

Sunway supercomputer architecture towards exascale computing: analysis and practice

Jiangang GAO*, Fang ZHENG, Fengbin QI, Yajun DING, Hongliang LI,
Hongsheng LU, Wangquan HE, Hongmei WEI, Lifeng JIN, Xin LIU,
Daoyong GONG, Fei WANG, Yan ZHENG, Honghui SUN,
Zhou ZHOU, Yong LIU & Hongtao YOU

National Research Center of Parallel Computer Engineering and Technology, Beijing 100190, China

Received 18 August 2020/Revised 4 October 2020/Accepted 3 November 2020/Published online 3 February 2021

Abstract In recent years, the improvements of system performance and energy efficiency for supercomputers have faced increasing challenges, which create more intensive demands on the architecture design for realizing exascale computing. This paper first analyzes the main requirements of exascale computing on the aspects of the parallel computing application and supercomputing center operation. Afterwards, a mapping scheme of “demands-challenges-architecture” is proposed. Then, the major challenges of exascale supercomputer, such as scalability, power consumption, data movement, programming and availability, are thoroughly analyzed, and the corresponding appropriate solutions are proposed. Moreover, this paper proposes the Sunway computer architecture towards exascale computing in which the many-core processor, network chipset and software system are all domestically-designed. The technology roadmap of Sunway supercomputer will hold the comprehensive design methods for the architecture, including the processor, interconnect network, assembly structure, power supply, cooling system, system software, parallel algorithm and application support, promising great advances for exascale supercomputing.

Keywords supercomputer, exascale, Sunway, scalability, power consumption, data movement, programming, availability

Citation Gao J G, Zheng F, Qi F B, et al. Sunway supercomputer architecture towards exascale computing: analysis and practice. *Sci China Inf Sci*, 2021, 64(4): 141101, <https://doi.org/10.1007/s11432-020-3104-7>

1 Introduction

Currently, the computation capability of No.1 supercomputer in TOP500 list¹⁾ reaches hundreds of Petaflops. It is expected that the first exascale supercomputer will be debuted in around 2021, which will open the exascale era for high performance computing (HPC). However, because of the slowdown of both Moore's law [1] and Dennard Scaling law [2], the improvements of system performance and efficiency are becoming increasingly difficult, bringing unprecedented challenges on architecture design for exascale supercomputer [3–12]. As one of the leading teams of supercomputer research and development in China, Sunway has solid foundations on both theoretical research and engineering practice on the supercomputer design, which has been proven by the successful implementations of Sunway BlueLight [13,14], Sunway TaihuLight [15–19] and Sunway exascale prototype. The successful developments of Sunway supercomputers demonstrate that the comprehensive co-design for system architecture, including the processor, interconnect network, assembly structure, power supply, cooling system, system software, parallel algorithm and application support, is crucial to achieve optimal system performance and efficiency.

From 1980s, most supercomputers in the world have been built using commercial devices (e.g., processors and network) and software to greatly accelerate the system development and reduce the cost. However, the commercial components are difficult to match supercomputing perfectly. For example, the

* Corresponding author (email: gaojgsig@163.com)

1) Top500. <http://top500.org>.

GPUs of NVIDIA and AMD have to take into account several applications including graphics, artificial intelligence, HPC, data center. Instead, the processors, system architecture, system software and applications in Sunway supercomputers are all customized for HPC, hence, making the system performance and energy efficiency world leading. This technology roadmap will be firmly held in Sunway exascale supercomputer.

This paper systematically demonstrates the collaborative design including the hardware, software and applications in Sunway supercomputer architecture. Also, the key challenges of scalability, energy consumption, data movement, programming and availability are deeply analyzed. Furthermore, the corresponding solutions are also proposed, some of which have been effectively validated on Sunway Taihu-Light [17] and Sunway exascale prototype system. This paper also presents the construction plan of Sunway exascale supercomputer, and provides an outlook for the performance of future exascale supercomputers.

The subsequent chapters are organized as follows. Section 2 introduces the main requirements for exascale computing including the parallel computing application and the supercomputing center operation. Section 3 presents the technical challenges and corresponding solutions for exascale supercomputing. Section 4 presents the architecture of Sunway exascale supercomputer. Section 5 analyzes the performance of the future exascale supercomputer and Section 6 concludes the paper.

2 Main requirements for exascale supercomputing

The main requirements for exascale supercomputing involve both the parallel computing application and supercomputing center operation, which are inherent motivations of the development of the exascale supercomputer architecture.

2.1 Application of parallel computing

Recently, the complex application systems in the areas of science and engineering are moving towards multi-scale, strong nonlinear coupling and three-dimensional computing. Through the analysis of computation and data movement features, it can be seen that complex applications exhibit novel features such as dynamic changes of computation and data movement with time [20], more prominent discrete and sparse features with non-localized data [21], complex task flow features at macroscale [22] and complex instruction flow features at microscale [22] (mixed [23] and/or variable precisions). Thus, the scalability, data movement efficiency and usability of supercomputer systems are facing great challenges.

For system scalability, complex applications contain a large number of multi-scale, multi-model computational problems and complex task flow features, which require a collaborative design in applications, algorithms, and architecture. This tight-bonding design is necessary to build multi-state and multi-scale systems in order to achieve effective mappings of sub-problems, and ensure the high application efficiency when the system scales up. For data movement, the dynamic and sparse features of some complex applications pose a great challenge to the architecture design. As the problem scales up, the amounts of data, memory access, and communication volume increase tremendously and data interaction becomes more complex. It is necessary to innovate on-chip interconnect and on-chip cache to alleviate the performance bottleneck of discrete memory access bandwidth. Furthermore, the supercomputer network needs to be greatly improved to solve the performance bottleneck of irregular communication. As regards usability, the development of exascale parallel applications is facing unprecedented difficulties. The contradiction between the high complexity of supercomputer architectures and that of application systems is becoming more severe, resulting in a more prominent programming wall.

2.2 Operation of supercomputing center

With the continuous improvement of supercomputers, energy consumption has become the most serious problem. The power consumption is particularly important for the supercomputing centers. Based on the current design methods, the power consumption of an exascale system is expected to reach tens of megawatts (MW) despite the general benefits from the new integrated circuit technology and architecture design. The huge power consumption of exascale supercomputer makes these supercomputing centers difficult to support such infrastructure and funding. Therefore, it is urgent to launch open researches on architectures to increase computing density, reduce data movement, and improve the energy efficiency.

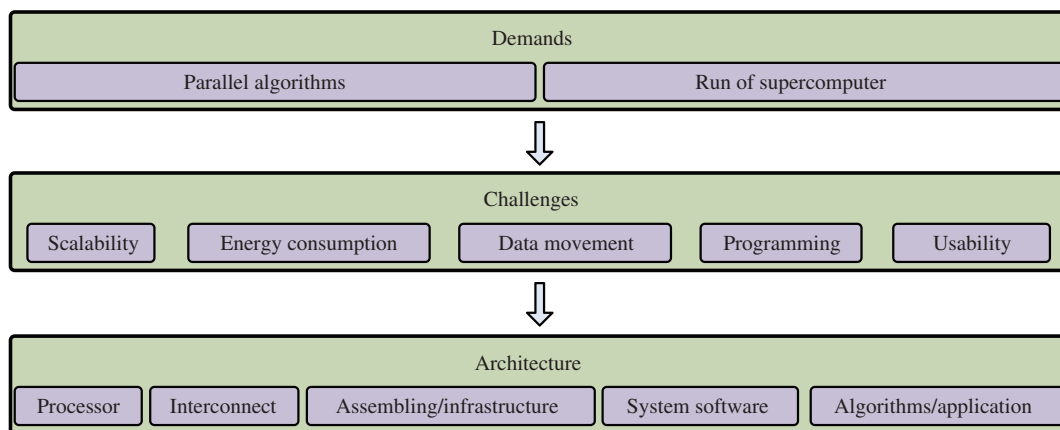


Figure 1 (Color online) The mapping from exascale demands to Sunway architecture.

For availability, the large amounts of processors, memories, storage units, networks, and power supplies make the random device failure a significant challenge. In addition, the error of large-scale software system is another important cause for system unavailability. The problem of unavailability becomes more severe as the system approaches the exascale level. Notably, the scale of some large applications has reached the level of million-cores, which puts higher demands on system availability.

2.3 Mapping of requirements to architecture

Based on the above analysis of the requirements of both the exascale parallel application and supercomputing center operations, the major challenges of supercomputer architecture design are listed as follows.

- Scalability: how to ensure the efficient operation and management as the system scales up.
- Energy consumption: how to reduce the system power consumption of both peak and running.
- Data movement: how to solve the “memory wall” and “communication wall” to ensure the application efficiency.
- Programming: how to improve the convenience of using large-scale systems and transplant the existing software systems to exascale system.
- Availability: how to ensure the high stability and high availability of the exascale system.

To improve the comprehensive performance of the Sunway supercomputer, we analyzed the above demands and challenges and then carried out customization and co-design including the processors, interconnect assembly structure, infrastructure, system software, parallel algorithm and application, and proposed an effective system architecture for Sunway exascale supercomputer, as shown in Figure 1.

3 Challenges and solutions

3.1 Scalability

3.1.1 Challenges

Stronger processors and larger-scale parallelism are imperative for the rapid improvements of supercomputer performance. Although the performance of single processor core has grown over the last decade by increasing frequency and single instruction multiple data (SIMD) width, the improvement of overall system performance relies much more on increasing the amount of processor cores.

The top 3 supercomputers in the TOP500 list basically represent the most advanced computer architectures from the perspective of scalability, which is the most significant reference for building an exascale system.

