B

APPENDIX

TH-2 High Performance Computing System

Li Shen
State Key Laboratory of High Performance Computing
P. R. China

B.1 Introduction

In June 2013, Tianhe-2 (TH-2) topped the TOP500 list of fastest supercomputers in the world. The computer beat out second place finisher Titan by nearly a 2-to-1 margin. Titan achieved 17.59 petaflops, while Tianhe-2 achieved 33.86 petaflops. In November 2013, Tianhe-2 remains the world's most powerful system.

TH-2 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. The complete system has a theoretical peak performance of 54.9 Pflop/s.

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, and the OpenMC programming model.

B.2 Compute Node

Each compute node is composed of 2 Intel Ivy Bridge sockets and 3 Intel Xeon Phi boards (Figure B.1). The system is built out of the nodes and is composed as follows: 2 nodes per blade, 16 blades per frame, 4 frames per rack, and 125 racks make up the system. In Figure B.2, the compute blade has two compute nodes and is composed of two halves: the CPM module and the APM module. The CPM portion of the compute blade contains the 4 Ivy Bridge processors, memory, and 1 Xeon Phi board and the APU half contains the 5 Xeon Phi boards. 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. There is also a PCI-E connection to the NIC.

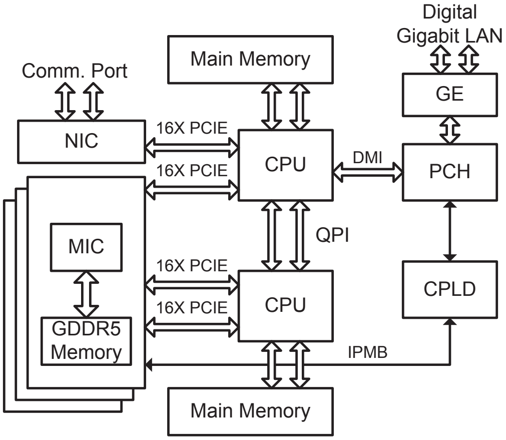

FIGURE B.1 TH-2 Compute Node. However, in implementation, the layout of computation units (i.e. Ivy Bridge CPUs and Xeon MICs) is not symmetrical, which can be seen in Figure B.2.

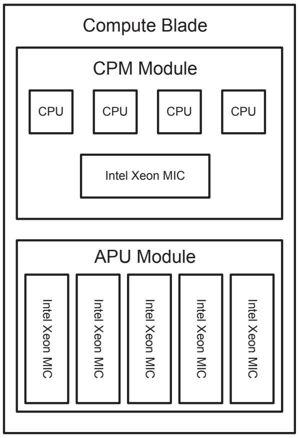

FIGURE B.2 TH-2 Compute Blade Each TH-2 compute blade has one CPM module which has four Intel Ivy Bridge CPUs and one Intel Xeon MIC, and one APU module which has five Intel Xeon MICs. Its theoretical peak performance is 6.862 Tflops.

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.

The Xeon Phi's used in the TH-2 each have 57 cores. 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*0.2112 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.

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.

B.3 The Frontend Processors

In addition to the compute nodes there is a frontend system composed of 4096 Galaxy FT-1500 CPUs (see Figure B.3). These processors were designed and developed at NUDT. They are not considered 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.

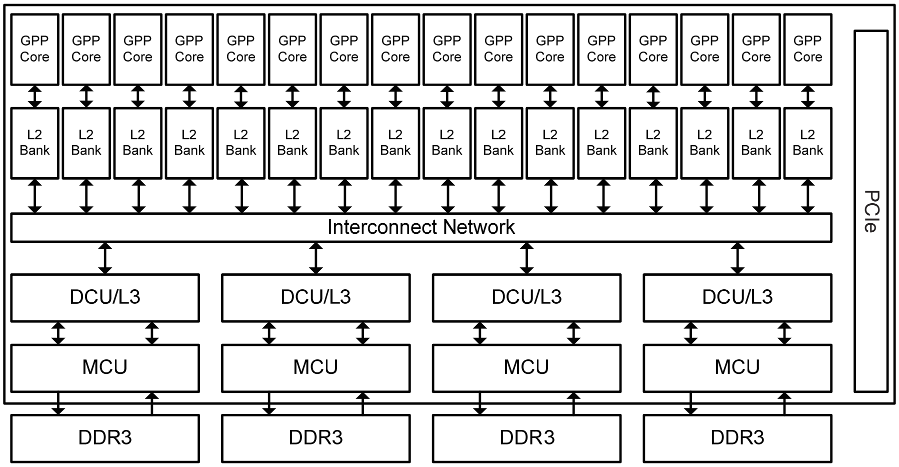

FIGURE B.3 Galaxy FT-1500 CPU. FT-1500 is a designed and produced in China. It is the third generation CPU in FeiTeng (FT) family.

B.4 The Interconnect

NUDT has built their own proprietary interconnect called the TH Express-2 interconnect network (Figure B.4). 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 μs.

