

Visit to the National University for Defense Technology Changsha, China

Jack Dongarra

University of Tennessee

Oak Ridge National Laboratory

6/2/13 9:42 PM

On May 28-29, 2013, I had the opportunity to attend an International HPC Forum (IHPCF) in Changsha China, which was organized by the National University of Defense Technology (NUDT) and sponsored by the Ministry of Science and Technology of China, the National Natural Science Foundation of China, and the National University of Defense Technology. NUDT is located in Changsha in Hunan province of China.

The IHPCF had a number of invited speakers for many countries:

- Taisuke Boko, University of Tsukuba, Japan
- Xuebing Chi CAS, China
- Jack Dongarra, University of Tennessee/ORNL, USA
- Wuchun Feng, Virginia Tech, USA
- William Gropp, UIUC, USA
- Rajeeb Hazra Intel, USA
- Xiangke Liao, NUDT, China
- Guangming Liu, NSCC Tianjin China
- Zeyao Mao, IAPC M, China , China
- Marek T. Michalewicz, A*Star, Singapore
- Sebastian Schmidt, Juelich Research Center, Germany
- Jiachang Sun, CAS, China
- Xianhe Sun, IIT, USA
- Jeffrey Vetter, ORNL, USA

At the IHPCF workshop Xiangke Liao from NUDT presented details on the new Tianhe-2 (TH-2) also called the Milkyway-2 supercomputer. The project is sponsored by the 863 High Technology Program of the Chinese Government, the Government of Guangdong province, and the Government of Guangzhou city. The system will be in the National Supercomputer Center in Guangzhou (NSCC-GZ) by the end of the year. It will provide an open platform for research and education and provide high performance computing service for southern China. At the end of the first day of IHPCF there was a tour of the TH-2 computer room. The IHPCF participants had an opportunity to see the system and receive addition information on various components of the computer and its operation.

Overview

The TH-2 was developed by NUDT and Inspur. Inspur is a Chinese multinational information technology company headquartered in Jinan, Shandong, China. Inspur's business activities include server manufacturing and software development. Inspur contributed to the manufacturing of the printed circuit boards and is also contributing to the system installation and testing. The TH-2 is undergoing assembly and testing at NUDT and will be moved to its permanent home at the National Supercomputer Center in Guangzhou (NSCC-GZ) by the end of the year. The complete system has a theoretical peak performance of 54.9 Pflop/s. It is based on Intel's Ivy Bridge and Xeon Phi components and a custom interconnect network. There are 32,000 Intel Ivy Bridge Xeon sockets and 48,000 Xeon Phi boards for a total of 3,120,000 cores. This represents the world's largest (public) installation of Intel Ivy Bridge and Xeon Phi's processors. The system will be located in Southwest China.

For comparison, the next large acquisition of a supercomputer for the US Department of Energy will not be until 2015.

While the TH-2 system is based on Intel multicore (Ivy Bridge) and coprocessors (Xeon Phi), there are a number of features of the TH-2 that are Chinese in origin, unique and interesting, including the TH-Express 2 interconnection network, the Galaxy FT-1500 16-core processor, the OpenMC programming model, their high density package, the apparent reliability and scalability of the system.

Compute Node

Each compute node is composed of 2 Intel Ivy Bridge sockets and 3 Intel Xeon Phi boards see figure 1. The system is build out of the nodes and is composed as follows: 2 nodes per board, 16 board per frame, 4 frames per rack, and 125 racks make up the system, see figures 3 - 5. The compute board has two compute nodes and is composed of two half's the CPM and the APM halves. The CPM portion of the compute board contains the 4 Ivy Bridge processors, memory, and 1 Xeon Phi board and the CPM half contains the 5 Xeon Phi boards. There are 5 horizontal blind push-pull connections on the edge. Connections from the Ivy Bridge processor to each of the coprocessors are made by a PCI-E 2.0 multi-board, which has 16 lanes and is 10 Gbps each. (The actual design and implementation of the board is for PCI-E 3.0, but the Xeon Phi only supports PCI-E 2.0.) There is also a PCI-E connection to the NIC.

