Yang XJ, Liao XK, Lu K *et al.* The TianHe-1A supercomputer: Its hardware and software. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 26(3): 344–351 May 2011. DOI 10.1007/s11390-011-1137-4

The TianHe-1A Supercomputer: Its Hardware and Software

Xue-Jun Yang (杨学军), *Senior Member*, *CCF*, *Member*, *ACM*, *IEEE*
Xiang-Ke Liao (廖湘科), *Senior Member CCF*, *Member*, *ACM*, Kai Lu (卢　凯), Qing-Feng Hu (胡庆丰)
Jun-Qiang Song (宋君强), and Jin-Shu Su (苏金树), *Senior Member*, *CCF*, *Member*, *IEEE*

School of Computer, National University of Defense Technology, Changsha 410073, China

E-mail: {xjyang, xkliao, kailu, qingfenghu, junqiangsong, jinshusu}@nudt.edu.cn

Received January 10, 2011; revised March 4, 2011.

Abstract This paper presents an overview of TianHe-1A (TH-1A) supercomputer, which is built by National University of Defense Technology of China (NUDT). TH-1A adopts a hybrid architecture by integrating CPUs and GPUs, and its interconnect network is a proprietary high-speed communication network. The theoretical peak performance of TH-1A is 4700 TFlops, and its LINPACK test result is 2566 TFlops. It was ranked the No. 1 on the TOP500 List released in November, 2010. TH-1A is now deployed in National Supercomputer Center in Tianjin and provides high performance computing services. TH-1A has played an important role in many applications, such as oil exploration, weather forecast, bio-medical research.

Keywords TianHe-1A supercomputer, hybrid architecture, Kylin operating system, power computing

1 Introduction

TianHe-1A (TH-1A for short) supercomputer is developed by National University of Defense Technology (NUDT). Its theoretical peak performance is 4.7 Peta-flops, and maximal Linpack performance achieves 2.566 Peta-flops. TH-1A is currently the fastest supercomputer in the world according to the latest TOP500 List[1] issued in November, 2010.

The research and development of TianHe-1 series system was launched in 2005 and has three milestones. The preliminary research was started in 2005. We focused on investigating a new high efficient computing architecture based on stream processing technology[2]. The first stream processor named FT-64 was designed and tested in 2006, and its prototype system of 1024 nodes was evaluated in 2007, using mixed general processors and FT-64 stream processors.

Based on the former evaluation, the first Peta-flops hybrid system TH-1 was accomplished in 2009. TH-1 system has 6250 nodes connected by DDR Infiniband. Each node has two Xeon processors and one ATI's GPGPU. The theoretical peak performance of TH-1 was 1206 TFlops and its Linpack test result reached 563.1 Tflops[3].

To address the requirement of NSCC-TJ (National Supercomputer Center in TianJin), the TH-1 system was upgraded and enhanced in August, 2010, named TH-1A. The TH-1A system is constructed by 7168 compute nodes, each of which has two general processors and one stream processor. Besides the promotion of performance, TH-1A system adopts a proprietary interconnect network with higher bandwidth than Infiniband QDR, and is partly equipped with FeiTeng-1000 (FT-1000 for short) CPUs designed by NUDT.

The requirements and challenges when designing TH-1A system are as follows:

1) The power consumption becomes the major bottleneck for supercomputers[4-5]. We need to achieve high power efficiency in the architecture design for such a large-scale system.

2) To match the enhancement of node performance and system scale, the communication bandwidth and network switch capability of TH-1A should be increased accordingly.

3) Services on demand and data security are very important for current public information infrastructure. We should take these requirements into account in system design.

The remainder of this paper is organized as follows. In Section 2, we describe the rationale of TH-1A architecture including its design principles and subsystem overview. In Section 3, we detail the system hardware including compute node, NIC, and NRC. In Section 4, we describe the system software including operating system, compiler system, and parallel developing

Regular Paper
Supported by the National High Technology Research and Development 863 Program of China under Grant No. 2009AA01A128.

environment. In Section 5, we present the mechanism to tune or reduce power consumption. Finally, we conclude this paper in Section 6.