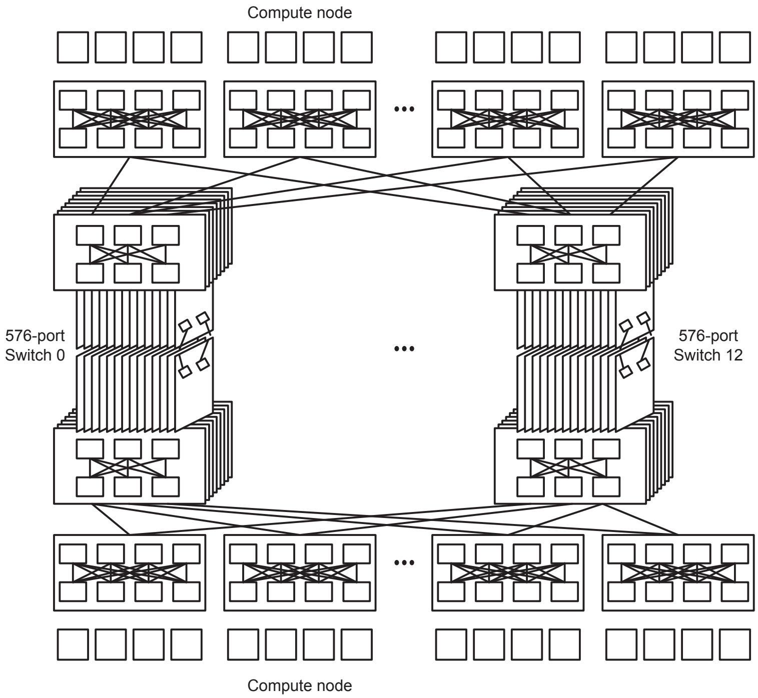

FIGURE B.4 TH Express-2 interconnection network. TH-Express 2 installed in TH-2 tries to avoid communications bottlenecks via its bidirectional bandwidth of 16GB/s, low latency and fat tree topology.

B.5 The Software Stack

The TH-2 uses Kylin Linux as the operating system. Kylin is an operating system developed by the National University for Defense Technology. (See http://en.wikipedia.org/wiki/Kylin_(operating_system) for additional 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 policies. 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 multichannel message data transfers, dynamic flow control and have offload collective operations. In addition, they 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 is based on Intel's MKL 11.0.0 and BLAS for the GPU based on Xeon Phi and optimization by the NUDT.

B.6 LINPACK Benchmark Run (HPL)

The fastest result shown in Figure B.5 was only using 90% of the machine. They are expecting to 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. Gflops/Watt efficiency of the Top 5 systems on the Top 500 list (November 2012) are 2.143 (Titan), 2.069 (Sequoia), 0.830 (K), 2.069 (Mira) and 2.102 (JUQUEEN), respectively.

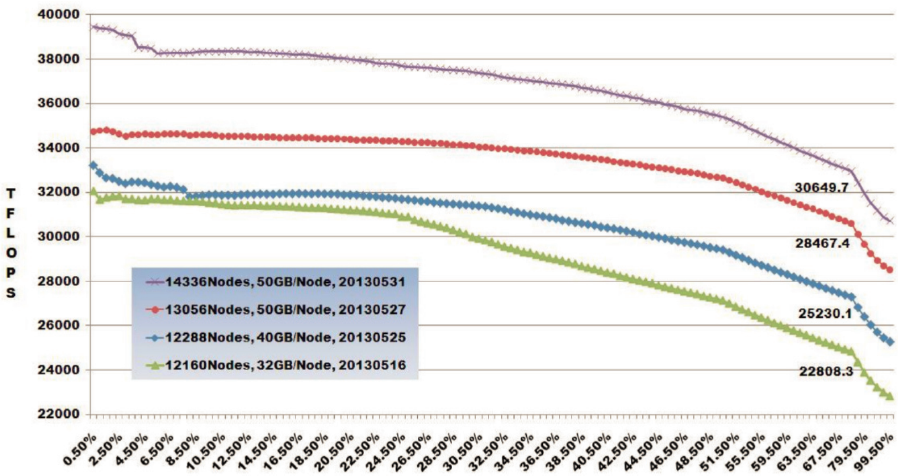

FIGURE B.5 HPL Perfomance. The purple line shows a run of HPL benchmark using 14,336 nodes. It was made using 50 GB of the memory of each node and achieved 30.65 Pflop/s out of a theoretical peak of 49.19 Pflop/s or an efficiency of 62.3% of theoretical peak performance taking a little over 5 hours to complete.

B.7 Concluding Remarks

The TH-2 system will be in the National Supercomputer Center in Guangzhou (NSCC-GZ). It will provide an open platform for research and education and provide high performance computing service for southern China.

Many-core coprocessors such as Intel Xeon MIC and NVidia GPGPU are widely used in current supercomputer systems. Other than TH-2, three systems using many-core coprocessors entered the Top 10 on the newest Top 500 List (November, 2013) also.

The theoretical peak performance of a computer system can be computed simply by accumulating the performance of all its processing units. However, its maintain performance is usually obtained via measuring the running of typical benchmark. For example, HPL is the most popular benchmark for supercomputer system performance evaluation.