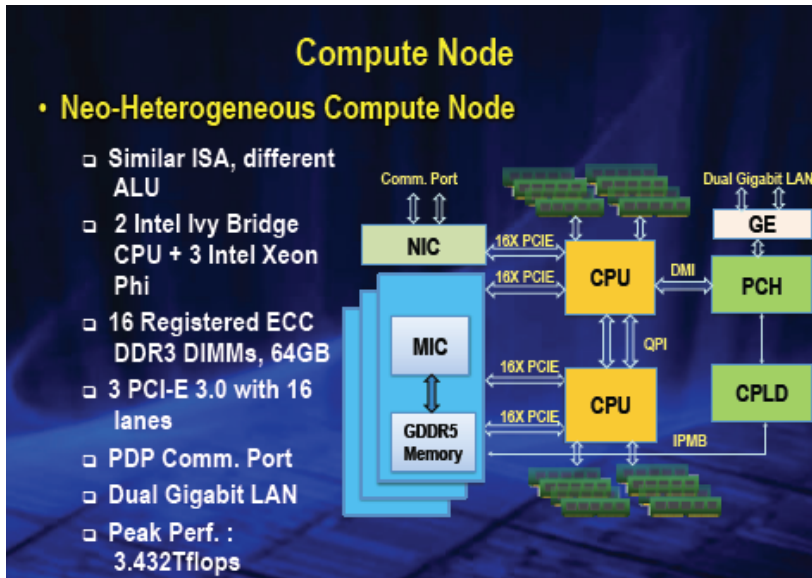


Figure 1: Compute Node

The Intel Ivy Bridge can perform 8 flops per cycle per core. Each socket has 12 cores*8 flops / cycle *2.2 GHz = 211.2 Gflop/s peak performance per socket. A node of the TH-2 has 2 Ivy Bridge sockets, so 422.4 Gflop/s is the theoretical peak from the Ivy Bridge processors on a node.

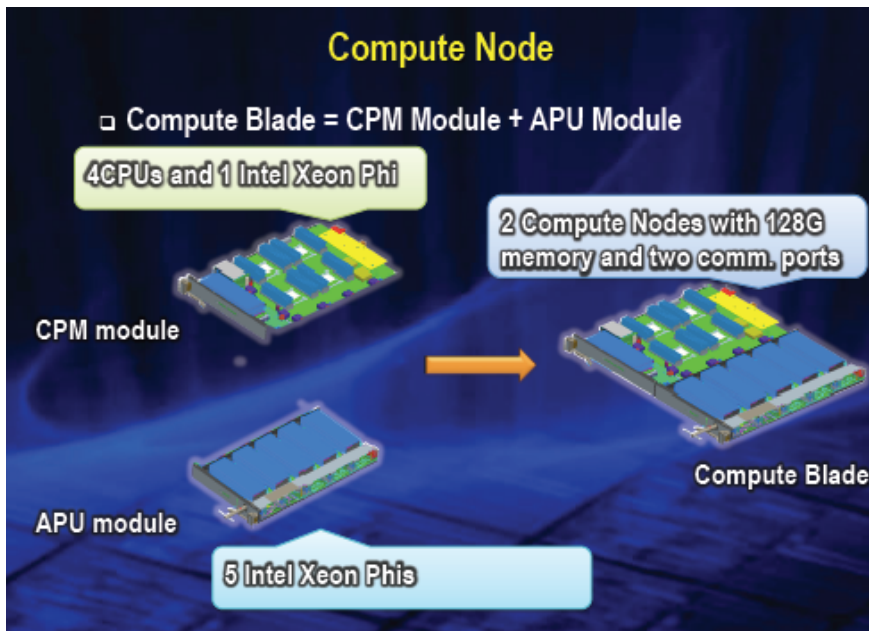


Figure 2: Compute blade's two halves

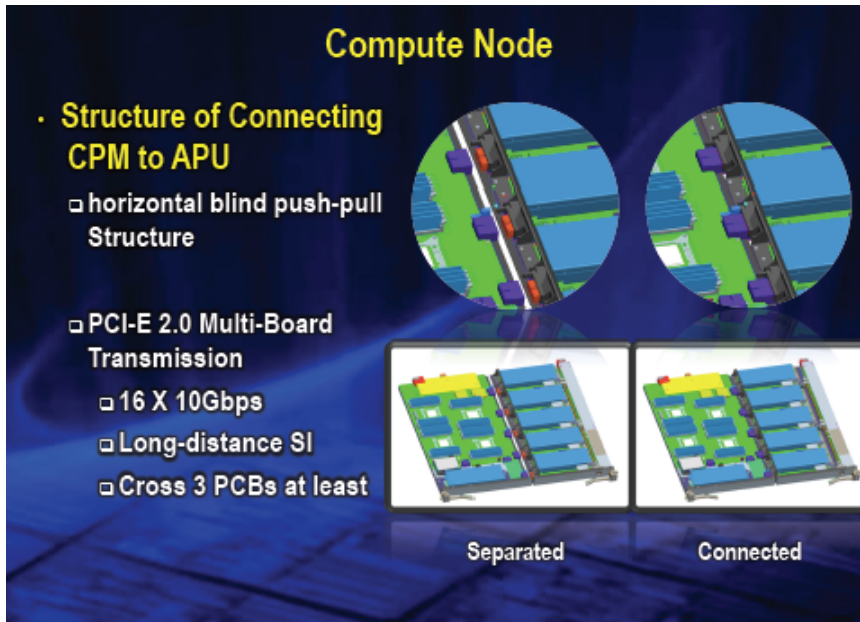


Figure 3: Compute blade halves joined



Figure 4: Compute frame

Compute Node

□ Structure of Compute Frame

- middle backplane double sides central symmetry assemblage
- MGH(Multi-Giga Hz) Signals on Multi-Boards Transmission
 - 10Gbps Backplane Transmission
 - 8 X 10Gbps or 8 X 14Gbps
 - Long-distance(Cross Backplane) SI

