REPORT ON THE TIANHE-2A SYSTEM

Tech Report No. ICL-UT-17-04

Jack Dongarra

University of Tennessee, Knoxville
Oak Ridge National Laboratory

September 24, 2017

Prepared by
UNIVERSITY OF TENNESSEE,
DEPARTMENT OF ELECTRICAL ENGINEERING AND COMPUTER SCIENCE,
INNOVATIVE COMPUTING LABORATORY

Overview

The TianHe-2A (TH-2A) compute system, designed by China's National University of Defense Technology (NUDT), is an upgrade of the TianHe-2 (TH-2) system.[1] TianHe is sometimes referred to as "Milkyway," and the latest iteration of this system is currently undergoing assembly and testing at China's National Supercomputer Center in Guangzhou (NSCC-GZ). At the time of this report, the system is 25% complete and should be fully functional by November 2017. The most significant enhancement to the system is the upgrade to the TianHe-2 nodes; the old Intel Xeon Phi Knights Corner (KNC) accelerators will be replaced with a proprietary accelerator called the Matrix-2000. In addition, the network has been enhanced, the memory increased, and the number of cabinets expanded. The completed system, when fully integrated with 4,981,760 cores and 3.4 PB of primary memory, will have a theoretical peak performance of 94.97 Pflop/s, which is roughly double the performance of the existing TianHe-2 system. NUDT also developed the heterogeneous programming environment for the Matrix-2000[2] with support for OpenMP and OpenCL.

Table 0-1. Comparison of the TianHe-2 and TianHe-2A.

Components	TianHe-2	TianHe-2A
Nodes and performance	16,000 nodes with Intel CPUs + KNC	17,792 nodes with Intel CPUs + Matrix-2000
	54.9 Pflop/s	94.97 Pflop/s
Interconnection	10 Gbps, 1.57 us	14 Gbps, 1 us
Memory	1.4 PB	3.4 PB
Storage	12.4 PB, 512 GB/s	19 PB, 1 TB/s (upgrading, maybe larger)
Energy efficiency	17.8 MW, 1.9 Gflop/s per Watt	16.9MW, >5 Gflop/s per Watt (predicted)
Heterogeneous software	MPSS for Intel KNC	OpenMP/OpenCL for Matrix-2000

Each of the 17,792 TH-2A compute nodes will use two of Intel's Ivy Bridge CPUs (12 cores clocked at 2.2 GHz) and two of the new NUDT-designed Matrix-2000 accelerators (128 cores

[1] A report on the TianHe-2 can be found here: http://bit.ly/tianhe-2.

[2] The name Matrix-2000 comes from Chinese name "迈创", meaning "taking a creative step", this word's Chinese pronunciation is "mai chuang", similar to Matrix.

clocked at 1.2 GHz). This combination results in a compute system with 35,584 Ivy Bridge CPUs, 35,584 Matrix-2000 accelerators, and a grand total of 4,981,760 compute cores.

Each of the Matrix-2000's 128 compute cores can perform 16 double precision FLOPs per cycle, which results in 2.4576 Tflop/s total peak performance for each accelerator (128 cores × 16 FLOPs per cycle × 1.2 GHz clock). Each of the Intel Ivy Bridge CPU's 12 compute cores can perform 8 (64 bit) FLOPs per cycle per core, which results in 211.2 Gflop/s total peak performance per socket (12 cores × 8 FLOPs per cycle × 2.2 GHz clock).

National Supercomputer Centre in Guangzhou

The NSCC-GZ was jointly sponsored by the provincial government of Guangdong, the city government of Guangzhou, the NUDT, and Sun Yat-Sen University and has a total investment of over 2.5 billion RMB.

The goal of NSCC-GZ is to become the central node of China's high-performance computing (HPC) network, to provide HPC service for big science and big technology, to build an advanced supporting platform for the development of key technologies, as well as to attract high-end industries and top intelligence to push forward the industrial upgrade of Guangzhou.

NSCC-GZ is located in the East Campus of Sun Yat-Sen University, and NSCC-GZ takes up a total area of 42,332 square meters. The computer rooms (the mainframe room and the room for storage cabinets), together with other equipment rooms (rooms for the high/low voltage distribution system and the cooling system), take up an area of about 17,500 square meters.

Workshop

The Third International High Performance Computing Forum (IHPCF2017), sponsored by the Ministry of Science and Technology (MOST) and the National Science Foundation of China (NSFC), organized by NUDT, and hosted by the NSCC-GZ, was held on September 18–20, 2017 in Guangzhou, China.

IHPCF2017 offered an opportunity for 160 distinguished scholars, scientists, and engineers from around the world to exchange expertise and information on the breakthrough technologies that impact the future of HPC. The presentations and discussion of this year's meeting centered on the challenges of, and possible solutions for, emerging HPC architectures, programming models, big data analysis, power management, benchmarks, and the convergence of big data and artificial intelligence.

One of the highlights of IHPCF2017 was Professor Kai Lu's (NUDT) presentation on the new TianHe-2A system, and what follows in this report is a deep dive into what was described in that presentation.

Matrix-2000 Accelerator

TH-2A is equipped with 35,584 proprietary Matrix-2000 accelerators (Figure 1). Each Matrix-2000 has 128 compute cores clocked at 1.2 GHz, achieving 2.4576 Tflop/s of peak performance. The peak power dissipation is about 240 Watts and the dimensions are 66mm by 66mm. Looking at Figure 2, which shows the Matrix-2000's conceptual structure, we see that the accelerator itself is configured with four supernodes (SNs) that are connected through a scalable on-chip communication network. Each SN has 32 compute cores and complies with the cache coherence. The accelerator supports eight DDR4-2400 channels and is integrated with a ×16 PCI Express 3.0 endpoint port. The compute core is an in-order 8~12 stage reduced instruction set computer (RISC) pipeline extended with a 256-bit vector instruction set architecture (ISA). Two 256-bit vector functional units (VFUs) are integrated into each compute core, resulting in 16 double precision FLOPs per cycle. Thus, the peak performance of the Matrix-2000 can be calculated as: 4 SNs × 32 cores × 16 FLOPs per cycle × 1.2 GHz clock = 2.4576 Tflop/s.

