termed as PetaShake earthquake problems, the SCEC HPC codes must scale up from the current terascale to future petascale machines to be developed and deployed in the next few years.

We have two primary goals on BGW. One is to demonstrate/investigate the scalability of our AWM code, a fourth-order finite differences application for earthquake wave propagation simulation, on a very large number of mesh points (32 billion) and a very high resolution of 100m, which is a factor of 18 larger than current TeraShake investigation, the largest and most detailed simulation on Southern San Andreas Fault today. The other is to perform the first of many anticipated PetaShake simulations, for a 'wall-to-wall' rupture scenario region (Parkfield to the Salton Sea, outer scales ~800km), with waveforms valid for high frequency of 0.75 Hz (inner scales ~150m) that are relevant to a higher percentage of multi-story buildings in Southern California.

2. Experiment Results of Scalability

For BGW day, we have achieved our first goal of demonstrating the scalability of AWM code up to 40,960 processors. This alone has been a breakthrough in the field of earthquake ground motion simulation. The ability to benchmark the code on this scale and prove that such a calculation can be done has redefined the benchmark in terms of the timescale for computational seismology and is of enormous benefit to the seismological community.

On BGW day, we began with a test run on 1 rack. The executable, compiled in advance at SDSC BG/L, was used for the benchmark simulations without recompilation. The successful test immediately allowed us to scale the application to run on 4 and 8 racks. Well-prepared benchmark package made it possible for us to run many cases within limited time, in both virtual node (VN) and co-processor (CO) modes. After the successful run on 16 racks during the morning session, we were given 20 minutes allocation on 20 racks in the early afternoon. We performed both strong scaling and weak scaling runs. The problem sizes for strong scaling runs are 2048^3 (8.59e9 mesh pts), PetaShake-100m (8000x4000x1000 mesh nodes (32e9 mesh pts), and PetaShake-150m (5312x2656x520 mesh nodes (7.33e9 mesh pts). The weak scaling are on 150^3 km and 800km x 400km x 100km with four different grid resolutions respectively. The timing results are presented in Table 1 and 2 respectively.

Table 1: BGW Strong Scaling Measurements (computing time in seconds)

Nr of processors	2048 ³	PS-150m	PS-100m
2048	-	-	-
4096	3.175	2.919	-
8192	1.68	1.417	-
10240	-	-	-
16384	0.84	0.717	3.132
20480	-	-	-
32768	0.425	0.375	1.56
40960	0.411	0.313	1.3

Table 2: BGW Weak Scaling Measurements

Problem Size 1	150x150x150km			
Nr of processors	Initial	computing	grid resolution	mode
512	1.644	1.286	200m	vn
4096	2.47	1.267	100m	co
13824	9.718	1.524	67m	vn
32768	45.18	1.296	50m	vn
Problem Size 2	800x400x100km			
512	1.919	1.613	400m	vn
4096	2.978	1.669	200m	co
13824	10.046	1.729	133m	vn
32768	46.39	1.62	100m	vn

Figure 1 demonstrated impressive parallel efficiency of 96% on full system with AWM code on PetaShake domain with a high resolution of 100m. The weak scaling also shows nearly linear up to 32,768 processors (Fig. 2). AWM code achieved 6.1 Teraflop/s on BGW on PetaShake domain with a resolution of 100m. We believe there is still room for further improvements through single-processor optimization.

Previous scaling results on different computing resources, including IBM Power4 Datastar, TeraGrid IA-64, and BG/L at SDSC, indicated that its good scalability will reach a limit at certain point, in particular when interconnect performance is poor. AWM requires point-to-point communication and uses MPI_Barrier operation. Blue Gene's 3-D torus communication network and extremely fast MPI_Barrier operation, assisted by hardware, help achieve perfect scaling. In addition to this, BGW dedicated compute nodes to running user application also made the

difference. BGW compute nodes run a very lightweight kernel, and implement only the strictly necessary functionality, which minimize any perturbation of running process.