2 System Outline

TH-1A (Massively Parallel Processing) is a hybrid MPP system with CPUs and GPUs. The hardware of TH-1A system consists of five subsystems: service subsystem, compute subsystem, communication subsystem, I/O storage subsystem, monitoring and diagnostic subsystem. Its software system is composed of the operating system, compiler system, and parallel developing environment.

There are 140 racks in the whole TH-1A system, including 112 compute racks, 8 service racks, 6 communication racks, and 14 I/O racks. The entire system occupies 700 m^2. TH-1A includes 7168 compute nodes and 1024 service nodes. Each compute node is configured with two Intel CPUs and one NVIDIA GPU. The power efficiency of each compute node is 785.7 MFlops/W. Each service node has two FT-1000 CPUs. There are totally 23 552 microprocessors, including 14 336 Intel Xeon X5670 CPUs, 2048 FT-1000 CPUs, and 7168 NVIDIA M2050 GPUs. The total memory of the system is 262 TB, and the disk capacity is 2 PB. The power consumption at full load is 4.04 MW, and the power efficiency is about 635.1 MFlops/W[6].

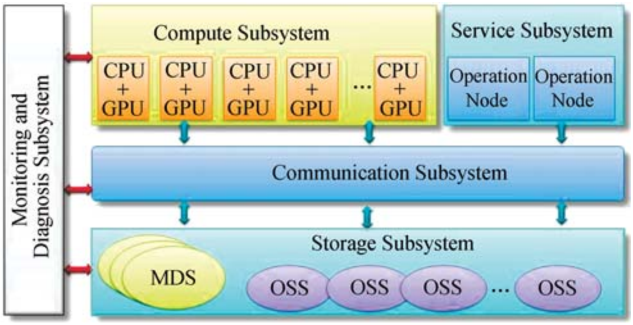

Fig.1. Structure of TH-1A system.

Fig.1 illustrates the structure of TH-1A system. All the subsystems are connected through two networks, a fat tree network for the data exchange among all the nodes and a Giga-Ethernet for monitoring all the parts of TH-1A system. The fat tree network is a proprietary high speed interconnect network, which connects all the compute nodes, service nodes, I/O management nodes, and I/O storage nodes with low latency and high bandwidth. The monitoring network provides functions of monitoring and diagnosing for the entire system.

The fat tree communication network is constructed by high-radix Network Routing Chips (NRC) and high-speed Network Interface Chips (NIC). Both NRC and NIC chips are designed by NUDT. The interconnect topology is an optic-electronic hybrid fat-tree structure with the bi-directional bandwidth of 160 Gbps and the latency of 1.57 μs.

The monitoring and diagnostic network is an Ethernet which connects all the nodes of TH-1A and the monitoring severs. It can monitor, control, diagnose, and debug the entire system in a real-time way.

The I/O storage system of TH-1A uses the Luster file system. It has 6 I/O management nodes, 128 I/O storage nodes, and 2 PB storage capacity.

Besides, the infrastructure design of TH-1A system is a good example of high power efficiency and low cost. TH-1A uses a high-density assembling technique, in which every eight mainboards can be plugged in a double-side backplane. In each rack, there are 128 XONE processors and 64 GPGPUs. The peak performance of one rack is about 42 TFlops. Compared with air cooling system in TH-1, the cooling mechanism used in TH-1A system is closed air cooling. Two liquid cooling air-conditioners are listed in one rack, the cooling air cycles in the rack to dissipate heat loads from chipsets. With these techniques, the power consumption of a compute rack is about 35 KW per rack, thus the cost and complexity of TH-1A racks are acceptable.

The parallel software stack of TH-1A includes operating system, compiler system, and parallel developing environment. The operating system of TH-1A is 64-bit Kylin Linux. Kylin Linux is designed for high-performance parallel computing optimization, which supports power management and high-performance virtual zone. Based on the high performance virtual zone technique, TH-1A can construct customized virtual running environments for various users, and enable data isolation between zones. The Kylin Linux operating system supports a broad range of the third-party application software.

The compiler system supports C, C++, Fortran and CUDA languages[7], as well as MPI and OMP parallel program model. In order to develop applications on hybrid system efficiently, TH-1A introduces a parallel heterogeneous programming framework to abstract the programming model on GPU and CPU.

