Report on the Sunway TaihuLight System

Jack Dongarra

University of Tennessee

Oak Ridge National Laboratory

June 24, 2016
University of Tennessee
Department of Electrical Engineering and Computer Science
Tech Report UT-EECS-16-742

Overview

The Sunway TaihuLight System was developed by the National Research Center of Parallel Computer Engineering & Technology (NRCPC), and installed at the National Supercomputing Center in Wuxi (a joint team with the Tsinghua University, City of Wuxi, and Jiangsu province), which is in China's Jiangsu province. The CPU vendor is the Shanghai High Performance IC Design Center. The system is in full operation with a number of applications implemented and running on the system. The Center will be a public supercomputing center that provides services for public users in China and across the world.

The complete system has a theoretical peak performance of 125.4 Pflop/s with 10,649,600 cores and 1.31 PB of primary memory. It is based on a processor, the SW26010 processor, that was designed by the Shanghai High Performance IC Design Center. The processor chip is composed of 4 core groups (CGs), see figure 1, connected via a NoC, see figure 2, each of which includes a Management Processing Element (MPE) and 64 Computing Processing Elements (CPEs) arranged in an 8 by 8 grid. 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).

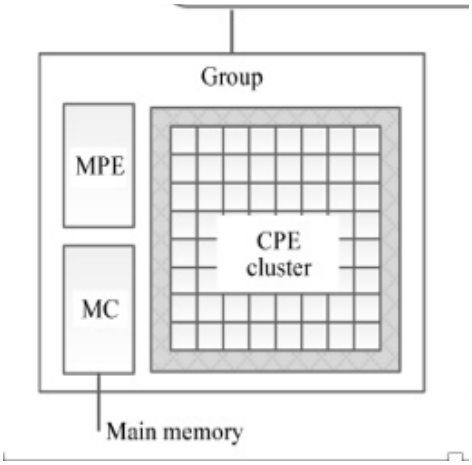


Figure 1: Core Group for Node

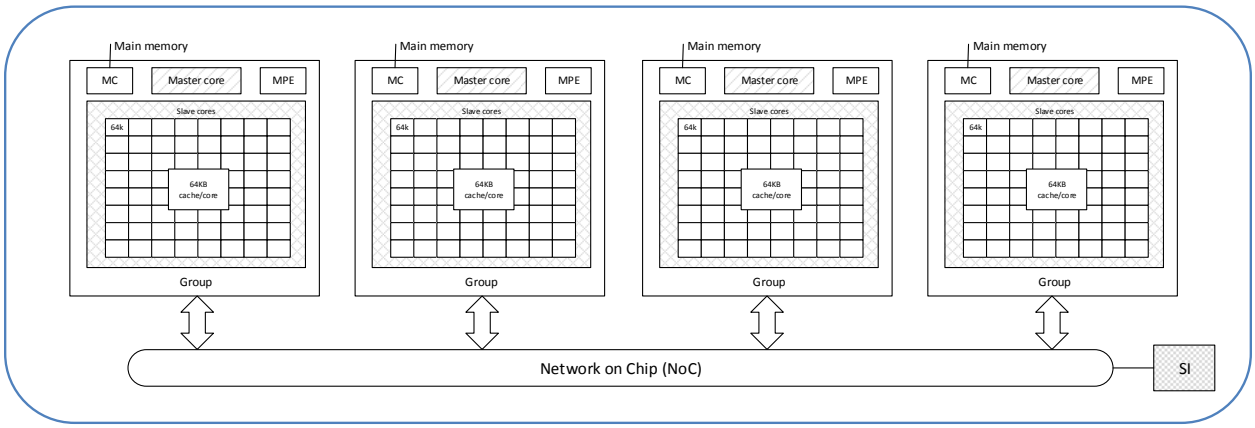


Figure 2: Basic Layout of a Node, the SW26010

Each CPE Cluster is composed of a Management Processing Element (MPE) which is a 64-bit RISC core which is supporting both user and system modes, a 256-bit vector instructions, 32 KB L1 instruction cache and 32 KB L1 data cache, and a 256KB L2 cache. The Computer Processing Element (CPE) is composed of an 8x8 mesh of 64-bit RISC cores, supporting only user mode, with a 256-bit vector instructions, 16 KB L1 instruction cache and 64 KB Scratch Pad Memory (SPM).

Figure 3: Image of SW26010 many-core (260 core) processor that makes up the node.

The SW26010 chip is a Chinese “homegrown” many-core processor. The Vendor is the Shanghai High Performance IC Design Center which was supported by National Science and Technology Major Project (NMP): Core Electronic Devices, High-end Generic Chips, and Basic Software.

Computer Node

A computer node of this machine is based on one many-core processor chip called the SW26010 processor. Each processor is composed of 4 MPEs, 4 CPEs, (a total of 260 cores), 4 Memory

Controllers (MC), and a Network on Chip (NoC) connected to the System Interface (SI). Each of the four MPE, CPE, and MC have access to 8 GB of DDR3 memory. The total system has 40,960 nodes for a total of 10,649,600 cores and 1.31 PB of memory.

The MPE's and CPE's are based on a RISC architecture, 64-bit, SIMD, out of order microstructure. Both the MPE and the CPE participate in the user's application. The MPE performance management, communication, and computation while the CPEs mainly perform computations. (The MPE can also participate in the computations.)

Each core of the CPE has a single floating point pipeline that can perform 8 flops per cycle per core (64-bit floating point arithmetic) and the MPE has a dual pipeline each of which can perform 8 flops per cycle per pipeline (64-bit floating point arithmetic). The cycle time for the cores is 1.45 GHz, so a CPE core has a peak performance of 8 flops/cycle * 1.45 GHz or 11.6 Gflop/s and a core of the MPE has a peak performance of 16 flops/cycle * 1.45 GHz or 23.2 Gflop/s. There is just one thread of execution per physical core.

A node of the TaihuLight System has a peak performance of (256 cores * 8 flops/cycle * 1.45 GHz) + (4 core * 16 flops/cycle * 1.45 GHz) = 3.0624 Tflop/s per node. The complete system has 40,960 nodes or 125.4 Pflop/s for the theoretical peak performance of the system.

Each CPE has a 64 KB local (scratchpad) memory, no cache memory. The local memory is SRAM. There is a 16KB instruction cache. Each of the 4 CPE/MPE clusters has 8 GB of DDR3 memory. So a node has 32 GB of primary memory. Each processor connects to four 128-bit DDR3-2133 memory controllers, with a memory bandwidth of 136.51 GB/s. Non-volatile memory is not used in the system.

The MPE/CPE chip is connected via a network-on-chip (NoC) and the system interface (SI) is used to connect the system outside of the node. The SI is a standard PCIe interface. The bidirectional bandwidth is 16 GB/s with a latency around 1 us.