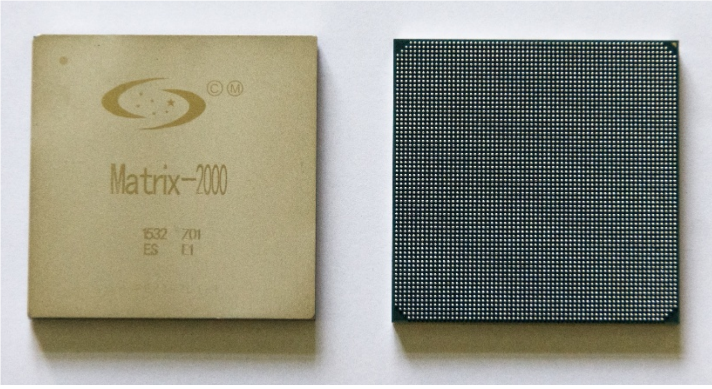


Figure 1. Matrix-2000.

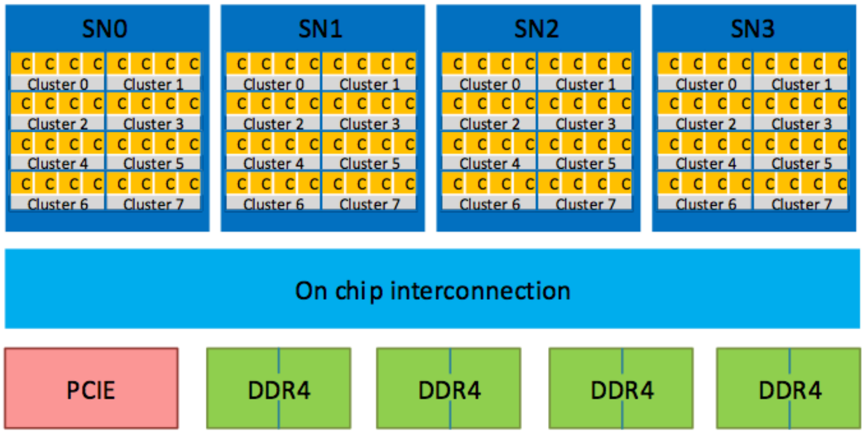


Figure 2. Conceptual structure of Matrix-2000.

The topology of the network on chip (NoC) in one SN is a 4 × 2 mesh. Figure 3 shows the basic network structure. In total, there are eight routers in the NoC. One cluster (made up of four compute cores) and a directory control unit (DCU) are connected to each router (Figure **3**). For each router port there are four physical channels to support the cache coherence protocol: (1) request, (2) response, (3) snoop, and (4) acknowledgement; each with an effective data width of 128 bits. Four compute cores are organized as a cluster. In addition, two DDR4-2400 memory control units are integrated into each SN.

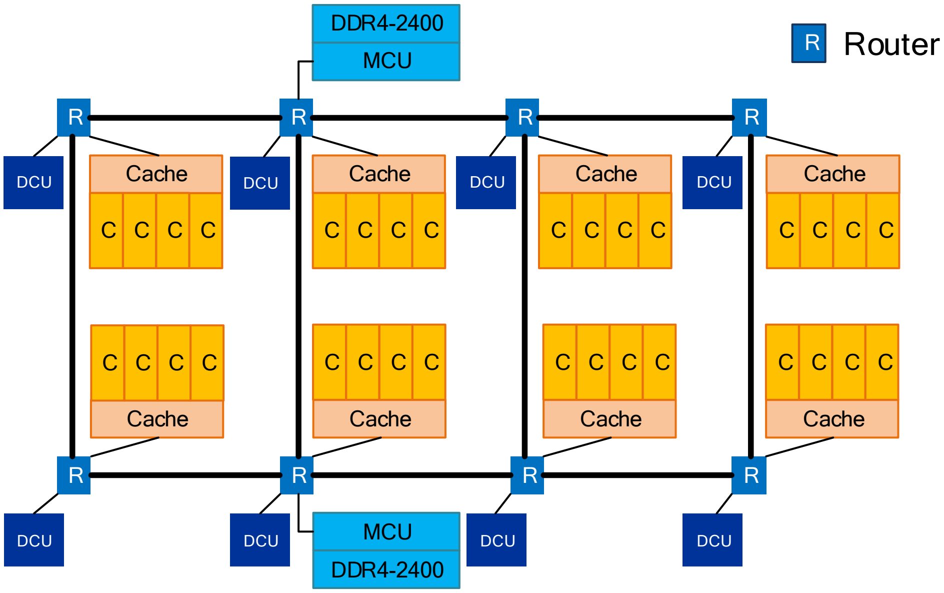


Figure 3. Topology of the NoC.

There are three Fast Interconnect Transport (FIT) link ports per SN, and each port is used to connect with another SN. On the Matrix-2000, four SNs are connected to each other with a point-to-point interconnect network. The bidirectional bandwidth of one FIT link is 25.6 GB/s, and the round-trip delay is about 20 ns. The cyclic redundancy check (CRC) and retry mechanism is supported to ensure correct data transmission. Direct memory access (DMA) between the SNs is supported to improve the bandwidth utilization of the FIT link. The bandwidth utilization is about 93.8% in DMA mode.

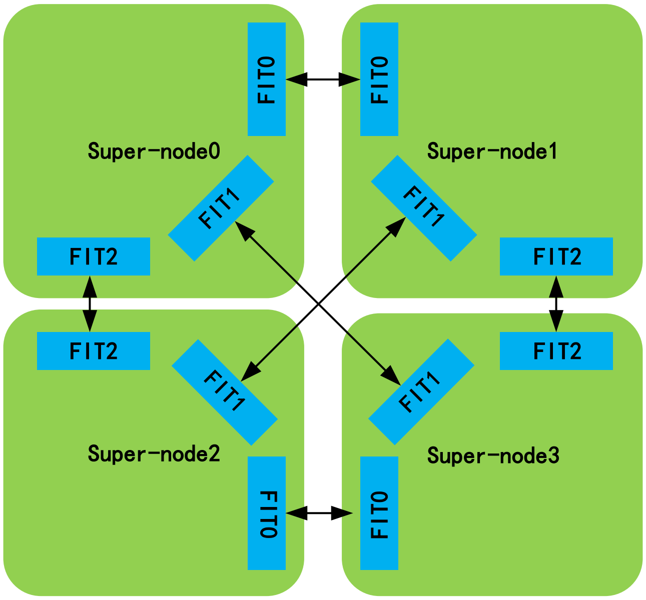


Figure 4. Matrix-2000 interconnect between SNs.

