



Extreme-scale earthquake simulations on Sunway TaihuLight

Haohuan Fu¹ · Bingwei Chen² · Wenqiang Zhang³ · Zhenguo Zhang⁴ · Wei Zhang⁴ · Guangwen Yang² · Xiaofei Chen⁴

Received: 10 November 2018 / Accepted: 28 February 2019 / Published online: 5 April 2019
© China Computer Federation (CCF) 2019

Abstract

Earthquakes, as one of the most disruptive natural hazards, have been a major research target for generations of scientists. Numerical simulation of earthquakes, as one of the few methods to verify and improve scientists' understanding about the earthquake process, and a key tool in various earthquake engineering applications, has long been both an important and challenging application on supercomputers. In this paper, we discuss the major challenges for developing an accurate earthquake simulation tool on supercomputers. Based on the discussion, we then demonstrate our efforts on performing extreme-scale earthquake simulations on Sunway TaihuLight, a 125-Pflops machine with over 10 million heterogeneous cores. With systematic approaches to resolve the memory bandwidth constraint, we manage to achieve 8% to 16% efficiency for utilizing the entire machine to simulate Tangshan and Wenchuan Earthquakes with an unprecedented spatial resolution.

Keywords Sunway TaihuLight · Computational seismology · Earthquake ground motions · Parallel scalability · Neural nets

1 Introduction

Earthquake, for both the tremendous damage that it can cause and the complexities involved in understanding its formation and triggering process, is one of the ultimate scientific puzzles that scientists have been working on for generations.

In Chinese history, the study of earthquakes dated back to a famous scholar, Zhang Heng, (AD78 to AD179) Stein and Wysession (2009), who designed a delicate seismoscope that consisted of eight dragons holding balls in their mouths, and eight toads sitting beneath. With a ball falling from the dragon mouth to the toad, the seismoscope was able to indicate an earthquake that happened thousands of miles away. While designed thousands of years ago, Zhang's seismoscope already exhibited technical features similar to modern seismic measurement instruments originated in the 1880s (Milne 1886).

This work was supported in part by the National Key R&D Program of China (Grant No. 2017YFA0604500), by the National Natural Science Foundation of China (Grant Nos. 91530323, 5171101179).

✉ Haohuan Fu
haohuan@tsinghua.edu.cn
Bingwei Chen
cbw15@mails.tsinghua.edu.cn
Wenqiang Zhang
wqzhang1@mail.ustc.edu.cn
Zhenguo Zhang
zhangzg@sustech.edu.cn
Wei Zhang
zhangwei@sustech.edu.cn
Guangwen Yang
ygw@tsinghua.edu.cn
Xiaofei Chen
chenxf@sustech.edu.cn

- ¹ Ministry of Education Key Laboratory for Earth System Modeling and Department of Earth System Science, Tsinghua University, National Supercomputing Center in Wuxi, Wuxi, China
- ² Department of Computer Science and Technology, Tsinghua University, National Supercomputing Center in Wuxi, Wuxi, China
- ³ School of Earth and Space Sciences, University of Science and Technology of China, Hefei, China
- ⁴ Department of Earth and Space Sciences, Southern University of Science and Technology (SUSTech), Shenzhen, China

Ever since then, science and technology have evolved along the way to provide significantly more accurate measurement devices to record the seismic events across the globe. However, on the other hand, direct detection of the subsurface is still only at the range of tens of thousands of meters, leaving the rest over 99% of the earth still unseen.

Therefore, numerical simulation of the earthquake, which provides an important experimental platform for seismologists, demonstrates important functions in at least three aspects:

- as a forward modeling engine, earthquake simulation provides an important tool to produce potential scenarios of specific earthquakes happening at specific locations,
- as a test engine, the forward simulation verifies, and potentially improves scientists' hypothesis about earthquake happening processes, as well as the underlying structures of the earth,
- when coupled with engineering tools, the simulation engine can serve as an important guidance on earthquake engineering-related design and policy making processes.

While the numerical simulation approach presents multi-fold scientific and engineering values, it is never an easy task, even on the most powerful supercomputers in the world. In this paper, we would first start the discussion with the major challenges for enabling high-fidelity earthquake simulation on parallel computers. We then demonstrate our efforts on designing a highly-scalable and highly-efficient earthquake simulation framework on Sunway TaihuLight, which efficiently utilizes the entire set of over 10 million heterogeneous cores to simulate extreme-scale earthquake scenarios with up to 8% to 16% efficiency of the peak. The toughest challenge comes from the memory bandwidth constraint, as Sunway TaihuLight provides relatively limited memory capacities for earthquake simulation, which is a memory-bound application.

The resulting software managed to simulate both the Tangshan and the Wenchuan Earthquakes with an unprecedented level of details, demonstrating great potential to enable more exciting seismology research in the next few years.

2 Simulating earthquakes: challenges

Historically, earthquake simulation has been a major challenge for different generations of supercomputers (Bao et al. 1996; Komatitsch et al. 2003; Cui et al. 2010). Numerical simulation of earthquakes demonstrate at least the following difficulties:

- **Memory requirements** A destructive earthquake normally covers a region in the scale of hundreds of kilo-

meters. Depending on the size and shape of the fault, the complete domain of the simulation scenario can vary significantly along different dimensions. A general case usually covers a plane area of a few hundred kilometers by a few hundred kilometers, and 50 to 100 km along the vertical axis. Take the Tangshan Earthquake for example, we are considering a 3D domain in the size of 320 km by 312 km by 40 km. To fulfill basic earthquake engineering analysis purposes, we need a spatial resolution of 20 m to support a frequency range up to 10 Hz. In such a scenario, the simulation involves 562.5 billion grids, 2.25 trillion unknowns, and roughly 150 TB of memory space. For the 1-PB memory system of Sunway TaihuLight, further improvement of the spatial resolution would only be possible with integration of compression schemes (Fu et al. 2017).

