SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

July 2016, Vol. 59 072001:1–072001:16
doi: 10.1007/s11432-016-5588-7

The Sunway TaihuLight supercomputer: system and applications

Haohuan FU[1,3], Junfeng LIAO[1,2,3], Jinzhe YANG[3], Lanning WANG[4], Zhenya SONG[6], Xiaomeng HUANG[1,3], Chao YANG[5], Wei XUE[1,2,3], Fangfang LIU[5], Fangli QIAO[6], Wei ZHAO[6], Xunqiang YIN[6], Chaofeng HOU[7], Chenglong ZHANG[7], Wei GE[7], Jian ZHANG[8], Yangang WANG[8], Chunbo ZHOU[8] & Guangwen YANG[1,2,3*]

[1]*Ministry of Education Key Laboratory for Earth System Modeling, and Center for Earth System Science, Tsinghua University, Beijing* 100084*, China;*
[2]*Department of Computer Science and Technology, Tsinghua University, Beijing* 100084*, China;*
[3]*National Supercomputing Center in Wuxi, Wuxi* 214072*, China;*
[4]*College of Global Change and Earth System Science, Beijing Normal University, Beijing* 100875*, China;*
[5]*Institute of Software, Chinese Academy of Sciences, Beijing* 100190*, China;*
[6]*First Institute of Oceanography, State Oceanic Administration, Qingdao* 266061*, China;*
[7]*Institute of Process Engineering, Chinese Academy of Sciences, Beijing* 100190*, China;*
[8]*Computer Network Information Center, Chinese Academy of Sciences, Beijing* 100190*, China*

Received May 27, 2016; accepted June 11, 2016; published online June 21, 2016

Abstract The Sunway TaihuLight supercomputer is the world's first system with a peak performance greater than 100 PFlops. In this paper, we provide a detailed introduction to the TaihuLight system. In contrast with other existing heterogeneous supercomputers, which include both CPU processors and PCIe-connected many-core accelerators (NVIDIA GPU or Intel Xeon Phi), the computing power of TaihuLight is provided by a homegrown many-core SW26010 CPU that includes both the management processing elements (MPEs) and computing processing elements (CPEs) in one chip. With 260 processing elements in one CPU, a single SW26010 provides a peak performance of over three TFlops. To alleviate the memory bandwidth bottleneck in most applications, each CPE comes with a scratch pad memory, which serves as a user-controlled cache. To support the parallelization of programs on the new many-core architecture, in addition to the basic C/C++ and Fortran compilers, the system provides a customized Sunway OpenACC tool that supports the OpenACC 2.0 syntax. This paper also reports our preliminary efforts on developing and optimizing applications on the TaihuLight system, focusing on key application domains, such as earth system modeling, ocean surface wave modeling, atomistic simulation, and phase-field simulation.

Keywords supercomputer, many-core, high performance computing, scientific computing, computer architecture

Citation Fu H H, Liao J F, Yang J Z, et al. The Sunway TaihuLight supercomputer: system and applications. Sci China Inf Sci, 2016, 59(7): 072001, doi: 10.1007/s11432-016-5588-7

* Corresponding author (email: ygw@mail.tsinghua.edu.cn)

© Science China Press and Springer-Verlag Berlin Heidelberg 2016

1 Introduction

Since the development of supercomputers in the 1970s, scientific computing has become a major scientific paradigm that is as important as the theoretical and experimental branches of the discipline [1]. The computational paradigm has been applied to various scientific domains, such as climate modeling [2], earth subsurface modeling and inversion [3], sky simulation [4, 5], and phase-field simulation [6], with significant contributions to the advancement of those fields.

With scientific advancements, the models that scientists simulate are becoming increasingly complex, and the temporal and spatial resolutions they require are also increasing rapidly. All these factors contribute to the demand for progressively greater computing power.

Moreover, the computing capability of supercomputers has also been growing rapidly in the last few decades. If we compare the top one system in 2005 (BlueGene/L, 183.5 TFlops [7]) and the top one system in 2015 (Tianhe-2, 54.9 PFlops [8]), computing capability has improved nearly three hundred fold. In terms of architecture, as constrained by the heat dissipation and power consumption issues, most of the large systems in the last decade came in the form of heterogeneous systems with both CPU resources and many-core accelerator resources, such as GPUs [9] and Intel Xeon Phi Coprocessors [8].

While previous TOP500 lists were mostly dominated by US and Japanese systems, in recent years, with the support of the National High Technology Research and Development Program (863 Program) of China, we have seen the swift development of Chinese supercomputer systems. The Tianhe-1A system [9], a CPU-GPU hybrid machine, was ranked the top one system on the list of November, 2010. The Tianhe-2 system [8] retained the top one position on the lists from June, 2013 to November, 2015.

As a successor of the Sunway BlueLight system, the Sunway TaihuLight system is also supported by 863 Program of China. Its peak performance is 125 PFlops, sustained Linpack performance is 93 PFlops, and performance per Watt is 6.05 GFlops/W. Compared with the TOP500 list of November, 2015, all three key performance results would rank the first in the world. To support both high performance computing and big data applications, the Sunway TaihuLight adopts key technologies, such as a highly-scalable heterogeneous architecture, high-density system integration, high-bandwidth multi-level network, highly efficient DC power supply, and customized water cooling system. The supercomputer is also equipped with highly efficient scheduling and management tools, and a rich set of parallel programming languages and development environments in order to support research and development of applications on the system. The Sunway TaihuLight system is the world's first supercomputer with a peak performance greater than 100PFlops, and also China's first top one system that is completely based on homegrown many-core processors.

In this paper, we provide an in-depth introduction to the Sunway TaihuLight system. Section 2 provides an overview of the Sunway TaihuLight supercomputer. Section 3 describes the architecture of the homegrown many-core SW26010 processor, which is the key component that provides the computing capacity of the system. Section 4 introduces the major subcomponent systems, particularly the computing, network, peripheral, maintenance and diagnostic, power supply and cooling systems, along with the software system that supports the development of parallel applications. In Section 5, we provide a preliminary progress report on the key scientific applications of this system, which thus far include earth system modeling, ocean surface wave modeling, atomic simulation, and phase-field simulation. Section 6 concludes the paper.

2 The Sunway TaihuLight supercomputer: an overview

The Sunway TaihuLight supercomputer is hosted at the National Supercomputing Center in Wuxi (NSCC-Wuxi), which operates as a collaboration center between the City of Wuxi, Jiangsu Province, and Tsinghua University. NSCC-Wuxi focuses on the development needs of technological innovation and industrial upgrading around Jiangsu Province and the Yangtze River Delta economic circle, as well as the demands of the national key strategies on science and technology development.

Table 1 Major parameters of the Sunway TaihuLight supercomputer

Peak performance	125 PFlops
Linpack performance	93 PFlops
CPU frequency	1.45 GHz
Peak performance of a CPU	3.06 TFlops
Total memory	1310.72 TB
Total memory bandwidth	5591.5 TB/s
Network link bandwidth	16 GB/s
Network bisection bandwidth	70 TB/s
Total storage	20 PB
Total I/O bandwidth	288 GB/s
Power consumption when running the Linpack test	15.371 MW
Performance power ratio	6.05 GFlops/W

Based on the TaihuLight system, NSCC-Wuxi will try to build both high performance computing (HPC) and big data analytic platforms, to serve key high performance computing applications (earth system modeling, oil & gas exploration, bioinformatics, novel drug development, advanced manufacturing, etc.), and also to provide public cloud service.