The total computing capability, total number of cores, and average computing capability per core of the top 3 supercomputers from 2008 to 2019 are shown in Figure 2. It should be noted that in each subfigure the value stands for the mean value of top 3 systems. It is obvious that the improvement of total computing capability increases more than $131.1\times$ (Figure 2(a)), which attributes mainly to the increase of core numbers ($44.2\times$, Figure 2(b)) yet merely to the contribution of computing capability

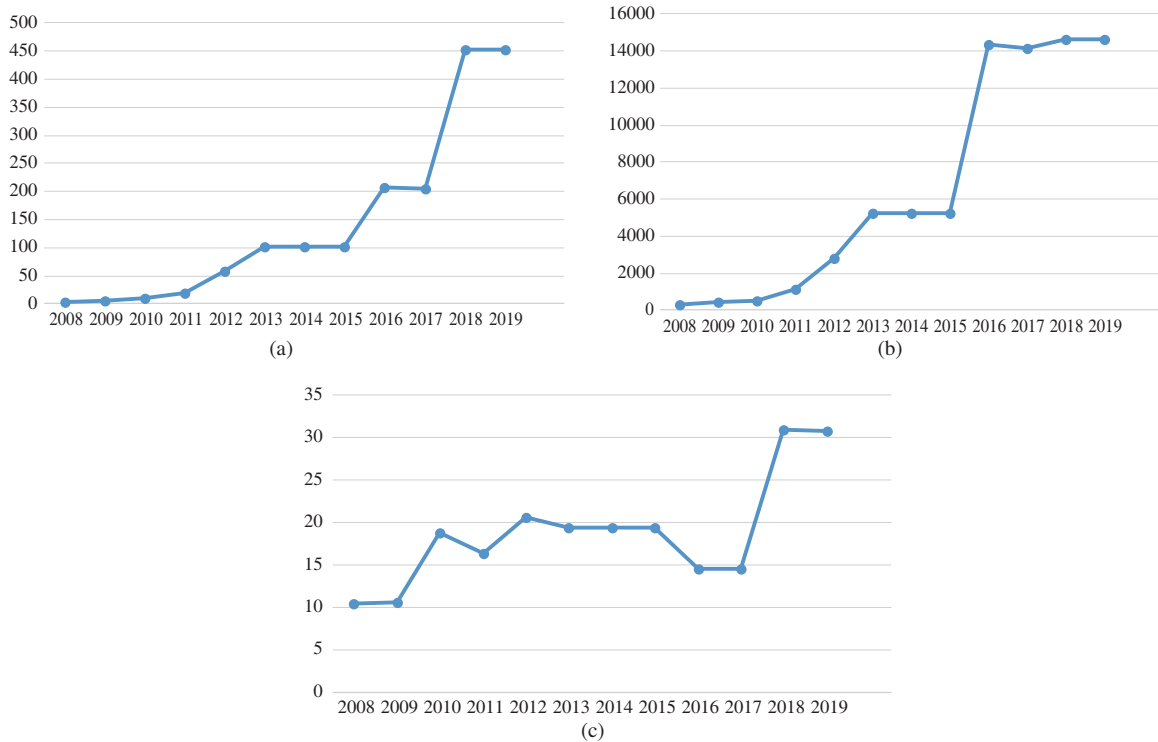


Figure 2 (Color online) The computing capabilities of top 3 supercomputers from 2008 to 2019. (a) Peak performance; (b) total number of cores; (c) peak performance of each core.

per core ($3\times$, Figure 2(c)). Thus, we can conclude that increasing the number of cores will still be an effective way to boost the construction of exascale systems in the coming year. The number of total cores of an exascale system is estimated to reach tens of millions. It is a huge challenge to ensure the system efficiency of operation and management as it expands to exascale.

3.1.2 Solutions

Scalability challenges need to be addressed from multiple technical dimensions as discussed below.

(1) High-performance many-core processor with on-chip heterogeneous integration. Shenwei (SW) many-core processors were used in several generations of Sunway supercomputers, of which the architecture is mainly customized for supercomputing. For example, SW26010 many-core processor has achieved the highest double-precision floating-point computing performance among the contemporaneous processors in spite of its relatively behindhand IC process. This advancement has greatly reduced the total number of processors in Sunway TaihuLight system [17]. Also, the more serious challenge of scalability caused by the increased computing nodes has been avoided.

The performance of SW many-core processor is greatly improved by integrating a large number of simplified computing cores based on the fact that the HPC applications are usually separable and regular. More complex general purpose core is also a necessary component to deal with the serial part of the program and meet the diversity of applications in supercomputing centers. Unlike the “CPU + accelerator” method, SW many-core processor heterogeneously integrates different types of cores in a single chip. In the heterogeneous architecture, a few powerful management processing elements (MPEs) are responsible for discovering the instruction-level parallelism and managing the chip, while the large amounts of computing processing elements (CPEs) aim to handle the thread-level parallelism, which greatly improves the chip performance. The heterogeneous property of this many-core processor can provide both the flexibility of the general purpose CPU and the high performance of the accelerator, increasing the computing density effectively. Notably, unified instruction sets are used to facilitate the design and compatibility of the software system.

The architecture of SW many-core processor is shown in Figure 3, in which the main components such as MPEs, CPE cluster, protocol processing unit (PPU), memory controller (MC) and system interface (SI) are connected by a high-bandwidth on-chip network. The computing elements in the CPE cluster

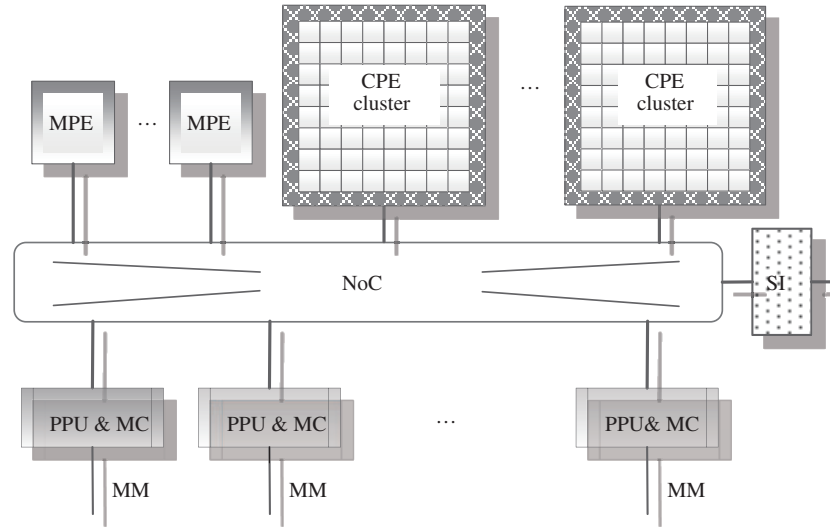


Figure 3 Heterogeneous architecture of SW many-core processor.

are arranged in a tight-bonded array. The percentage of each component and the number of elements in a CPE cluster can be flexibly adjusted based on the demands and the difficulty of physical implementation. The computation capability of SW many-core processor can achieve over 10 TFlops to support the construction of an exascale system.

(2) Interconnect topology based on tight-coupled elastic super-nodes. To better cater for the characteristics of communication and input/output (I/O) of HPC, tight-coupled elastic super-node is used as the basic unit of system. In each super-node, the computing node adopts a cable-free tight-coupled full-cross interconnect structure to maximize the efficiency of data exchange for local communication-intensive applications. On the one hand, the super-node with more processors can effectively reduce the degradation of system performance when system scales up. On the other hand, the scaling up of super-node will inevitably bring complexity in implementation. The two factors should be well balanced when designing architecture. Figure 4 schematically shows the super-node in Sunway TaihuLight which is composed of 256 processors. Through high-speed and high-density engineering technology, as well as the integrated design of power supply, cooling and high-speed interconnection, the efficient and reliable operation of the system is acquired.

The central interconnect network based on the tree topology among super-nodes provides sufficient communication bandwidth and low latency. At the same time, each super-node is directly connected to the redundant hot backup of the system computing resource pool and I/O resource pool through the shared resource network, which can flexibly allocate redundant computing resources and I/O resources to a specific super-node according to the running application to achieve efficient and flexible resource sharing. The interconnect architecture of Sunway supercomputer is shown in Figure 5, which adopts three-level composite interconnect of super-node network, central exchange network and shared resource network.

Table 1 lists the Linpack performance on computing nodes, super-nodes and the whole system of Sunway TaihuLight in actual operation. It is obvious that the Linpack efficiency remains in the same level when the system scales up, indicating that the architecture has good scalability.

(3) Multi-dimensional, multi-grained parallel software and algorithms for tens of millions of cores. In the face of the scalability challenge, innovations of system management, parallel language environment and parallel algorithm are also performed to achieve efficient and scalable parallel application.

For large-scale system management, we have adopted the hierarchical parallel control method, which decomposes the large-scale one-to-many control of a single parallel level into tree-type multi-level parallel control model composed of multiple small-scale one-to-many control. Intra-level parallelism and inter-level pipeline dramatically prompt the improvement of the scalability of the system management. The start-up overhead of parallel applications on tens of millions of cores is further reduced to less than 1 min.

For parallel language design, a wide range of communications are supported, which effectively utilize the architecture advantage of the Sunway supercomputer to limit the most frequent communication within the chip to improve the scalability of extreme scale parallel applications. The combination of the dynamic

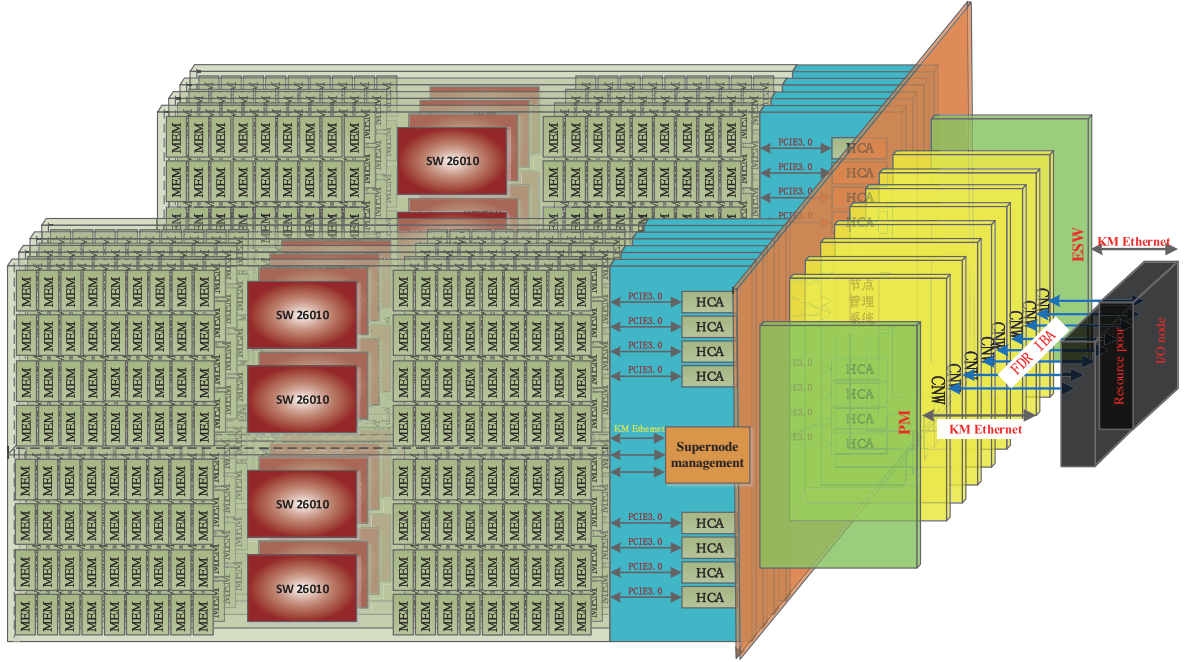


Figure 4 (Color online) The computing super-node of Sunway supercomputer.