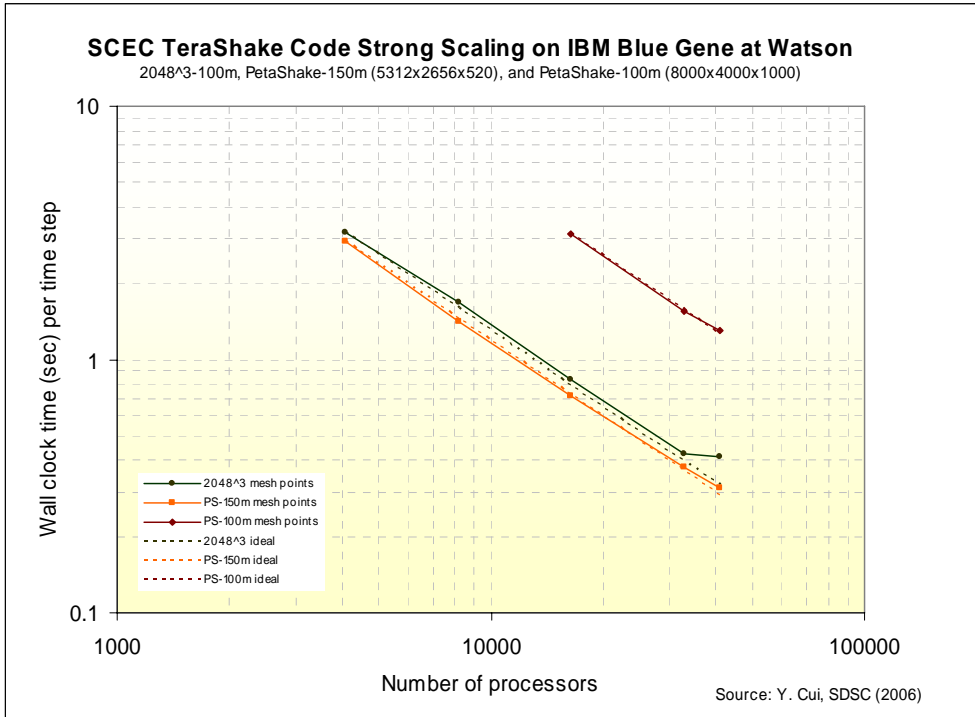


Figure 1: SCEC TeraShake AWM Code Strong Scaling on BGW

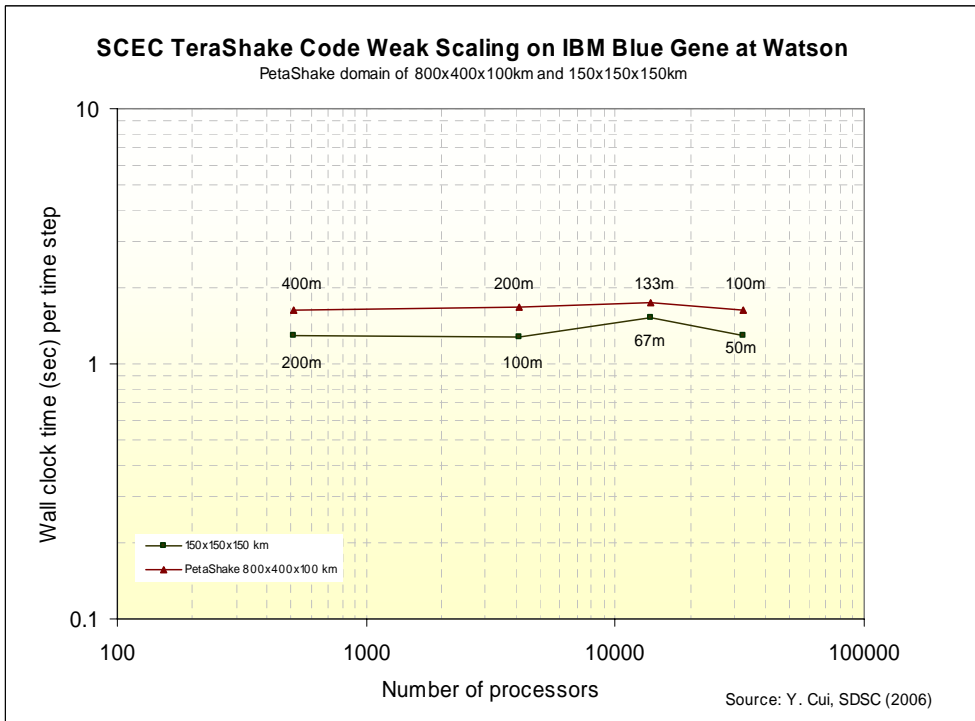


Figure 2: SCEC TeraShake AWM Code Weak Scaling on BGW

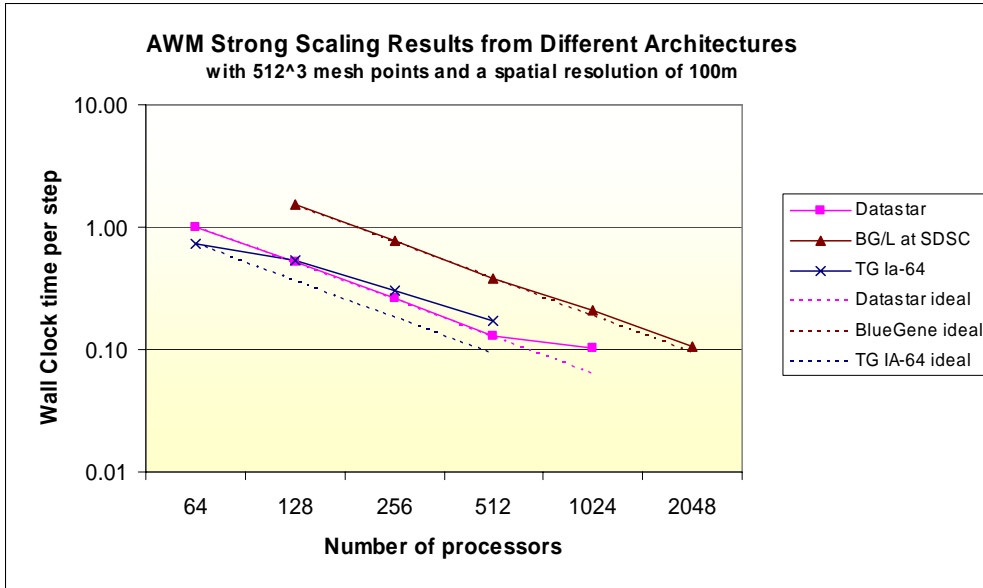


Figure 3: SCEC TeraShake AWM Code Scaling Comparisons on Different Architectures

3. I/O Performance and Issues on BGW

AWM testing on BGW was designed to determine whether current initialization approach will be feasible for future production simulations of very large earthquakes using large number of processors. The planned PetaShake 150m science simulation on BGW was designed to archive three 161 MB surface velocity outputs every 20 time steps. The outputs are accumulated and dumped out every 2000 time steps (16 GB written from 512-1024 processors), and the total size of the outputs is 1.2 TB for 48,193 time steps.

Table 3: BGL Initialization (I/O read) Times with point source and homogenous media

Nr of processors	2048^3	PS-150m	PS-100m
2048	-	-	-
4096	-	2.47	-
8192	3.80	3.02	-
10240	-	-	-
16384	10.70	10.872	-
20480	-	-	-
32768	42.733	43.138	43.003
40960	67.234	66.187	66.537

The initialization using point source on BGW (Tab. 3) showed that the read performance is comparable to those measured on SDSC BG/L GPFS, even though BGW has smaller I/O bandwidth. After 1-rack run with MPI-IO velocity outputs successfully written, the later runs on 4 or more racks didn't produce any velocity outputs. We managed to narrow down the source of failure to MPI-IO, after recompiling the code in debugging mode and tracking at run-time. This leads to speculation that there is compiler issue or file system issue on BGW. We provided BGW scientists with a test code after BGW day for debugging purpose. We hope to see this problem resolved soon to enable the real PetaShake simulation at BGW.

Prior to BGW day, SCEC scientists prepared necessary inputs for the PetaShake 150m science simulation, including 40,960 sub-domain velocity model files and sub-domain source files. This was to target the science run completing within 5-7 hours, reducing initialization time from many tens of hours to 1-2 hours on BGW.