The Sunway TaihuLight supercomputer system is based on high-density flexible super nodes and a high-bandwidth multi-level network. Table 1 shows the major parameters of the TaihuLight supercomputer. The peak performance of TaihuLight is 125 PFlops, and the sustained Linpack performance is 93 PFlops. When performing the Linpack test, the system's measured power consumption is 15.371 MW, giving a performance per Watt of 6.05 GFlops/W.

Figure 1 shows the general architecture of the Sunway TaihuLight system. The computing node is the basic element that forms the computing system. A set of computing nodes (in the current configuration of TaihuLight, 256 computing nodes) form a super node, which then connect to each other through the central switch network.

The management network connects the central computing system to the management servers, i.e. the directory, database, system control, web, and application servers. The storage network connects the central computing system to the storage system, the import/export nodes, and the management nodes.

The core component of the computing node is the SW26010 many-core processor, which will be introduced in Section 3. The major sub-component systems of TaihuLight include the computing, network, peripheral, maintenance and diagnostic, power supply, cooling, and software systems, which will be described in more detail in Section 4.

3 The SW26010 many-core processor

One major technology innovation of the Sunway TaihuLight supercomputer is the homegrown SW26010 many-core processor. The general architecture of the SW26010 processor [10] is shown in Figure 2. The processor includes four core-groups (CGs). Each CG includes one management processing element (MPE), one computing processing element (CPE) cluster with eight by eight CPEs, and one memory controller (MC). These four CGs are connected via the network on chip (NoC). Each CG has its own memory space, which is connected to the MPE and the CPE cluster through the MC. The processor connects to other outside devices through a system interface (SI).

The MPE is a complete 64-bit RISC core, which can run in both the user and system modes. The MPE completely supports the interrupt functions, memory management, superscalar processing, and out-of-order execution. Therefore, the MPE is an ideal core for handling management and communication functions.

In contrast, the CPE is also a 64-bit RISC core, but with limited functions. The CPE can only run

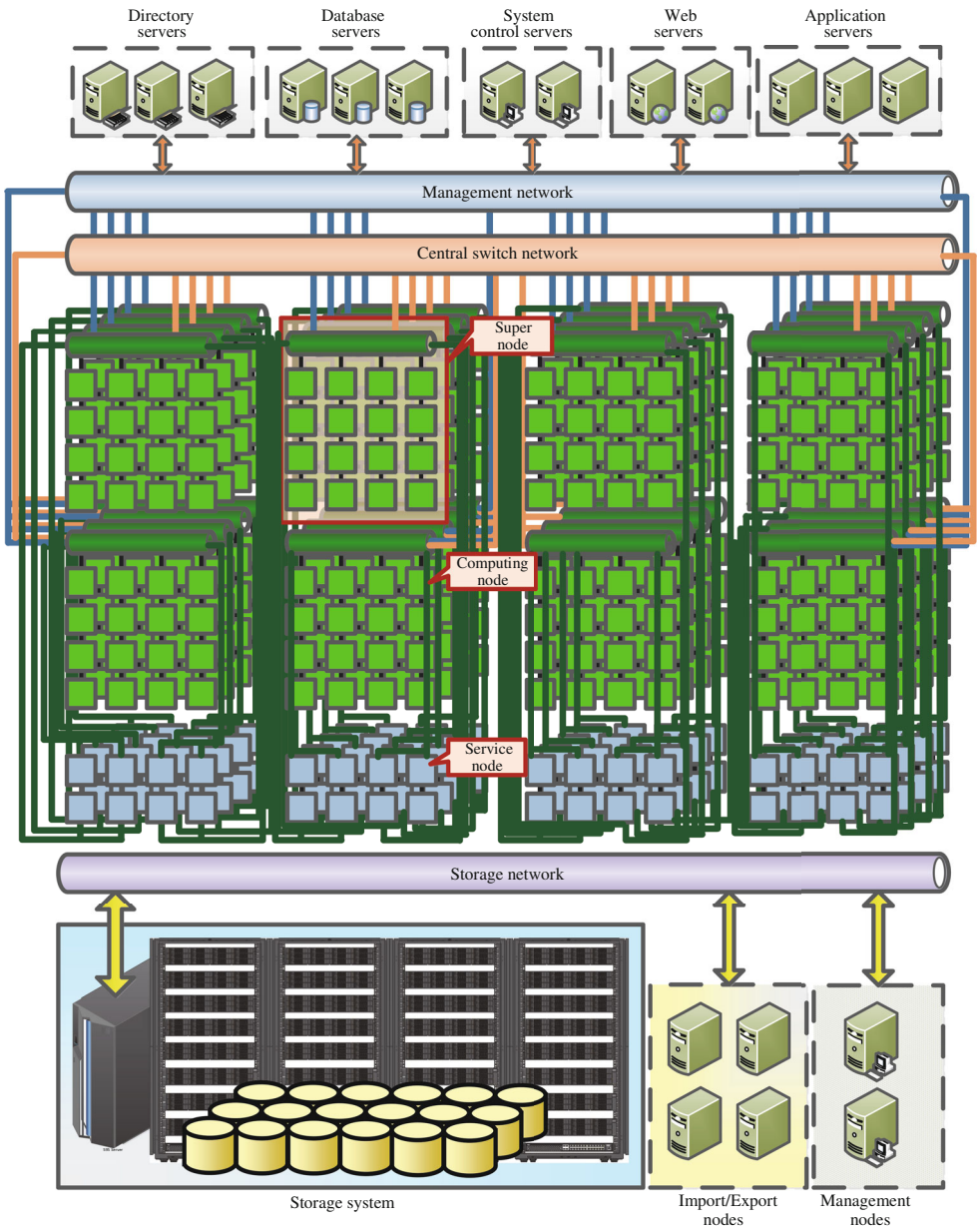

Figure 1 General architecture of the Sunway TaihuLight system.

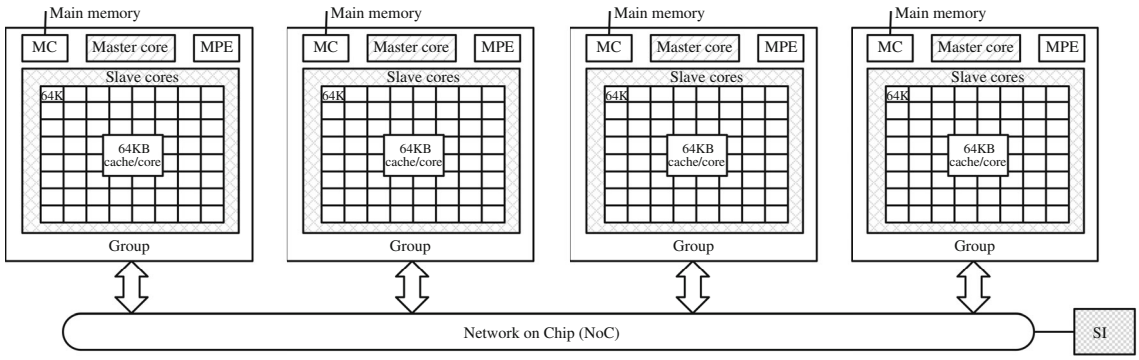

Figure 2 General architecture of the new Sunway processor.

in user mode and does not support interrupt functions. The design goal of this element is to achieve the maximum aggregated computing power, while minimizing the complexity of the micro-architecture. The CPE cluster is organized as an eight by eight mesh, with a mesh network to achieve low-latency register data communication among the eight by eight CPEs. The mesh also includes a mesh controller that handles interrupt and synchronization controls. Both the MPE and CPE support 256-bit vector instructions.