The parallel application developing environment provides a component-based network developing platform to program, compile, debug, and submit jobs through the network. Besides the tools provided by TH-1A, users can integrate different tools into this platform dynamically, such as Intel Vtune[8] and TotalView[9].

3 Hardware System

3.1 Chipsets

The key technical breakthroughs of TH-1A supercomputer lie in three VLSI chips: FT-1000 CPU, high-radix routing chip NRC and high-speed networking interface chip NIC.

3.1.1 FT-1000 CPU

FT-1000 CPU adopts the Parallel System on Chip (PSoC) multi-core architecture. Eight multi-thread cores are integrated on the chip. Each core can execute eight independent threads. The instruction set of FT-1000 is compatible with SPARC V9. The core works at 1 GHz frequency and the memory access frequency is 400 MHz. A large capacity multi-bank L2 cache is shared by the cores via crossbar switches. The L2 cache can access the off-chip high speed DDR3 DRAM via four on-chip memory controller units (MCU). The coherence inter-chip direct connect interface is integrated on the chip, which can be used to construct an SMP system with multiple processors. The chip provides efficient I/O access supports by integrated PCIE 2.0 standard interfaces. The architecture of FT-1000 CPU is illustrated in Fig.2.

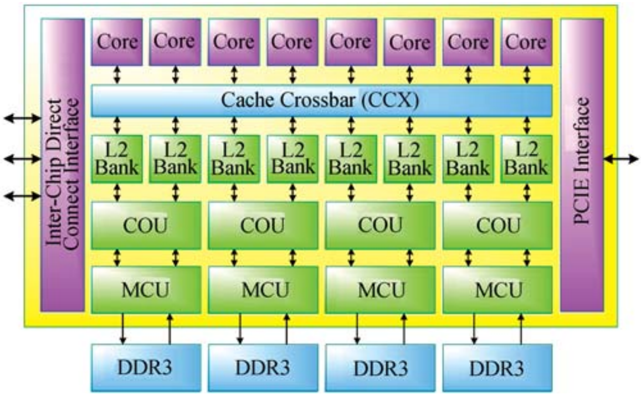

Fig.2. Architecture of FT-1000.

3.1.2 High-Radix Routing Chip (NRC)

In order to reduce the communication latency in largescale interconnect network, the high-radix routing technique is proposed for the routing chip. The high-radix architecture is able to degrade the number of the maximal interconnect hops, thus improve the bandwidth, reliability, and flexibility of the interconnect communication system. As shown in Fig.3, each NRC chip consists of 16 symmetric interconnect switch interfaces. The NRC uses a tile-based switching structure to simplify the difficulty of designing a 16 × 16 crossbar. In addition, smaller crossbar design can localize the communication, and avoid long-distance transmissions. The transmission rate of each serial link in NRC is 10 Gbps, and the bi-directional bandwidth of each port is 20 GB/s.

For fault tolerance, the NRC supports automatic frequency adjusting operation. The links of NRC support working at full-speed mode or half-speed mode, and can negotiate the bit-width statically or dynamically. Based on these functions, the NRC is able to isolate the failure links and improve the reliability and usability of the interconnect network.

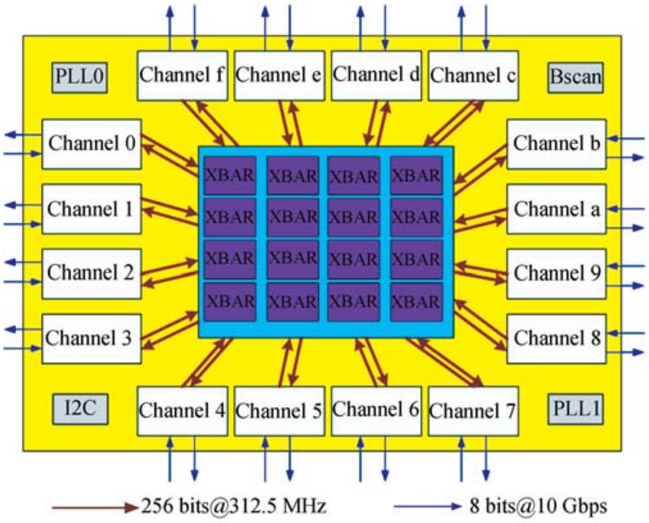

Fig.3. Architecture of NRC.