Compute Blade

The original TH-2 compute blade consisted of two nodes split into two modules: (1) the Computer Processor Module (CPM) module and (2) the Accelerator Processor Unit (APU) module (Figure 5). The CPM contained four Ivy Bridge CPUs, memory, and one Xeon Phi KNC accelerator, and the APU contained five Xeon Phi KNC accelerators. Connections from the Ivy Bridge CPUs to each of the KNC accelerators are made through a ×16 PCI Express 2.0 multi-board with 10 Gbps of bandwidth. The actual design and implementation of the board supports PCI Express 3.0, but the Xeon Phi KNC accelerator only supports PCI Express 2.0. There was also a PCI Express connection for the network interface controller (NIC).

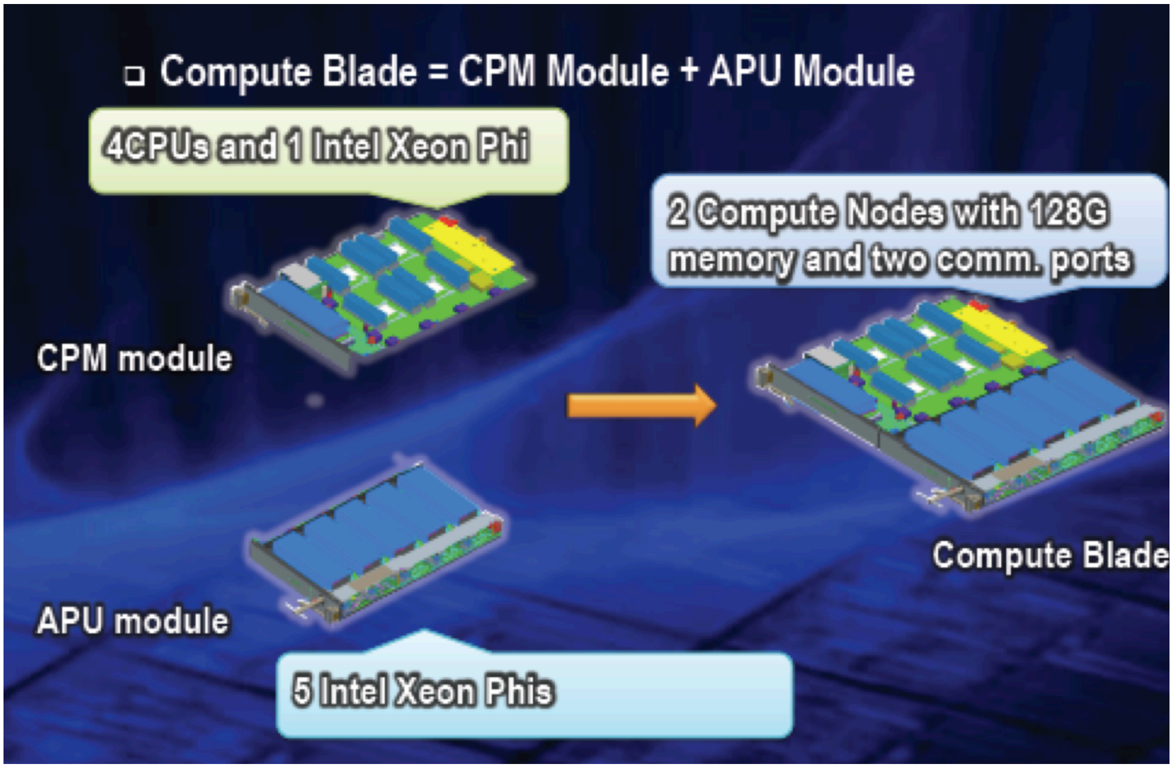

❑ Compute Blade = CPM Module + APU Module
4CPUs and 1 Intel Xeon Phi
2 Compute Nodes with 128G memory and two comm. ports
CPM module
Compute Blade
APU module
5 Intel Xeon Phis

Figure 5. Original TH-2 compute blade.

With the upgraded TH-2A, the Intel Xeon Phi KNC accelerators have been replaced. The CPM module still has four Ivy Bridge CPUs but is no longer housing an accelerator. The APU now houses four Matrix-2000 accelerators instead of the five Intel Xeon Phi KNC accelerators. So, in the TH-2A, the compute blade has two heterogeneous compute nodes, and each compute node is equipped with two Intel Ivy Bridge CPUs and two proprietary Matrix-2000 accelerators. Each node has 192 GB memory, and a peak performance of 5.3376 Tflop/s. The Intel Ivy Bridge processors have not been changed and are the same as in the original TH-2. Each of the Intel Ivy Bridge CPU's 12 compute cores can perform 8 FLOPs per cycle per core, which results in 211.2 Gflop/s total peak performance per socket (12 cores × 8 FLOPs per cycle × 2.2 GHz clock).

The logical structure of the compute node is shown in Figure 6. The two Intel Ivy Bridge CPUs are linked using two Intel Quick Path Interconnects (QPI). Each CPU has four memory channels with eight dual in-line memory module (DIMM) slots. CPU_0 expands its I/O devices using Intel's Platform Controller Hub (PCH) chipset and connects with a 14G proprietary NIC through a ×16 PCI Express 3.0 connection. Each CPU also uses a ×16 PCI Express 3.0 connection to access the Matrix-2000 accelerators. Each accelerator has eight memory channels. In a compute node, the CPUs are equipped with 64 GB of DDR3 memory, while the accelerators are equipped with 128 GB of DDR4 memory. With 17,792 compute nodes, the total memory capacity of the whole system is 3.4 PB.