In terms of the memory hierarchy, each MPE has a 32 KB L1 instruction cache and a 32 KB L1 data cache, with a 256 KB L2 cache for both instruction and data. Each CPE has its own 16 KB L1 instruction cache, and a user-controlled scratch pad memory (SPM). The SPM can be configured as either a fast buffer that supports precise user control or a software-emulated cache that achieves automatic data caching. However, as the performance of the software-emulated cache is low, in most cases we need a user-controlled buffering scheme to achieve good performance.

Combining the four CGs of the MPE and CPE clusters, each Sunway processor provides a peak performance of 3.06 TFlops, with a performance-to-power ratio of over 10 GFlops/Watt. While the computing performance and power efficiency are among the top when compared with existing GPU and MIC chips, the on-chip buffer size and the memory bandwidth are relatively limited.

The SW26010 processor uses SoC technology, and integrates four DDR3 memory controllers, the PCIe3.0, Gigabit Ethernet, and the JTAG interfaces. Its memory bandwidth is 136.51 GB/s; its bidirectional network interface bandwidth is 16 GB/s. The CPU frequency is 1.45 GHz, and it uses the FCBGA3832 package.

4 Subcomponent systems of the Sunway TaihuLight

In this section, we provide more detail about the various subcomponent systems of the Sunway TaihuLight, specifically the computing, network, peripheral, maintenance and diagnostic, power and cooling, and the software systems.

4.1 The computing system

Aiming for a peak performance of 125 PFlops, the computing system of the Sunway TaihuLight is built using a fully customized integration approach with a number of different levels: (1) computing node (one CPU per computing node); (2) super node (256 computing nodes per super node); (3) cabinet (4 super nodes per cabinet); and (4) the entire computing system (40 cabinets).

The computing nodes are the basic units of the computing system, and include one SW26010 processor, 32 GB memory, a node management controller, power supply, interface circuits, etc. Groups of 256 computing nodes, are integrated into a tightly coupled super node using a fully-connected crossing switch, so as to support computationally-intensive, communication-intensive, and I/O-intensive computing jobs.

4.2 The network system

The network system consists of three different levels, with the central switching network at the top, super node network in the middle, and resource-sharing network at the bottom. The bisection network bandwidth is 70 TB/s, with a network diameter of 7.

Each super node includes 256 Sunway processors that are fully connected by the super node network, which achieves both high bandwidth and low latency for all-to-all communications among the entire 65536 processing elements.

The central switching network is responsible for building connections and enabling data exchange between different super nodes.

The resource-sharing network connects the sharing resources to the super nodes, and provides services for I/O communication and fault tolerance of the computing nodes.

4.3 The peripheral system

The peripheral system consists of the network storage system and peripheral management system. The network storage system includes both the storage network and storage disk array, providing a total storage of 20 PB and a high-speed and reliable data storage service for the computing nodes. The peripheral management system includes the system console, management server, and management network, which enable system management and service.

4.4 The maintenance and diagnostic system

The maintenance and diagnostic system provides comprehensive online maintenance management, status and environment monitoring, fault location and recording, and security services for the entire system.

Adopting a multi-level management architecture, the maintenance and diagnostic system integrates embedded modules for the testing and maintenance of the nodes, super nodes, and the entire computing system. By using a distributed dynamic data collection approach, the system achieves runtime monitoring and visualization of the computing system status.

4.5 The power supply system and cooling system

The TaihuLight supercomputer uses a mutual-backup power input of 2×35 KV. The cabinets of the system use a three-level (300 V-12 V-0.9 V) DC power supply mode. The front-end power supply output is 300 V, which is directly linked to the cabinet. The main power supply of the cabinet converts 300 V DC to 12 V DC, and the CPU power supply converts 12 V into the voltage that the CPU needs.

The cabinets of the computing and network systems use indirect water cooling, while the peripheral devices use air and water exchange, and the power system uses forced air cooling. The cabinets use closed-loop, static hydraulic pressure for cavum, indirect parallel flow water cooling technology, which provides effective cooling for the full-scale Linpack run.

4.6 The software system

The software system of Sunway TaihuLight provides support for applications in different scientific domains and industries. The major components include the basic software for the homegrown many-core CPU, parallel operating system environment, high-performance storage management system, parallel programming language and compilation environment, and parallel development environment.

The basic software for the many-core processor includes basic compiler components, such as C/C++, and Fortran compilers, an automatic vectorization tool, and basic math libraries.

The parallel operating system environment, which includes the parallel operating system, network management system, and availability and power management systems, provides computing, application, and management services, in addition to other system services to the users.

The high-performance storage management system includes the parallel file system, lightweight file system, and storage management platform, which provide storage support for running system software and large-scale parallel I/O applications.

In addition to the basic software, the TaihuLight system also has a parallel programming language and compilation environment to support parallelization at different levels. For parallelization at the node level, MPI is generally applied. For the four CGs within the same processor, we can either use MPI or OpenMP. For parallelization within a CG, we use Sunway OpenACC, a customized parallel compilation tool that supports OpenACC 2.0 syntax and targets the CPE clusters. The customized Sunway OpenACC tool supports parallel task management, heterogeneous code extraction, and data transfer descriptions. Moreover, based on the specific features of the Sunway processor architecture, the Sunway OpenACC tool has also made a number of syntax extensions from the original OpenACC 2.0 standard, such as a finer control over multi-dimensional array buffering, and packing distributed variables for data transfer.

Table 2 A brief comparison between the Sunway TaihuLight and other large-scale systems

System	Peak performance	Linpack performance	Node architecture
Sunway TaihuLight	125 PFlops	93 PFlops	One 260-core Sunway CPU 4 MPEs and 256 CPEs
Tianhe-2	55 PFlops	34 PFlops	Two 12-core Intel CPUs and three 57-core Intel Xeon Phi Coprocessors
Titan	27 PFlops	18 PFlops	One 16-core AMD CPU and one K20x NVIDIA GPU (2688 CUDA cores)
Sequoia	20 PFlops	17 PFlops	One 16-core PowerPC CPU
K	10 PFlops	11 PFlops	One 8-core SPARC64 CPU

4.7 Comparing the Sunway TaihuLight with other large-scale systems

Table 2 provides a brief comparison between the Sunway TaihuLight and other four top systems of the TOP500 list in November, 2015. Compared with the previous top one sytem, Tianhe-2, the TaihuLight system doubles its peak performance, and almost triples its sustainable Linpack performance. The MPE-CPE hybrid architecture of the new Sunway system is also largely different from previous heterogeneous systems. While the MPE is like a CPU core, and the CPE cluster is like a many-core accelerator, both the CPU and accelerator are now fused into one chip with a unified memory space, which is different from both the homogeneous clusters with only multi-core CPUs (such as Sequoia, and K) and the heterogeneous clusters with both CPUs and PCIe-connected accelerators (such as Tianhe-2, and Titan).

5 Preliminary progress of scientific computing applications on the TaihuLight

Although the system has only been installed for a few months, efforts to develop highly efficient and highly scalable parallel software started with the pilot system around a year ago. In this section, we provide preliminary results for some of the key scientific computing applications on the TaihuLight platform.