3.1.3 High-Speed Network Interface Chip (NIC)

The network interface chip named NIC is used to transmit data between compute nodes. As shown in Fig.4, NIC integrates a 16-lane PCIE 2.0 interface, and connects with the high-radix router via 8 × 10 Gbps links. The bi-directional bandwidth is up to 20 GB/s.

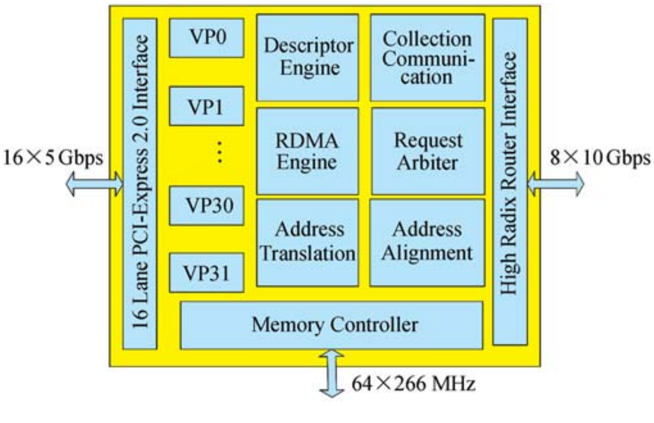

Fig.4. Architecture of NIC.

The NIC supports both remote direct memory access (RDMA) for long message and packet transfer for short message. Communications in NIC are controlled by descriptors, which is an ID of a message. Since long message is bandwidth-sensitive and short message is latency-sensitive, we adaptively deal with these two types of messages. For short message, data is packed in the descriptor. For long message, the descriptor contains the buffer address and length of the data. Descriptors are queued in descriptor engine and processed by NIC automatically, thus the communication operations can be overlapped with computing.

Each NIC has 32 virtual ports (VPs) to provide deadlock-free communication and increase throughput. VP is an abstract structure of NIC hardware resources, which enables several processes to use NIC simultaneously in user space. Schedule and arbitration between VP are implemented by NIC hardware, and communication operations can bypass the operating system, thus the software overhead is reduced significantly.

The NIC also implements virtual-physical address translation from user space to kernel space. This translation is byte-aligned, therefore data transfer between user space and destination buffer can use RDMA directly without any copy. Besides, this translation can avoid the software processing overhead.

Moreover, the NIC provides support for optimizing collective communication.

3.1.4 Fabrication

The FT-1000 chip adopts 65 nm CMOS fabrication and FCBGA package. Its substrate contains 0.35 billion transistors and 1517 pins. Both NRC and NIC adopt the 90 nm CMOS fabrication and array flip-chip package. The substrate of the NRC contains 0.46 billion transistors and 2577 pins, while the substrate of the NIC contains 0.15 billion transistors and 657 pins.

3.2 Compute Node

TH-1A consists of 7168 compute nodes. Each compute node is configured with two Intel CPUs and one NVIDIA GPU. The CPU is Xeon X5670 (2.93 GHz, six-core) and the GPU is Tesla M2050 (1.15 GHz, 14 cores/448 CUDA cores)[10]. Each compute node has 655.64 GFlops peak computing performance (i.e., CPU has 140.64 GFlops and GPU has 515 GFlops) and 32 GB total memory.

The peak double precision performance of each NVIDIA Tesla M2050 GPU is 515 GFlops. The M2050 GPU integrates 3 GB GDDR5 memory, with the bit width of 384 bits and peak bandwidth of 148 GB/s. All the register files, caches and specific memories support ECC, which improves the reliability of GPU computation.

The responsibilities of CPU in compute node are running operation system, managing system resources, and executing general purpose computation. GPU mainly performs large scale parallel computation. By cooperating CPU and GPU, the compute nodes could effectively accelerate many typical parallel applications, for example, sparse-matrix computation.

3.3 Interconnect Network

Interconnect network is the core of the massive parallel processing system. TH-1A consists of 8192 compute nodes and service nodes. These nodes are installed in 120 racks. The topological structure of the TH-1A interconnect network is a hierarchical fat tree, as shown in Fig.5.