- **Compute requirements** To capture the complex non-linear behavior of ground motions in earthquakes, the current methods also involve a high complexity. For each spatial grid, we normally have 30 to 40 variables and around 500 floating-point arithmetic operations. For a similar scenario (300 km by 300 km by 50 km in problem size, and 20-m spatial resolution) as mentioned above, a complete modeling process (100,000 time steps to finish a simulation of 100 s) means a total operation count of 100 exa floating-point operations. Even on the most powerful systems in the world, such a computation requires weeks to months to accomplish (Fig. 1).
- **I/O requirements** For simulations that happen at the scale of the full machine, normal I/O operations can also become tough challenges. A complete earthquake simulation process involves hundreds of thousands of timesteps, which translate to weeks or months of run time even on some of the largest supercomputers in the world.

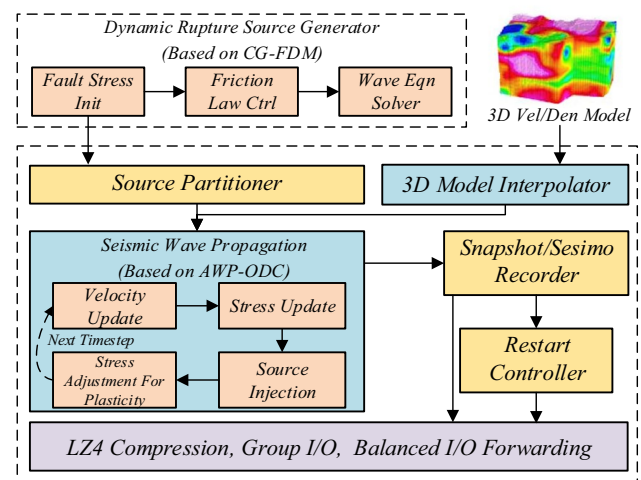


Fig. 1 The general structure of our earthquake simulation framework on Sunway TaihuLight (Fu et al. 2017)

On the other hand, when running at full scale or half scale of the large supercomputers, the MTBF (mean time between failures, usually hardware failures) is usually a day or several days (in our case, when running earthquake simulation on Sunway TaihuLight, we observe a MTBF of 18 to 19 h for full-scale runs, and a MTBF of 4 to 5 days for half-scale runs). Checkpoints that enable restart of the simulation are therefore needed. When running at the full scale of Sunway TaihuLight, the entire memory space is around 1 petabytes. Storing a minimum set that can restart the simulation is around 100 terabytes. The size and scale present both capacity and throughput challenges for I/O experts.

- **Multi-disciplinary brainpower requirements** While the above challenges bring exciting research questions for computer science researchers, the biggest challenge in the process of achieving scientifically-better earthquake simulation lies in the complexity of the problem and the necessity to involve people from many different domains. While the study of earthquakes and the related interior structure of the earth is a grand scientific problem, developing a computational platform that can enable simulation of the process and analysis of the data, on the other hand, is a grand engineering challenge (a lot of aspects touched in the previous items). Therefore, any progress made along the road really requires efforts from different domains and different disciplines. The SCEC (Southern California Earthquake Center) program shows a good example for forming such a multi-disciplinary team and performing long-term research and development efforts. In our case of modeling earthquakes on Sunway TaihuLight, the supercomputing platform becomes the glue that attracts people from different organizations to contribute to a same scientific goal. However, in a general case, we are still in great demand of brains that can understand both science and engineering, both earth-

quakes and supercomputers. Such a shortage of inter-disciplinary brainpower also exist for many other domains, such as climate, life science, astrophysics, etc.

3 Sunway TaihuLight: the hardware context

While the architectural discussion is already done many times in previous literature (Fu et al. 2016a, b, 2017; Yang et al. 2016), we still present the key information here, to provide a hardware context for the following technical sections.

The major difference to consider in the context of Sunway TaihuLight is the micro-architecture of the SW26010 many-core processor, shown in Fig. 2. Each SW26010 CPU includes 4 core-groups (CGs), each of which includes one management processing element (MPE), one computing processing element (CPE) cluster with 8 by 8 CPEs, and one memory controller (MC).

While the MPE in each CG takes a similar memory hierarchy design with both L1 instruction and data caches, each of the 64 CPEs uses a 64-KB local data memory (LDM) as the scratch-pad buffer instead of data cache. Replacing the L1 data cache with a 64-KB LDM brings a completely different memory hierarchy for programmers, and requires, in many cases, a completely rethink of the memory scheme to enable any meaningful utilization of the system.

4 Framework design

While the earthquake simulation framework project on Sunway TaihuLight only started from early 2017, we are fortunate to start our development based on the established codes and collaborations with domain scientists. The first part of the framework is the source generator that produces the dynamic rupture source file for the following simulation of