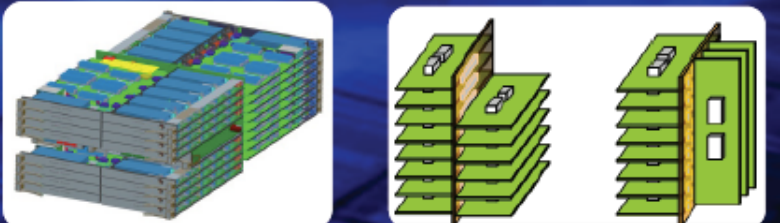


Figure 5: Compute Frame from Node Boards

The Xeon Phi's used in the TH-2 each have 57 cores. (Normally an Intel Xeon Phi has 61 cores. When asked why 57 cores, Intel said these were early chips in the production cycle and yield was an issue.) Each of the 57 cores can have 4 threads of execution and the cores can do 16 double precision flops per cycle per core. With a cycle time of 1.1 GHz this yields a theoretical peak performance of 1.003 Tflop/s for each Xeon Phi. On a node there are 2 Ivy Bridge*.2112 Tflop/s + 3 Xeon Phi*1.003 Tflop/s or 3.431 Tflop/s per node. The complete system has 16,000 nodes or 54.9 Pflop/s for the theoretical peak performance of the system. The footprint is 720 square meters, see figure 6. The system was constructed in a very confined space and hence not laid out optimally. When the system is moved to Guangzhou it will be laid out in a more uniform fashion, see figure 7.