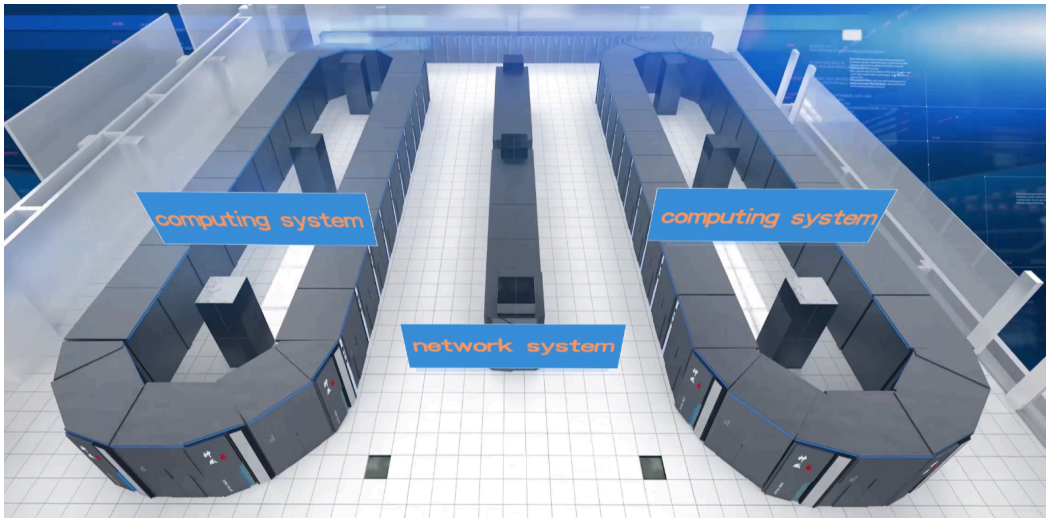


Figure 4: Overview of the Sunway TaihuLight System

Figure 5: Picture of the Sunway TaihuLight System computer room

Sunway TaihuLight System is based exclusively on processors designed and built in China.

For comparison, the next large acquisition of supercomputers for the US Department of Energy will not be until 2017 with production beginning in 2018. The US Department of Energy schedule is for a planned 200 Pflop/s machine called Summit at Oak Ridge National Lab by early 2018, a planned 150 Pflop/s machine called Sierra at Lawrence Livermore National Lab by mid-2018, and a planned 180 Pflop/s machine called Aurora at Argonne National Lab in late 2018.

Power Efficiency

The peak power consumption under load (running the HPL benchmark) is at 15.371 MW or 6 Gflops/W. This is just for the processor, memory, and interconnect network. The cooling system used is a closed-coupled chilled water cooling with a customized liquid water-cooling unit.

The Power System

The power system is made up of a mutual-backup power input of 2x35 KV which go to a Front-end power supply with output of DC 300V. The Cabinet power supply is DC 12V and the CPU power supply is DC 0.9V.

The Interconnect

Sunway has built their own interconnect. There is a five-level integrated hierarchy, connecting the computing node, computing board, super-nodes, cabinet, to the complete system. Each card

has two nodes, see figure 6.

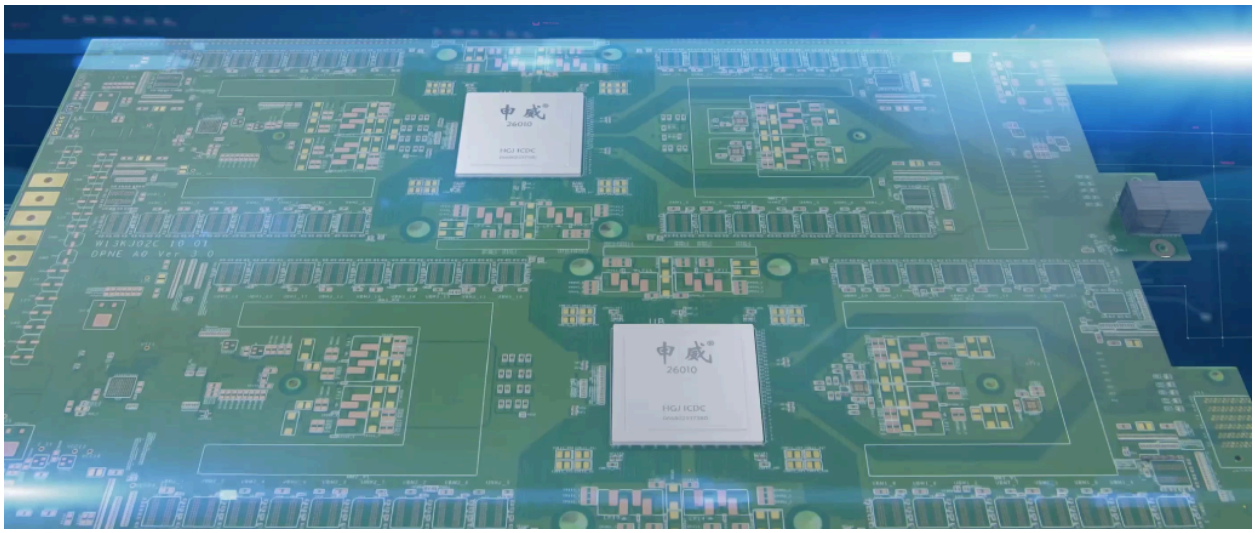


Figure 6: Two nodes on a card.

Each board has four cards, one facing up and one facing down, so each board has 8 nodes, see figure 7. There are 32 boards in a supenode or 256 nodes in a supernode, see figure 8.

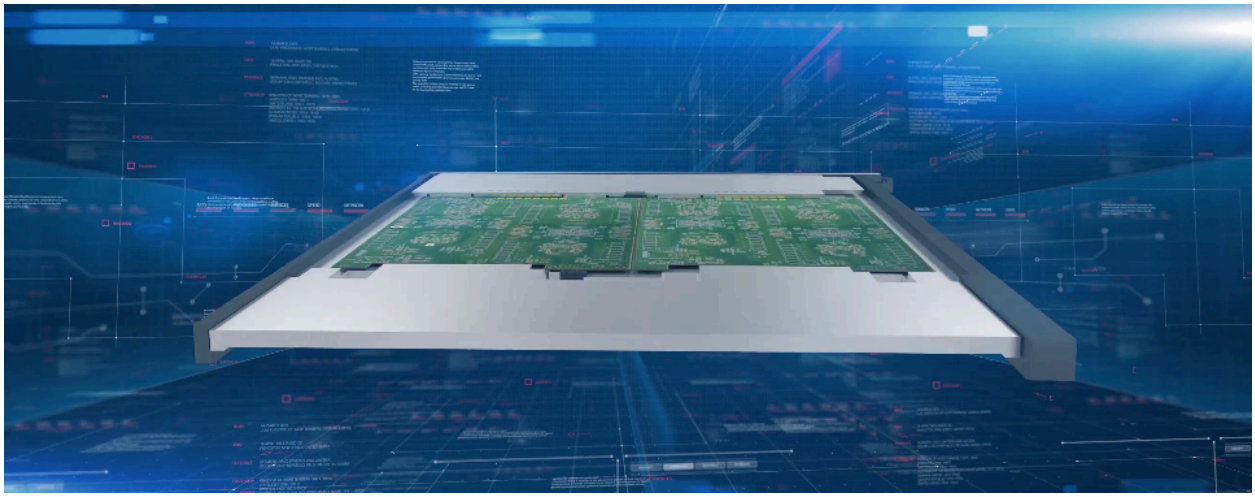


Figure 7: Four cards on a board, two up and two down (on the other side).