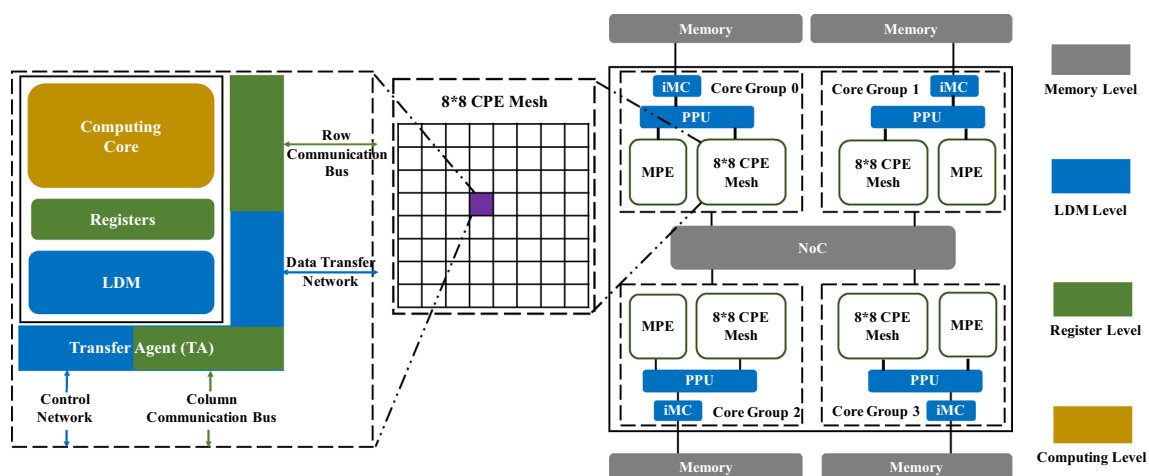


Fig. 2 The architecture of the SW26010 many-core processor

the earthquake, detailed in Sect. 4.1. The second part, which is also the more computationally-intensive part, is the forward modeling engine that simulates the propagation of seismic waves across large geographic regions, detailed in Sect. 4.2. Our current development is mainly based on two existing codes. One is the AWP-ODC (Cui et al. 2010) developed by SCEC, which is one of the most widely used simulation engine on some world-leading supercomputer facilities. The other one is the curved-grid finite-difference method (CG-FDM) (Zhang et al. 2014) developed by researchers from Southern University of Science and Technology (SUSTech).

Based on the existing scientific codes, we build a unified software framework that integrates different functions, as shown in Fig. 1, including different components that range from dynamic rupture source generation, mesh generation, to the most time-consuming wave propagation part.

4.1 Source generator

In this work, we generate the source by dynamic rupture modeling on non-planar Tangshan fault with a curved-grid finite-difference method (CG-FDM) (Zhang et al. 2014). While keeping the computational efficiency and easy implementation of conventional FDM, the CG-FDM also is flexible in modeling the complicated fault model by using general curvilinear grids. Thus, this method can model the rupture dynamics of a fault with complex geometry, such as non-planar fault, fault with step-over, even if irregular topography exists. This method has been proved to be an efficient tool in the dynamic rupture modeling by benchmarking problem test (Harris et al. 2018) and has been used in scenario earthquake simulations (Zhang et al. 2017).

The CG-FDM maps the real coordinates (x, y, z) in to a computational one (ξ, η, ζ) , and discretizes the curved fault plane by splitting nodes in such a way each one on the fault plane has two sets of values with symbols “+” and “−” representing the two sides of the fault. According the momentum equation in curvilinear system and the continuity of traction across fault plane, we have trial traction vector on fault plane:

$$\tilde{T}_v = 2 \frac{\Delta t^{-1} M^+ M^- (v_v^+ - v_v^-) + M^- R_v^+ - M^+ R_v^-}{\Delta h^2 (M^+ + M^-) J |\nabla \xi|} + T_v^0. \quad (1)$$

where

$$M_{j,k}^\pm = \frac{1}{2} J \rho_{j,k}^\pm \Delta h^3, \quad (2)$$

$$[R_v^\pm]_{j,k} = \frac{\Delta h^2}{2} \left\{ \pm [\hat{T}_{1v}(t)_{i_0 \pm 1, j, k}] + \left[\frac{\partial \hat{T}_{2v}^\pm(t)}{\partial \eta} \right]_{j,k} \Delta h + \left[\frac{\partial \hat{T}_{3v}^\pm(t)}{\partial \zeta} \right]_{j,k} \Delta h \right\}, \quad (3)$$

and Δh is the grid step in the computational domain (ξ, η, ζ) . The trial traction (Eq. 1), if applied, would lock the split nodes, and the fault would then heal. The tangential component of this trial traction is:

$$\tilde{T}_s = (\mathbf{I} - \mathbf{nn}) \cdot \tilde{\mathbf{T}}, \quad (4)$$

where $\mathbf{I}$ is the identity tensor and $\mathbf{nn}$ is the outer product of the unit vector of fault's normal direction $\mathbf{n}$ with itself, and the normal traction:

$$\tilde{T}_n = \mathbf{n} \cdot \tilde{\mathbf{T}} \cdot \mathbf{n}. \quad (5)$$

which will ensure zero separation of the split nodes. Considering that the fault may open, in which case the normal traction will disappear and will not increase to above zero, we enforce the follow condition:

$$T_n = \begin{cases} 0, & \tilde{T}_n > 0 \\ \tilde{T}_n, & \tilde{T}_n \leq 0 \end{cases}, \quad (6)$$

According to the fault boundary condition, the shear traction is bounded by the frictional strength τ_c :

$$\mathbf{T}_s = \begin{cases} \tilde{\mathbf{T}}_s, & |\tilde{T}_s| < \tau_c \\ \tau_c \frac{\tilde{\mathbf{T}}_s}{|\tilde{T}_s|}, & |\tilde{T}_s| \geq \tau_c \end{cases}. \quad (7)$$

Finally the total traction on the fault can be assembled as:

$$\mathbf{T} = T_n \mathbf{n} + \mathbf{T}_s. \quad (8)$$

With this traction, we can update slip velocity $(v_x^\pm, v_y^\pm, v_z^\pm)$ of the split nodes. The stress components also can be calculated according the velocity fields. Note that the Runge–Kutta scheme discussed will be used for the time integration. More details about the theory and running examples can refer to previous works (Zhang et al. 2014, 2017).

4.2 Forward modeling engine

In our current framework, we include two sets of forward modeling engines to support the part of wave propagation simulation.

One is the AWP-ODC code, developed by the research groups from SCEC Cui et al. (2010). The other one is the CG-FDM code developed by the SUSTech research team Zhang et al. (2014), which is also used for the source generation part.

Both the AWP-ODC and the CG-FDM code solves the same set of elastic equations:

$$\rho \mathbf{v}_{,t} = \nabla \cdot \boldsymbol{\sigma} + \mathbf{f}, \quad (9)$$

$$\boldsymbol{\sigma}_{,t} = \mathbf{c} : \boldsymbol{\epsilon}_{,t}, \quad (10)$$

where $\mathbf{f}$ is the source term. In the above equations, t is time, ρ is density, $\mathbf{v}$ is the velocity vector, $\boldsymbol{\sigma}$ and $\boldsymbol{\epsilon}$ are the stress and strain tensor respectively, and matrix $\mathbf{c}$ is the fourth order

stiffness tensor. A comma followed by t denotes the first-order derivation with respect to time.