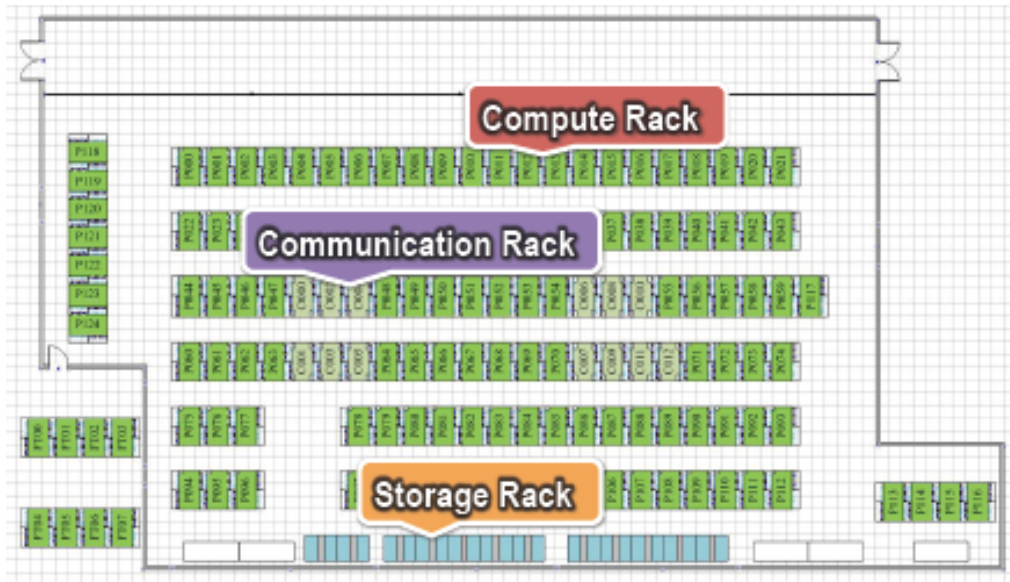


Figure 6: Layout of TH-2 at NUDT

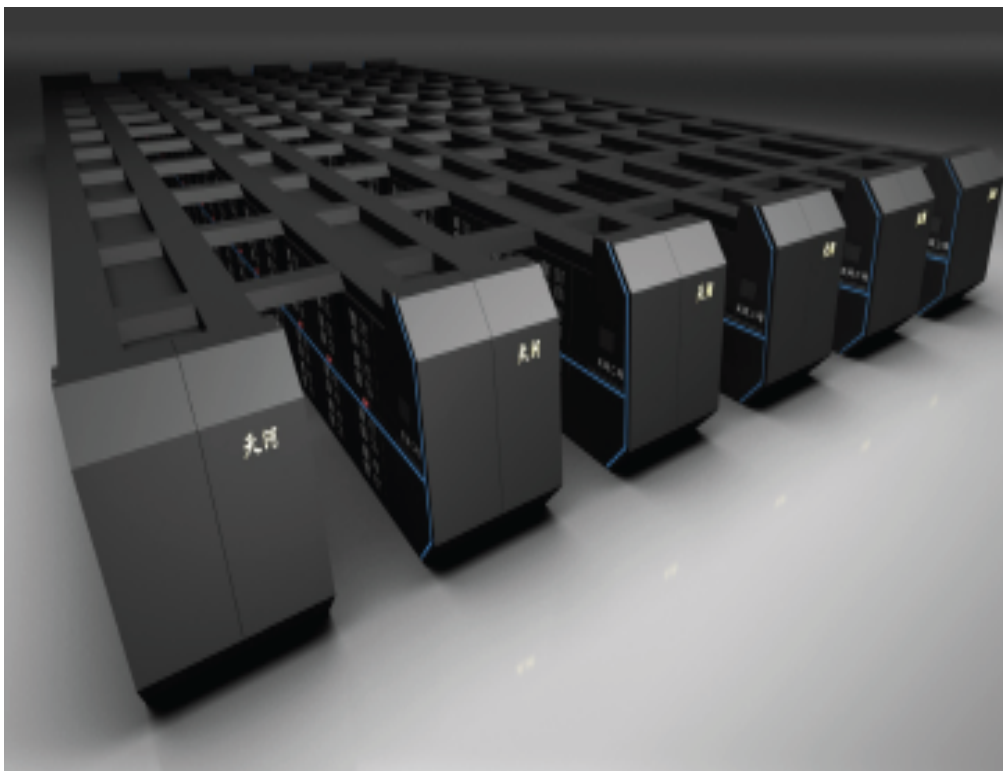


Figure 7: Artist Rendering of Computer Room in Guangdong

Each node has 64 GB of memory and each Xeon Phi has 8 GB of memory for a total of 88 GB of memory per node. With 16,000 nodes the total memory for the Ivy Bridge part is 1.024 PB and the Xeon Phi Coprocessors contributed 8 GB per board or a total of 24 GB per node or .384 PB for the Coprocessors. Bringing the total memory to 1.404 PB for the system.

Power and Cooling

The peak power consumption under load for the system is at 17.6 MWs. This is just for the processors, memory and interconnect network. If cooling is added the total power consumption is 24 MWs. The cooling system used is a closed-coupled chilled water-cooling with a customized liquid water-cooling unit. It has a high cooling capacity of 80 kW. When the machine is moved to the NSCC in Guangzhou it will use city water to supply cool water for the system.

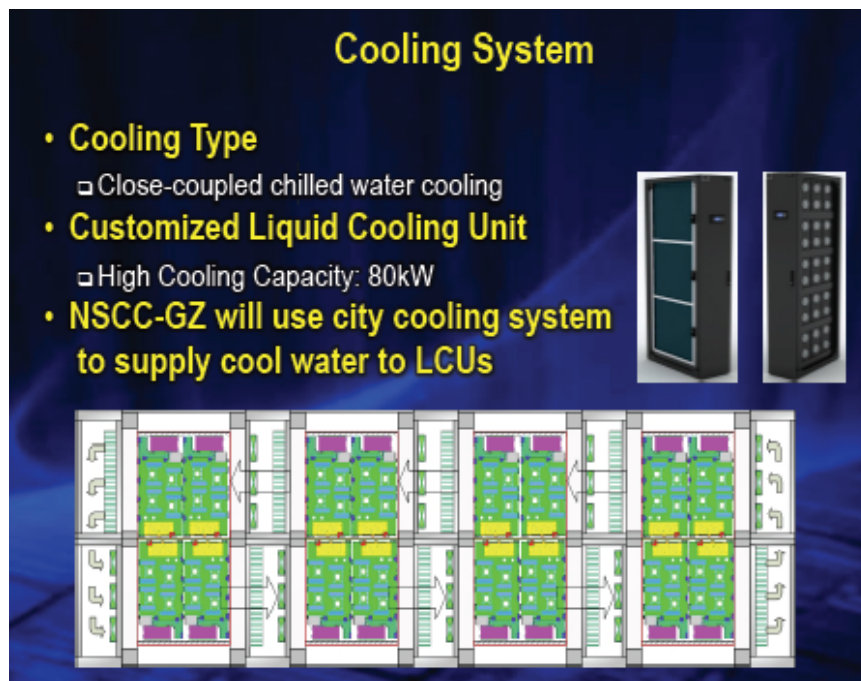


Figure 8: System Cooling

Each cabinet (frame) has a series of lights on the frame door. During operation the lights on the frame blink showing activity. There is also a horizontal strip of lights on the cabinet door which changes color depending on the power load on the frame, see Figure 9 and 10.