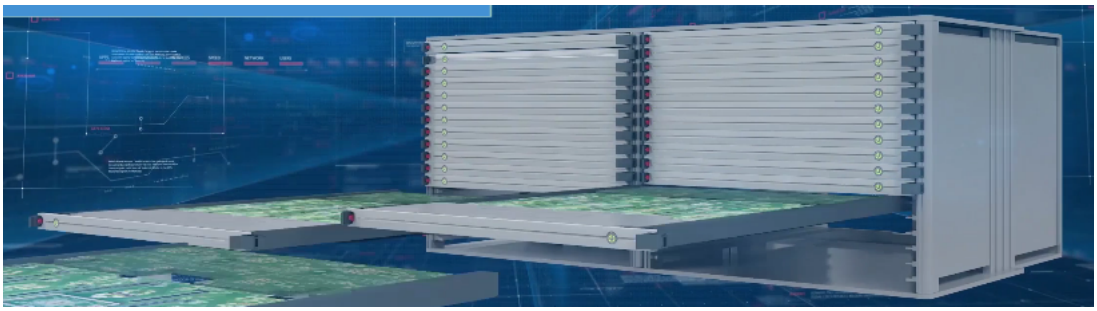


Figure 8: A Supernode composed of 32 boards or 256 nodes.

In a cabinet there are 4 Supernodes for a total of 1024 nodes, see figure 9.

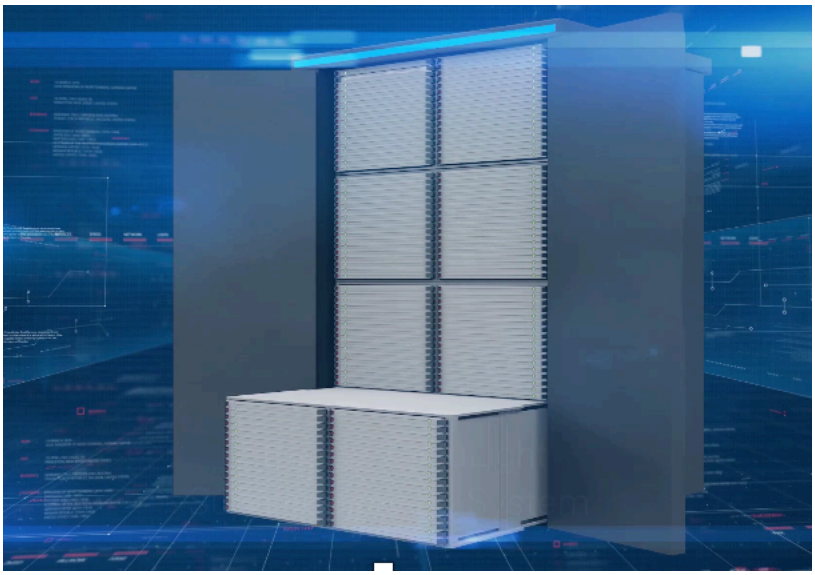


Figure 9: A cabinet composed of 4 supernodes or 1024 nodes.

Nodes are connected using PCI-E 3.0 connections in what's called a Sunway Network. Sunway custom network consists of three different levels, with the central switching network at the top connecting different supernodes, the super node network in the middle which fully connects all 256 nodes within each supernode providing high bandwidth and low latency for communication within the supernode, and the resource-sharing network at the bottom connecting the computing system to other resources, such as I/O services. The bisection network bandwidth is 70TB/s, with a network diameter of 7. Mellanox supplied the Host Channel Adapter (HCA) and switch chips for the Sunway TaihuLight.

Communication between nodes via MPI is at 12 GB/second and a latency of about 1 us.

Complete System

The complete system is composed of 40 Cabinets, see figure 10. Each Cabinet contains 4 Supernodes and each Supernode has 256 Nodes, see figure 11. Each node has a peak floating point performance of 3.06Tflop/s.

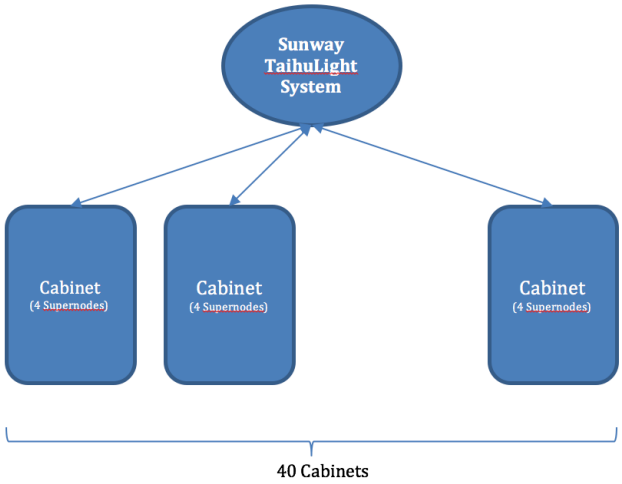


Figure 10: Sunway system with 40 cabinets

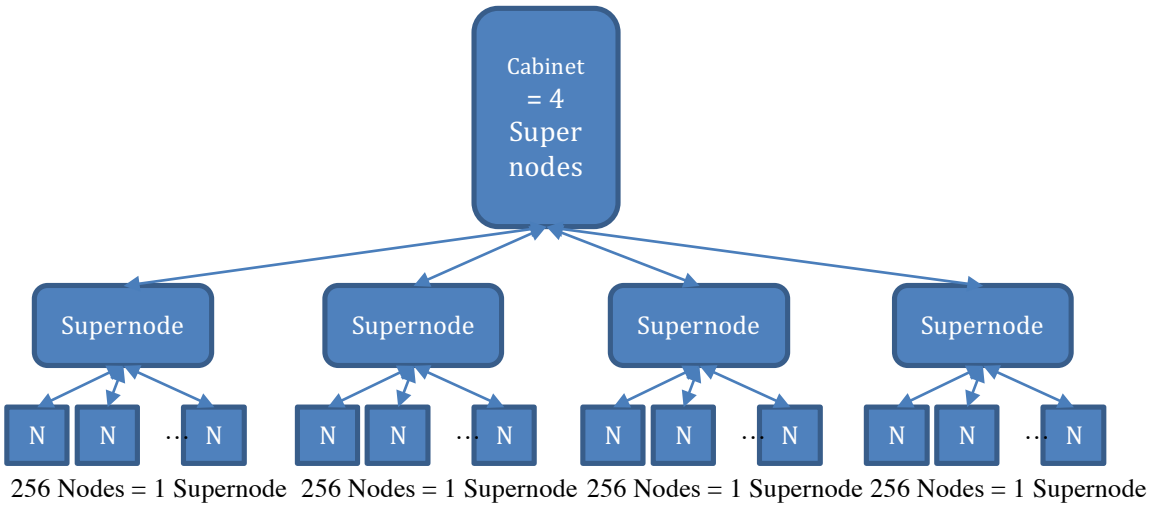


Figure 11: One cabinet of the system.