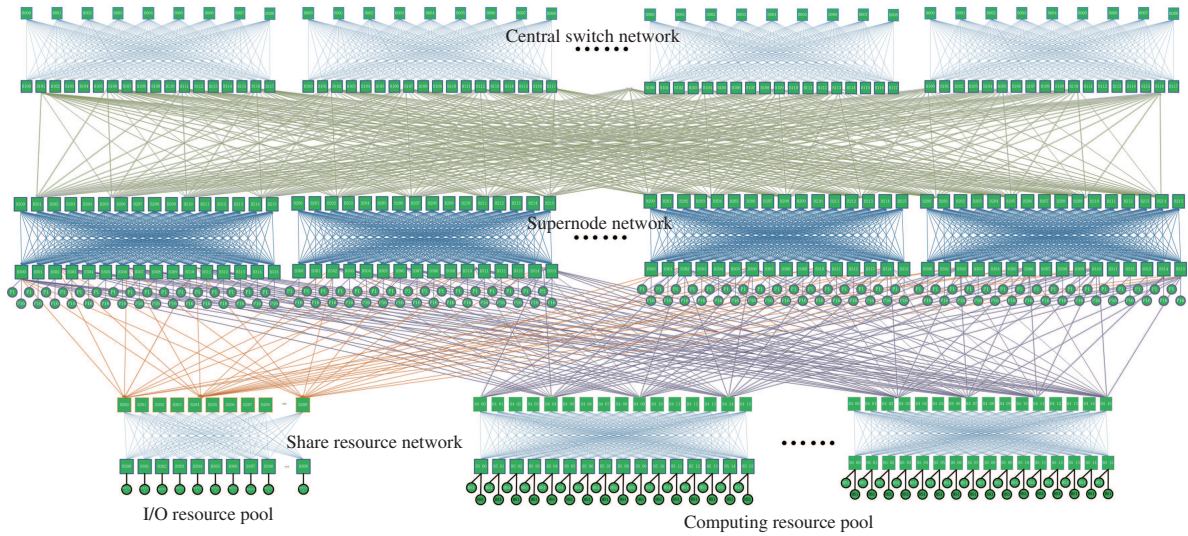


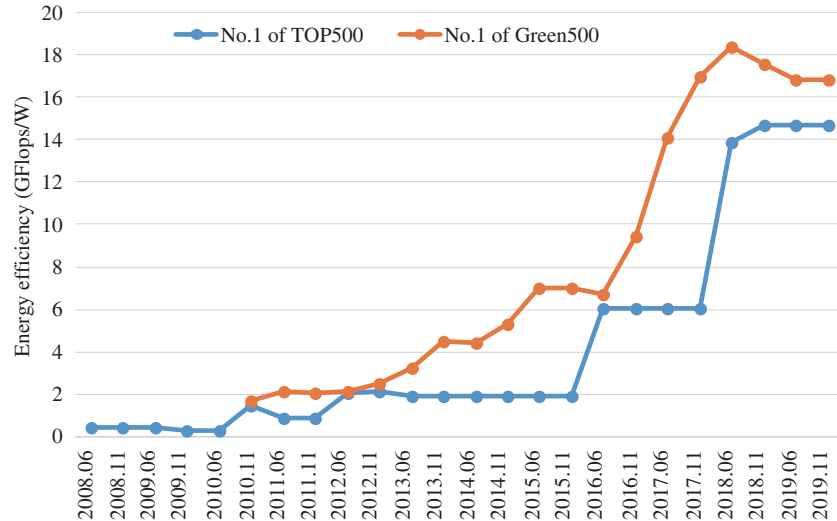
Figure 5 (Color online) The interconnect architecture of Sunway supercomputer.

and static partition connection optimization technique is adopted to reduce the expansion of the parallel language memory management overhead, and realize the efficient scalability of common communication modes.

For parallel algorithm design, a highly scalable parallel approach for cutting the tasks and data has been innovatively designed to collaborate with the unique architecture of heterogeneous many-core and high density integration. Based on various parallel models of the accelerated, asynchronous, collaborative and dynamic parallelisms of manage-computing cores, we have proposed a multi-dimensional pipeline parallelism method and multi-grain dynamic task evaluation mapping algorithm has been proposed to solve the challenges of irregular and complex applications, and improve the parallelism efficiency of some typical applications on the whole system over 80%.

Table 1 The efficiency of computing node, super-node and the whole system of Sunway TaihuLight

	Super-node (256 CPUs)	1 cabin (1024 CPUs)	4 cabins (4096 CPUs)	Total system (40960 CPUs)
Linpack efficiency (%)	82.52	80.8	77.9	74.15

**Figure 6** (Color online) The energy efficiency of the first system in TOP500 lists and the Green500 lists from 2008 to 2019.

3.2 Energy consumption

3.2.1 Challenges

Power wall is one of the biggest challenges for the development of HPC. The system reliability and stability will deteriorate severely with the dramatic increase of system power consumption.

The energy efficiency of the No.1 systems in the TOP500 and Green500²⁾ lists is depicted in Figure 6. With the rapid development of supercomputers, the peak performance of the No.1 system in TOP500 list gained 145.9× increases while the energy efficiency gained only 33.65× and 34.6×, respectively, for the No.1 systems in TOP500 and Green500 lists, indicating the improvement mismatch between system performance and energy efficiency. The energy efficiency of the No.1 systems in TOP500 and Green500 lists in 2019 is 14.72 GFlops/W and 16.9 GFlops/W, respectively. Thus, the energy consumption of exascale supercomputer is estimated to be at least 59.17 MW based on the conservative value (i.e., 16.9 GFlops/W). Thus, it is a huge challenge to keep the energy consumption at a reasonable level.

3.2.2 Solutions

(1) Low-power technologies for many-core processors. The processor power consumption is the largest part of the total power consumption of a HPC system. Low power designs in three different levels are performed in SW many-core processor to improve the energy efficiency.

At the microstructure level, signal toggle rate control, reduced structure and locality development are adopted. An architecture with global asynchronization and local synchronization is designed for SW many-core processor. The decoding width, issue width, execution sequence, speculative mechanism and SIMD width of the computing core are all optimized according to the high energy efficiency requirements. L0 Cache and locking technique of operands are implemented. Special instruction sets are designed to improve the execution energy efficiency, by analyzing the typical applications.

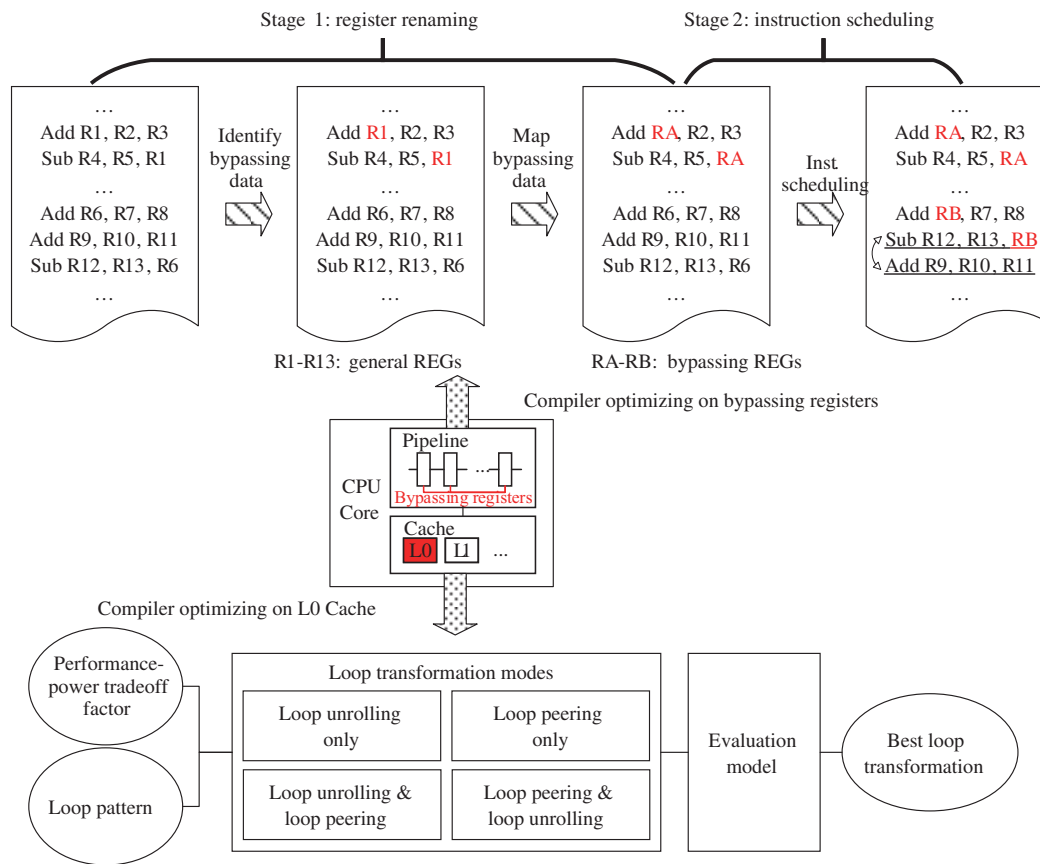
At the circuit level, the high energy-efficient circuits are used to implement the functions. For example, technologies such as gated clocking, operand isolation and finite state machine coding are fully implemented, while the use of low threshold voltage transistors is strictly controlled in order to reduce processor static power consumption.

At the logic design level, an iterative design approach that prioritizes energy efficiency is used. The module-level value change dump (VCD) vector and toggle rate are used as the design input to conduct a comprehensive optimization of power and frequency. The iteration design is also carried out when

2) Green500. <http://green500.org>.

Table 2 Comparison of energy efficiency ratios of several mainstream processors

Mainstream processors	SW 26010	Intel Xeon Phi Knight	NVIDIA Kepler-GK110B	AMD GCN 2 nd gen Grenada XT	Intel Xeon E7-8890 v4
Type	CPU	CPU	GPU	GPU	CPU
Core Num	260	72	2880	2816	24
Frequency (GHz)	1.5	1.05	0.875	0.93	2.4
Peak FP6 (TFlops)	3.618	1.01	1.68	2.62	0.92
Power efficiency (GF/W)	10.559	4.49	7.15	9.53	5.58
Time (year)	2014	2012	2014	2015	2016

**Figure 7** (Color online) Schematic illustration of low-power compilation.

designing the depth of pipeline, the multi-bit flip-flop and the clock tree to achieve optimal energy efficiency.