Figure 9: Lights on the TH-2



Figure 10: Lights on the TH-2

The Frontend processors

In addition to the compute nodes there is a frontend system composed of 4096 Galaxy FT-1500 CPUs. These processors were designed and developed at NUDT. They are not considered as part of the compute system. The FT-1500 is 16 cores and based on SparcV9. It uses 40 nm technology and has a 1.8 GHz cycle time. Its performance is 144 Gflop/s and each chip runs at 65 Watts. By comparison the Intel Ivy Bridge has 12 cores uses 22 nm technology and has a 2.2 GHz cycle time with a peak performance of 211 Gflop/s.

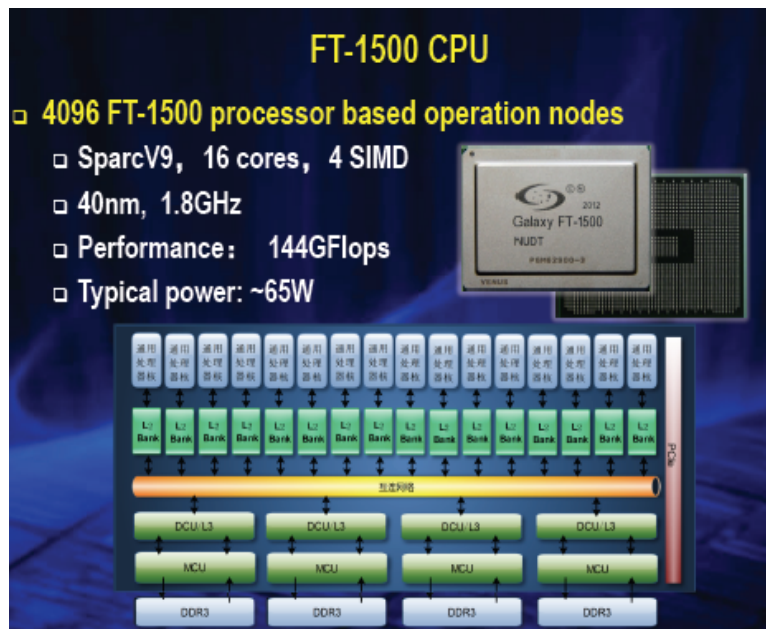


Figure 11: Frontend Processor

The Interconnect

NUDT has built their own proprietary interconnect called the TH Express-2 interconnect network. The TH Express-2 uses a fat tree topology with 13 switches each of 576 ports at the top level. This is an optoelectronics hybrid transport technology. Running a proprietary network. The interconnect uses their own chip set. The high radix router ASIC called NRC has a 90 nm feature size with a 17.16x17.16 mm die and 2577 pins. The throughput of a single NRC is 2.56 Tbps. The network interface ASIC called NIC has the same feature size and package as the NIC, the die size is 10.76x10.76 mm, 675 pins and uses PCI-E G2 16X. A broadcast operation via MPI was running at 6.36 GB/s and the latency measured with 1K of data within 12,000 nodes is about 9 us.