5.1 Refactoring the community atmospheric model (CAM) on the Sunway TaihuLight

Since the very first generation of supercomputer systems (e.g. CDC 6600, and Cray-I), atmospheric models have been among the major users of computing resources [11], and have evolved along with the development of supercomputer systems.

While climate models call for more computing power to support higher resolution and more complex physics [12], the millions of lines of legacy code designed for multi-core CPUs makes it difficult to exploit increasing computing power of many-core accelerators. To fill this gap, in our work, we performed an extensive refactoring and optimization of the CAM atmospheric model for the TaihuLight system. We chose CAM [13] as our target application, as it is one of the most widely used advanced atmospheric models in the world. Note that, to achieve high scalability over the system with millions of cores, we chose to use the SE dynamic core of CAM [14]; the other dynamic core options are not considered in this work.

We use the Sunway OpenACC compiler (a customized version that expands on the OpenACC standard) as the key tool for achieving a suitable mapping of CAM onto the new Sunway heterogeneous many-core processors. Due to the large code base developed over the last few decades and the general demand from climate scientists to maintain the same source, we tried to minimize manual refactoring efforts (to only the dynamic core part) and mainly relied on source-to-source translation tools to achieve automated and efficient porting. Additionally, compared with GPUs and other many-core accelerators, both the on-chip fast buffer and available memory bandwidth of the Sunway processor are relatively limited (as detailed in Section 3), which made our porting significantly more challenging. Many of our tools and optimization strategies focused on minimizing the memory footprints.

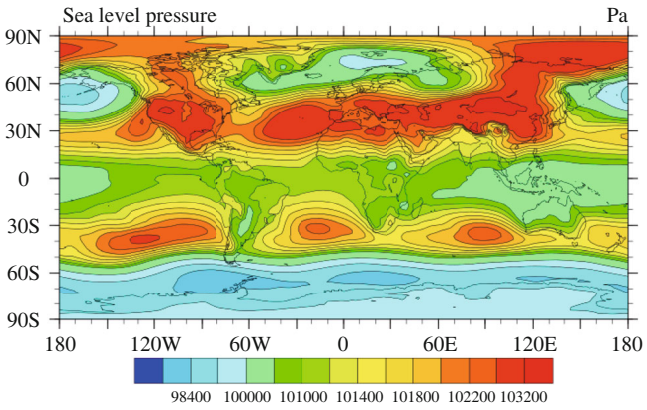

Figure 3 Surface pressure results of a coupled CAM and CLM run on the new Sunway system.

As CAM has not been run on the Sunway architecture before, the first step of porting was to verify the modeling results. As the current version of CAM has to be run coupled with the Common Land Model (CLM), we port both CAM and CLM onto the Sunway system, using only the MPE to perform the computation.

After running the coupled CAM and CLM models on the Sunway system for the duration of three years, we compare results of major variables, and the conservation of mass and energy, to verify the correctness of our ported version. Figure 3 demonstrates the surface pressure results (describing the total mass of the air in each column) of our ported Sunway version. Compared with results on Intel clusters using the same modeling parameter configuration, we see a nearly identical distribution of values and an average relative error in the range of 10^{-6}.

Using the ported MPE-only version as the starting point, we then performed refactoring and optimization of both the dynamic core and the physics schemes to utilize the CPE clusters. During the process of expanding each kernel from MPE-only to MPE-CPE hybrid mode, we took the numerical result of the MPE-only mode as the true value, and ensured that the MPE-CPE hybrid mode generated identical values.

Comparing the refactored hybrid version using both the MPEs and CPE clusters against the ported version using only the MPEs, we can achieve up to 22 times speedup for computationally-intensive kernels, and two to seven times speedup for kernels involving both computation and memory operations. For the entire CAM model, we achieve a performance improvement of around two times.

Figure 4 shows the simulation speed of the CAM model, measured in Model Year per Day (MYPD), on the new Sunway supercomputer, with the number of CGs increasing from 1024 to 24000. Similar to previous reported results on other systems, the CAM model demonstrates good scalability on the new Sunway supercomputer system, with the simulation speed increasing steadily with the number of CGs. For large-scale cases, we demonstrated the performances for using the MPEs only, and using both the MPEs and CPE clusters. As shown in the last two points in Figure 4, by using both the MPEs and CPE clusters, we can further improve the simulation speed by another two times. When scaling the CAM model to 24000 CGs (24000 MPEs, and 1536000 CPEs), we can achieve a simulation speed of 2.81 MYPD.

While the increase in speed is not significant, this work provides an important base for us to continue optimizing the performance of large and complicated scientific programs, such as CAM.

5.2 A fully-implicit nonhydrostatic dynamic solver for cloud-resolving atmospheric simulation on eight million cores

While the work in the previous section focused on porting CAM, one of the most widely used atmospheric models, onto the Sunway TaihuLight system, the work in this section focuses on the study of a highly scalable fully implicit solver for three-dimensional nonhydrostatic atmospheric simulations. We use fully compressible Euler equations with the moist process as the governing equation set, which is accurate at

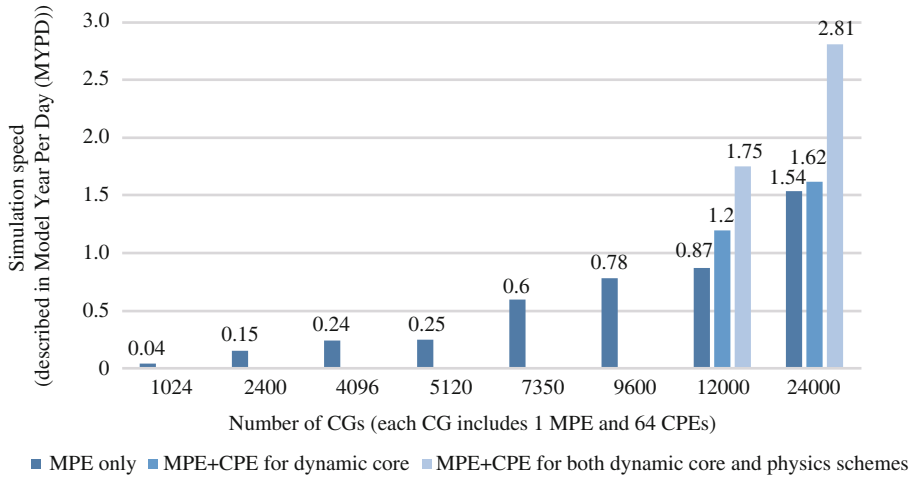

Figure 4 Simulation speed of the CAM model, measured in Model Year per Day (MYPD), on the new Sunway supercomputer, with the number of CGs increasing from 1024 to 24000. For large-scale runs with 12000 and 24000 CGs, we show the performances of the model in three scenarios: (1) using the MPEs only; (2) using both the MPEs and CPEs for the dynamic core; (3) using both the MPEs and CPEs for both the dynamic core and physics schemes.