Further, the collaborated power management of both hardware and software is realized to dynamically adjust frequency, voltage, and operating speed of the processor. It also adopts multi-level and multi-state low-power sleep mechanism of core, low-power sleep mechanism of memory, low-power management of device, and power management of other fine-grained modules in the processor to minimize unnecessary power consumption.

Based on the above technologies, the energy efficiency of SW many-core processors used in Sunway supercomputers has been the best among the contemporary commercial processors. The comparison of energy efficiency of several contemporary commercial processors is shown in Table 2. It is obvious that the performance of SW 26010 many-core processor is the best. Further optimization of energy efficiency of microprocessor is planned to be carried out during the construction of Sunway exascale supercomputer.

(2) Low-power technologies for compilation. Power consumption is one of the biggest technical obstacles to build an exascale computer system. The compilation system of Sunway supercomputer executes a transition from the performance-oriented optimization to the performance & power balance-oriented optimization. Figure 7 shows the various techniques of low-power compilation. Firstly, the bypass of

Table 3 Comparison of various sleep measures

Means	Shallow core hibernation	Array sleep	Full chip sleep
Granularity	Single core	Computing elements array	Full chip
Control mode	OS independent control	Out-of-band control	Out-of-band control
Control overhead	ms	s	About 1 min
Power saving (%)	2	80	90

Table 4 A brief comparison between the Sunway TaihuLight and other large-scale systems (June, 2016)

System	Sunway TaihuLight	Tianhe-2	Titan	Sequoia	K
Peak performance (PFlops)	125.436	54.90	27.11	20.13	11.28
Linpack performance (PFlops)	93.015	33.86	17.59	17.17	10.51
Performance per watt (MFlops/W)	6051.3	1901.54	2142.77	2176.58	1062.69
Performance per cubic meter (TFlops/M3)	523.1	174.1	69.9	67.8	10
Node architecture	One 260-core SW CPU with 4 MPEs and 256 CPEs	Two 12-core Intel CPUs and three 57-core Intel Xeon Phi Coprocessors	One 16-core AMD CPU and one K20x NVIDIA GPU (2688 CUDA cores)	One 16-core PowerPC CPU	One 8-core SPARC64 CPU

register file will reduce the power consumption of register access. Secondly, L0 Cache is employed to reduce the power consumption of instruction fetching and decoding through multi-mode iterative optimization and benefits evaluation. Based on these techniques, a benefit of 12% power decrease is achieved at the expense of 2% execution time increase, bringing great efficiency to this compilation technique.

(3) Dynamic control of operation power. System power consumption is closely related to real-time workload. Sunway supercomputers are designed with various energy saving measures for different resource conditions and usage patterns, and a combination of various means is used to reduce system energy consumption. The comparison of three different sleep measures of Sunway TaihuLight is shown in Table 3. For idle resources, sleep measures for different resource granularity such as shallow hibernation of core, sleep of array and sleep of all, combined with the idle time threshold control mechanism, are used for demand-oriented progressive sleep to improve the energy efficiency. For operational resources, there exists remarkable imbalance among all tasks. DVFS and other methods are employed to reduce frequency and voltage of those waiting tasks for energy saving. For increasing fragmentation of idle system resources caused by scheduling, the topological aggregation of idle resources can be realized through time migration integration to form a relatively complete topology and thorough energy saving control.

(4) Efficient power supply and cooling technologies. Efficient power supply and cooling technologies are important safeguard for supercomputers under high power peaks and fast power fluctuations. In Sunway supercomputers, the high-voltage phase shift rectification system is self-developed to simplify the power conversion process and the silicon controlled rectifier (SCR) technology is also introduced to improve the power conversion efficiency for about 98.46% or more. Further, a scheme of multiphase voltage regulator with the synchronous rectification technology for processor cores is independently developed, of which the pulse width modulation (PWM) double circuits are used to achieve dynamic current balance between phases and fast transient response during the load current change. The high frequency switching and interphase magnetic field coupling technology are used to reduce the size of the passive components and improve the power density and efficiency furthermore.

An advanced natural cooling system has been designed. It directly exchanges heat between external circulation cooling water and internal circulation chilled water. This system automatically adjusts the cooling system operation mode according to the ambient temperature to reduce (or even shut down) the operating chillers to achieve energy saving. We also developed a large-size double-sided reinforced heat-exchange cold plate, which adopts convection enhancement and flow-field equalization technologies to effectively solve the thermal problems of computing nodes and power supply devices.

Based on the above low-power technologies for processor, compilation, system operation power supply and cooling, the energy efficiency of Sunway TaihuLight outperforms other large-scale systems (see Table 4) in the TOP500 list of June 2016. These advanced technologies will be adopted and further developed in Sunway exascale supercomputer to achieve better energy efficiency.

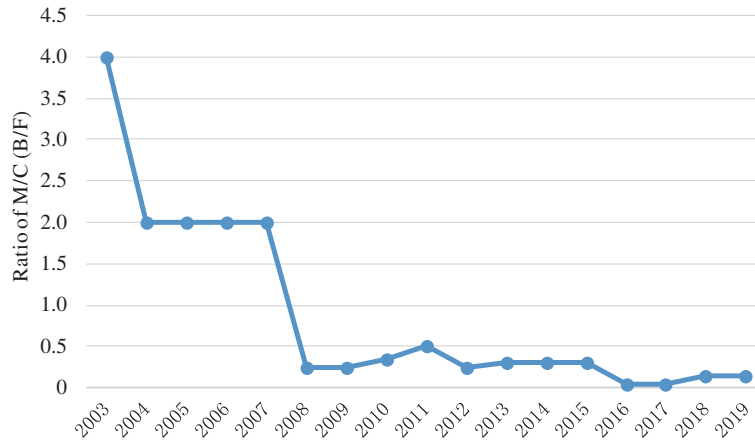


Figure 8 (Color online) The ratio of M/C of the No.1 system from 2003 to 2019.

3.3 Data movement

3.3.1 Challenges

The increasing mismatch between the improvements of data movements (e.g., memory access and communication) and computing performance creates a huge obstacle for the performance of HPC. In the last 20 years, the performance of microprocessor has kept a rapid development which followed the Moore's law, thanks to the multi/many-core techniques. However, the improvements of dynamic random access memory (DRAM) access bandwidth and latency stay only at about $20\times$ and $1.3\times$, respectively. Owing to this mismatch, the ratio of memory access and computation of No.1 supercomputer has been decreasing distinctly since 2003, as shown in Figure 8. According to the Roofline model [24], an application will become memory-bound when its arithmetic intensity is smaller than the ratio of its peak performance and peak bandwidth. Therefore, the memory access wall has become one of the key barriers against building an exascale supercomputer.

The performance of the communication between computing nodes has become another obstacle as supercomputers continue to scale up. The bandwidth of physical link in interconnect networks increases only about 26% per year, which is far behind the performance improvement of computing nodes and systems. Thus, the system efficiency is greatly limited for communication-bonded applications. Further, the communication latency deteriorates dramatically by the increasing physical links between computing nodes as the system enlarges.

In addition, the performance gap between data storage system and computing capability is also expanding, which makes storage system to become another bottleneck for system development. In recent years, with the rapid reduction in the cost of solid state drive (SSD) storage media, hybrid storage is becoming a trend. In this architecture, SSD serves as the burst buffer storage while disk serves as the large capacity section. However, this causes more frequent data migration overhead between SSD and disk, and poses greater challenge for the storage system design and application data management [25,26].

3.3.2 Solutions

(1) Optimized design for on-chip memory hierarchy. SW many-core processors make good use of on-chip resources to design a specific hierarchical memory architecture to alleviate the performance loss caused by data movement, as shown in Figure 9.

A reconfigurable local data memory technology is used inside the CPEs. Based on this technology, the local data memory can be managed by software as scratchpads, or hardware as caches. Different modes can be used simultaneously to support the dynamic division of capacity. The scratchpad method is used for the predictable regulated data to achieve the accurate utilization of space, and cache is used for the data which are difficult to be efficiently managed by software to reduce the complexity of software use.

In order to improve the data reuse and cooperation within the CPE cluster and expand the size of on-chip task sets, an on-chip multi-dimensional parallel data communication architecture was designed. This system utilized direct register-level communication, configurable multi-grained data sharing, multi-pattern data stream transfer, and fast synchronization technique, to efficiently realize the on-chip data

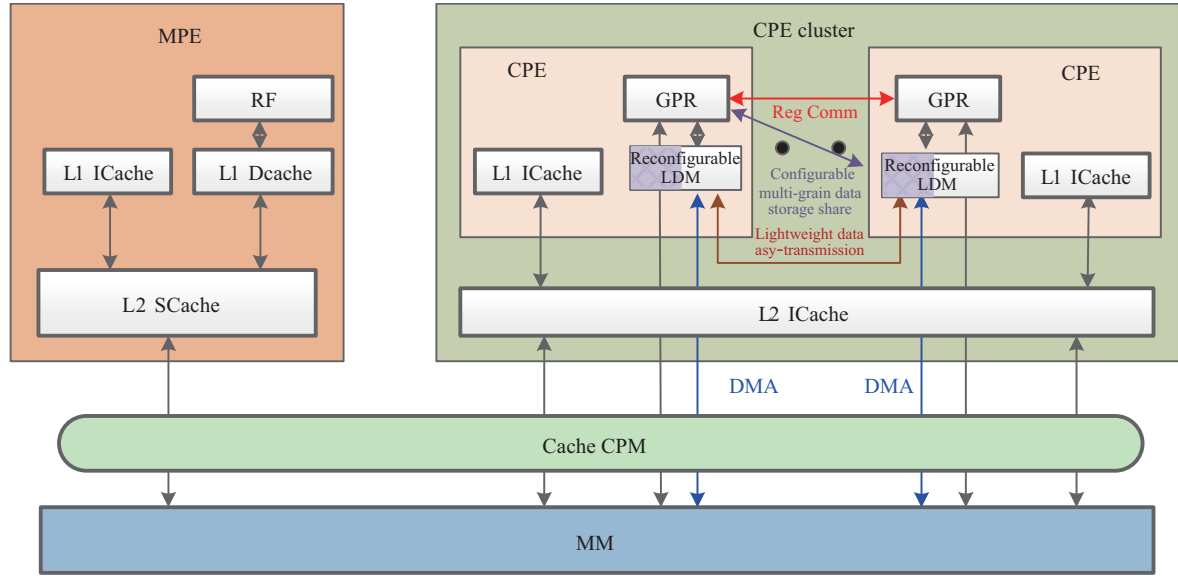


Figure 9 (Color online) The on-chip hierarchical memory architecture of SW many-core processors. GPR: general purpose register, LDM: local data memory, DMA: direct memory access, CPM: coherence process module, MM: main memory.