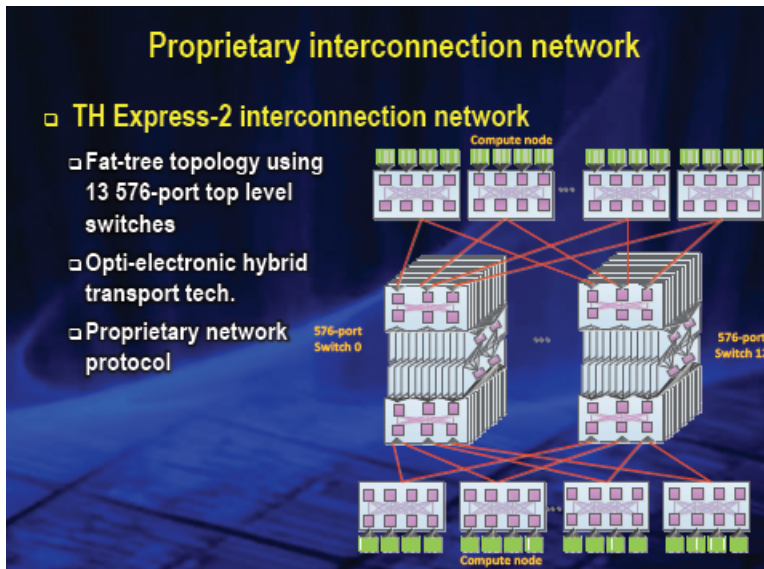


Figure 12: Interconnect

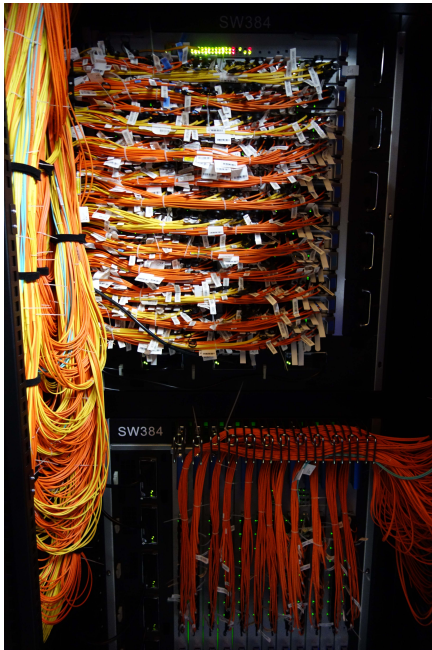


Figure 13: Picture of Interconnect

Proprietary interconnection network

- High radix router ASIC: NRC
 - Feature size: 90nm
 - Die size: 17.16mm x 17.16mm
 - Package: FC-PBGA
 - 2577 pins
 - Throughput of single NRC: 2.56Tbps
- Network interface ASIC: NIC
 - Same Feature size and package
 - Die size: 10.76mm x 10.76mm
 - 675 pins, PCI-E G2 16X




Figure 14: Chips for interconnect network

The Software Stack

The Tianhe-2 is using Kylin Linux as the operating system. Kylin is an operating system developed by the National University for Defense Technology, and successfully approved by China's 863 Hi-tech Research and Development Program office in 2006. See [http://en.wikipedia.org/wiki/Kylin_\(operating_system\)](http://en.wikipedia.org/wiki/Kylin_(operating_system)) for addition details. Kylin is compatible with other mainstream operating systems and supports multiple microprocessors and computers of different structures. The Kylin packages all include standard open source and public packages. This is the same OS used in the Tianhe-1A. Resource management is based on SLURM. They have a power-aware resource allocation and use multiple custom scheduling polices.

There are Fortran, C, C++, and Java compilers, OpenMP, and MPI 3.0 based on MPICH version 3.0.4 with custom GLEX (Galaxy Express) Channel support. They can do multi-channel message data transfers, dynamic flow control and have offload collective operations. In addition, they are developed something called OpenMC. It is a directive based intra-node programming model. Think of it as a way to use OpenMC instead of Open-MP and either CUDA, OpenACC, or OpenCL. This new abstraction for hardware and software provides for a unified logical layer above all computing including CPU cores and Xeon Phi processors but could be extended to architectures with similar ISA and heterogeneous processors. They provide directives for high efficient SIMD operations and directives for high efficiency data locality exploitation and data communication. Open-MC is still a work in progress.

They are using the Intel ICC 13.0.0 compiler. They claim to have a math library, which are based on Intel's MKL 11.0.0 and BLAS for the GPU based on Xeon Phi and optimization by the NUDT.

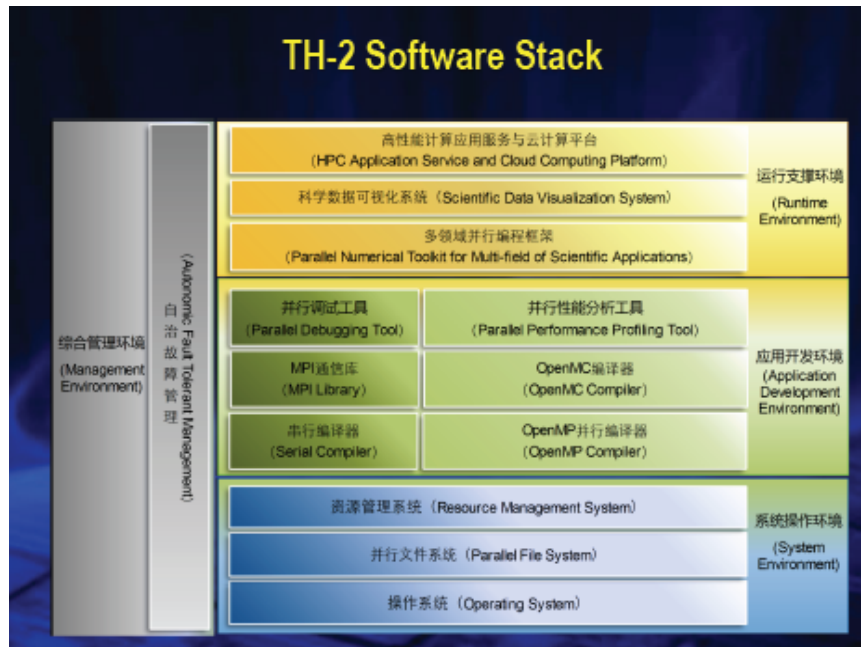


Figure 15: Software Stack

Storage

There is a global shared parallel storage system containing 12.4 PB and uses the H2FS hybrid hierarchy file system.