mesocale with almost no assumptions made [15]. In particular, we consider a channel domain above a rotating sphere with possibly nonsmooth bottom topography [16]:

$$\frac{\partial Q}{\partial t}+\frac{\partial F}{\partial x}+\frac{\partial G}{\partial y}+\frac{\partial H}{\partial z}+S=0, \tag{1}$$

$$\begin{aligned}
Q&=\left(\rho', \rho u, \rho v, \rho w, (\rho e_T)', (\rho q)'\right)^{\mathrm{T}},\\
F&=\left(\rho u, \rho uu+p', \rho uv, \rho uw, (\rho e_T+p)\,u, \rho uq\right)^{\mathrm{T}},\\
G&=\left(\rho v, \rho vu, \rho vv+p', \rho vw, (\rho e_T+p)\,v, \rho vq\right)^{\mathrm{T}},\\
H&=\left(\rho w, \rho wu, \rho wv, \rho ww+p', (\rho e_T+p)\,w, \rho wq\right)^{\mathrm{T}},\\
S&=\left(0, \partial\bar{p}/\partial x-f\rho v, \partial\bar{p}/\partial y+f\rho u, \rho' g, 0, 0\right)^{\mathrm{T}},
\end{aligned}$$

where ρ, $\boldsymbol{v}=(u,v,w)$, p, θ and q are the density, velocity, pressure, virtual potential temperature, and moisture of the atmosphere, respectively. The Coriolis parameter is provided as f and all other variables such as g and γ are given constants. The values of $\rho'=\rho-\bar{\rho}$, $(\rho e_T)'=\rho e_T-\bar{\rho}\bar{e_T}$, and $p'=p-\bar{p}$ have been shifted according to a hydrostatic state that satisfies $\partial\overline{p}/\partial z=-\overline{\rho}g$. The system is closed with the equation of state $p=(\gamma-1)\rho(e_T-gz-\|\boldsymbol{v}\|^2/2)$. Note that we choose the total energy density instead of the traditional pressure- or temperature-based values as a prognostic variable to fully recover the energy conservation law and avoid the repeated calculation of powers that may introduce a long latency and substantially degrade the overall performance.

In this work, an efficient and scalable solver is developed with three key features. First, a hybrid multigrid domain decomposition preconditioner is proposed to greatly accelerate the convergence of the solver while efficiently exploiting inter-node level parallelism. Second, a physics-based multi-block asynchronized incomplete LU factorization method is customized to solve the subproblems of each overlapped subdomain to further exploit intra-node level concurrency. Third, implementation and optimization are done to achieve high-performance with considerations toward the process, thread, and instruction levels. By incorporating into a Jacobian-Free Newton-Krylov framework, the solver enables fast and accurate atmospheric simulations on leading heterogeneous systems such as the Sunway TaihuLight.

We use the baroclinic instability test in a β-plane 3D channel [17] to validate the correctness and examine the performance of the proposed fully implicit solver. The test is initiated by adding a confined perturbation in the zonal wind field to a geostrophically balanced background flow. This setup is useful in examining the correct response a numerical scheme produces to interact with the unbalanced trigger.

The size of the 3D channel in the baroclinic instability test is 40000 km $\times$ 6000 km $\times$ 30 km. We use a relatively small mesh with a horizontal resolution of 100 km and a vertical resolution of 1 km to verify

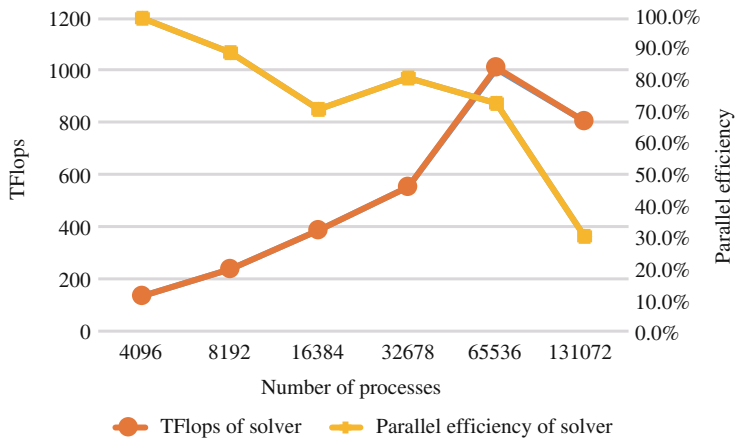

Figure 5 Strong scaling results of the non-hydrostatic dynamic solver on the Sunway TaihuLight supercomputer.

Table 3 Hardware, case configurations, and time to solutions (solver part, and the entire program) for strong-scaling tests

Parallelism		4096	8192	16384	32768	65536	131072
Process	NX	128	256	256	512	512	1024
	NY	32	32	64	64	128	128
		NZ = 1					
Sockets		1024	2048	4096	8192	16384	32768
Cabinets		1	2	4	8	16	32
Total cells		$16384 \times 2048 \times 64 = 2.15$ B					
Total unknowns		12.88 B					
Total solver time (h)		4.744	2.665	1.675	0.588	0.407	0.493
Total time (h)		4.771	2.679	1.682	0.591	0.409	0.494

the model and compare with reference results.

In the strong scaling tests, we fix the total problem size to be 2.15 billion mesh cells ($16384 \times 2048 \times 64$) and increase the number of computing nodes from 4096 processes to 131072 MPI processes. The test results are provided in Figure 5. Detailed configurations of the hardware resources (CGs, processors, and cabinets) and the simulation case (number of cells, number of unknowns, etc.) are shown in Table 3. From the figure, we observe that when we increase the number of processes from 4096 (1 cabinet) to 8192 (2 cabinets), the parallel efficiency of the solver part drops to 89%, due to the introduction of inter-cabinet communication. For the scale of 16384 (4 cabinets) to 65532 (16 cabinets), parallel efficiency persisted at around 70%, as the number of iterations remains almost the same for these different configurations. When the scale increases to 131072 processes (32 cabinets, 8.52 million cores in total), the efficiency drops to 30%, as the communication becomes the major overhead. The computational performance of the solver part changes from 130 TFlops to 806 TFlops (the highest performance, 1.01 PFlops, is achieved with the configuration of 65536 processes). Compared with the solver part, the performances for the entire program (including the I/O and initialization) are quite similar.

In the weak scaling tests, we observe better scalability when using 32 cabinets, with the sustained performance further improved to 1.5 PFlops.

As shown in Table 3, the time to solution of the solver part for the simulation of 1000 time steps reduces from 17080 s to 1775 s (the shortest time to solution, 1465 s, is achieved with the configuration of 65536 processes).

In terms of the numerical methodology, our work has demonstrated that fully-implicit methods could be an important option for performing numerical weather and climate simulations, especially for high-resolution scenarios.

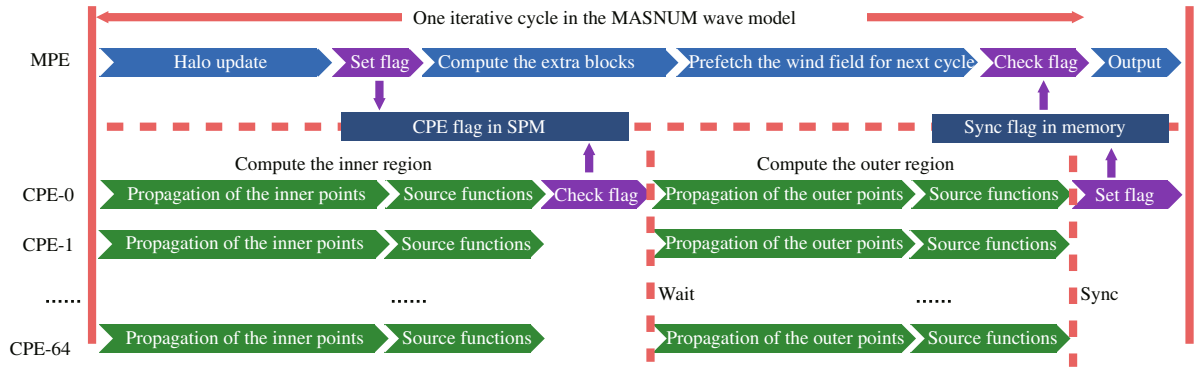

Figure 6 An optimized algorithm for one iterative cycle of the MASNUM wave model. The halo update, wind input, and result output are handled by the MPE. Through setting and checking the CPE flag in SPM, CPE-0 knows the status of the halo update. A flag in memory is used for synchronizing the MPE and CPE-0. All CPEs compute the propagation and source functions for the inner points first, and for the outer points when the halo update is finished. All CPEs except CPE-0 communicate with CPE-0 only.