sharing and memory access overlap and improve the cooperative execution efficiency of the CPEs. The fine-grained, low-latency and handshake-free movements between the CPEs are realized using register communication. The configurable multi-grained data memory sharing technology can reconfigure the local data memory within the cores into different capacities and various shared ranges of memory spaces, allowing other cores within the CPE cluster to use different granularity for discrete access, which effectively adapt to the applications with irregular memory sharing access. The multi-mode asynchronous data stream transmission technology includes data asynchronous transmission between the CPE's data memory and main memory, and data asynchronous transmission among local data memory is employed in the CPE cluster to effectively achieve the parallelization of computing and data movement for regular applications. Through the comprehensive use of the above data optimization techniques, Sunway supercomputers can achieve high computational efficiency in case of a weak memory access/computation ratio. For example, for the world's fastest system, Summit, the performance of the computing node, which consists of two CPUs and six GPUs, is about 45.6 TFlops; the memory access bandwidth is about 5.6 TB/s; the memory access/computation ratio is 0.123 B/F. In comparison, the computing node of the Sunway TaihuLight is composed of a CPU with a performance of 3.168 TFlops, the memory access bandwidth of 136 GB/s, and the memory access/computation ratio of 0.043 B/F. It is noted that the Linpack efficiency of Sunway TaihuLight is higher than that of Summit based on a much smaller access/computation ratio, indicating the important role of the on-chip multi-dimensional parallel data communication architecture.

In terms of instruction stream, we used an instruction merging technology to design an L2 instruction cache shared by the CPE clusters. The merge of missing instruction in L1 instruction Cache was implemented by hardware and returned to the CPEs in multi-broadcast form to improve the utilization of memory bus bandwidth. Meanwhile, each CPE cluster integrates a shared L2 instruction cache with larger capacity, which further decreases the instruction miss delay.

Another important feature of SW many-core processor is the support for coherence sharing between the MPEs and the CPEs with directory based protocol. This not only significantly reduces the data movement between cores, but also efficiently supports fine-grained interactions between different cores, which is especially important for applications with irregular data sharing access. For example, although the peak performance of Sunway TaihuLight supercomputer was surpassed by Summit system in 2018, Sunway TaihuLight still has better rank in Graph 500 owing to this feature.

(2) High-throughput multi-track pan-tree network. In the aspect of communications, Sunway supercomputers adopt a self-developed network chipset to comprehensively improve the interconnection network performance and reliability based on the innovations from topology to the messaging mechanism design.

In the Sunway exascale prototype system, SW interconnection network introduces a whole new ar-

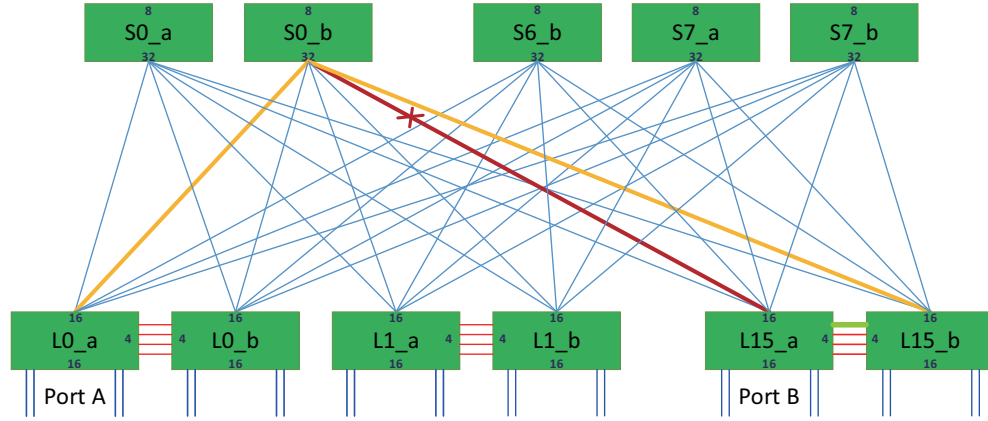


Figure 10 (Color online) The pan-tree structure of interconnect.

Table 5 Comparison of aggregated message latency implemented by the hardware and the software (unit: μs)

Nodes	256	256	512	512	1024	1024
Implementation	Software algorithm	Hardware implementation	Software algorithm	Hardware implementation	Software algorithm	Hardware implementation
8B full reduction (AND synchronization)	54.55	7.71	79.64	10.30	107.97	13.49
1 kB broadcast	28.00	9.29	33.93	9.66	37.58	11.67

chitecture of high throughput multi-track pan-tree network. The network employs a self-developed SW high-order routing chip (SWHRC) to implement the same leaf-spine architecture, and the adopted self-developed high-performance network interface chip uses a shared dual-engine dual-port architecture to realize an efficient out-of-order communication mechanism based on a fine-grained packet-level dynamic network ports selection. Each packet can dynamically select the final destination port based on the link states (i.e., quality, busy or idle) of the port to increase the network port utilization and improve communication reliability. The network layer utilizes the redundant paths of the multi-track pan-tree network structure, combining adaptive routing policies, route reconfiguration policies and transport layer message retransmission techniques, to maintain uninterrupted message service for an application in the event of single port failure, single chip failure, or even a whole switch failure.

The problem of downlink paths being determined by uplink paths in fat tree networks, which causes downlink path failures affecting node communication, is solved in the pan-tree structure. As shown in Figure 10, when a downlink path from A to B fails, the downstream packet can still reach B smoothly by first reaching the brother switch, which B is connected to and then arriving at the switch connected to B. The pan-tree structure has better fault tolerance feature than the standard fat-tree structure.

The network system of Sunway Supercomputers also implements a hardware aggregation communication mechanism, which adopts a fusion mode of software-defined logic trees and hardware chain-tables to realize efficient and flexible synchronization, multi-broadcast, protocol and other aggregation messages. The mechanism breaks through the constraints of network topology, enhances the adaptability under different network topologies, and solves the contradiction between the limited hardware resources of aggregation tree and the huge requirements of software. Meanwhile, it also simplifies the hardware design and software usage patterns, thus effectively improving the processing power, scalability and practicality of the aggregated communication. Table 5 shows the comparison of messaging latency between hardware and software implementations, from which the aggregated communication latency of hardware obviously improves and increases less with scale.

(3) Software and algorithms. The software and algorithms of Sunway supercomputers are designed with a strong focus on localization, limiting the data movement to the local range as small as possible for more data exchange. Within a single CPE cluster of the chip, relying on the high-speed on-chip network, we can design an efficient peer-to-peer and aggregated communication mechanism between threads, as well as fine-grained software parallel pipelines between threads and other optimization technologies. These advanced techniques can support the frequent exchange of critical data. The efficient data movement is achieved by sharing the main memory within a chip. The parallel languages and algorithms achieve

efficient data exchange between nodes based on the high-speed full-exchange network within a super-node.

As for parallel algorithm and parallel program design, various methods are proposed to reduce data movement and fully exploit processor performance. Based on the efficient on-chip array communication mechanism, an efficient implementation method for solving sparse-class problems is developed; a butterfly-optimized many-core implementation technique for global collection communication is also proposed; a hierarchical protocol communication strategy for different communication requirements is designed to support the efficient global communication at the scale of the whole machine; and a large-scale multi-level flexible grouping parallel I/O method is proposed to achieve the efficient utilization of resources such as computing, memory access, communication, and I/O processing.

Specific design is performed in software level for the burst buffer storage system architecture. Firstly, flexible resource management is achieved to make full use of SSD resources to improve the I/O performance of applications. Then, automatic data migration from SSD to disk is realized to improve the utilization efficiency of SSD storage. Moreover, I/O conflict awareness and fine-grained QoS among multiple applications are designed to dynamically adjust the in-time allocation strategy of multi-layer storage resources to reduce the impact of local hot spots and predict the performance reduction [27].

3.4 Programming

3.4.1 Challenges

To maximize the computing performance of many-core processors, a reduced design scheme for architecture is adopted, which is different from that of multi-core processors. For example, software-managed scratch pad memory (SPM), rather than the hardware-supported cache, serves as the main choice for high-speed on-chip memory. This will, to some extent, put limitation on the communication between cores and make compiling and programming more difficult. Based on the stronger computing capability of many-core processors, the mismatch between computing capability and data movements, such as memory access and communication, becomes more serious than that of multi-core processors. The challenges are effectively overcome by using on-chip network and local communication in Sunway architecture while putting forward higher requirements for programming and optimization.

In general, the complexity of many-core parallel systems (e.g., parallel levels and memory levels) and varieties of many-core processor structures make it particularly difficult for programming and incompatible among each other. Furthermore, the migration and optimization of legacy codes and development of new programs are more difficult than those of multi-core systems. Therefore, parallel programming has become a great challenge for many-core systems.

3.4.2 Solutions

To meet the programming challenges of heterogeneous many-core systems, the comprehensive design of programming languages, application frameworks, and high-performance function libraries is performed to shield the underlying details as much as possible, so as to reduce the programming's complexity.

(1) MPI³⁾+X programming environment. MPI is the most famous programming environment for parallel applications of high-performance computing. MPI+X has become the mainstream in many-core systems, in which MPI is used to exchange data between the nodes and the programming languages, such as OpenMP⁴⁾, OpenACC⁵⁾, CUDA⁶⁾. Sunway TaihuLight and the Sunway exascale prototype are compatible with OpenACC 2.0 standard. Programmers only need to add a small amount of compilation instructions to the core code to achieve efficient multi-core parallelism. Generally, the changing amount of the program is less than 1%.

(2) New model and language of many-core programming. Based on the characteristics of heterogeneous paralleling and main memory sharing of SW many-core processor, we have developed the accelerated computing model with heterogeneous fusion and parallel C programming language [28] with unified architecture.

The accelerated computing model of heterogeneous fusion is shown in Figure 11. The model has the following features. (i) The process is responsible for management, control, communication, I/O and other complex operations, while a large number of accelerated threads are responsible for accelerating the core

3) MPI documents. <http://www.mpi-forum.org>.

4) OpenMP. <http://www.openmp.org>.

5) OpenACC. <http://www.openacc.org>.

6) NVIDIA CUDA. http://www.nvidia.com/object/cuda_home_new.html.