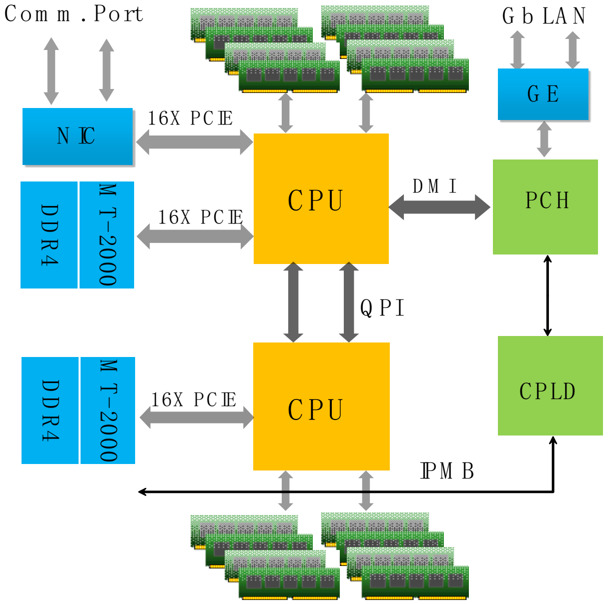


Figure 6. Logical structure of the compute node.

As shown in Figure 7, a TH-2A compute blade is composed of two parts: the CPM (left) and the APU (middle). The CPM integrates four Ivy Bridge CPUs, and the APU integrates four Matrix-2000 accelerators. Each compute blade contains two heterogeneous compute nodes.

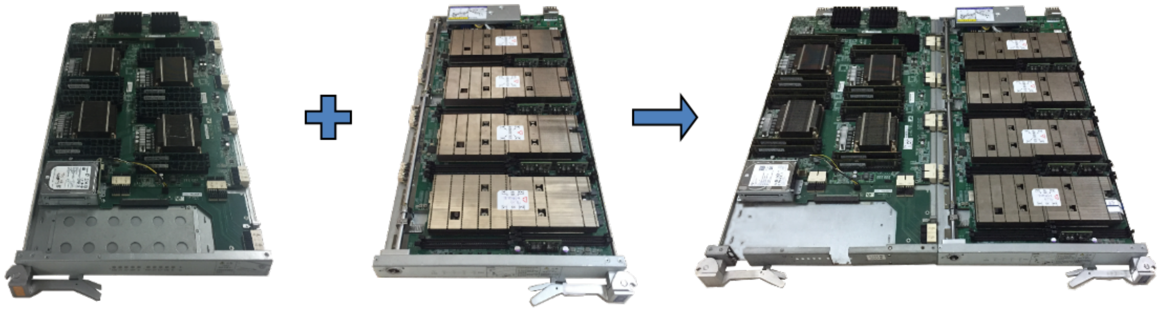


Figure 7. Two-part compute blade structure.

As stated earlier, the peak performance of each Ivy Bridge CPU is 211.2 Gflop/s, and the peak performance of each Matrix-2000 accelerator is 2.4576 Tflop/s. Thus, the peak performance of each compute node can be calculated as (0.2112 Tflop/s × 2) + (2.4576 Tflop/s × 2) = 5.3376 Tflop/s. With 17,792 compute nodes, the peak performance of the whole system is 94.97 Pflop/s (5.3376 Tflop/s x × 17,792 nodes = 94.97 Pflop/s total).

Proprietary Interconnect

The TH-2A also has an upgraded TH Express-2 interconnection network, originally used in the TH-2. All network logic was developed and implemented in two application-specific integrated circuits (ASICs): the NIC and the network router chip (NRC). Both the NIC and the NRC adopt efficient mechanisms to achieve high-performance communications with regard to bandwidth, latency, reliability, and stability. The link rate is upgraded from 10 Gbps to 14 Gbps, and the collective offload accelerates collective operations, effectively improving the throughput of a single chip.

The NIC provides the software-to-hardware interface for accessing the high-performance network and implements the proprietary MP/BLT (Mini-Packet/Block Transfer) communication and collective offload mechanisms. The NIC contains a ×16 PCI Express 3.0 interface and connects through interconnect fabric and network ports. Each port has an eight-lane, 14 Gbps SerDes interface. The NRC is an evolutionary upgrade of its predecessor used in TH-2 and offers improved performance. The NRC switches data among 24 network ports, and each port has an 8-lane, 14 Gbps SerDes interface with a bi-directional bandwidth of 224 Gbps. The throughput of a single NRC chip is 5.37 Tbps.

TH-2A's interconnection network adopts a hierarchical topology, where 32 compute nodes are packaged in one compute frame and connected using a 32 × 32 switch board, and the compute frames are connected through 24 top-level 576 port switches using active optical cables (Figure 5).

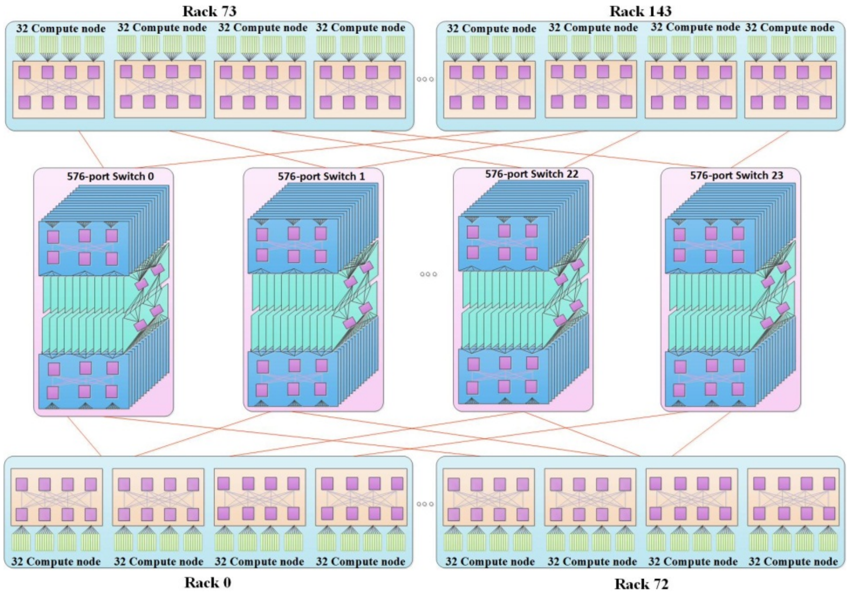


Figure 5. TH Express-2 network architecture and topology.

TH-2A's interconnection network enhances network management design by combining in-band and out-of-band management to improve the remote access services (RAS) capability of the TH Express-2 network. Advanced functions like route tracing, link testing, fault reporting, and topology discovery are supported by in-band management functions. Chip configuration and status queries are achieved through the IIC bus interface. These RAS features enable both real-time and historical status monitoring and alerts to facilitate fault locating, diagnosis, and recovery.