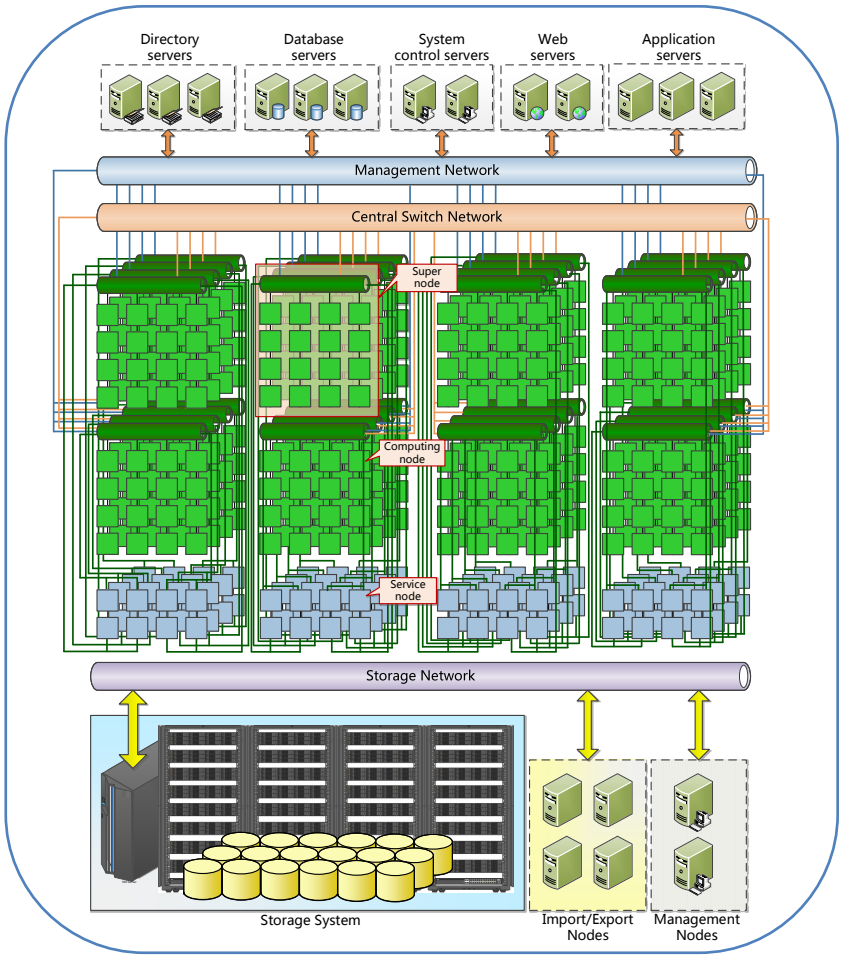


Figure 12: General Architecture of the Sunway TaihuLight

Each Supernode then is 256*3.06 Tflop/s and a Cabinet of 4 Supernodes is at 3.1359 Pflop/s.

All number are for 64-bit Floating Point Arithmetic.

1 Node = 260 cores
1 Node = 3.06 Tflop/s

1 Supernode = 256 Nodes
1 Supernode = 783.97 Tflops

1 Cabinet = 4 Supernodes
1 Cabinet = 3.1359 Pflops

1 Sunway TaihuLight System = 40 Cabinets = 160 Supernodes = 40,960 nodes = 10,649,600 cores.
1 Sunway TaihuLight System = 125.4359 Pflop/s

Assuming 15.311 MW for HPL using 40 cabinets, each cabinet is at 382.8 KW. Each cabinet has 4*256 nodes or 373.8 W/node.

The Flops/W for the theoretical peak is at 8 Gflops/W and for HPL the efficiency is 6.074 Gflops/W (93 Pflops/15.311MW).

The Software Stack

The Sunway TaihuLight System is using Sunway Raise OS 2.0.5 based on Linux as the operating system.

The basic software stack for the many-core processor includes basic compiler components, such as C/C++, and Fortran compilers, an automatic vectorization tool, and basic math libraries. There is also the Sunway OpenACC, a customized parallel compilation tool that supports OpenACC 2.0 syntax and targets the SW26010 many-core processor.

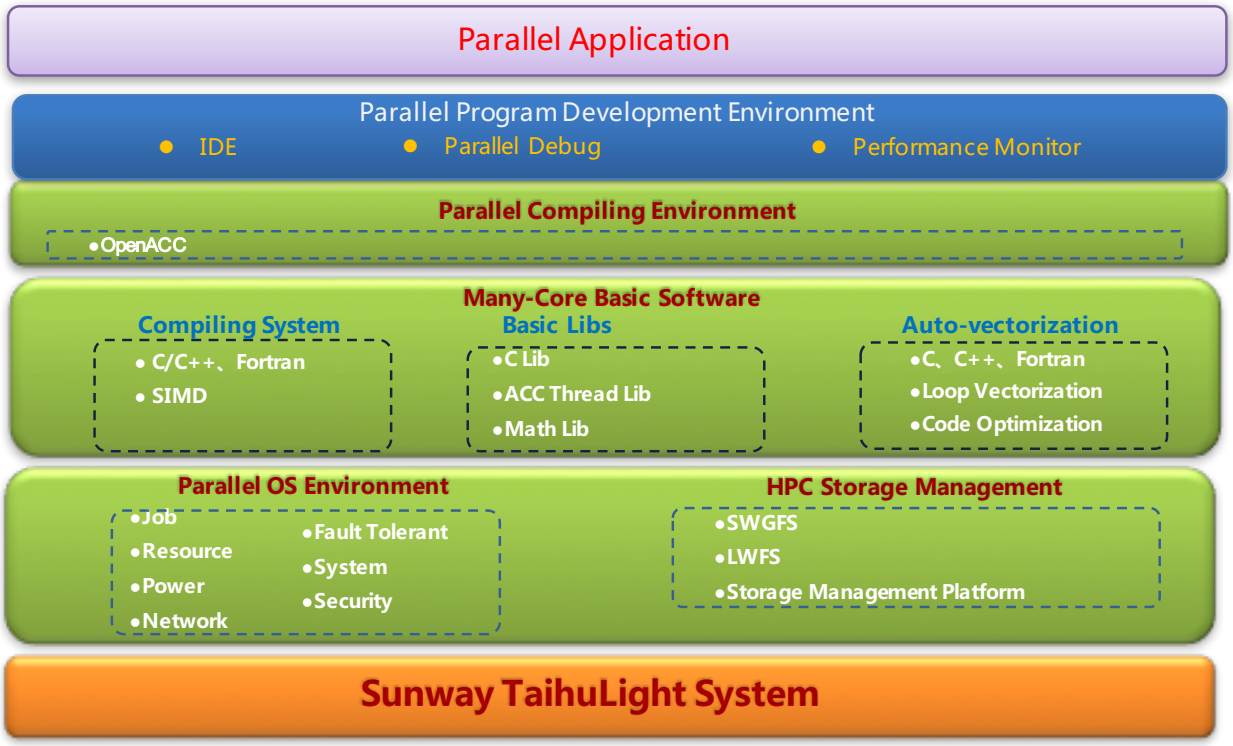


Figure: Sunway TaihuLight Software Stack

Cooling

To satisfy the need of 28 MW cooling system Climaveneta delivered 15 TECS2-W/H water-cooled chillers equipped with magnetic levitation, oil free VFD compressors, with the best Seasonal Energy Efficiency Ratio (ESEER), close to 10. The Climaveneta cooling system,

combined with further eco sustainable technologies adopted, such as free cooling and VPF, has contributed to cut the entire energy consumption of the data center by 45%.