LINPACK Benchmark Run (HPL)

I was sent results showing a run of HPL benchmark using 14,336 nodes, that run was made using 50 GB of the memory of each node and achieved 30.65 Pflop/s out of a theoretical peak of 49.19 Pflop/s or an efficiency of 62.3% of theoretical peak performance taking a little over 5 hours to complete, see figure 16.

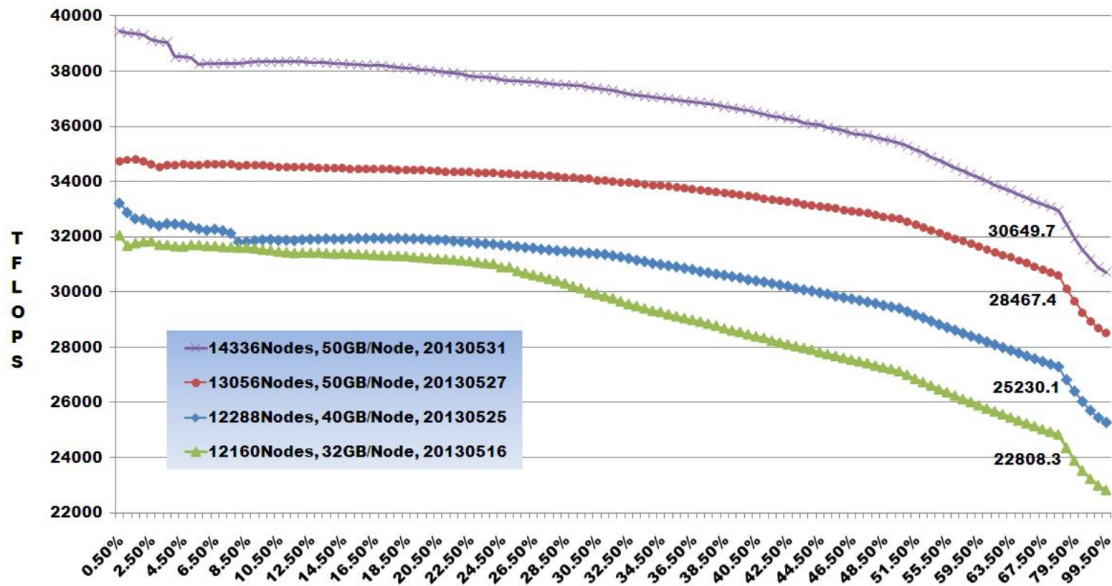


Figure 16: HPL Performance

The fastest result shown in Figure 16 was only using 90% of the machine. They are expecting the make improvements and increase the number of nodes used in the test. To compute the flops/watt one can take the power under load for the whole system (processors, memory and interconnect) at 17.6 MW and divide by the percent of the machine used to run the benchmark, in this case 14,336 nodes of the total 16,000 nodes or 90% of the machine. The performance achieved was 30.65 Pflop/s or 1.935 Gflop/Watt. The Top 5 systems on the Top 500 list have the following Gflops/Watt efficiency.

Rank	Site	Manufacture	Name	System	Gflops/Watt
1	DOE/SC/Oak Ridge National Laboratory	Cray Inc.	Titan	Cray XK7, Opteron 6274 16C 2.200GHz & NVIDIA K20x, Cray Gemini interconnect	2.143
2	DOE/NNSA/LLNL	IBM	Sequoia	BlueGene/Q, Power BQC 16C 1.60 GHz, Custom	2.069
3	RIKEN Advanced Institute for Computational Science (AICS)	Fujitsu	K	K computer, SPARC64 VIIIfx 2.0GHz, Tofu interconnect	0.830
4	DOE/SC/Argonne National Laboratory	IBM	Mira	BlueGene/Q, Power BQC 16C 1.60GHz, Custom Interconnect	2.069
5	Forschungszentrum Juelich (FZJ)	IBM	JUQUEEN	BlueGene/Q, Power BQC 16C 1.600GHz, Custom Interconnect	2.102

Applications

NUDT claims to have a number of applications that are being ported to the TH-2. The list includes:

- High-order CFD Simulation : HostA
 - Complex flow simulation of the full plane : C919
 - WCNS - Weighted Compact Nonlinear Scheme
 - 10 billion grid points
 - Over 1000 compute nodes
 - 1 MIC= 70% of 2 Ivy Bridge
- Gyrokinetic Toroidal Code : GTC
 - Porting to TH-2 using CPU+MIC (TH-1A: CPU+GPU)
 - Over 4096 compute nodes
 - 1 MIC= 80% of 2 Ivy Bridge
- Business Opinion Analysis
 - Store and process 600TB structured/non structured data with Hadoop on 1024 nodes
 - Process 100 Million data records per day
- Security e-Government Cloud
 - 512 FT-1500 nodes
 - Increase server utility from 30% to 71%

Summary of the Tianhe-2 (TH-2) or Milkyway-2	
Items	Configuration
Processors	32,000 Intel Xeon CPU's + 48,000 Xeon Phi's (+ 4096 FT-1500 CPU's frontend) Peak Performance 54.9 PFlop/s (just Intel parts)
Interconnect	Proprietary high-speed interconnection network, TH Express-2
Memory	1 PB
Storage	Global Shared parallel storage system, 12.4 PB
Cabinets	125 + 13 + 24 = 162 compute/communication/storage cabinets
Power	17.8 MW
Cooling	Closed air cooling system

Summary of the Tianhe-2 (TH-2) Milkyway-2

Model	TH-IVB-FEP
Nodes	16000
Vendor	NUDT, Inspur
Processor	Intel Xeon IvyBridge E5-2692
Speed	2.200 GHz
Sockets per Node:	2
Cores per Socket:	12
Accelerator/CP:	Intel Xeon Phi 31S1P
Accelerators/CP per Node:	3
Cores per Accelerator/CP:	57
Operating System:	Kylin Linux
Primary Interconnect:	Proprietary high-speed interconnecting network (TH Express-2)
Peak Power (MW):	17.8
Size of Power Measurement (Cores)	3,120,000
Memory per Node (GB)	64

Summary of all components

CPU Cores	384,000
Accelerators/CP	48,000
Accelerator/CP Cores	
Memory	1,024,000 GB

About NUDT

The National University for Defense Technology is one of the highly rated universities in China for Computer Science. In the 2012 Chinese university-ranking list, NUDT was ranked No.1 at the subject software engineering before Peking University, and No.2 at the subject computer science and Technology behind the Tsinghua University.

Ranking of Key Subjects in China by Chinese (lower number is better)				
		NUDT		NUDT
Subject		2009 Ranking		2004 Ranking
Computer Science and Technology		2		1
Information and Communication Engineering		3		4
Optical Engineering		4		7
Management Science and Engineering		4		7
Control Science and Engineering		10		8

Top Chinese universities with computer science and technology programs		
Overall level		
University Code / Name	Ranking	Score
	1	100
10003 Tsinghua University		
90002 NUDT	2	94
10006 Beihang University	3	88
10335 Zhejiang University	3	88
10213 Harbin Institute of Technology	5	87

The tables above courtesy of Asian Technology Information Program (<http://www.atip.org/>).

The Ministry of National Defense and the Ministry of Education jointly run the National University of Defense Technology (NUDT), based in Changsha, Hunan province. NUDT has long history in research and development on supercomputers, developed the first GFlops, TFlops, and Pflops supercomputer in China. As one of the leading research institutions on HPC in China, the NUDT has established itself for research on processors, compilers, parallel algorithms, and systems. The NUDT developed the Galaxy FT Series Stream Processor called YHFT64, a novel stream programming language called Stream FORTRAN95 (SF95), and its corresponding

compiler. The compiler uses stream architecture-oriented optimization techniques, including loop streaming, vector streaming, stream reusing, etc. The Galaxy series HPC systems developed by the NUDT have been installed for meteorological and national defense applications. The NUDT is also known for in-depth research on numerical weather forecasting parallel algorithms, remote sensing image parallel algorithms, molecular dynamics related parallel algorithms, and parallel algorithms for classical mathematics.

The series of supercomputers developed at the NUDT follows. In 1983, the Galaxy-I was developed. It included vector processing and ran at 100 Mflop/s, becoming the first computer with that capability in China; in 1992, Galaxy-II was China's first computer achieving 1 Gflop/s; in 1997, the Galaxy-III parallel computer was developed with a peak performance of 13 GFlop/s; in 2000, a version of the Galaxy was developed with a peak performance of 1 TFlop/s. A few Galaxy computers were used for national defense applications. In 2010 NUDT built the Tianhe-1A, which became the world's fastest computer at 2.6 PFlop/s. Today NUDT has constructed the Tianhe-2 with a theoretical peak performance of 54.9 Pflop/s using over 3 million processors.



Figure 17 Pictures of the Galaxy (Yinhe) I - VI and the Tianhe-1, Tianhe-1A, and Tianhe-2