System Layout

TianHe-2A is composed of 168 total cabinets: 139 compute cabinets, 4 of ION cabinets, and 24 communication cabinets; 1 cabinet is empty(Figure 6).

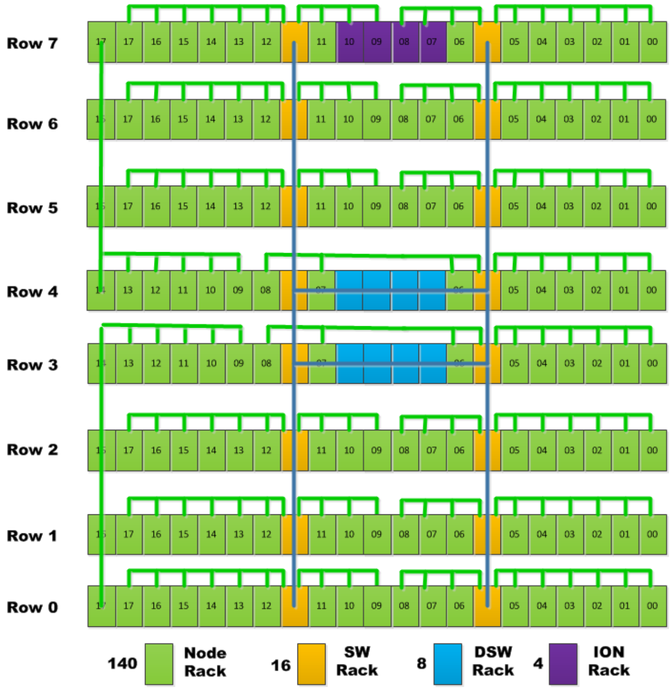


Figure 6. TianHe-2A system layout.

The Software Stack

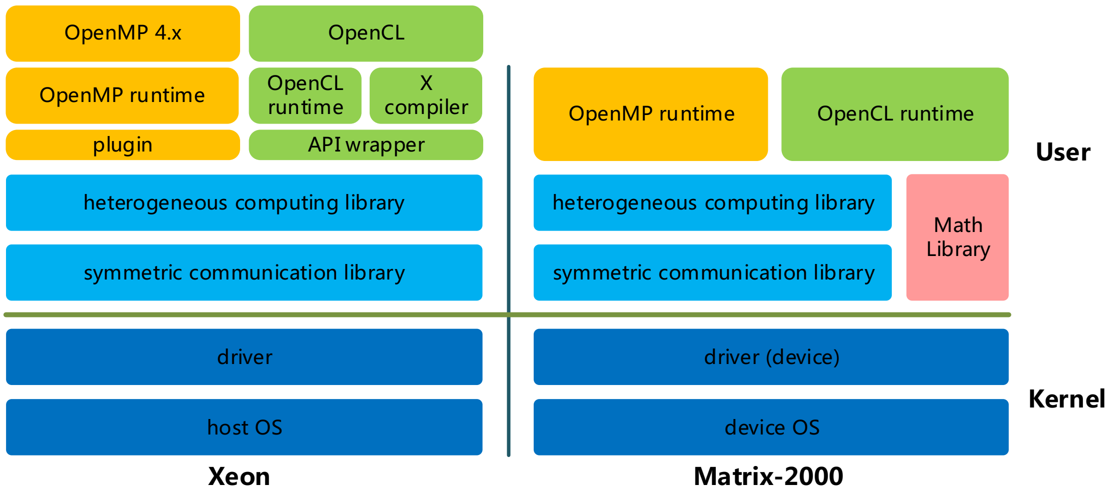


Figure 7. Runtime software stack for heterogeneous computing.

The TH-2A upgrades required the design and implementation of a heterogeneous computing software stack for the Matrix-2000 accelerator. This software stack provides a compiling and execution environment for OpenMP 4.5 and OpenCL 1.2. The runtime software stack is illustrated in Figure 7. In kernel mode, there is a light-weight Linux-based operating system (OS), with the accelerator device driver embedded within it, running on the Matrix-2000 that provides device resource management and data communication with the host CPU through the PCI Express connection. The OS manages the computing cores through an elaborately designed thread pool mechanism, which enables task scheduling with low overhead and high efficiency. In user mode, there is a symmetric communication library acting as the application programming interface (API) wrapper for the device driver, which constructs a virtual, symmetric data communication network between the processors. The major functions that the library provides include connection (establishing socket-like connections between processes on different nodes in the network), messaging (exchange of short, latency-sensitive messages such as commands and synchronization operations), and RMA operations (one-sided communication for transferring large amounts of data, optimized for bandwidth-sensitive tasks).

The heterogeneous computing library that sits above the communication library is a common middleware that serves the runtime of high-level API implementations, including OpenMP and OpenCL. It provides heterogeneous computing APIs, including device management, remote data management, and task offload management.

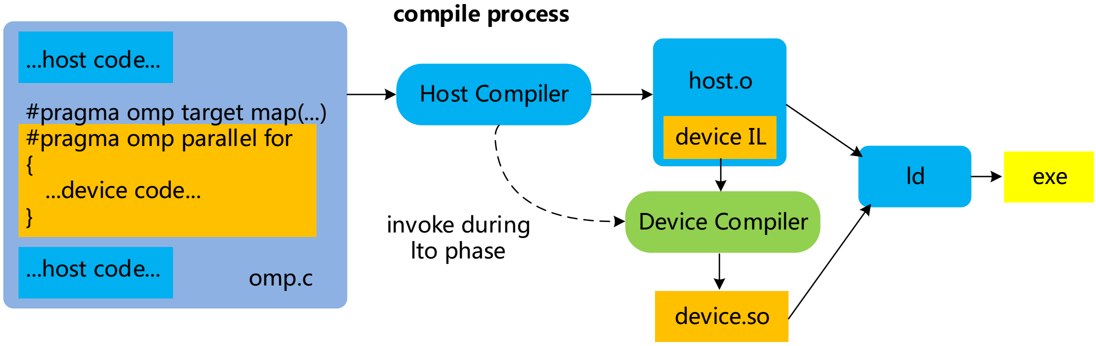


Figure 8. Compile process of an OpenMP program with target directives.