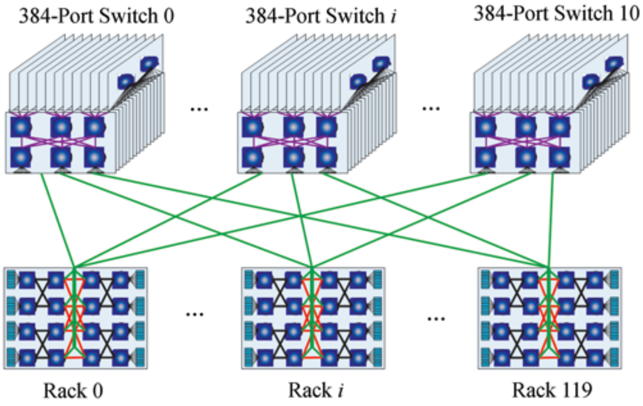

Fig.5. Architecture of the interconnect network.

The first layer consists of 480 switching boards. In each rack, every 16 nodes connected with each other through the switch board in the rack. The main boards of the nodes are connected with the switch board via a back board. The communication on switching boards uses electrical transmission.

The second layer contains 11 384-port switches, connected with QSFP optical fibers. There are 12 leaf switch boards and 12 root switch boards in each 384-port switch. A high density back board connects them in an orthotropic way. The switch boards in the rack are connected to 384 port switches through optical fibers.

For the requirement of high performance communication between kernel modules or user processes, such as parallel applications, global parallel file system and resource management system, we implement Galaxy Express (GLEX) communicating system. GLEX uses user-level communication technique. Based on the NIC virtual port, GLEX encapsulates the communication interfaces of NIC, and provides both user space and kernel space programming interfaces which can fulfill the function and performance requirement from other software modules.

GLEX consists of kernel module for managing NIC and providing kernel programming interfaces, network driver module for supporting TCP/IP protocol, user-level communication library for providing user space programming interfaces, collective communication library for optimizing collective operations, and management tools for system administration and maintenance.

Based on the fault tolerance supported in NRC and NIC, GLEX implements reduced communication protocol, which provides zero-copy RDMA transfer between user space and data buffers. By utilizing virtual memory protection and memory mapping in CPU, GLEX permits several processes to communicate simultaneously and securely in user space and bypass the interference of operating system kernel and data copy in communication path.

The interconnect network supports the collective operations, such as multicast and broadcast collectives, based on the automatic message switching mechanism implemented by hardware.

4 Software System

4.1 Kylin Linux Operating System

The operating system of TH-1A supercomputer is Kylin Linux. According to the architecture of TH-1A and user demands, the kernel of the Kylin Linux is optimized to support multi-core and multi-thread processors, heterogeneous computing synchronization, power management, system fault tolerance, and data security protection, which provides users an execution supporting environment with high efficiency, reliability, security, and usability.

To improve the usability and security of the system, Kylin Linux utilizes virtualization techniques, i.e., the High Performance User Container (HPUC), which supports dynamic environment customization. Fig.6 shows the architecture of HPUC, which consists of three parts, virtual computing zone on the service nodes, high performance computing zone on the compute nodes and HPUC task management.

The virtual computing zone on the service nodes uses the high performance virtual technique to support multiple users to run in independent runtime environments on a single OS simultaneously. To reduce performance loss, the HPUC uses light weighted security file system to reduce the overhead of constructing multiple running environments, and localize the file position to reduce the overhead of remote file access. The task management system oriented virtual zone is the linkage between the virtual computing zone on the service nodes and the high performance computing zone on the compute nodes, which couple the virtual zones on compute nodes and service nodes according to the configuration of user runtime environment.