5.3 A highly effective global surface wave numerical simulation with ultra-high resolution

Surface waves are one of the most energetic motions in the global ocean, and understanding them is crucially important to marine safety and climate change. A high resolution global wave model is the key to accurate wave forecasting. However, parallel efficiency with a large amount of computation is currently a significant barrier to such a model. In this work, by adopting a new design of irregular quasi-rectangular grid decomposition, master-slave cooperative computing workflow and pipelining schemes for high resolution global wave model, the MASNUM (Key laboratory of MArine Science and NUmerical Modeling) wave model [18], has been successfully run with an ultra-high horizontal resolution of $(1/60)^\circ$ by $(1/60)^\circ$ in a global scale. The results show that peak performance of our model reaches 30.07 PFlops with 8519680 cores (close to the full scale of the system). These innovations provide good scalability and high efficiency for an ultra-high resolution global wave model.

In the original MPI implementation of the MASNUM wave model, the workflow in one iterative step mainly consists of five stages: input of the wind field, source functions, wave propagation, halo update, and output. In order to fully utilize the heterogeneous resources in the new Sunway processor, we design a new master-slave cooperative computing workflow for the MASNUM wave model.

After decomposing the whole computing domain into blocks, we divide each block into an inner region and an outer region. As shown in Figure 6, we first exchange the outer region in a wave propagation block through MPI ISEND and IRECV functions in the MPE. When the halo update is finished, we set one flag in the SPM of CPE-0. As the computing of source functions and wave propagation of inner points is not dependent on the halo region, these two procedures can be run on the CPEs first. After that, CPE-0 checks the CPE flag to detect whether the halo update has finished, while the other CPEs wait on CPE-0. If the flag has been set, the wave propagation of the outer region can be subsequently executed. Then, all the CPEs must synchronize and CPE-0 sets a sync flag in memory to prepare for data output. In order to avoid the idling of the MPE during the period in which the CPE flag is set and the sync flag is checked, we compute the extra blocks and prefetch the wind field for the next cycle. The extra blocks are defined as the residual data blocks when the grid points cannot be divided exactly by the number of MPI processes. Note that there is one wait operation among all the CPEs and one synchronous operation between the MPE and CPE-0 in each integration cycle. Through the reordering of computation sequences, we implement a master-slave cooperative computing workflow for the MASNUM wave model. Since we exchange flag values between the MPE and CPEs via SPM and main memory instead of MPI communication, the implementation of the set flag and check flag functions are more efficient. In short, the new master-slave workflow is helpful for fully utilizing the MPE and CPEs' computing resources provided by the Sunway many-core processor.

The peak performance for the simulations with horizontal resolution of $(1/60)^\circ$ by $(1/60)^\circ$ is 56.6-

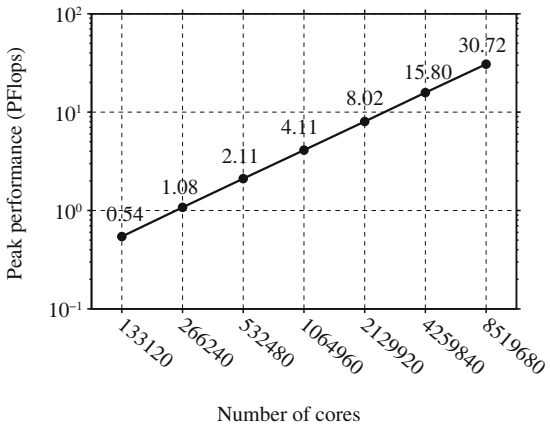

Figure 7 Peak performance results for simulations with a horizontal resolution of (1/60)° by (1/60)°. Numbers along the y-axis are peak performance for simulations with different numbers of cores. The peak performance with 8519680 cores is up to 30.7 PFlops.

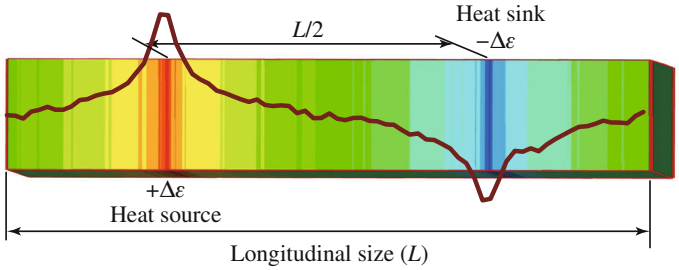

Figure 8 Calculated temperature field in the longitudinal direction of silicon nanowires. The color variation from blue to red represents an increase in temperature. Energy $\Delta\epsilon$ is added in the heat source, and removed in heat sink at each time step.

fold from 2048 processors with 133120 cores up to 131072 processors with 8519680 cores (Figure 7). The computing performance is proportional to the increase in core number, and the peak performance obtained is 30.7 PFlops.

5.4 Peta-scale atomistic simulation of silicon nanowires

Covalent materials are both common and important in modern industries and technologies. Typical examples include carbon for diamonds, graphite, carbon nanotubes, and graphene; silicon for semiconductors in the information technology (IT) industry and solar energy; and other elements, such as germanium, for similar purposes. For the application of these materials, a central problem is the relationship between the micro-scale (atomistic) structures and their macro-scale properties and behaviors, or functions. However, it is a grand challenge to both experimental work and numerical simulations in that it is almost unattainable to meet the requirements for large scale, high resolution and accuracy simultaneously.

To use the example of silicon, silicon nanowires have diverse potential applications in field-effect transistors, photovoltaic and thermoelectric devices, and biological and chemical sensors. However, their performance and stability are sensitive to thermal properties and heat dissipation processes from the atomistic scales up to micron scales. In simulations, a highly efficient, scalable, and general purpose algorithm was developed for non-equilibrium molecular dynamics simulation of thermal conductivity (Figure 8); its highest efficiency on a single many-core processor has reached 15.1% of its theoretical peak performance with various optimizations such as atom ordering [19], SIMD vectorization, global memory access reduction, and efficient use of local memory and nonlinear mathematical functions.

Strong and weak scaling of the algorithm are also analyzed on the Sunway TaihuLight supercomputer. As shown in Figure 9, weak scaling for the algorithm is ideal until the full-scale simulation, and the parallel efficiency is kept above 82%, without evident decrease with increasing scale. The communication

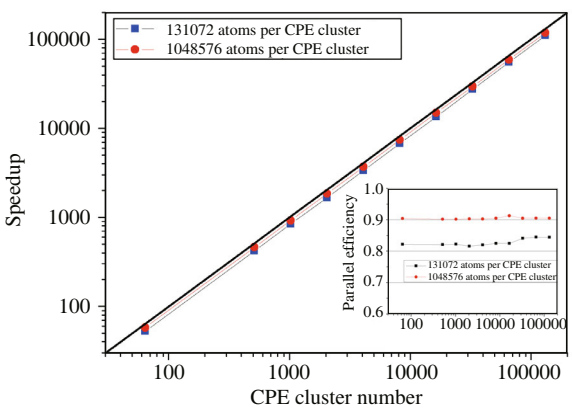

Figure 9 (Color online) Weak scaling for the silicon nanowire simulations. A CPE cluster represents a computing processing element cluster that includes 64 hardware processing cores.