High Performance LINPACK Benchmark

Figure 9 shows a run of the High Performance LINPACK (HPL) benchmark on 4,096 nodes of TH-2A. This run uses about 50.7 GB of the memory of each node and achieves 13.987 Pflop/s out of a theoretical peak of 21.86 Pflop/s, resulting in a 63.98% efficiency of the peak performance.

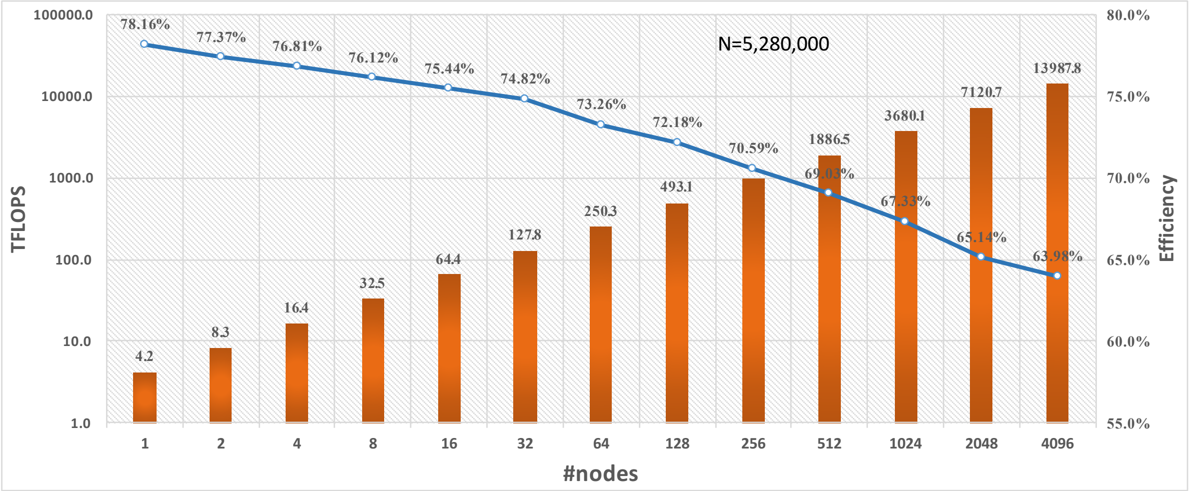


Figure 9. TianHe-2A performance on HPL.

Summary

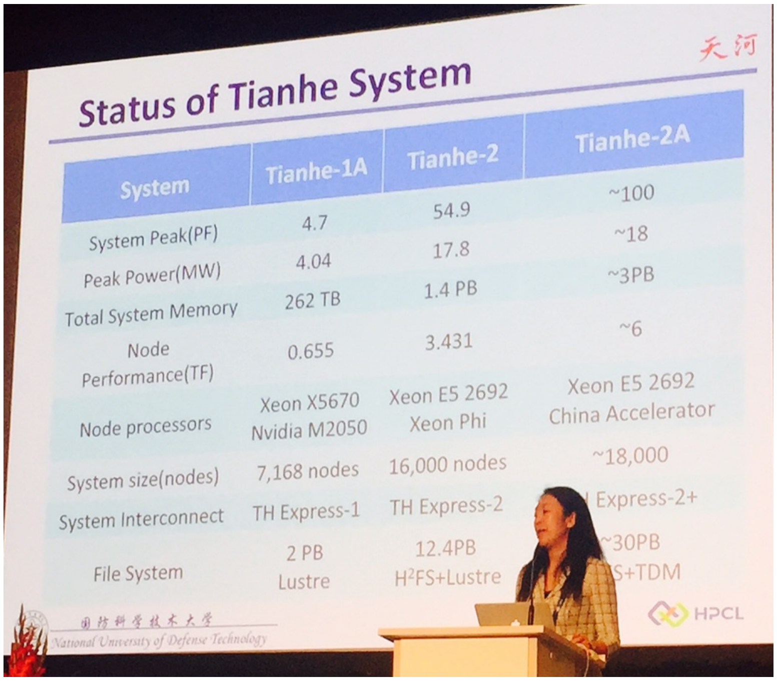
In February 2015, the US Department of Commerce prevented some Chinese research groups from receiving Intel technology. The department cited concerns about nuclear research being performed on compute systems equipped with Intel components. The research centers affected include: NSCC-G, site of Tianhe-2; the National SC Center Tianjin, site of Tianhe-1A; the NUDT, developer; and the National SC Center Changsha, location of NUDT.

At the 2015 International Supercomputing Conference (ISC) in Frankfurt, Yutong Lu, the director of the NSCC-G, described the TianHe-2A system (Figure 10). Most of what was shown in her slide in 2015 has been realized in the Matrix-2000 accelerator. They had hoped to replace the Intel KNC accelerator in their TH-2 with the Matrix-2000 by 2016. However, because of delays that has not happened until very recently.

After the embargo on Intel components by the US Department of Commerce, it has taken NUDT about two years to design and implement a replacement for the Intel Xeon Phi KNC accelerator. Their replacement is about the same level of performance as the current generation of Intel's Xeon Phi, known as Knights Landing (KNL). Equaling the performance of the state-of-the-art KNL chip and developing the accompanying software stack in such a short time is an impressive result.

The system fully supports IEEE-standard double precision (64-bit) and single precision (32-bit) arithmetic, although there is no support for half precision (16-bit) arithmetic at this time.

The bottom line is that after just two years the team at NSCC-G has replaced the Xeon Phi KNC—not with the embargoed Intel Xeon Phi KNL—but with their own, equally capable accelerator, the Matrix-2000.



天河

Status of Tianhe System

System	Tianhe-1A	Tianhe-2	Tianhe-2A
System Peak(PF)	4.7	54.9	~100
Peak Power(MW)	4.04	17.8	~18
Total System Memory	262 TB	1.4 PB	~3PB
Node Performance(TF)	0.655	3.431	~6
Node processors	Xeon X5670 Nvidia M2050	Xeon E5 2692 Xeon Phi	Xeon E5 2692 China Accelerator
System size(nodes)	7,168 nodes	16,000 nodes	~18,000
System Interconnect	TH Express-1	TH Express-2	Express-2+
File System	2 PB Lustre	12.4PB H²FS+Lustre	~30PB S+TDM

国防科学技术大学
National University of Defense Technology

HPCL

Figure 10. Yutong Lu describing the TH-2A at the 2015 ISC meeting in Frankfurt.