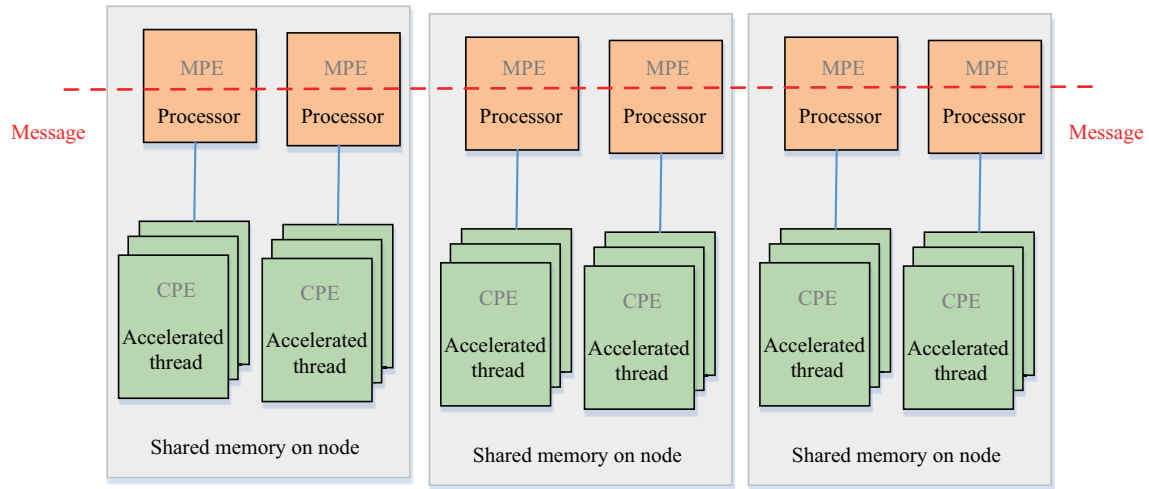


Figure 11 (Color online) Accelerated computing model with heterogeneous fusion.

code. (ii) The process and the accelerated threads share main memory, which is more convenient and efficient than the programming models on x86+GPU or x86+MIC, because it does not need to transfer data between two kinds of memories. (iii) By sharing the extended description of keywords within node, shared programming and improved efficiency of data exchange can be both accomplished. (iv) By exchanging the data through messaging and supporting multi-level local descriptions between processes, the parallel language can be efficiently extended to the large-scale environment.

The random combination is supported among messaging, sharing and accelerated computing by the model and language discussed above, which can effectively describe the parallelism of the heterogeneous systems in multiple dimensions. Also, it provides a much better global perspective for programming and compiling. Notably, there is no need to master two or three programming languages for MPI+X model, thus improving the programming efficiency and compiler optimization.

(3) Framework for key applications. We have developed an application supported framework for scientific and engineering computing with many-core architectures. The common requirements of polymorphic applications are extracted to closely match the characteristics of architecture and application. A multilevel parallel model of heterogeneous perception is adopted to redesign and thoroughly optimize the application algorithms. It supports the calculation of finite volume, finite differences, and finite element methods. It also supports the calculation of both structured and unstructured grids. These could effectively accelerate the development and optimization of scientific and engineering computing applications, and support the multiple deployments of large and complex applications.

(4) High-performance libraries for many-core processors. High-performance libraries can be used to effectively shield the underlying details of many-core systems, so as to reduce the complexity of programming and improve the performance of the applications. Efficient basic math libraries, dense linear algebra libraries, sparse linear algebra libraries, fast Fourier transformation libraries, etc., have been developed on Sunway TaihuLight and Sunway exascale prototype. The acceleration of the common functions in parallel applications can be acquired by adopting the interfaces of related high performance function libraries to effectively decrease the programming complexity.

3.5 Reliability and availability

3.5.1 Challenges

With the rapid developments of supercomputers, the prominent increase of device number, hardware complexity and software complexity raises great challenges for reliability and availability, which is critical for exascale supercomputers.

The number of components of Sunway TaihuLight has exceeded 150 million, which is expected to reach 300 million of Sunway exascale supercomputer. The dramatic increase is expected to make mean time between failure (MTBF) fall to few hours or even less than the time interval required for the checkpoint updating, resulting in the system being unavailable [29–31].

3.5.2 Solutions

To overcome the above challenges, we mainly enhance the basic reliability and system availability during the whole development process of the Sunway supercomputer.

(1) Enhancing the basic reliability. It is important to predict and allocate the reliability of supercomputers. The reliability of components can be improved through device selection, reliability tests, derating design, hardware redundancy and so on. To meet the needs of the HPC domain, the reliability enhancement techniques that can be adopted are as follows.

Real-time result-verification for the computing units. The computing unit is one of the most important parts of the SW many-core processors, which is used frequently and covers a large area on the chip. The real-time verification based on the remainder algorithm is capable of covering the error rate up to 93% for floating-point operation and integer multiplication, effectively preventing the silent error propagation.

Enhanced memory reliability by efficient correction for random/burst errors. To meet the high reliability of supercomputers on the memory system, we have designed a new error-correcting code of memory that can accommodate both random errors and burst errors. By optimizing the decoding circuits, a 576-bit RS code is implemented in the memory controller. And the memory access transaction retransmission is created based on full error detection.

Interconnect network reliability. The two-track pan-tree structure is used to increase the redundant ports. Combined with the forward error verification codes, link degradation, link retransmission, credit recovery, path reconstruction and other reliability techniques, high reliability of the interconnect network is realized in large-scale systems at high link transmission rates.

Double-sided reinforced cold plate for heat exchange with a multi-material composite three-dimensional channel. Local enhanced heat transfer technique is adopted on the cold plate design for high heat flux parts such as computing node. Combined with the rigid and flexible composite contact, the board temperature of the computing nodes and the junction temperature of the components are well controlled. As a result, the junction temperature of the CPU in the whole machine is controlled below 50°, which prominently improves the basic reliability of the system (i.e., the reliability of the components will decrease twice with 10° increase of component temperature).

(2) Improving the system availability. Highly available systems are achieved through the collaboration between hardware and software, the active fault tolerance based on prediction and the passive fault tolerance with multiple strategies.

High availability by the collaboration between hardware and software. The system availability is decomposed into various parts of software and hardware, and then fault-tolerance technologies are selected for them. The hardware uses correction, retransmission, redundancy, rollback, and other mechanisms, while the software design includes fault monitoring, on-line diagnosis, fault tolerance, and repair processing. The capacities and effects of fault-tolerance are improved by combining the fault tolerance control mechanism with active/passive fault tolerance and multi-scenario individualized fault tolerance measures (see Figure 12).

Active fault tolerance based on fault prediction. The system fault prediction analysis model is established by utilizing theoretical modeling, machine learning and other methods. Further, reliable fault prediction is achievable by combining historical and real-time data mining, training, learning, etc. According to the results of fault prediction, the status of system resource, the characteristics and cost of subject fault tolerance, and the fault prediction driver, the active migration of local hidden hazard and active control perform fault can be employed based on the checking points, respectively to improve the availability of the system.

Passive fault tolerance with multi-strategy. Once a fault occurs, a variety of methods is used to analyze and assess the cause and loss of the fault to match the most appropriate measures of fault tolerance. Under the triggering of a fault event, message retransmission, route reconstruction, resource takeover, rollback recovery, local degradation and other methods are chosen, respectively, for targeted fault tolerance to reduce the loss of fault tolerance and improve the availability of the system.

Through the availability design in multi-level (i.e., system level and hardware/software level), high reliability and availability are achieved on the Sunway supercomputer system. The actual availabilities of Sunway TaihuLight and Sunway exascale prototype have both exceeded 99%.

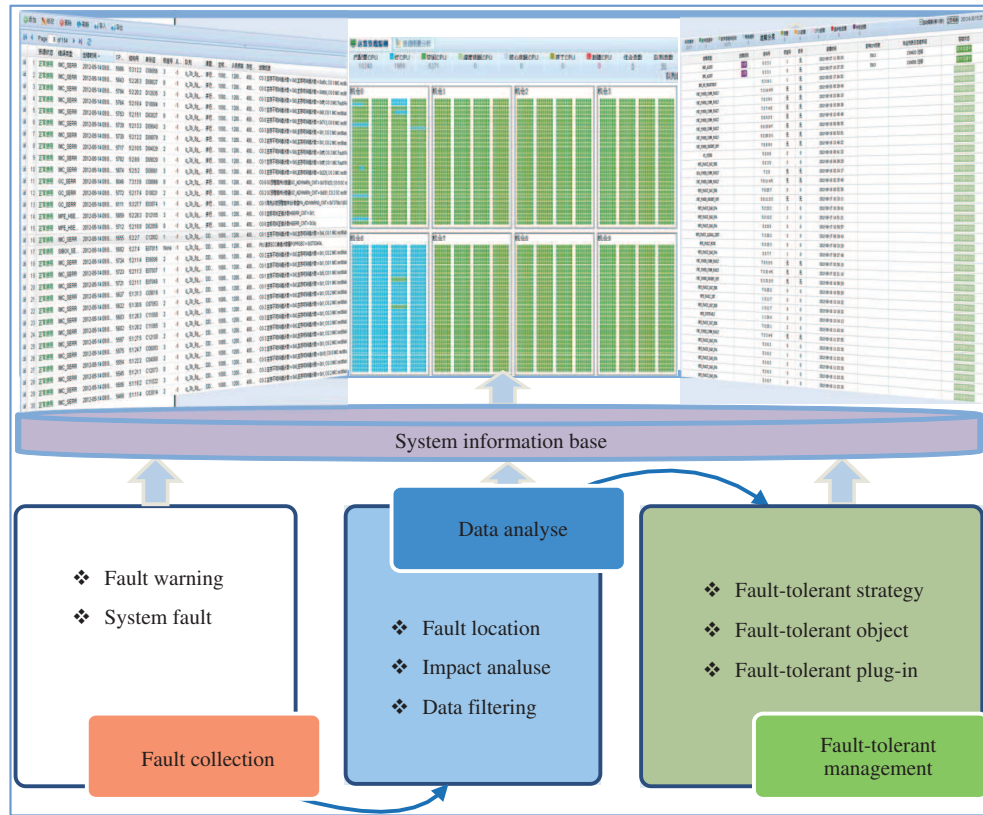


Figure 12 (Color online) Schematic of initiative-passive fault-tolerance control mechanism.