The CG-FDM, which is recently designed to solve the seismic wave propagation in media with complex geometry, discretizes the computational volume by collocated grid in curvilinear coordinates. Because of flexibility of discretization, the CG-FDM can model the response from real features of an earthquake, such as the topography, complex fault system, and so on.

In contrast, the AWP-ODC solves the wave equations with a staggered grid system in the Cartesian coordinate. The AWP-ODC has been a popular tool with high efficiency and well optimization after developing for a long time.

Our efforts reported in this paper focus on the redesign and tuning of AWP-ODC for Sunway TaihuLight. AWP-ODC, which stands for Anelastic Wave Propagation by Olsen, Day, and Cui, develops over the year from the original finite difference code developed by Kim Olsen at University of Utah (Olsen 1994). The code was the major tool for the Community Modeling Environment (CME) of SCEC, and has scaled to various parallel computing platforms, such as TeraGrid (Cui 2007), Jaguar (Cui et al. 2010), and Titan (Cui et al. 2013; Roten et al. 2016).

The current version of AWP-ODC has already built up the plasticity simulation capabilities for nonlinear effects, which can largely improve the simulation accuracy but at the cost of more variable arrays and more computation. The next sections demonstrate the major approaches that we take to achieve AWP-Sunway, an inherited version of AWP-ODC that is completely redesigned and tuned for the Sunway TaihuLight system.

5 Parallelization and optimization

5.1 A customized parallelization design

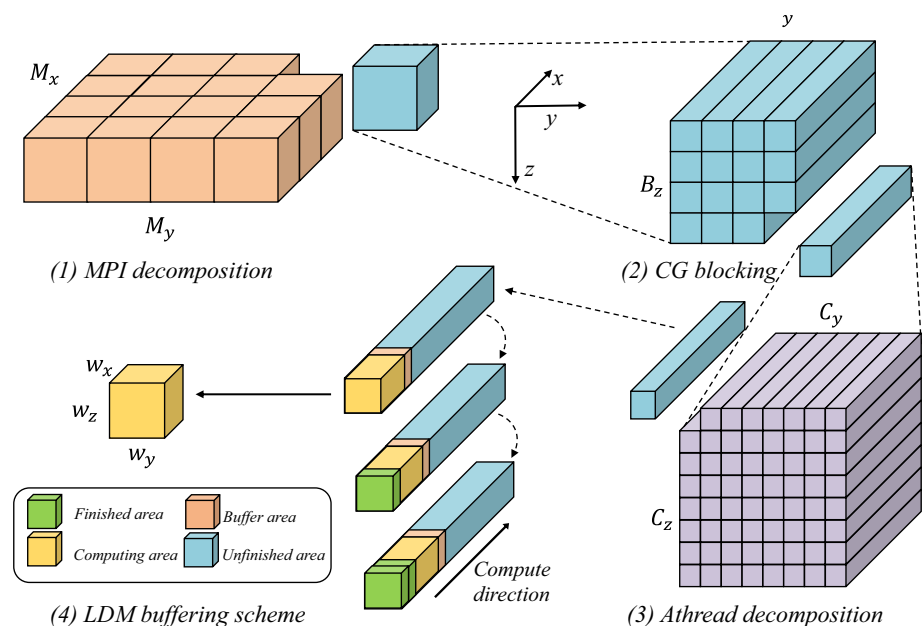
The first challenge of achieving a full-scale application on Sunway TaihuLight, is, as mentioned above, to identify a suitable mapping scheme that translate the physics into the compute instructions, data movements, and message passings among the 10 million cores in the system.

A large part of the mapping scheme is about decomposition, i.e. how we decompose the large problem into parts that sit in different nodes, and further down, in different cores. One specific complexity in the case of earthquake simulation is that the computational kernels normally involve reading and writing of over 20 variable arrays that covers the entire mesh grid. In such cases, many previous optimization techniques, such as the 3.5D blocking scheme (Nguyen et al. 2010), becomes impractical due to the extremely high memory volume requirement. As a result, our solution is a customized domain decomposition scheme that exposes enough parallelism for the 10 million cores and minimizes related memory costs at the same time.

Figure 3 shows our multi-level approach that decomposes the domain into different partitions for each MPI process, and further into different regions for each CPE thread, detailed as follows:

- (1) 2D decomposition for MPI processes: For the storage of all the 3D arrays, we take the z axis (the vertical direction) as the fastest axis, the y axis as the second, and the x axis as the slowest axis. In typical earthquake

Fig. 3 The multi-level domain decomposition scheme: **1** MPI decomposition, **2** CG blocking, **3** athread decomposition, and **4** LDM buffering scheme (Fu et al. 2017)



simulation scenarios, we generally have significantly larger dimensions for x and y (hundreds of kilometers) than the dimension for z (tens of kilometers). Therefore, to minimize communication among the different processes, at the first level, instead of taking a 3D approach, we decompose the horizontal plane into M_x by M_y different partitions, each corresponding to a specific MPI process. With the well-designed MPI scheme to hide halo communication in computation inherited from AWP-ODC (Cui et al. 2010), in extreme cases, we can have up to 160,000 (400 by 400) MPI processes to scale to the full machine.

- (2) Blocking for each CG: At the second level, instead of assigning all the mesh points to different cores within a CG, we add a blocking mechanism along y and z axes to assign a suitable size of block to the CG, so as to achieve a more efficient utilization of the 64-KB LDM of each CPE. Each CG would iterate over these different blocks to finish the processing.
- (3) 2D decomposition for Athreads: We further partition the block into different regions for each CPE, but along the y and z dimensions (with each thread iterating along the direction of x), so as to achieve fast memory accesses for the different threads.
- (4) LDM buffering scheme: For each CPE, we load a suitable size of the computing domain (both the central and the halo parts) into the LDM using DMA operations, so as to perform the computation afterwards. The DMA operations are designed to be asynchronous, so as to overlap with the computation part.