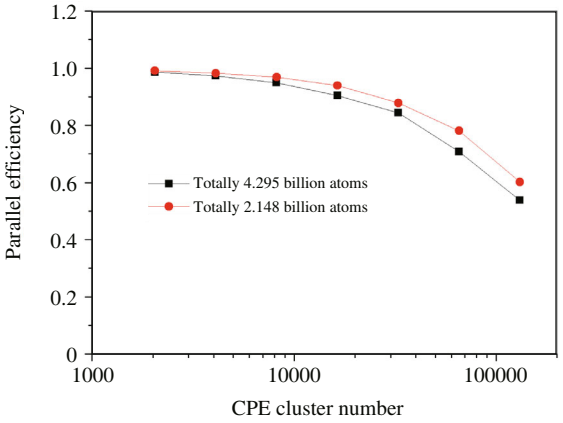

Figure 10 (Color online) Strong scaling for the silicon nanowire simulations. A CPE cluster represents a computing processing element cluster that includes 64 hardware processing cores.

overhead is reasonably low and does not change much with increasing processor number or decreasing atom number on each processor. Strong scaling is also favorable for decreasing atom number to 16384 per CG; the lowest parallel efficiency is above 54% (see Figure 10). The two series of simulations used up to 131072 CGs. A sustainable performance of 6.62 PFlops in double precision was achieved, and a highest performance of 14.7 PFlops is estimated with 24576 atoms per CG and 3.22 billion atoms in total. This is, to our knowledge, the highest performance of a many-body MD simulation of covalent crystals reported so far. The nearest work is from our work on the Tianhe-1A supercomputer [20]. The realized simulations enable virtual experiments on real-world materials and devices for predicting macro-scale properties and behaviors from micro-scale structures directly, bringing about many exciting new possibilities in nanotechnology, information technology, electronics, and renewable energies.

5.5 Large-scale phase-field simulation for coarsening dynamics based on Cahn-Hilliard equation with degenerated mobility

In material science, coarsening refers to the changes in spatial scales over time associated with mesoscale morphological patterns. Such patterns are often referred to as the microstructure of the material, which has a critical role in determining many important material properties, such as strength, hardness, and conductivity. In order to predict and optimize the resulting macroscopic material parameters, it is important to obtain a deeper insight into this microscopic structure evolution. Diffuse-interface or phase-field based modeling and simulations have become a widely used methodology for describing the microstructure formation of materials. We work with the Cahn-Hilliard equation with the double-well potential and a degenerate mobility. Figure 11 shows a typical scenario that we simulate.

We have designed a highly scalable, large time-step integrating algorithm, the Scalable compact Localized Exponential Time Differencing (ScLETD) algorithm, and incorporated various optimization techniques exploiting the computing power and data moving capability of the Sunway TaihuLight supercomputer. These innovations are essential for the successful simulation of extreme-scale coarsening dynamics on an extreme-scale supercomputer.

We present a large-scale phase-field simulation on the new Sunway TaihuLight supercomputer. The highly nonlinear and severely stiff Cahn-Hilliard equations with degenerated mobility for microstructure evolution are solved at extreme scale, demonstrating that the latest advent of high performance computing platform and the new advances in algorithm design are now offering us the possibility to accurately simulate coarsening dynamics at unprecedented spatial and temporal scales. Figures 12 and 13 show the weak scaling and strong scaling results. The code has demonstrated good strong and weak scalability and achieved 39.687 PFlops in double precision for the largest configuration, using over 9 million cores.

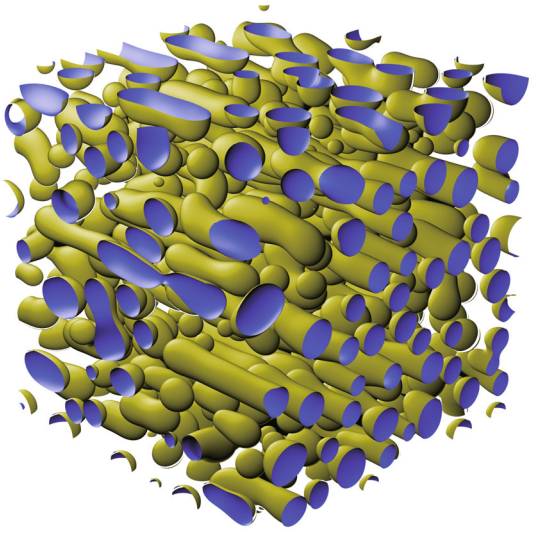

Figure 11 Grains in various sizes and shapes coexist, revealing a transition to columnar-like morphologies with unprecedented precision and detail (previously reported experiments have been largely limited to resolutions with at most a handful of cylinder-like structure in the computational domain).

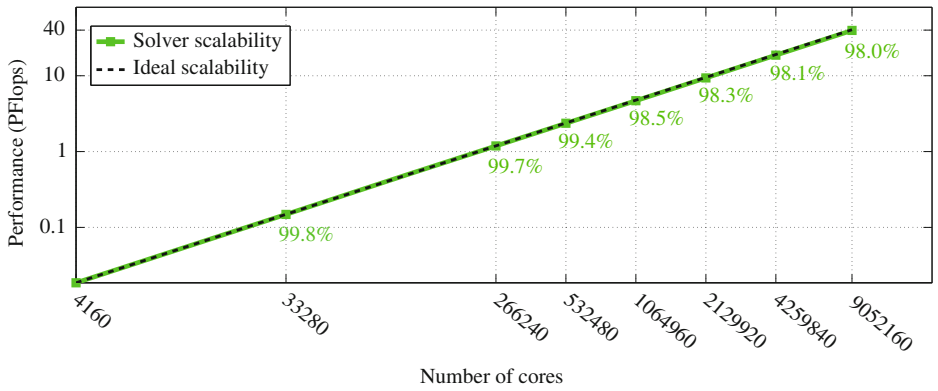

Figure 12 Weak scalability results of the phase-field simulation. Numbers along the y-axis indicate efficiency with respect to the ideal scaling (the efficiency baseline is the 4160-core result). The largest configuration has 18.3 trillion grid points and the highest performance is 39.678 PFlops.

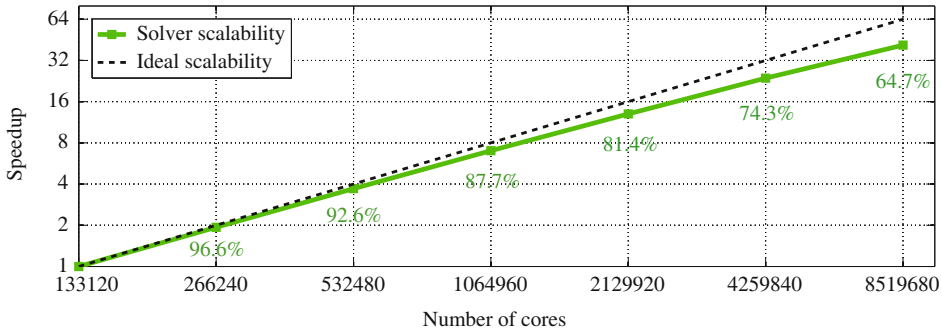

Figure 13 Strong scalability results of the phase-field simulation. Numbers along the y-axis indicate efficiency with respect to the ideal speedup (the efficiency baseline is the 133120-core result).