4 The Sunway exascale architecture

Based on the above analysis of challenges and corresponding solutions, we propose the Sunway exascale architecture which gives instructions to the construction of the Sunway exascale supercomputer. In this system, the key components, such as many-core processors, network chipsets, and software systems, are all self-designed to ensure that the key technologies can be independently controlled. The hardware system consists of more than 80000 processors which will provide 1 and 4 EFlops computation performance of double-precision and half-precision, respectively. Combined with the configurations of the bi-bandwidth of interconnect (more than 1000 TB) and the message passing delay (less than $1.5 \mu\text{s}$), these important technical indicators are expected to reach a world-leading level. The software system mainly includes: the operating system and compiler for many-core processors, large-scale scheduling management and massive storage system, OpenACC, debugger, support environment for big data and development environment for artificial intelligence.

4.1 Hardware system

Figure 13 schematically demonstrates the hardware system of Sunway exascale supercomputer which consists of a self-designed high-performance many-core processor, computing system, interconnection system, storage system, maintenance system, power supply system and cooling system. It is worth to note that the application-specific systems, such as artificial intelligence acceleration system, can be flexibly connected for dedicated needs.

The new generation of domestic SW many-core processor in Sunway exascale supercomputer still adopts the highly efficient scalable architecture (as demonstrated in Figure 3) in which the main components are connected by an on-chip torus network. The scalable processor includes eight core-groups (CGs). Each CG includes one MPE and one CPE cluster with 8×8 CPEs. The high-performance many-core processor is designed to provide more than 12 TFlops computing capability of double precision floating point.

The MPE is a complete 64-bit RISC core, which can run in both the user and system modes. It supports superscalar processing and out-of-order execution. These features make MPE an ideal choice for dealing

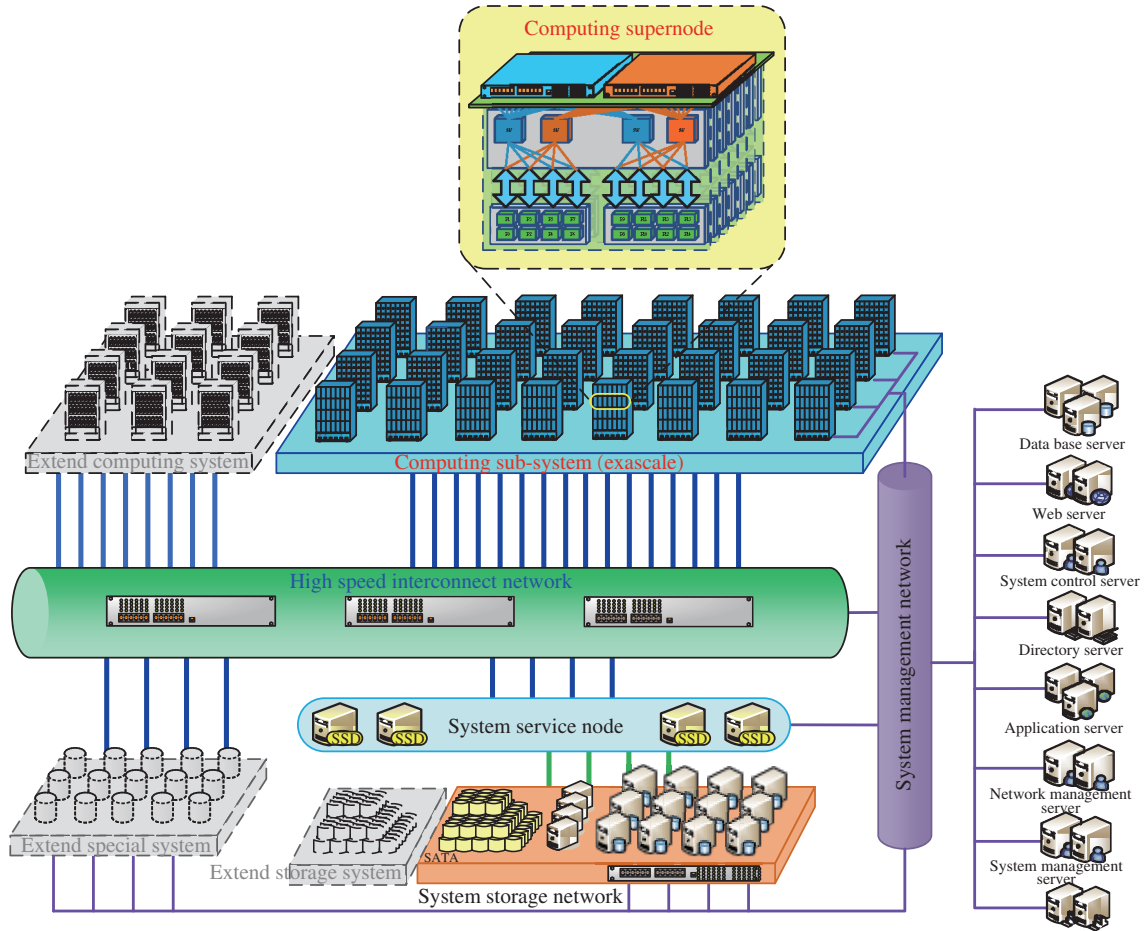


Figure 13 (Color online) Hardware architecture of Sunway exascale supercomputer.

with the tasks of management and communication. The CPE is also a 64-bit RISC core and supports 512-bit vector instructions. Different from MPE, the CPE supports limited functions and can only run in user mode. This promises the maximum aggregated computing power, while minimizing the complexity of the micro-architecture. The CPE cluster is organized by a 4×4 concentrated mesh (CMESH) network on chip (NoC), with each node including 4 CPEs. This design can achieve low-latency data communication among the 64 CPEs. The torus network tackles the message transmission between the components on the chip and realizes the distributed shared memory consistency.

The computing system is the core unit of a supercomputer to achieve ultrahigh computing capability. There will be more than 80000 independently designed computing nodes in Sunway exascale system, which can be configured as standard or fat computing nodes based on various memory capacity.

The interconnect system, consisting of computing network, storage network and management network, implements data transmission between different types of nodes. The computing network is based on the self-developed Sunway network chipset, which realizes high-bandwidth and low-latency communication; the storage network adopts the Sunway network technology for storage; the management network uses Ethernet protocol to connect all nodes and management units of this system.

The computing network is constructed based on the high-throughput multi-track pan-tree network topology and adopts the high-efficient hardware aggregation communication mechanism and uninterrupted communication failure processing mechanism. Each network interface chip (NIC) supports two $16 \times$ PCIE 4.0 interfaces and $4 \times$ 56 Gbps network interfaces, which supplies more than 400 Gbps bandwidth for the corresponding connected CPU. The maximum step size of the whole computing network is constrained to only 6, which is attributed to the proposed topology and the MPI latency of the point-to-point communication is less than $1.5 \mu\text{s}$. Besides, transmission control protocol (TCP) is compatible except for traditional MPI to enhance the system availability.

The storage system, including a global storage system and local storage system, provides high-capacity global I/O service with unified namespace and high-performance local I/O service. The global storage system mainly supports online scalability and consists of metadata server cluster, data server cluster, storage disk array, and supporting service nodes; the local storage system uses SSD in system service nodes and provides high-performance I/O services that can be dynamically deployed for the host; it also supports storage system expansion when necessary.

The maintenance system provides system configuration and management, real-time state monitoring, real-time monitoring and diagnoses of the operating environment, which covers computing, interconnect, storage, cooling, power, system service, operating environment.

The power system provides stable, reliable and efficient power supply for the computing, interconnect and the storage systems. The power system adopts a three-stage DC power conversion mode based on high voltage rectification.

The cooling system which is responsible for providing good cooling conditions for the main computer, storage, power systems, etc., provides efficient cooling technology that contains liquid cooling, heat pipe conduction, air cooling and hybrid cooling.

4.2 Software system

The software system of the Sunway exascale system mainly consists of basic software, parallel management software, parallel language environment and parallel development environment.

The basic software is designed for the SW many-core processor, and supports not only the basic compiler and toolset with C/C++/Fortran, but also the C library of the many-core processor, the basic/extended math library, and auto-vectorization tools.

The Sunway Raise OS operating system environment is the parallel management software for the Sunway supercomputer, which is responsible for the efficient management of computing, network, and storage resources; it also supports virtualization, power consumption management, system fault tolerance and on-demand resource customization. The self-designed OS provides users with an efficient, reliable, and friendly basic platform.

The parallel language environment supports the mainstream MPI+X programming method that supports the inter-node MPI 3.0 standard, and intra-node OpenCL and the OpenACC* language that are compatible with OpenACC 2.0 standard. In addition, it supports Parallel C parallel language with unified architecture, multi-level space sharing, many-core multi-level parallel description within nodes, and a concise and efficient communication mechanism between nodes.

The parallel development environment provides HPC developers with an integrated development environment, parallel debugging, and the optimization tools to solve the challenges of programming, optimization and debugging.

5 Analysis of the system performance

According to the classification criteria of science and engineering computing applications by UC Berkeley [32], the applications are divided into thirteen kinds of topics such: dense linear algebras, sparse linear algebras, spectral methods, N-body methods, structural grids, unstructured grids, MapReduce, combinational logic, graphs, traversal, dynamic programming, backtrack and branch+bound, construct graphic models, and finite state machine. The practical applications are classified into two categories based on the analysis of the computing and data movement characteristics (i.e., time complexity, space complexity, and communication complexity) of the above topics: (1) the applications with regular computing and data migration; (2) the applications with irregular computing and data migration. The analysis and estimates of these two applications on Sunway exascale supercomputer are discussed as follows.

Application with regular computing and data migration [16, 18, 33–38] has the characteristics of regular programming, large computational amount, good parallelism, and regular memory access. With the expansion of the problem scale, the time complexity, space complexity and communication complexity increases in a close linear fashion, preserving good scalability and parallel efficiency. This characteristic can be well scaled to exascale system, promising the linear performance improvement.