Input data transfer from SDSC/NCSA to BGW has been an issue during BGW day. We note that BGW has somewhat limited support for file transfer. The SCEC science run requires 235 GB input data to be pre-loaded on the BGW staging server. The staging server has a USB connection (v1.1) (~1.5MB/s) which made it difficult to use USB hard drive for data transfer. BGW only accepts LTO Ultrium3 tape, but was unable to read the tape created on SGI, shipped from NCSA after the data had been transferred from SDSC to NCSA. We propose to work with IBM Watson to develop a LTO format description that we can provide to our TeraGrid support staff so that our NCSA collaborators can write a tape format that is readable by the BGW system. This description of how to write a BGW compatible LTO tape may then be useful to other research groups trying to use the BGW system.

4. Conclusions and Future Work

Thanks to IBM TJ at Watson research center for providing access to the fastest open supercomputer in the world and a tremendous service, SCEC AWM application was able to demonstrate excellent scaling on BGW, achieving parallel efficiency of 96% on 40,960 processors.

Our scaling study clearly illustrates that it is possible to efficiently utilize 40k processors for Geoscience applications. The study also proves that the performance of BlueGene/L is promising. These results suggest that SCEC AWM application could benefit tremendously from access to BGW resource with a large number of processors for PetaShake simulations.

We were unable to perform PetaShake science simulation during BGW day. Based on the strength of our results, we propose that 40,960 processors of the BGW be dedicated for approximately 7 hours to the PetaShake 150m simulations in the near future, once BGW compiler or file system issue is resolved.

5. References

- [1] Olsen, K., Day, S.M., Minster, J.B., Cui, Y., Chourasia, A., Faerman, M., Moore, R., Maechling, P. and Jordan, T., 2006, Strong Shaking in Los Angeles Expected From Southern San Andreas Earthquake, *Geophysical Research Letters*, 33(7), 1-4.
- [2] Cui, Y., K. Olsen, Hu, Y., Day, S.M., Dalguer, L., Minster, J.B., Moore, R., Zhu, J., Maechling, P. and Jordan, T., 2006, Optimization and Scalability of a Very-large Scale Earthquake Simulation Application, *Eos, Trans., AGU*, 87(52), *Fall Meeting Suppl.*, Abstract S41C-1351.