The heat exchange is at the level of the computing boards. The system is able to recycle the cooling water.

Applications

There are currently four key application domains for the Sunway TaihuLight system:

- Advanced manufacturing: CFD, CAE applications.
- Earth system modeling and weather forecasting.
- Life science.
- Big data analytics.

There are three submissions which are finalists for the Gordon Bell Award at SC16 that are based on the new Sunway TaihuLight system. These three applications are: (1) a fully-implicit nonhydrostatic dynamic solver for cloud-resolving atmospheric simulation; (2) a highly effective global surface wave numerical simulation with ultra-high resolution; (3) large scale phase-field simulation for coarsening dynamics based on Cahn-Hilliard equation with degenerated mobility. Here's a link to a paper describing the applications:
http://engine.scichina.com/downloadPdf/rdbMLbAMd6ZiKwDco

All these three applications have scaled to around 8 million cores (close to the full system scale). The applications that come with an explicit method (such as wave simulation and phase-field simulation) have achieved a sustained performance of 30 to 40 PFlops. In contrast, the implicit solver achieves a sustained performance of around 1.5 PFlops, with a good convergence rate for large-scale problems. These performance number may be improved before the SC16 Conference in November 2016.

The Gordon Bell Prize is awarded each year to recognize outstanding achievement in high-performance computing. The purpose of the award is to track the progress over time of parallel computing, with particular emphasis on rewarding innovation in applying high-performance computing to applications in science, engineering, and large-scale data analytics. Prizes may be awarded for peak performance or special achievements in scalability and time-to-solution on important science and engineering problems. Financial support of the $10,000 award is provided by Gordon Bell, a pioneer in high-performance and parallel computing.

LINPACK Benchmark Run (HPL)

The results for the Linpack Benchmark showing a run of the HPL benchmark program using 165,120 nodes, that run was made using 1.2 PB total or 7.2 TB of the memory of each node and achieved 93 Pflop/s out of a theoretical peak of 125 Pflop/s or an efficiency of 74.15% of

theoretical peak performance taking a little over 3.7 hours to complete, with an average power consumption of 15.37 MW, See figure 9.

Summary of HPL Benchmark run:

- HPL number = 93 Pflop/s
  - 74.15% efficient (peak at 125 Pflop/s)
  - Size of the matrix, n = 12,288,000 (1.2 PB)
  - Logical process grid of pxq = 256 x 640
  - 163,840 MPI processes, which corresponds to 4 x 40,960 CGs in the system.
  - Each CG has one MPE, and 64 CPEs. So within the MPI process, 64 threads to use the 64 CPEs.
  - The number of cores is 163,840 x 64 = 10,649,600 cores for the HPL run.
  - Time to complete benchmark run: 13,298 seconds (3.7 hours)
  - Average 15.371 MW
  - 6 Gflops/W

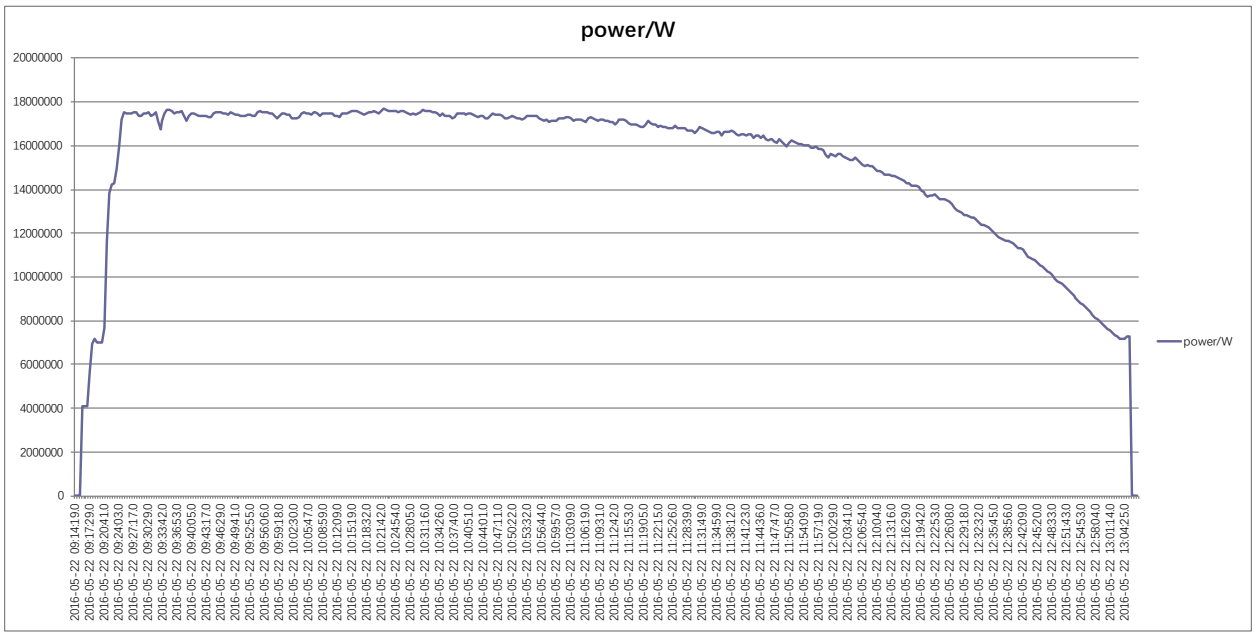


Figure 9: Power consumption for the HPL benchmark run.

Funding for the System

The system was funded from three sources, the central Chinese government, the province of Jiangsu, and the city of Wuxi. Each contributed approximately 600 million RMBs or a total of 1.8 billion RMBs for the system or approximately $270M USD. That is the cost of the building, hardware, R&D, and software costs. It does not cover the ongoing maintenance and running of the system and center.