Appendix A.

Table A-1. TianHe-2A system summary

CPU	Intel Xeon Ivy Bridge plus Matrix-2000 accelerator
Developer	NUDT
Chip fab	N/A
Instruction set	x86 + Matrix-2000 ISA
Node processor cores	Intel's Ivy Bridge Processor (12 cores @ 2.2 GHz) plus Matrix-2000 (128 cores @ 1.2 GHz)
Node peak performance	5.3376 Tflop/s
Process technology	N/A
Power	16.9 MW
Peak performance of system	94.97 Pflop/s
Targeted application	HPC
Nodes	17,792
Total memory	3.4 PB
Compute cabinets	139
Nodes per compute cabinet	128 Nodes
Cores per node	280 cores
Total system core count	4,981,760

Table A-2. Comparison with top machines on the TOP500

	ORNL Titan	**NUDT TianHe-2**	**TianHe-2A**	**Sunway TaihuLight**
Theoretical Peak	27.1 Pflop/s = (2.6 CPU + 24.5 GPU)	54.9 Pflop/s = (6.75 CPU + 48.14 Coprocessor)	94.97 Pflop/s = (7.52 CPU + 87.45 Accelerator) Pflop/s	125.4 Pflop/s = CPEs + MPEs Cores per Node = 256 CPEs + 4 MPEs

	Pflop/s	Pflop/s		Supernode = 256 Nodes
				System = 160 Supernodes
				Cores = 260 * 256 * 160 = 10.6M

HPL Benchmark flop/s	17.6 Pflop/s	30.65 Pflop/s	4,096 nodes This run uses 50.7 GB of memory of each node, and achieves 13.987 Pflop/s out of a theoretical peak of 21.86 Pflop/s.	93 Pflop/s
HPL % peak	65.19%	55.83%	63.98%	74.16%
HPCG benchmark	0.322 Pflop/s	0.580 Pflop/s	4096 nodes .0798 Pflop/s	.371 Pflop/s
HPCG % peak	1.2%	1.1%	0.365%	0.30%
Compute nodes	18,688	16,000	17,792	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)	2 – Intel Ivy Bridge (12 cores, 2.2 GHz) plus 2 Matrix-2000, 1.2 GHz)	256 CPEs + 4 MPEs
Sockets	18,688 Interlagos + 18,688 Nvidia boards	32,000 Ivy Bridge + 48,000 Xeon Phi accelerators	Complete system: 17,792 compute nodes Each node contains 2 Intel Ivy Bridge sockets and 2 Matrix-2000 accelerators	40,960 nodes with 256 CPEs and 4 MPEs per node
Node peak performance	1.4508 Tflop/s =	3.431 Tflop/s =	5.3376 Tflop/s = (2 × 211.2 CPU + 2 × 2.4576	12

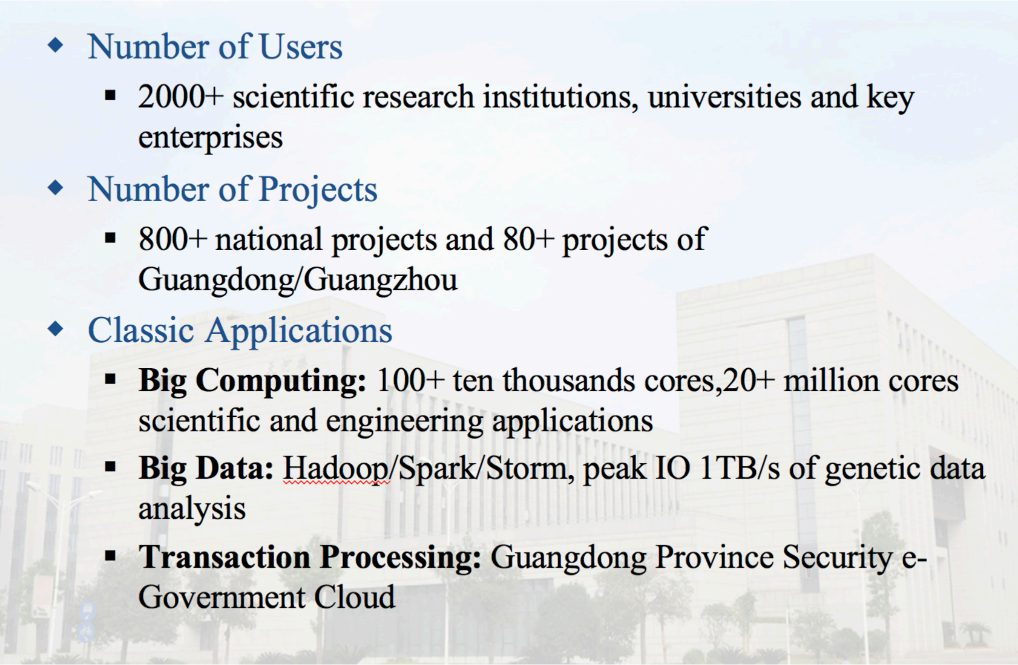
	(.1408 CPU + 1.31 GPU) Tflop/s	(2 × .2112 CPU + 3 × 1.003 Coprocessor) Tflop/s	Accelerator) Tflop/s	
Node memory	32 GB CPU + 6 GB GPU	64 GB CPU + 3*8 GB Coprocessor	64 GB CPU + 128 GB Accelerator	32 GB per node
System memory	.710 PB = (.598 PB CPU and .112 PB GPU)	1.4 PB = (1.024 PB CPU and .384 PB Coprocessor)	3.4 PB = 17,792 × (64 GB + 128 GB)	1.31 PB (32 GB × 40,960 nodes)
Configuration	4 nodes per blade, 24 blades per cabinet, and 200 cabinets in the system	2 nodes per blade, 16 blades per frame, 4 frames per cabinet, and 125 cabinets in the system	2 Nodes per blade, 16 blades per frame and 4 frames per cabinet and 139 cabinets in the system.	Node peak performance is 3.06 Tflop/s, or 11.7 Gflop/s per core. 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 SNs = cabinet Each cabinet ~3.164 Pflop/s 256 nodes per SN 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)	4,981,760 cores = (17,792 × 2 Ivy Bridge with 12 cores) + (2 × Matrix-2000 × 128)	10,649,600 cores = Node (260) × supernodes(256 nodes) × 160 supernodes
Power (processors, memory, interconnect)	9 MWatts	17.8 MW	16.9 MW for full system	15.3 MW