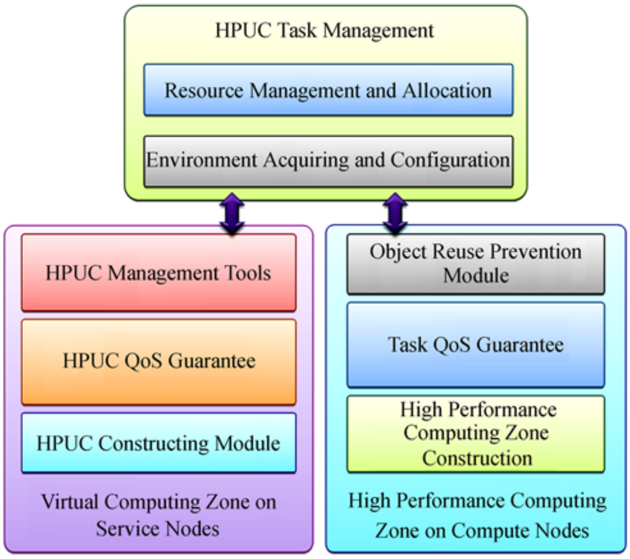
HPUC Task Management
Resource Management and Allocation
Environment Acquiring and Configuration
HPUC Management Tools
HPUC QoS Guarantee
HPUC Constructing Module
Virtual Computing Zone on Service Nodes
Object Reuse Prevention Module
Task QoS Guarantee
High Performance Computing Zone Construction
High Performance Computing Zone on Compute Nodes

Fig.6. Architecture of HPUC.

To facilitate users, HPUC in TH-1A system is divided into two-levels, the common HPUC (CUC) and the specialized HPUC (SUC). The CUC environment is constructed during system initialization. It is the main working environment for most common users, that provides basic tools for application programming, compiling, and debugging, as well as task submitting. The SUC environment is designed for the users with special demands. When constructing an SUC, the system manager provides the users a basic SUC template. Based on this template, the users can construct their own running environment according to their own demands in this SUC.

The resource management of OS is in charge of global resource management, partition management, resource allocation and task scheduling. The resource management of TH-1A provides several resource allocating strategies for different types of jobs. Based on these custom resource schedule algorithms, TH-1A system can improve the resource utilization and throughput of the system efficiently.

4.2 Compiler System

The compiler system of TH-1A supercomputer supports serial programming languages such as C/C++, Fortran77/90/95, Java, and the parallel programming languages such as OpenMP, MPI, and OpenMP/MPI. TH-1A uses OpenCL and CUDA for GPU programming.

In order to use CPU and GPU cooperatively, and to improve the performance, some techniques are presented in TH-1A. For the inter-node computing, homogeneous parallel programming is adopted. The optimizations of zero-copy communication and dynamic load balancing are exploited to improve the performance of the inter-node computing. The zero-copy optimization can improve the communication performance by about 40%. The intra-node computing uses heterogeneous parallel programming to dig the performance of CPU and GPU sufficiently. Many optimizations are proposed during the development of TH-1A. For example, the adaptive partitioning method is put forward to balance the workload between CPU and GPU; the asynchronous data transfer/computing method is presented to overlap CPU operations with GPU operations; the software pipelining method is proposed to overlap GPU computing with the data transfer between host and GPU device memory.

In order to simplify the programming on the hybrid system, TH-1A introduces a heterogeneous programming infrastructure named TianHe Hybrid Programming Infrastructure (TH-HPI). Aiming at the problems of large scale heterogeneous parallel programming, such as program segmentation, data distribution, processes synchronization, load balancing, performance optimization, TH-HPI uses three layers: parallel system supporting layer, common parallel algorithm layer and application interface layer.

The parallel system supporting layer provides the packages of managing patch data structure to divide user program area into many data patches. Because of these packages, the details of parallel computation and the communication between nodes can be hidden from users. Users only need to emphasize on inner-patch calculations.

The common parallel algorithm layer supplies common solvers which can be executed on CPUs or GPUs efficiently.

The application interface layer provides developing interfaces for user applications. These interfaces guide user to make full use of the infrastructure. They also guide user to parallelize the calculation kernel by creating nested threading model to control the cores of CPUs and GPUs on each node.

Based on this three-layer structure, TH-HPI reduces the difficulty of heterogeneous parallel programming, and improves the application developing efficiency of heterogeneous parallel computers.

4.3 Parallel Developing Environment

TH-lA provides a flexible parallel developing environment for the users on network, called Flexible Service Environment (FSE). By integrating techniques such as visualization, object orientation, and working flow, TH-1A has implemented a new high performance computing environment. This environment has changed the traditional way of using high performance computer and enhanced the security and usability of the supercomputer.

The FSE includes parallel task scheduler, network integrated development platform and parallel application tool suite.

Through the parallel task scheduler, users only need to tell their demands on the parallel processing capability. There is no need to care about which service node or computing/storage array provides the service. In other words, the parallel task scheduling mechanism implements the goal of providing high performance computing capacity as services.