Summary

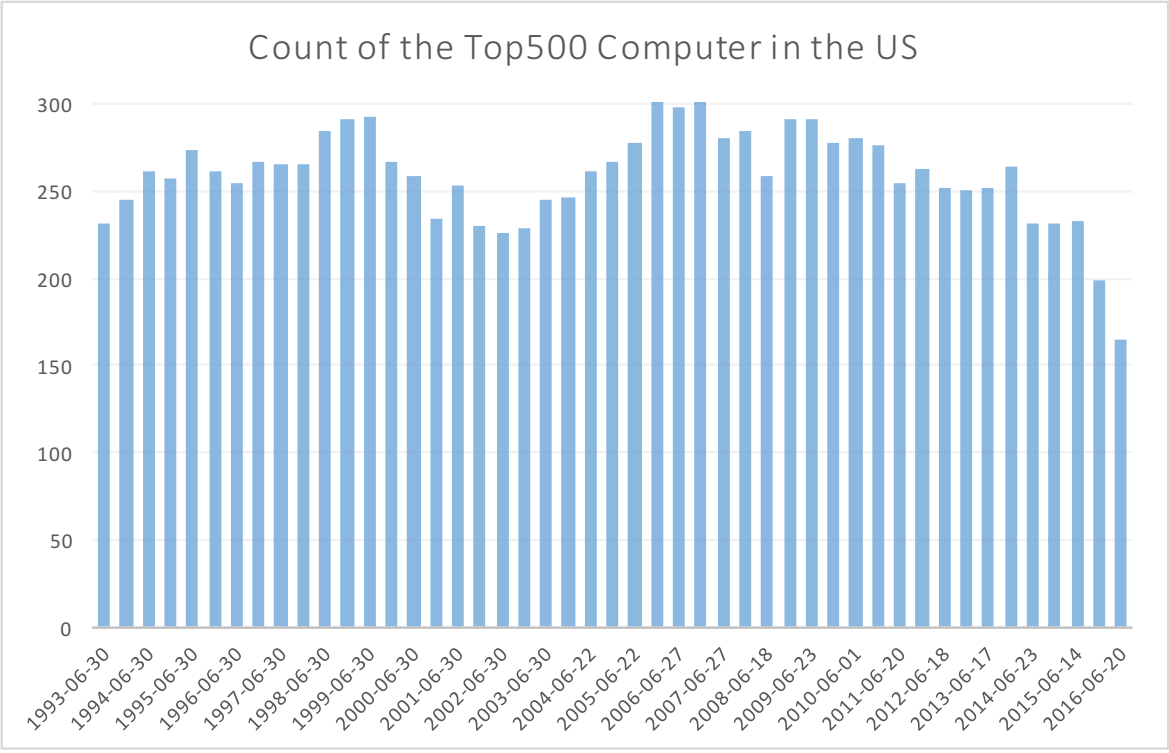
The Sunway TaihuLight System is very impressive with over 10 million cores and a peak performance of 125 Pflop/s. The Sunway TaihuLight is almost three times (2.75 times) as fast and three times as efficient as the system it displaces in the number one spot. The HPL Benchmark results at 93 Pflop/s or 74% of theoretical peak performance is also impressive, with an efficiency of 6 Gflops per Watt. The HPCG performance at only 0.3% of peak performance shows the weakness of the Sunway TaihuLight architecture with slow memory and modest interconnect performance. The ratio of floating point operations per byte of data from memory on the SW26010 is 22.4 Flops(DP)/Byte transfer, which shows an imbalance or an overcapacity of floating point operations per data transfer from memory. By comparison the Intel Knights Landing processor with 7.2 Flops(DP)/Byte transfer. So for many "real" applications the performance on the TaihuLight will be no where near the peak performance rate. Also the primary memory for this system is on low side at 1.3 PB (Tianhe-2 has 1.4 PB and Titan has .71 PB).

The Sunway TaihuLight system, based on a homegrown processor, demonstrates the significant progress that China has made in the domain of designing and manufacturing large-scale computation systems.

The fact that there are sizeable applications and Gordon Bell contender applications running on the system is impressive and shows that the system is capable of running real applications and not just a "stunt machine".

China has made a big push into high performance computing. In 2001 there were no supercomputers listed on the Top500 in China. Today China has 167 systems on the June 2016 Top500 list compared to 165 systems in the US. This is the first time the US has lost the lead. No other nation has seen such rapid growth. See Graphs 1 and 2. According to the Chinese national plan for the next generation of high performance computers, China will develop an exascale computer during the 13th Five-Year-Plan period (2016-2020). It is clear that they are on a path which will take them to an exascale computer by 2020, well ahead of the US plans for reaching exascale by 2023.

Graph 1: Number of Top500 Computers in China over time



Graph 2: Number of Top500 Computers in the US over time

Table 1: Sunway TaihuLight System Summary

CPU	Shenwei-64
Developer	NRCPC
Chip Fab	CPU vendor is the Shanghai High Performance IC Design Center
Instruction set	Shenwei-64 Instruction Set (this is NOT related to the DEC Alpha instruction set)
Node Processor cores	256 CPEs (computing processing elements) plus 4 MPEs (management processing elements)
Node Peak Performance	3.06 TFlop/s
Clock Frequency	1.45 GHz
Process Technology	N/A
Power	15.371 MW (average for the HPL run)
Peak Performance of system	125.4 Pflop/s system in Wuxi
Targeted application	HPC
Nodes	40,960
Total memory	1.31 PB
Cabinets	40
Nodes per cabinet	1024 Nodes
Cores per node	260 cores
Total system core count	10,649,600

Table 2: Comparison with Top 3 Machines

	ORNL Titan	NUDT Tianhe-2	Sunway TaihuLight
Theoretical Peak	27 Pflop/s = (2.6 CPU + 24.5 GPU) Pflop/s	54.9 Pflop/s = (6.75 CPU + 48.14 Coprocessor) Pflop/s	125.4 Pflop/s = CPEs +MPEs Cores per Node = 256 CPEs + 4 MPEs Supernode = 256 Nodes System = 160 Supernodes Cores = 260 * 256 * 160 = 10.6M
HPL Benchmark Flop/s	17.6 Pflop/s	30.65 Pflop/s	93 Pflop/s
HPL % Peak	65.19%	55.83%	74.16%
HPCG Benchmark	0.322 Pflop/s	0.580 Pflop/s	.371 Pflop/s
HPCG % Peak	1.2%	1.1%	0.30%
Compute Nodes	18,688	16,000	40,960
Node	AMD Optron Interlagos (16 cores, 2.2 GHz) plus Nvidia Tesla K20x (14 cores, .732 GHz)	2 – Intel Ivy Bridge (12 cores, 2.2 GHz) plus 3 - Intel Xeon Phi (57 cores, 1.1 GHz)	256 CPEs + 4 MPEs
Sockets	18,688 Interlagos + 18,688 Nvidia boards	32,000 Ivy Bridge + 48,000 Xeon Phi boards	40,960 nodes with 256 CPEs and 4 MPEs per node
Node peak performance	1.4508 Tflop/s = (.1408 CPU + 1.31 GPU) Tflop/s	3.431 Tflop/s = (2*.2112 CPU + 3*1.003 Coprocessor) Tflop/s	3.06 Tflop/s CPE: 8 flops/core/cycle (1.45 GHz*8*256 = 2.969 Tflop/s) MPE (2 pipelines) 2*4*8 flops/core/cycle (1.45 GHz*1= 0.0928Tflop/s)
Node Memory	32 GB CPU + 6 GB GPU	64 GB CPU + 3*8 GB Coprocessor	32 GB per node
System Memory	.710 PB = (.598 PB CPU and .112 PB GPU)	1.4 PB = (1.024 PB CPU and .384 PB Coprocessor)	1.31 PB (32 GB*40,960 nodes)
Configuration	4 nodes per blade, 24 blades	2 nodes per blade, 16 blades per	Node peak performance is 3.06 Tflop/s, or 11.7 Gflop/s per core.