5.2 Memory-oriented optimization

To break the memory constraints, a key part of the solution is to efficiently utilize the memory hierarchy of the SW26010 processor.

The challenges are clear, a relatively low memory bandwidth, and the absence of automated hardware cache in the 64 CPEs. While the absence of automated cache brings extra efforts for programmers to make efficient utilization of the bandwidth, the introduction of user-controlled 64-KB LDM also brings the option to explore a customized memory scheme for the given algorithm or application.

The key philosophy is shown in Fig. 4. While the major volume of data items reside in the main memory, we need to load the corresponding parts into the LDM, perform the computation, and store the results back to the main memory. As a result, the programmers are exposed to the design of a specific DMA load and store scheme.

Another hardware feature that we can take advantage of is the two instruction issuing ports of each CPE. One port is specifically for compute instructions, while the other one is for DMA instructions. Therefore, a large part of design

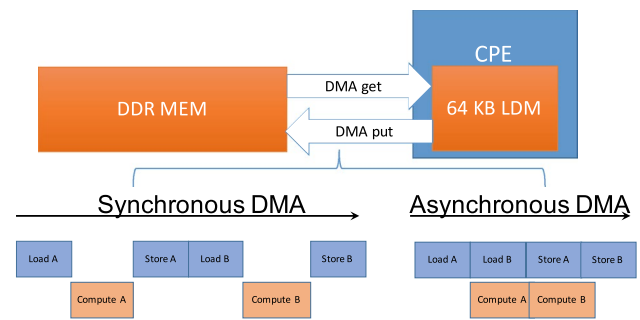


Fig. 4 The memory model for Sunway TaihuLight: designing a customized DMA memory scheme

is to achieve an asynchronous design that can overlap the compute and the DMA instructions to the maximum possible extent.

One unique feature of SW26010 is the register communication among the 64 CPEs in each CG, which provides a perfect solution for data reuse in stencil-like computations. Using register communication based halo exchange, inside each CG, the CPE thread only needs to load its corresponding central region, and can acquire the halo regions from the neighboring threads through register communication operations. Only the boundary CPE threads that need to communicate across different CGs still need to initialize DMA loads for the corresponding halo regions.

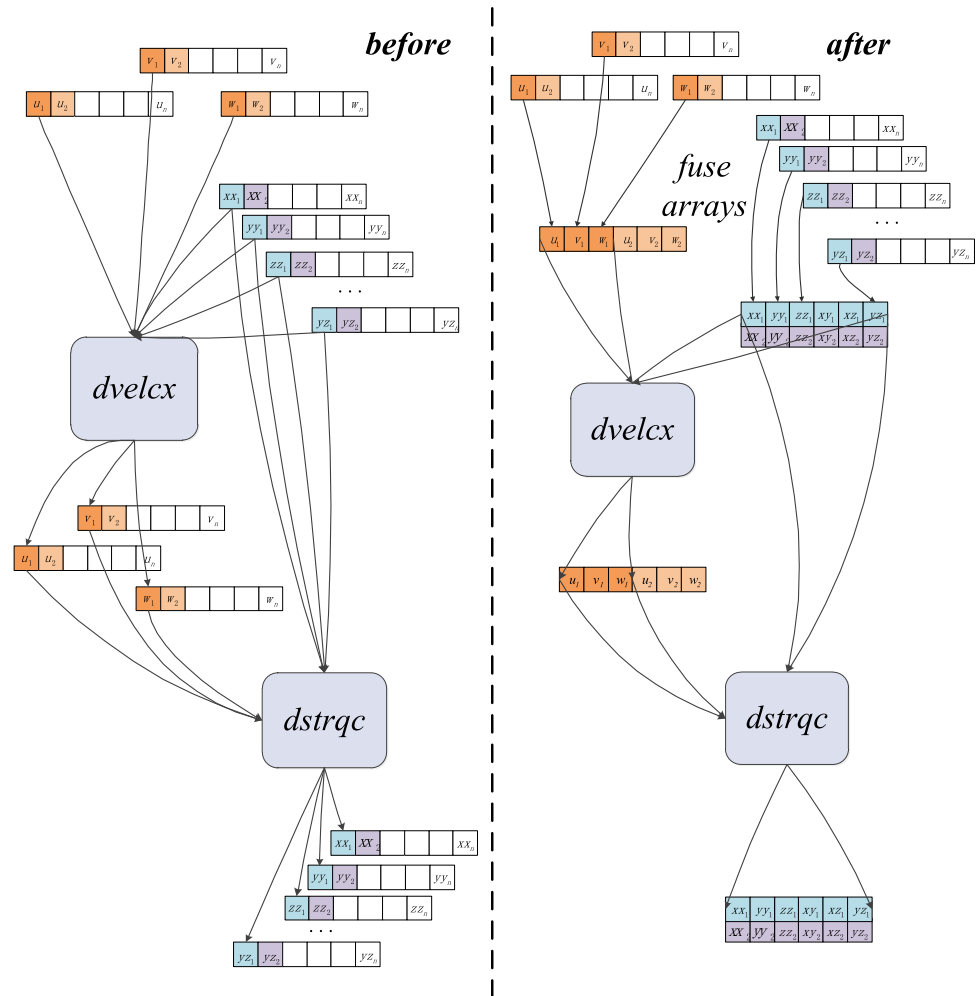
As a result, tuning the dimension parameters for the blocking scheme (such as the parameters in Fig. 3) become another key step for achieving good utilization of bandwidth. We propose an optimized blocking configuration guided by an analytic model to ensure: (1) minimize the number of DMA loads required for redundant halo region reads; (2) maximize the effective memory bandwidth by using a large chunk size.

Even after adopting the optimal parallelization scheme mentioned above, in most cases, due to the large number of arrays that we need to access, the 64 KB LDM would limit us to a too small portion of the array, and a low efficiency of the DMA reads.