The applications with irregular computing and data migration [20, 21, 37, 39] create tough challenges, such as the proliferation of memory accessing and communication. The complexity of memory accessing

Table 6 Scales and performances of ten types of scientific applications

Type	Typical application and representative algorithm	Application scale of Sunway TaihuLight	Application performance of Sunway TaihuLight	Rank of Application Performance of Sunway TaihuLight	Prediction of exascale application scale	Prediction of exascale application performance
Dense linear algebra	LINPACK	12.288 millions	93 PFlops	1	About 20 millions	About 700 PFlops
Sparse linear algebra	HPCG	343.5 billion nonzero elements	480 TFlops	3	About 10^{12} nonzero elements	Over 3 PFlops
Spectral methods	FFT	16384^3	97 s/step	Maximum scale	32768^3	About 90 s/step
Multi-body problem	Universe evolution	11.2 thousand billion particles	21.3 PFlops	Maximum scale	Hundreds of thousands of billions of particles	About 160 PFlops
Structured grids	Stencil computing	5.1×10^{11} grids	25.96 PFlops	2016 Gordon Bell Prize [33]	About 10^{12} grids	About 200 PFlops
Unstructured grids	Throughput computing	10^{10} grids	—	Parallelism of tens of millions of cores	About 10^{11} grids	Parallelism of tens of millions of cores
MapReduce	MapReduce	10^7 task number	12.5 PFlops	Maximum scale [36]	About 2×10^7 task number	About 100 PFlops
Traversal of graphs	BFS	2^{40} vertexs	23755.7 GTEPS [20]	2	2^{43} vertexs	About 14000 GTEPS
Dynamic planning	Sequence comparison	800 GB gene sequences	Fixed-point computing	Parallelism of tens of millions of cores	Parallelism of tens of millions of cores	Fixed-point computing
Graphic models	Convolutional neural network	—	Core FP computing efficiency of 94% [16]	—	—	Core FP computing efficiency of over 90%

and communication can lead to bad scalability, with the expansion of problem scale. Consequently, multi-level optimization is very important to ensure the effective expansion to exascale scale level. Compared to Sunway TaihuLight, the estimated capabilities of computing, memory accessing and communication in the Sunway exascale prototype system are greatly improved, for example, the processor performance increased by 4 times, the memory access bandwidth increased by 6.8 times and the network bandwidth increased by 8 times. Owing to the increasing ratio of memory access bandwidth compared to computing performance and the increasing ratio of network bandwidth compared to computing performance, the performance of the irregular applications is improved significantly. This has been well proven by the testing results of HPCG⁷⁾ and GRAPH500⁸⁾, which indicates the effective expansion of system performance of the exascale supercomputer.

Table 6 lists the scales and performances of ten common large-scale applications implemented on Sunway TaihuLight and estimated on exascale system, respectively. Applications of combinational logic, backtrack, branch+bound and finite state machine are excluded because of their small parallel scale. For dense linear algebra applications, LINPACK is used for the evaluations as the typical application and representative algorithm, which got 93 PFlops performance on Sunway TaihuLight with 12.288 million variables and can be linearly extended to exascale level with 700 PFlops performance of 20 million variables. For sparse linear algebra applications, HPCG is used for the evaluations as the representative algorithm, which got 480 TFlops on Sunway TaihuLight with 343.5 billion nonzero elements and can be linearly extended to exascale level with over 3 PFlops performance of a thousand billion nonzero elements. For structured grids applications, stencil computing is used for the evaluations as the representative algorithm, which got 25.96 PFlops on Sunway TaihuLight with 5.1×10^{11} grids and can be linearly extended to exascale level with over 200 PFlops performance of 10^{12} grids. For Traversal of graphs, breadth first search (BFS) is used for the evaluations, which got 23755.7 GTEPS on Sunway TaihuLight

7) High Performance Conjugate Gradient Benchmark (HPCG). <http://hpcg-benchmark.org>.8) Graph500. <http://graph500.org>.

with 2^{40} vertices and can be linearly extended to exascale level with about 140000 GTEPS of 2^{43} vertices.

6 Conclusion

The developments of Sunway supercomputers is a proof that the customized design of HPC is a successful method. The Sunway supercomputers can achieve better performance and energy efficiency than contemporary supercomputers using the integrated design of system architecture, such as the processor, interconnect network, assemble structure, power supply, cooling system software, parallel algorithm and application support. This unique roadmap will be insisted on and optimized continuously for the implementation of world-leading Sunway exascale supercomputer.

Acknowledgements This work was supported by National Key Research and Development Project of China (Grant No. 2016YFB-0200500).

References

- Moore G E. Cramming more components onto integrated circuits, reprinted from electronics, volume 38, number 8, April 19, 1965, pp.114 ff. IEEE Solid-State Circuits Soc Newsl, 2006, 11: 33–35
- Dennard R H, Gaensslen F H, Yu H N, et al. Design of ion-implanted MOSFET's with very small physical dimensions. IEEE J Solid-State Circ, 1974, 9: 256–268
- Agerwala T. Challenges on the road to exascale computing. In: Proceedings of the 22nd Annual International Conference on Supercomputing, 2008. 2
- Alvin K, Barrett B, Brightwell R, et al. On the path to exascale. Int J Distrib Syst Technol, 2010, 1: 1–22
- Beckman P. Looking toward exascale computing. In: Proceedings of the 9th International Conference on Parallel and Distributed Computing, Applications and Technologies, 2008. 3
- Balaprakash P, Buntinas D, Chan A, et al. Exascale workload characterization and architecture implications. In: Proceedings of IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS), 2013. 120–121
- Dally B. Power, programmability, and granularity: the challenges of exascale computing. In: Proceedings of IEEE International Test Conference, 2011. 12
- Hluchy L, Bobák M, Müller H, et al. Heterogeneous exascale computing. In: Recent Advances in Intelligent Engineering. Cham: Springer, 2020. 81–110
- Kogge P M, Shalf J. Exascale computing trends: adjusting to the “new normal” for computer architecture. Comput Sci Eng, 2013, 15: 16–26
- Lu Y. Paving the way for China exascale computing. CCF Trans HPC, 2019, 1: 63–72
- Shalf J, Dosanjh S S, Morrison J P. Exascale computing technology challenges. In: Proceedings of the 9th International Conference on High Performance Computing for Computational Science, 2010. 1–25
- Vijayaraghavany T, Eckert Y, Loh G H, et al. Design and analysis of an APU for exascale computing. In: Proceedings of IEEE International Symposium on High Performance Computer Architecture (HPCA), 2017. 85–96
- Feng J Q, Gu W D, Pan J S, et al. Parallel implementation of BP neural network for traffic prediction on Sunway Blue Light supercomputer. Appl Mech Mater, 2014, 614: 521–525
- Tian M, Gu W, Pan J, et al. Performance analysis and optimization of PalaBos on petascale Sunway BlueLight MPP supercomputer. In: Proceedings of International Conference on Parallel Computing in Fluid Dynamics, 2013. 311–320
- Chen Y, Li K, Yang W, et al. Performance-aware model for sparse matrix-matrix multiplication on the Sunway TaihuLight supercomputer. IEEE Trans Parallel Distrib Syst, 2019, 30: 923–938
- Fang J, Fu H, Zhao W, et al. swDNN: a library for accelerating deep learning applications on Sunway TaihuLight. In: Proceedings of IEEE International Parallel and Distributed Processing Symposium (IPDPS), 2017. 615–624
- Fu H H, Liao J F, Yang J Z, et al. The Sunway TaihuLight supercomputer: system and applications. Sci China Inf Sci, 2016, 59: 072001
- Zhang J, Zhou C, Wang Y, et al. Extreme-scale phase field simulations of coarsening dynamics on the Sunway TaihuLight supercomputer. In: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, 2016. 4
- Zheng F, Xu Y, Li H L, et al. A homegrown many-core processor architecture for high-performance computing. Sci Sin Inform, 2015, 45: 523–534
- Lin H, Zhu X, Yu B, et al. ShenTu: processing multi-trillion edge graphs on millions of cores in seconds. In: Proceedings of International Conference for High Performance Computing, Networking, Storage and Analysis, 2018. 56
- Meng D-L, Wen M-H, Wei J-W, et al. Porting and optimizing OpenFOAM on Sunway TaihuLight system. Comput Sci, 2017, 44: 64–70
- Fu H, Liu W, Wang L, et al. Redesigning CAM-SE for peta-scale climate modeling performance and ultra-high resolution on Sunway TaihuLight. In: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, 2017. 1
- Fu H, Yin W, Yang G, et al. 18.9-PFlops nonlinear earthquake simulation on Sunway TaihuLight: enabling depiction of 18-Hz and 8-meter scenarios. In: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, 2017. 2
- Williams S, Patterson D A, Olicker L, et al. The roofline model: a pedagogical tool for auto-tuning kernels on multicore architectures. In: Proceedings of Symposium on High Performance Chips, Stanford, 2008
- Oral S, Vazhkudai S S, Wang F, et al. End-to-end I/O portfolio for the summit supercomputing ecosystem. In: Proceedings of the International Conference for High Performance Computing, Networking, Storage, and Analysis, 2019. 1–14
- Shi X, Li M, Liu W, et al. SSDUP: a traffic-aware ssd burst buffer for HPC systems. In: Proceedings of the International Conference on Supercomputing, 2017. 1–10
- Shi X, Liu W, He L, et al. Optimizing the SSD burst buffer by traffic detection. ACM Trans Archit Code Opt, 2020, 17: 1–26

- 28 He W Q, L Y, Fang Y F, et al. Design and implementation of Parallel C programming language for domestic heterogeneous many-core systems. *J Softw*, 2017, 28: 764–785
- 29 Schroeder B, Gibson G A. A large-scale study of failures in high-performance computing systems. *IEEE Trans Dependable Secure Comput*, 2010, 7: 337–350
- 30 Cappello F. Resilience: One of the Main Challenges for Exascale Computing. Technical Report of the INRIA-Illinois Joint Laboratory, 2011
- 31 Kusnezov D. DOE exascale Initiative. 2013. <https://www.energy.gov/downloads/doe-exascale-initiative>
- 32 Asanovic K, Bodik R, Catanzaro B C, et al. The Landscape of Parallel Computing Research: A View from Berkeley. Technical Report Uc Berkeley. eecs-2006-183. 2006
- 33 Chao Y, Wei X, Fu H, 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, 2016. 6
- 34 Qiao F, Zhao W, Yin X, et al. A highly effective global surface wave numerical simulation with ultra-high resolution. In: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, 2016. 5
- 35 Fu H, Liao J, Xue W, et al. Refactoring and optimizing the community atmosphere model (CAM) on the Sunway TaihuLight supercomputer. In: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, 2016. 83
- 36 Liu J, Qin H, Wang Y, et al. Largest particle simulations downgrade the runaway electron risk for ITER. 2016. ArXiv: 1611.02362
- 37 Dong W, Kang L, Quan Z, et al. Implementing molecular dynamics simulation on Sunway TaihuLight system. In: Proceedings of IEEE 18th International Conference on High Performance Computing and Communications; IEEE 14th International Conference on Smart City; IEEE 2nd International Conference on Data Science and Systems (HPCC/SmartCity/DSS), 2016. 443–450
- 38 Duan X, Xu K, Chan Y, et al. S-Aligner: ultrascaleable read mapping on Sunway TaihuLight. In: Proceedings of IEEE International Conference on Cluster Computing (CLUSTER), 2017
- 39 Yao W J, Chen J S, Su Z-C, et al. Porting and optimizing of NAMD on SunwayTaihuLight system. *Comput Eng Sci*, 2017, 39: 1022–1030