5.6 Application: summary and comparison

Table 4 shows a summary of a number of the applications mentioned above, compared with similar applications that have been implemented and optimized on other large-scale systems. In most applications,

Table 4 Summary of the major applications on the Sunway TaihuLight, compared with similar applications on other large-scale systems

Category	System	Application summary	Scale of run	Performance
Non-linear solver	Sunway TaihuLight	A fully-implicit nonhydrostatic dynamic for cloud-resolving atmospheric simulation	131072 MPEs and 8388608 CPEs	1.5 PFlops
	Sequoia	An implicit solver for complex PDEs in highly heterogeneous flow in Earth's mantle [3]	1572864 cores	687 TFlops
Molecular dynamics	Sunway TaihuLight	Atomic simulation of silicon nanowires	131072 MPEs and 8388608 CPEs	14.7 PFlops
	Tianhe-1A	Molecular dynamics simulation of crystalline silicon [20]	7168 GPUs (3211264 CUDA cores)	1.87 PFlops
Phase-field simulation	Sunway TaihuLight	Coarsening dynamics based on Cahn-Hilliard equation with degenerated mobility	131072 MPEs and 8388608 CPEs	39.678 PFlops
	Tsubame 2.0	Dendritic solidification [6]	16000 CPU cores and 4000 GPUs (1792000 CUDA cores)	1.017 PFlops

we see a comparable application efficiency at a significantly larger parallel scale. For computationally-intensive applications, such as molecular dynamics, and phase-field simulation, with the improvement in the computing power in TaihuLight, we achieve high performances of around 40 PFlops for phase-field simulation, and 15 PFlops for molecular dynamics, compared with the previously reported performances of around 1 to 2 PFlops [6, 20]. For implicit solvers, we achieve a highly efficient solver that targets atmospheric dynamics, which can efficiently scale up to 8 million cores and provide a performance of 1.5 PFlops, compared to the 687 TFlops performance of the Earth's mantle solver on Sequoia [3].

6 Conclusion

In this paper, we describe the newly-installed Sunway TaihuLight supercomputer, at the National Supercomputing Center in Wuxi. Based on the homegrown SW26010 many-core processor, the TaihuLight supercomputer has adopted other key technologies, including the high-density integration of millions of cores, and fully customized water cooling system, and is the first system in the world to provide a peak performance over 100 PFlops (with a peak performance of 125 PFlops, a sustained Linpack performance of 93 PFlops, and a performance power ratio of 6.05 GFlops/W).

A number of key scientific applications have been ported and optimized for the Sunway TaihuLight supercomputer. Various contributions have been made, including the refactoring of a widely-used atmospheric model with a half-million lines of code, the design and optimization of a fully-implicit solver that can scale to 8 million cores, high-resolution surface wave modeling with a sustained performance of 30 PFlops, highly scalable atomic simulation of silicon nanowires with a performance of 14.7 PFlops, and a large-scale phase-field simulation with a performance of 40 PFlops. These initial results have demonstrated the significant performance improvement that the new TaihuLight system could bring to various scientific computing applications, and the great potential for advancing scientific discoveries in corresponding domains.

Conflict of interest The authors declare that they have no conflict of interest.

References

1 Hey A J, Tansley S, Tolle K M, et al. The Fourth Paradigm: Data-Intensive Scientific Discovery. Vol. 1. Redmond: Microsoft Research, 2009

2 Shingu S, Takahara H, Fuchigami H, et al. A 26.58 TFlops global atmospheric simulation with the spectral transform method on the Earth Simulator. In: Proceedings of the ACM/IEEE Conference on Supercomputing. Los Alamitos: IEEE, 2002. 1–19

3 Rudi J, Malossi A C I, Isaac T, et al. An extreme-scale implicit solver for complex PDEs: highly heterogeneous flow in earth's mantle. In: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis. New York: ACM, 2015. 5

4 Ishiyama T, Nitadori K, Makino J. 4.45 PFlops astrophysical N-body simulation on K computer: the gravitational trillion-body problem. In: Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis. Los Alamitos: IEEE, 2012. 5

5 Habib S, Morozov V A, Finkel H, et al. The universe at extreme scale: multi-petaflop sky simulation on the BG/Q. In: Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis. Los Alamitos: IEEE, 2012. 4

6 Shimokawabe T, Aoki T, Takaki T, et al. Peta-scale phase-field simulation for dendritic solidification on the TSUBAME 2.0 supercomputer. In: Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis. New York: ACM, 2011. 3

7 Adiga N R, Almasi G, Aridor Y, et al. An overview of the BlueGene/L supercomputer. In: Proceedings of the 2002 ACM/IEEE Conference on Supercomputing. Los Alamitos: IEEE, 2002. 1–22

8 Liao X K, Xiao L Q, Yang C Q, et al. MilkyWay-2 supercomputer: system and application. Front Comput Sci, 2014, 8: 345–356

9 Yang X J, Liao X K, Lu K, et al. The TianHe-1A supercomputer: its hardware and software. J Comput Sci Technol, 2011, 26: 344–351

10 Zheng F, Li H L, Lv H, et al. Cooperative computing techniques for a deeply fused and heterogeneous many-core processor architecture. J Comput Sci Technol, 2015, 30: 145–162

11 Drake J, Foster I, Michalakes J, et al. Design and performance of a scalable parallel community climate model. Parallel Comput, 1995, 21: 1571–1591

12 Dennis J M, Vertenstein M, Worley P H, et al. Computational performance of ultra-high-resolution capability in the Community Earth System Model. Int J High Perform Comput Appl, 2012, 26: 5–16

13 Neale R B, Chen C-C, Gettelman A, et al. Description of the NCAR Community Atmosphere Model (CAM 5.0). The National Center for Atmospheric Research, Boulder. Note NCAR/TN-4861STR

14 Dennis J M, Edwards J, Evans K J, et al. CAM-SE: a scalable spectral element dynamical core for the Community Atmosphere Model. Int J High Perform Comput Appl, 2012, 26: 74–89

15 Lauritzen P H, Jablonowski C, Taylor M A. Numerical Techniques for Global Atmospheric Models. Berlin: Springer, 2011

16 Ogura Y, Phillips N A. Scale analysis of deep and shallow convection in the atmosphere. J Atmos Sci, 1962, 19: 173–179

17 Ullrich P, Jablonowski C. Operator-split Runge-Kutta-Rosenbrock methods for nonhydrostatic atmospheric models. Mon Weather Rev, 2012, 140: 1257–1284

18 Zeng Y Y, Li Q F, Wei Z, et al. MASNUM ocean wave numerical model in spherical coordinates and its application. Acta Oceanol Sin, 2005, 27: 1–7

19 Hou C F, Xu J, Wang P, et al. Efficient GPU-accelerated molecular dynamics simulation of solid covalent crystals. Comput Phys Commun, 2013, 184: 1364–1371

20 Hou C F, Xu J, Wang P, et al. Petascale molecular dynamics simulation of crystalline silicon on Tianhe-1A. Int J High Perform Comput Appl, 2013, 27: 307–317