To resolve the above issue, we analyze all the different related kernels, and identify a set of co-located arrays that demonstrate a common behavior for a majority of kernels. Such as shown in the example in Fig. 5, we identify the arrays (u , v , w , xx , yy , zz , xy , xz , and yz) to be the co-located arrays that demonstrate identical memory access patterns among different kernels. Therefore, we make the design strategy to fuse u , v and w , into an array of vectors with three elements, and to fuse xx , yy , zz , xy , xz , and yz into an array of vectors with six elements.

After the fusion of the arrays, with only 3 separate arrays to read, we can afford a DMA block size of 432 bytes, improving the memory bandwidth utilization to around 80%.

Fig. 5 For the kernel *dvelcx* and *dstrqc*, we fuse 3 velocity arrays and 6 stress arrays so as to increase the block size of DMA load to $\times 3$ and $\times 6$ respectively



In the extreme case of the *dstrqc* kernel, the array fusion technique could increase the DMA block size from 84 bytes to 512 bytes, improving the effective memory bandwidth from 50.47 to 104.82GB/s.

5.3 On-the-fly compression

After a systematic optimization from both the compute and memory aspects, we achieve a design that is close to the point of squeezing out the system's hardware capabilities. As the Sunway TaihuLight system, similar to many other supercomputers in the world, demonstrates an unbalanced ratio between compute and memory capacities, our method for further improvement is to shift the balancing point slightly by trading off a portion of the compute cycles to enable a compressed form of the data items when storing and moving them in memory.

Considering the features of different variables, we propose three different lossy compression schemes, with different levels of complexities and information loss, but a fixed compression ratio from 32-bit to 16-bit numbers.

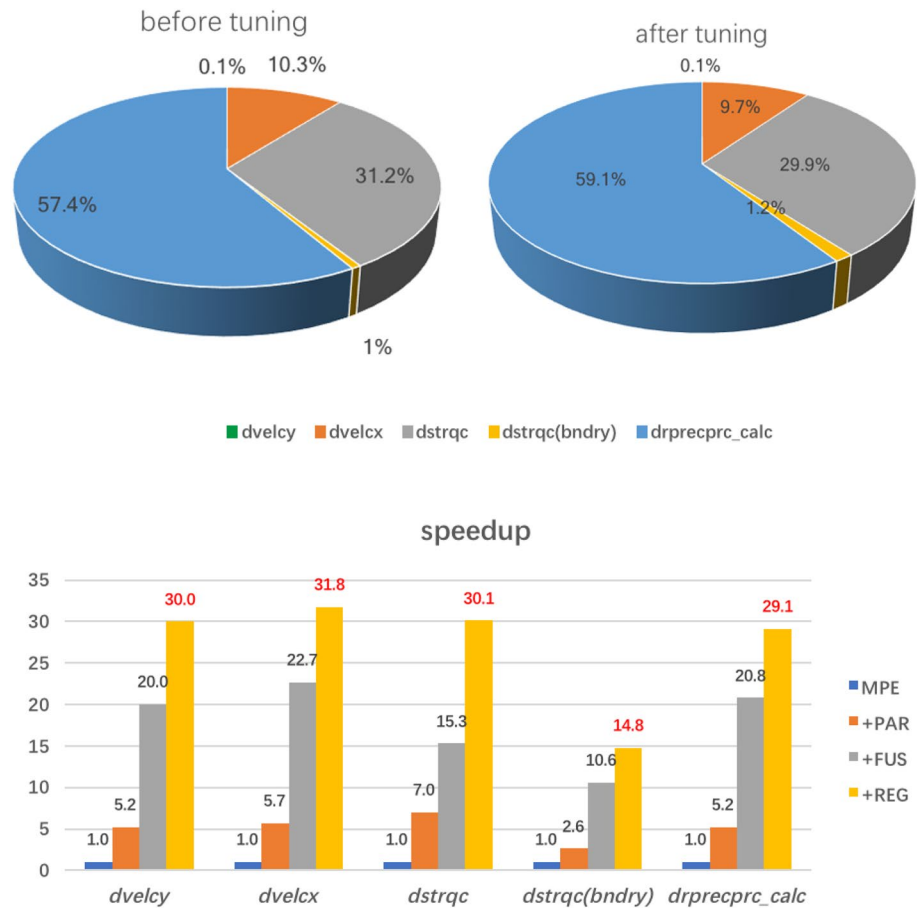
While the scheme can effectively reduce both the memory capacity and bandwidth requirement by two times, the challenge shifts to the efficiency of the scheme, so as to improve the simulation speed even though significantly more complexities are introduced to accommodate the compression operations.

With a carefully designed blocking scheme and re-scheduling of the compute and DMA instructions, we finally manage to improve the computing performance by another 24% (processing the same scenario 24% faster after adding the compression scheme), and to enable scenarios that require two times more memory space (more details in Fu et al. (2017)).

6 Results

For the earthquake simulation program, most of the cycles are consumed to update the velocity, stress, and plasticity for both the central region and the halos. Figure 6 demonstrates the performance improvements for these

Fig. 6 The speedups of different major kernels in AWP-Sunway, when applying different optimization techniques. ‘MPE’ stands for the original version that uses the MPE only. ‘+ PAR’ refers to the version that applies our specific parallelization scheme and uses both the MPE and the 64 CPEs for the computation. ‘+ FUS’ refers to the version that uses the fusion of co-located arrays. ‘+ REG’ refers to the version that further applies the register communication and CPE ID remapping techniques



different kernels when applying different approaches. We can observe that the speedups for almost all the different kernels are in the same range of around 30×. Among the different optimization schemes, the fusion of co-located arrays seems to play an important part, improving the performance by up to 4 times for the most time consuming kernels.

For both weak and strong scaling cases, we were able to scale from 520,000 cores (8000 MPI processes) to 10,400,000 cores (160,000 MPI processes) of the system. For weak scaling, we can maintain a parallel efficiency from 80 to 98% in different linear or non-linear cases. For strong scaling, the efficiency drops from 67 to 80%. For the performance at the largest scale, Table 1 demonstrates a comparison between previous work. We manage to improve the

performance to 18.9 Pflops when using over 10 million cores and the on-the-fly compression at such a parallel scale.