	per cabinet, and 200 cabinets in the system	frame, 4 frames per cabinet, and 162 cabinets in the system	260 cores / node CPE: 8 flops/core/cycle (1.45 GHz*8*256 = 2.969 Tflop/s) MPE (2 pipelines) 2*4*8 flops/core/cycle (1.45 GHz*1= 0.0928Tflop/s) Node peak performance: 3.06 Tflop/s 1 thread / core Nodes connected using PCI-E The topology is Sunway network 256 nodes = a supernode (256*3.06 Tflop/s = . 783 Pflop/s) 160 supernodes make up the whole system (125.4PFlop/s) The network system consists of three different levels, with the central switching network at the top, the super node network in the middle, and the resource-sharing network at the bottom. 4 supernodes = cabinet Each cabinet ~ 3.164 Pflop/s 256 nodes per supernode 1,024 nodes (3 Tflop/s each) per cabinet 40 cabinets ~ 125 Pflop/s
Total System	560,640 cores = (299,008 AMD cores + 261,632 Nvidia cores)	3,120,000 Cores = (384,000 Ivy Bridge cores + 2,736,000 Xeon Phi cores)	10,649,600 cores = Node (260) * supernodes(256 nodes) * 160 supernodes
Power (processors, memory, interconnect)	9 MWatts	17.8 MWatts	15.3 MWatts
Size	404 m^2	720 m^2	605 m^2

Table 3: Comparison of Top 6 Systems

Rank	Site	Manufacture	Name	System	Gflops/ Watt
1	National Supercomputing Center in Wuxi	Shanghai High Performance IC Design Center	Sunway TaihuLight system	ShenWei -64 10,649,600 cores = Node (260) * supernodes(256 nodes) * 160 supernodes Connected with Infiniband interconnect	6
2	National Supercomputer Center in Guangzhou	NUDT	Tianhe-2	32,000 Intel Xeon CPU's + 48,000 Xeon Phi's (+ 4096 FT-1500 CPU's frontend) Connected with Infiniband interconnet	1.95
3	DOE/SC/Oak Ridge National Laboratory	Cray Inc.	Titan	Cray XK7, Opteron 6274 16C 2.200GHz & NVIDIA K20x, Cray Gemini interconnect	2.143
4	DOE/NNSA/LLNL	IBM	Sequoia	BlueGene/Q, Power BQC 16C 1.60 GHz, Custom	2.069
5	RIKEN Advanced Institute for Computational Science (AICS)	Fujitsu	K	K computer, SPARC64 VIIIfx 2.0GHz, Tofu interconnect	0.830
6	DOE/SC/Argonne National Laboratory	IBM	Mira	BlueGene/Q, Power BQC 16C 1.60GHz, Custom Interconnect	2.069

Table 4: Comparison to Intel's KNC and KNL

FEATURE	INTEL® XEON PHI™ COPROCESSOR 7120P	Intel® Xeon Phi™ Processor (codename Knights Landing)	Sunway TaihuLight Node
Processor Cores	Up to 61 enhanced P54C Cores	Up to 72 enhanced Silvermont cores	260 cores / node
Key Core Features	In order, 4 threads / core, 2 wide	Out of order, 4 threads / core, 2 wide	1 thread / core
High Speed Memory	Up to 16 32-bit channels GDDR5 @ up to 5.5GT/s	Eight 128-bit channels MCDRAM @ 7.2 GT/s	Up to 4 128-bit channels
Off Package Memory	None	6 channels DDR4 2400MHz	4*128 channels DDR3 at 2133 MHz
Memory Bandwidth	Up to 181 GB/s STREAM Triad (GDDR5)	~ 490 GB/s STREAM Triad (to MCDRAM) + ~ 90GB/s STREAM Triad (to DDR4)	136 GB/s *128-bit DDR3-2133*
Memory Capacity	Up to 16 GB on-package GDDR5	16 GB on package memory (MCDRAM) + Up to 384 GB off package DDR4	32 GB off package DDR3
Peak FLOPS	SP: 2.416 TFLOPs; DP: 1.208 TFLOPs	Up to SP 6.912 TFs (at 1.5GHz TDP freq) Up to DP 3.456 TFs (at 1.5GHz TDP freq)	DP: 3.06 Tflop/s
FLOPS/Byte (from memory)	1.208 Tflop/s / 181 GB/s = 6.67 Flops/Byte	3.456 TFLOP/s at 490 GB/s = 7.05 Flops/Byte	3.06 Tflop/s / 136.51 GB/s = 22.4 Flops/Byte
Scalar Performance	1X	Up to 3x higher	
Power Efficiency	Up to 3.5 GF/watt, DGEMM	Up to 9.6 GF/watt, DGEMM (CPU only)	6.074 Gflops/W
Fabric I/O	No integrated fabric and accessed through host	Up to 50 GB/s with integrated fabric (bidirectional BW, both ports)	??
Configurations	Coprocessor only	Stand-alone host processor, stand-alone host processor with integrated fabric (and coprocessor)	Coprocessor, stand-alone processor, stand-alone processor with integrated fabric
On-die Interconnect	Bidirectional Ring Interconnect	Mesh of Rings Interconnect	Mesh and NoC interconnect
Vector ISA	x87, (no Intel® SSE or MMX™), Intel IMIC 16 floating point operations per cycle per core (64 bit floating point)	x87, SSE, SSE2, SSE3, SSSE3, SSE4.1, SSE4.2, Intel® AVX, AVX2, AVX-512 (no Intel® TSX), and AVX-512 extensions 32 double precision floating point operations per cycle per core	8 floating point operations per cycle per core (64 bit floating point) for the CPEs 16 floating point operations per cycle per core (64 bit floating point) for the MPEs

Table 5: #1 System on the Top500 Over the Past 24 Years (17 machines)

Top500 List (# of times)	Computer	r_max (Tflop/s)	n_max	Hours For Benchmark	MW under load
6/93 (1)	TMC CM-5/1024	.060	52224	0.4	
11/93 (1)	Fujitsu Numerical Wind Tunnel	.124	31920	0.1	1.
6/94 (1)	Intel XP/S140	.143	55700	0.2	
11/94 - 11/95 (3)	Fujitsu Numerical Wind Tunnel	.170	42000	0.1	1.
6/96 (1)	Hitachi SR2201/1024	.220	138,240	2.2	
11/96 (1)	Hitachi CP-PACS/2048	.368	103,680	0.6	
6/97 - 6/00 (7)	Intel ASCI Red	2.38	362,880	3.7	.85
11/00 - 11/01 (3)	IBM ASCI White, SP Power3 375 MHz	7.23	518,096	3.6	
6/02 - 6/04 (5)	NEC Earth-Simulator	35.9	1,000,000	5.2	6.4
11/04 - 11/07 (7)	IBM BlueGene/L	478.	1,000,000	0.4	1.4
6/08 - 6/09 (3)	IBM Roadrunner –PowerXCell 8i 3.2 Ghz	1,105.	2,329,599	2.1	2.3
11/09 - 6/10 (2)	Cray Jaguar - XT5-HE 2.6 GHz	1,759.	5,474,272	17.3	6.9
11/10 (1)	NUDT Tianhe-1A, X5670 2.93Ghz NVIDIA	2,566.	3,600,000	3.4	4.0
6/11 - 11/11 (2)	Fujitsu K computer, SPARC64 VIIIfx	10,510.	11,870,208	29.5	9.9
6/12 (1)	IBM Sequoia BlueGene/Q	16,324.	12,681,215	23.1	7.9
11/12 (1)	Cray XK7 Titan AMD + NVIDIA Kepler	17,590.	4,423,680	0.9	8.2
6/13 – 11/15(6)	NUDT Tianhe-2 Intel IvyBridge & Xeon Phi	33,862.	9,960,000	5.4	17.8
6/16 –	Sunway TaihuLight System	93,014.	12,288,000	3.7	15.4