Footprint	404 m^2	720 m^2	400 m^2 (50 m^2/line*8) or 720 m^2 total room	605 m^2

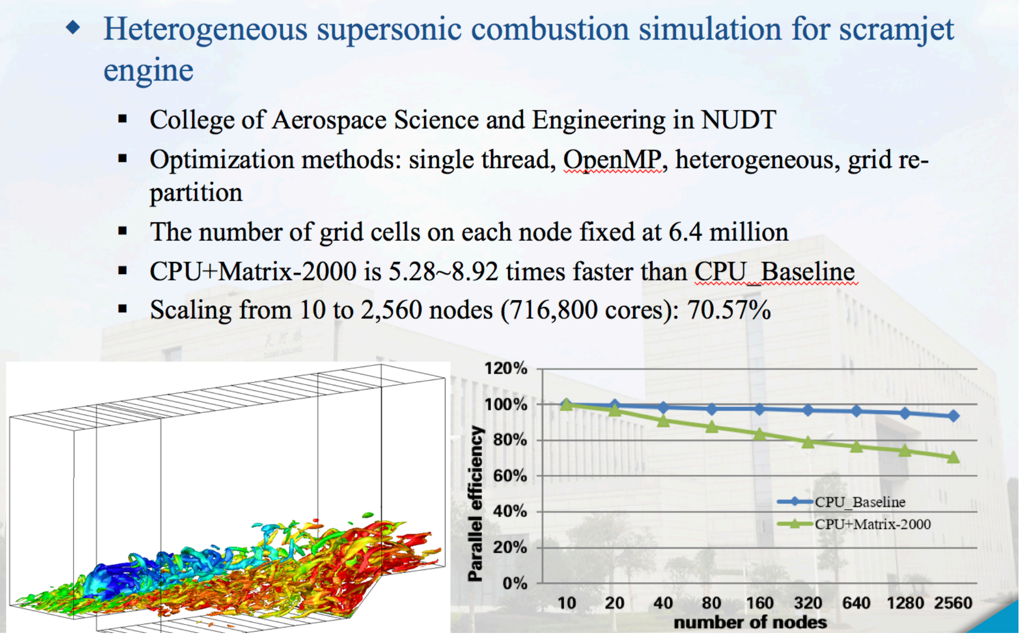
Appendix B.

Applications and Users of the TianHe-2 System

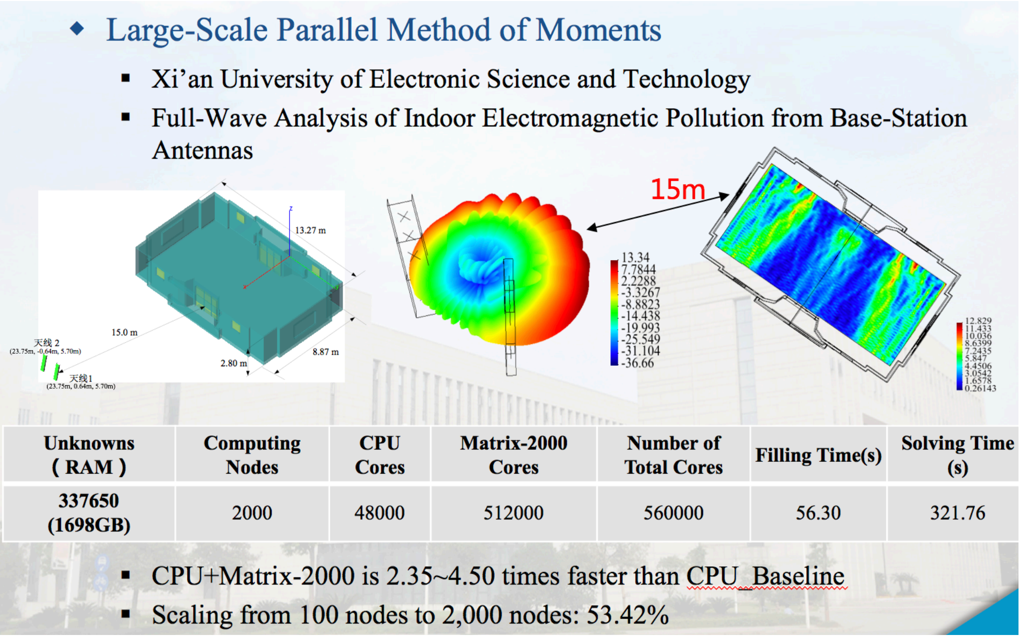


- Number of Users
  - 2000+ scientific research institutions, universities and key enterprises
- Number of Projects
  - 800+ national projects and 80+ projects of Guangdong/Guangzhou
- Classic Applications
  - **Big Computing:** 100+ ten thousands cores,20+ million cores scientific and engineering applications
  - **Big Data:** Hadoop/Spark/Storm, peak IO 1TB/s of genetic data analysis
  - **Transaction Processing:** Guangdong Province Security e-Government Cloud

Case study for Tianhe-2A system



- Heterogeneous supersonic combustion simulation for scramjet engine
  - College of Aerospace Science and Engineering in NUDT
  - Optimization methods: single thread, OpenMP, heterogeneous, grid re-partition
  - The number of grid cells on each node fixed at 6.4 million
  - CPU+Matrix-2000 is 5.28~8.92 times faster than CPU_Baseline
  - Scaling from 10 to 2,560 nodes (716,800 cores): 70.57%



- Large-Scale Parallel Method of Moments
  - Xi’an University of Electronic Science and Technology
  - Full-Wave Analysis of Indoor Electromagnetic Pollution from Base-Station Antennas

Unknowns (RAM)	Computing Nodes	CPU Cores	Matrix-2000 Cores	Number of Total Cores	Filling Time(s)	Solving Time (s)
337650 (1698GB)	2000	48000	512000	560000	56.30	321.76

- CPU+Matrix-2000 is 2.35~4.50 times faster than CPU_Baseline
- Scaling from 100 nodes to 2,000 nodes: 53.42%