Using the optimized simulation software, we are able to perform a series of simulations for the 1976 Tangshan earthquake with a problem domain of 320 km by 312 km by 40 km, with the spatial resolution increasing from 500 to 8 m, supporting a frequency range up to 18 Hz. To our best knowledge, this is the first time of performing a nonlinear plasticity earthquake simulation at such a scale, and with such a high frequency and high resolution in the world. The plasticity ground motion simulation for the Tangshan earthquake allows us for the first time to quantitatively estimate the hazards of the Tangshan earthquake in the affected area, and to provide guidance for designing proper seismic engineering standards for buildings in North China.

Table 1 The largest-scale results on Sunway TaihuLight compared with previous works

	Cui et al. (2010)	Cui et al. (2013)	Roten et al. (2016)	This work
Platform	Jaguar	Titan	Titan	Sunway TaihuLight
Wave propagation	Linear	Linear	Non-linear	Non-linear
Sustained performance	220 Tflops	2.33 Pflops	1.61 Pflops	18.9 Pflops
Parallel scale	223,074 cores	16,384 GPUs 229,376 SMXs	8,192 GPUs 114,688 SMXs	40,000 SW26010 CPUs 10,400,000 cores

7 Conclusion and outlook

In this paper, based on our recent experience on Sunway TaihuLight, we summarize the major challenges for performing extreme-scale earthquake simulations on leadership supercomputers of nowadays. One key message is that memory bandwidth and capacity are the major constraints that stop scientists from performing larger or higher-resolution simulation in a faster manner. As a result, memory-related design strategies and tuning techniques become a major of our work. Using a customized parallelization scheme and a set of memory-oriented optimization methods, even on TaihuLight's relatively modest memory system (a byte-to-flop ratio that is only 1/5 of Titan), we can achieve 15.2-Pflops nonlinear earthquake simulation by using the 10,400,000 cores of Sunway TaihuLight, up to 12.2% of the peak. Our compression scheme expands our computational performance to the level of 18.9 Pflops (15% of the peak), and enables us to support 18-Hz, 8-meter simulations, which is a big jump for the previous state of the art.

While these are exciting progresses by taking advantage of cutting-edge supercomputer systems, we are still far from the ideal state that scientists would demand for a more complete simulation system. On the science side, the current large-scale simulations are usually only focused on one part of the earth, such as the scenario simulation of a specific earthquake in a specific region introduced in this paper. There are still many other efforts that focus on different scales (city-oriented scenario simulation) and different parts of the earth (geo-dynamic simulation that focus on the mantle and the core). A more complete simulation platform would need to couple these different processes at different spatial and temporal scales for a more accurate picture. On the engineering side, we would also need a coupled system with not only the ground motion simulation, but also the behaviors of buildings, hills, and other elements that could be affected. For both frontiers, there will be interesting directions that demand for decades of efforts to decipher these grand scientific challenges or to achieve major engineering breakthroughs. Along the way, both hardware and software supercomputing technologies would always be an important foundation.

Acknowledgements This work was supported in part by the National Key R&D Program of China (Grant No. 2017YFA0604500), by the National Natural Science Foundation of China (Grant No. 51761135015), and by Center for High Performance Computing and System Simulation, Pilot National Laboratory for Marine Science and Technology (Qingdao).

Compliance with ethical standards

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

- Bao, H., Bielak, J., Ghattas, O., Kallivokas, L.F., O'hallaron, D.R., Shewchuk, J.R., Xu, J.: Earthquake ground motion modeling on parallel computers. In: Proceedings of the 1996 ACM/IEEE conference on Supercomputing, IEEE Computer Society, p. 13. (1996)
- Cui, Y., Moore, R., Olsen, K., Chourasia, A., Maechling, P., Minster, B., Day, S., Hu, Y., Zhu, J., Majumdar, A., et al.: Enabling very-large scale earthquake simulations on parallel machines. In: International Conference on Computational Science, pp. 46–53. Springer (2007)
- Cui, Y., Olsen, K.B., Jordan, T.H., Lee, K., Zhou, J., Small, P., Roten, D., Ely, G., Panda, D.K., Chourasia, A., et al.: Scalable earthquake simulation on petascale supercomputers. In: High Performance Computing, Networking, Storage and Analysis (SC), 2010 International Conference for IEEE, pp. 1–20. (2010)
- Cui, Y., Poyraz, E., Olsen, K.B., Zhou, J., Withers, K., Callaghan, S., Larkin, J., Guest, C., Choi, D., Chourasia, A., et al.: Physics-based seismic hazard analysis on petascale heterogeneous supercomputers. In: Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis, ACM, p. 70. (2013)
- Fu, H., He, C., Chen, B., Yin, Z., Zhang, Z., Zhang, W., Zhang, T., Xue, W., Liu, W., Yin, W., et al.: 18.9-pflops nonlinear earthquake simulation on Sunway TaihuLight: Enabling depiction of 18-Hz and 8-m scenarios. In: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, ACM, p. 2. (2017)
- Fu, H., Liao, J., Xue, W., Wang, L., Chen, D., Gu, L., Xu, J., Ding, N., Wang, X., He, C., et al.: Refactoring and optimizing the community atmosphere model (CAM) on the sunway taihuLight supercomputer. In: High Performance Computing, Networking, Storage and Analysis, SC16: International Conference for IEEE, pp. 969–980.
- Fu, H., Liao, J., Yang, J., Wang, L., Song, Z., Huang, X., Yang, C., Xue, W., Liu, F., Qiao, F., et al.: The Sunway TaihuLight supercomputer: system and applications. *Sci. China Inf. Sci.* **59**(7), 072001 (2016)
- Harris, R.A., Barall, M., Aagaard, B., Ma, S., Roten, D., Olsen, K., Duan, B., Liu, D., Luo, B., Bai, K., et al.: A suite of exercises for verifying dynamic earthquake rupture codes. *Seismol. Res. Lett.* **89**(3), 1146–1162 (2018)
- Komatitsch, D., Tsuboi, S., Ji, C., Tromp, J.: A 14.6 billion degrees of freedom, 5 teraflops, 2.5 terabyte earthquake simulation on the earth simulator. In: Supercomputing, 2003 ACM/IEEE Conference, IEEE, pp. 4–4. (2003)
- Milne, J.: Earthquakes and other earth movements, vol. 56. D. Appleton and company, New York, USA (1886)
- Nguyen, A., Satish, N., Chhugani, J., Kim, C., Dubey, P.: 3.5-d blocking optimization for stencil computations on modern cpus and gpus. In: High Performance Computing, Networking, Storage and Analysis (SC), 2010 International Conference for IEEE, pp. 1–13. (2010)
- Olsen, K.B.: Simulation of three-dimensional wave propagation in the salt lake basin. Ph.D. thesis, Department of Geology and Geophysics, University of Utah (1994)
- Roten, D., Cui, Y., Olsen, K.B., Day, S.M., Withers, K., Savran, W.H., Wang, P., Mu, D.: High-frequency nonlinear earthquake simulations on petascale heterogeneous supercomputers. In: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, p. 82. IEEE Press (2016)
- Stein, S., Wysession, M.: An Introduction to Seismology, Earthquakes, and Earth Structure. Wiley, New York (2009)