Table 6: The Top 10 Machines on the Top500

Rank	Site	Computer	Country	Cores	Rmax [Pflops]	% of Peak	Power [MW]	GFlops/Watt
1	National Super Computing Center in Wuxi	Sunway TaihuLight System SW26010 (260c) + Custom	China	10,649,600	93.0	74	15.3	6.07
2	National Super Computer Center in Guangzhou	Tianhe-2 NUDT, Xeon 12C + IntelXeon Phi (57c) + Custom	China	3,120,000	33.9	62	17.8	1.905
3	DOE / OS Oak Ridge Nat Lab	Titan, Cray XK7, AMD (16C) + Nvidia Kepler GPU (14c) + Custom	USA	560,640	17.6	65	8.3	2.120
4	DOE / NNSA L Livermore Nat Lab	Sequoia, BlueGene/Q (16c) + custom	USA	1,572,864	17.2	85	7.9	2.063
5	RIKEN Advanced Inst for Comp Sci	K computer Fujitsu SPARC64 VIIIfx (8c) + Custom	Japan	705,024	10.5	93	12.7	.830
6	DOE / OS Argonne Nat Lab	Mira, BlueGene/Q (16c) + Custom	USA	786,432	8.59	85	3.95	2.176
7	DOE / NNSA / Los Alamos & Sandia	Trinity, Cray XC40,Xeon 16C + Custom	USA	301,056	8.10	80		
8	Swiss CSCS	Piz Daint, Cray XC30, Xeon 8C + Nvidia Kepler (14c) + Custom	Swiss	115,984	6.27	81	2.3	2.726
9	HLRS Stuttgart	Hazel Hen, Cray XC40, Xeon 12C+ Custom	Germany	185,088	5.64	76		
10	KAUST	Shaheen II, Cray XC40, Xeon 16C + Custom	Saudi Arabia	196,608	5.54	77	2.8	1.954

Table 7: HPCG Performance for Top 10 Systems

Rank	Site	Computer	Cores	HPL Rmax Pflops	HPCG Pflops	HPCG / HPL	% of Peak
1	NSCC / Guangzhou	Tianhe-2 NUDT, Xeon 12C 2.2GHz + Intel Xeon Phi 57C + Custom	3,120,000	33.86	0.580	1.7%	1.1%
2	RIKEN Advanced Institute for Computational Science	K computer, SPARC64 VIIIfx 2.0GHz, Tofu interconnect	705,024	10.51	0.550	5.2%	4.9%
3	NSCC / Wuxi	Sunway TaihuLight System 1.45 GHz + Custom	10,649,600	93.02	0.371	0.4%	0.3%
4	DOE/SC/Oak Ridge Nat Lab	Titan - Cray XK7 , Opteron 6274 16C 2.200GHz, Cray Gemini interconnect, NVIDIA K20x	560,640	17.59	0.322	1.8%	1.2%
5	DOE/NNSA/LANL/SNL	Trinity - Cray XC40, Intel E5-2698v3, Aries custom	301,056	8.10	0.182	2.3%	1.6%
6	DOE/SC/Argonne National Laboratory	Mira - BlueGene/Q, Power BQC 16C 1.60GHz, Custom	786,432	8.58	0.167	1.9%	1.7%
7	NASA / Mountain View	Pleiades - SGI ICE X, Intel E5-2680, E5-2680V2, E5-2680V3, Infiniband FDR	185,344	4.08	0.156	3.8%	2.7%
8	HLRS/University of Stuttgart	Hazel Hen - Cray XC40, Intel E5-2680v3, Infiniband FDR	185,088	5.64	0.138	2.4%	1.9%
9	Swiss National Supercomputing Centre / CSCS	Piz Daint - Cray XC30, Xeon E5-2670 8C 2.600GHz, Aries interconnect , NVIDIA K20x	115,984	6.27	0.124	2.0%	1.6%
10	KAUST / Jeda	Shaheen II - Cray XC40, Intel Haswell 2.3 GHz 16C, Cray Aries	196,608	5.53	0.113	2.1%	1.6%

Wuxi Supercomputer Center

Operated together by Jiangsu Province, Wuxi City and Tsinghua University, the National Supercomputing Center in Wuxi (NSCC-Wuxi) hosts the new generation of Sunway Taihu-Light Supercomputer: a supercomputer that is able to provide a peak performance of over 100 PetaFlops. The NSCC-Wuxi now mainly focuses on providing sufficient computing power for the Technological innovation and industrial upgrading of many areas.

Guangwen Yang is the director of the National Supercomputer Center at Wuxi, and a professor in the Department of Computer Science and Technology. He is also the director of the Institute of High Performance Computing at Tsinghua University, and the director of the Ministry of Education Key Lab on Earth System Modeling. His research interests include parallel algorithms, cloud computing, and the earth system model.

Wuxi is one of six National Supercomputing Centers in China:

- Guangzhou (formerly known as Canton), the site of Tianhe-2;
- Wuxi, not far from Shanghai, which is the location of the Sunway TaihuLight System based on the ShenWei processor;
- Changsha, the capital of Hunan Province in south-central China, which hosts a Tianhe-1A machine;
- Tianjin, one of China's largest cities, sited near the coast to the south-east of Beijing, which also hosts a Tianhe-1A machine;
- Jinan, the capital of Shandong province in Eastern China, south of Tianjin and south-east of Beijing, where the current ShenWei Bluelight is located; and
- Shenzhen, in Guangdong Province just north of Hong Kong, where Nebulae, the Dawning TC3600 Blade System (also known as the Dawning-6000) operates.

Other ShenWei Processors

ShenWei SW-1

- First generation, 2006
- Single-core
- 900 MHz

ShenWei SW-2

- Second generation, 2008
- Dual-core
- 1400 MHz
- SMIC 130 nm process
- 70–100 W

ShenWei SW-3

- Third generation, 2010
- 16-core, 64-bit RISC
- 975–1200 MHz
- 65 nm process
- 140.8 GFLOPS @ 1.1 GHz
- Max memory capacity: 16 GB
- Peak memory bandwidth: 68 GB/s
- Quad-channel 128-bit DDR3

Acknowledgement:

The images in this report come from the following paper: “The Sunway TaihuLight Supercomputer: System and Applications”, by Fu H H, Liao J F, Yang J Z, et al. that will appear in Sci. China Inf. Sci., 2016, 59(7): 072001, doi: 10.1007/s11432-016-5588-7.
http://engine.scichina.com/downloadPdf/rdbMLbAMd6ZiKwDco