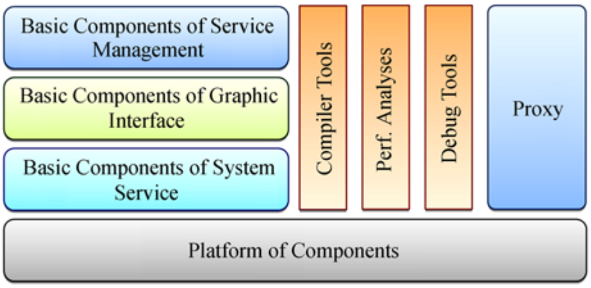
Basic Components of Service Management
Basic Components of Graphic Interface
Basic Components of System Service
Compiler Tools
Perf. Analyses
Debug Tools
Proxy
Platform of Components

Fig.7. Architecture of network integrated development platform.

Fig.7 shows the architecture of network integrated development platform. The network integrated development platform uses component structure, and communicates with the service proxies on remote service nodes through parallel service interface. Based on this platform, users can access and use the resources and services of TH-1A easily. The platform is based on Eclipse. It integrates multiple parallel programming tools in an Eclipse+plugin fashion and fully supports program editing, compiling, linking, running, debugging, as well as performance optimizing. The integrated development platform provides a unified programming interface and greatly simplifies the interaction between users and various tools of the supercomputer, thus remarkably improves the efficiency of parallel application development. The development framework runs on the service nodes and manipulates the resource management system to load parallel tasks as well as collect task and system state information.

The parallel application tool suite mainly consists of a parallel debugger and a parallel performance analysis tool. The parallel debugger can show the visualized executing and controlling information for the parallel applications, and provides visualized source-code level debug function for MPI parallel applications. The performance analysis tool provides profiling and event-tracing-based performance analysis. It helps users to locate the performance bottleneck and hotspot code regions of parallel applications. The analysis result of the tool is very useful for users to perform optimizations on their parallel applications.

Based on the FSE, login, programming and compiling services are distributed among multiple service nodes through a load-balance scheduler. FSE provides users a single system image environment via the HPUC technique. The single system image service of TH-1A system promises that users can see the same environment and data of the system at any service node. The load-balance service monitors the load of the service nodes, and guarantees that a new login request is distributed to the service node of light load.

5 Power-Aware Computing

Power consumption is a key issue in large scale supercomputers, which may impact the system usability. To save power consumption during system idle time, TH-1A system develops a software-hardware integrated power control frame, as depicted in Fig.8.

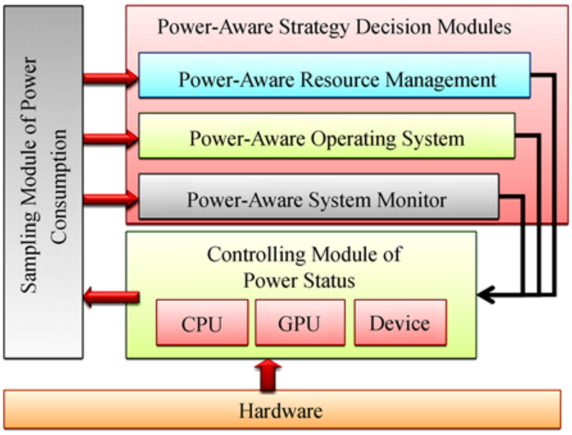
Sampling Module of Power Consumption
Power-Aware Strategy Decision Modules
Power-Aware Resource Management
Power-Aware Operating System
Power-Aware System Monitor
Controlling Module of Power Status
CPU
GPU
Device
Hardware

Fig.8. Frame of power control in TH-1A.

The power control is implemented on three layers, including power status sampling module, power-aware strategy decision modules and power status control module. The sampling module of power consumption collects the information of power consumption for different parts of TH-1A. The information includes the usability of CPU and GPU, temperature of racks and so on. Based on this information, the decision modules will control the implementation module to turn the CPU, GPU and other devices into suitable power status. To control the power consumption, the system uses ACPI compatible board-level design, and implements a feedback cooling control based on the temperature of the compute nodes.