- Yang, C., Xue, W., Fu, H., You, H., Wang, X., Ao, Y., Liu, F., Gan, L., Xu, P., Wang, L., et al.: 10M-core scalable fully-implicit solver for nonhydrostatic atmospheric dynamics. In: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, IEEE Press, p. 6. (2016)
- Zhang, Z., Zhang, W., Chen, X.: Three dimensional curved grid finite-difference method modelling for non-planar rupture dynamics. *Geophys. J. Int.* **199**(2), 860–879 (2014). <https://doi.org/10.1093/gji/ggu308>
- Zhang, Z., Zhang, W., Chen, X., Li, P., Fu, C.: Rupture dynamics and ground motion from potential earthquakes around Taiyuan. *China Bull. Seismol. Soc. Am.* **107**(3), 1201–1212 (2017). <https://doi.org/10.1785/0120160239>



Haohuan Fu is a professor in the Ministry of Education Key Laboratory for Earth System Modeling, and Department of Earth System Science in Tsinghua University, where he leads the research group of High Performance Geo-Computing (HPGC). He is also the deputy director of the National Supercomputing Center in Wuxi, and leads the research and development efforts on Sunway TaihuLight, the current fastest supercomputer in China. Fu has a PhD in computing from Imperial College London.

Since joining Tsinghua in 2011, Dr. Fu has been working towards the goal of providing both the most efficient simulation platforms and the most intelligent data management and analysis platforms for geoscience applications. His research has, for example, led to efficient designs of atmospheric dynamic solvers for both Tianhe-1A, Tianhe-2, Sunway TaihuLight supercomputers, and the reconfigurable computing platforms. The work based on reconfigurable technology can be both faster and more energy efficient than the Tianhe-1A and Tianhe-2 supercomputer, leading to a publication selected as one of the 27 most Significant Papers of the FPL conference in 25 years (27 out of 1765). He has received the 2016 and 2017 ACM Gordon Bell Prize.

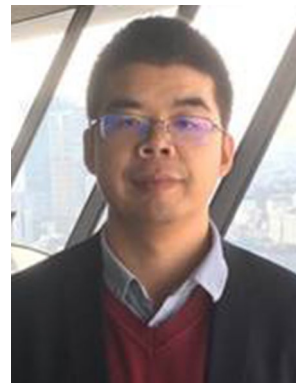


Bingwei Chen is a fourth-year PhD candidate from the Department of Computer Science and Technology in Tsinghua University. He is also a Research and Development Engineer in National Supercomputer Center in Wuxi, China. He received the 2017 ACM Golden Bell Prize, which recognizes outstanding achievement in high-performance computing applications. He does research in supercomputer-based real-world applications, including earthquake simulation, seismic exploration,

climate modeling and AI framework, on different platforms, such as Sunway TaihuLight, FPGA and GPU.



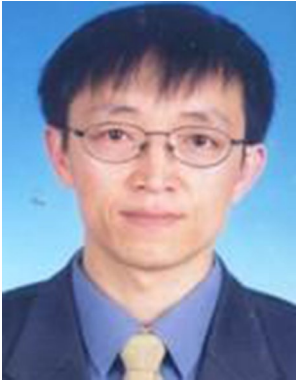
Wenqiang Zhang is a Ph.D. candidate majored Geophysics in School of Earth and Space Sciences, University of Science and Technology of China. He is working on numerical simulation of destructive earthquakes, seismic hazard estimation, and is particularly interested in the physics of earthquake source including spontaneous rupture process and earthquake cycling.



Zhenguo Zhang is an assistant professor working in Department of Earth and Space Sciences, Southern University of Science and Technology. His researches include modeling seismic wave propagation in complex media, rupture dynamics and earthquake disasters.



Wei Zhang is a professor working in Department of Earth and Space Sciences, Southern University of Science and Technology. He has been engaged in seismic wave forward modeling and inversion algorithms for a long time, and has developed new methods and technologies for strong ground motion prediction, earth structure detection and seismic exploration.



Guangwen Yang is a professor in the Department of Computer Science and Technology at Tsinghua University and the director of the National Supercomputing Center in Wuxi. His research interests include parallel algorithms, cloud computing, and the earth system model. Yang received a PhD in computer science from Tsinghua University. He has received the 2016 and 2017 ACM Gordon Bell Prize. He is a member of IEEE. Contact him at ygw@tsinghua.edu.cn.



Xiaofei Chen is a professor working in Department of Earth and Space Sciences, Southern University of Science and Technology. His main research interests are theoretical and computational geophysics and its applications in earthquake prevention and disaster mitigation and resource exploration.