Based on these techniques, the power consumption of the system fans could be adaptively reduced when the system temperature is low, while the ventilation and cooling of the system are effective when the system load is heavy.

The operating system supports dynamic frequency scaling for processors. By integrating CPU and main memory frequency control into proc file system, users can configure the frequency of processor and main memory through the interface of proc file system.

The resource management system can get the state information of all jobs in queue. TH-1A system implemented an adaptive power status switch algorithm for compute nodes according to job status, which can switch an idle node into sleep mode dynamically. The evaluation shows that in sleep mode, the compute node can save power by 90%.

6 Conclusions

TH-1A supercomputer adopts the hybrid architecture of heterogeneous integration of CPUs and GPUs. The communication network of TH-1A is proprietary high-speed network designed by NUDT. Many optimizations are adopted to improve the computing efficiency and reduce power consumption for TH-1A. Benchmark test and parallel application execution have shown that TH-1A exploits a new way to build supercomputers with high power efficiency and high performance price ratio.

TH-1A system has been installed to provide services at NSCC-TJ, and successfully applied to many fields, such as oil exploration, high-end equipment development, bio-medical research, animation design, exploitation of new energy sources, research of new materials, engineering design and simulation analysis, weather forecast, remote sensing data processing, and financial risk analysis.

The next generation of TH system is under design now. The new TH system will use domestic FT-2000 processors. Compared with TH-1A, the new system ill make breakthroughs in the power efficiency and system autonomy for the increasing application demands.

References

[1] http://www.top500.org/lists/2010/11, Dec. 1, 2010.

[2] Yang X, Yan X, Xing Z, Deng Y, Jiang J, Zhang Y. A 64-bit stream processor architecture for scientific applications. In *Proc. ISCA 2007*, San Diego, USA, June 9-13, 2007, pp.210-219.

[3] http://www.top500.org/lists/2009/11, Dec. 1, 2010.

[4] Rountree B, Lowenthal D K. Bounding energy consumption in largescale MPI programs. In *Proc. SC 2007*, Nevada, USA, Nov. 10-16, 2007, pp.1-9.

[5] A Berl, E Gelenbe, M Di Girolamo, G Giuliani, H De Meer, M Dang, K Pentikousis. Energy-efficient cloud computing. *The Computer Journal*, 2009, 53(7): 1045-1051.

[6] http://www.green500.org/lists/2010/11/top/list.php?from=1&to=100, Dec. 1, 2010.

[7] Kirk D. NVIDIA CUDA software and GPU parallel computing architecture. In *Proc. ISMM 2007*, Montreal, Canada, Oct. 21-22, 2007, pp.103-104.

[8] http://software.intel.com/en-us/articles/intel-vtune-amplifier-xe/, Dec. 1, 2010.
[9] http://www.totalviewtech.com/home/, Dec. 1, 2010.
[10] http://www.nvidia.com/docs/IO/43395/NV_DS_Tesla_M2050_M2070_Apr10_LowRes.pdf, Dec. 1, 2010.

Xue-Jun Yang received his Ph.D. degree from National University of Defense Technology, China in 1999. Currently he is a professor at the university. His interests include parallel processing and system software. He is the chief designer of TH-1A supercomputer system.

Xiang-Ke Liao received his Master's degree from National University of Defense Technology, China in 1999. Currently he is a professor at the university. His interests include parallel system architecture and system software. He is the associate chief designer of TH-1A supercomputer system.

Kai Lu received his Ph.D. degree from National University of Defense Technology, China in 1999. Currently he is a professor at the university. His interests include parallel processing and system software. He is the associate chief designer of TH-1A supercomputer system.

Qing-Feng Hu received his Master's degree from National University of Defense Technology, Hunan, China in 1987. Currently he is a professor at the university. His interests include parallel algorithm and parallel application. He is the associate chief designer of TH-1A supercomputer system.

Jun-Qiang Song received his Master's degree from National University of Defense Technology, China in 1986. Currently he is a professor at the university. His research interests include parallel algorithm and parallel application. He is the associate chief designer of TH-1A supercomputer system.

Jin-Shu Su received his Ph.D. degree from National University of Defense Technology, China in 2000. Currently he is a professor at the university. His research interests include parallel processing and high speed network. He is the associate chief designer of TH-1A supercomputer system.