MegaProto/E: Power-Aware High-Performance Cluster with Commodity Technology

Taisuke Boku* Mitsuhisa Sato* Daisuke Takahashi* Hiroshi Nakashima†
Hiroshi Nakamura‡ Satoshi Matsuoka§ Yoshihiko Hotta*

* University of Tsukuba, {taisuke,msato,daisuke,hotta}@hpcs.cs.tsukuba.ac.jp
† Toyohashi University of Technology, nakasima@tutics.tut.ac.jp
‡ University of Tokyo, nakamura@hal.rcast.u-tokyo.ac.jp
§ Tokyo Institute of Technology, matsu@is.titech.ac.jp

Abstract

In our research project named "Mega-Scale Computing Based on Low-Power Technology and Workload Modeling", we have been developing a prototype cluster not based on ASIC or FPGA but instead only using commodity technology. Its packaging is extremely compact and dense, and its performance/power ratio is very high. Our previous prototype system named "MegaProto" demonstrated that one cluster unit, which consists of 16 commodity low-power processors, can be successfully implemented on just 1U height chassis and it is capable of up to 2.8 times higher performance/power ratio than ordinary high-performance dual-Xeon 1U server units.

We have improved MegaProto by replacing the CPU and enhancing the I/O performance. The new cluster unit named "MegaProto/E" with 16 Transmeta Efficeon processors achieves 32 GFlops of peak performance, which is 2.2-fold greater than that of the original one. The cluster unit is equipped with an independent dual network of Gigabit Ethernet, including dual 24-port switches. The maximum power consumption of the cluster unit is 320 W, which is comparable with that of today's high-end PC servers for high performance clusters.

Performance evaluation using NPB kernels and HPL shows that the performance of MegaProto/E exceeds that of a dual-Xeon server in all the benchmarks, and its performance ratio ranges from 1.3 to 3.7. These results reveal that our solution of implementing a number of ultra low-power processors in compact packaging is an excellent way to achieve extremely high performance in applications with a certain degree of parallelism. We are now building a multi-unit cluster with 128 CPUs (8 units) to prove that this advantage still holds with higher scalability.

1. Introduction

The PC cluster solution continues to be a highly attractive method for achieving high performance computing (HPC) with its high performance/cost ratio supported by commodity technology. However, there is concern about its applicability to highly scalable systems with tens of thousands or millions of processors. In order to allay this concern, we at least have to solve the problems associated with power consumption, space and dependability. Our research project named "Mega-Scale Computing Based on Low-Power Technology and Workload Modeling" aims to establish fundamental technologies for this purpose. Our research investigates the feasibility and dependability of million-scale parallel systems as well as the programmability on such a large scale.

For the feasibility study, we especially focused on the performance/power/space ratio problem. To this end, we have developed the prototype system based on commodity-only technology, that is, we only used commodity processors and network elements. In order to achieve a high density implementation we developed cluster chassis units which can house a number of processors in a small space. The recent trend of dual-core CPUs such as Pentium-D or Opteron-D clearly demonstrates that the best way to improve total performance is to introduce multi-processors having relatively low-performances, instead of increasing the CPU clock frequency on a single processor. Based on this concept, an ideal platform is a cluster of ultra low-power processors implemented in a small space with high density.

Green Destiny[1] is a successful example of the above concept. While it consists of a commercial blade-style processor card, we have designed and implemented a higher density collection of processors in a 1U height chassis containing 17 Transmeta Crusoe processors. This prototype

1-4244-0054-6/06/$20.00

unit was named "MegaProto"[2, 6] which provides 14.9 Gflops of peak performance. When we designed the first prototype of MegaProto, we also intended to build an enhanced version incorporating more powerful processor and I/O bus. This enhanced version has now been completed with the name of "MegaProto/E" ('E' stands for Efficeon, the name of the new processor). In this paper, we describe the design, implementation and performance evaluation of the MegaProto/E cluster unit.

The rest of the paper consists of the following. Section 2 gives an overview of our Mega-Scale computing project along with the conceptual design of the MegaProto series cluster unit. In Section 3, we describe in detail the design and implementation of MegaProto/E. After describing the performance evaluation in Section 4, the cluster design with higher scalability based on MegaProto/E is described in Section 5. Finally, conclusions and future works are given in Section 6.

2. Mega-Scale Project and MegaProto Cluster Unit

The overview of our Mega-Scale Computing Project and the conceptual design were described in more detail in [2]. In this section, we give just a brief description of them.

2.1. Overview of the Mega-Scale Computing Project

Today, Peta-Flops computing is not just a dream anymore, and several projects have been launched aiming to achieve this level of computational performance. In these projects, the common key issues include (i) how to implement ultra large-scale systems, (ii) how to reduce the power consumption per Flop, and (iii) how to control ultra large-scale parallelism. From the viewpoint of hardware technology, the first two issues are critical. As demonstrated by BlueGene/L[3], the state-of-the-art MPP today, one promising way to achieve Peta-Flops computing is to build an MPP having very low-power processors and a simple switch-less network to reduce both space and power consumption. However, such a system requires a dedicated hardware platform including specially designed processor chips and network and system racks, which are expensive to implement and require a long time to design and implement.

On the other hand, the rapid progress of computation and communication technologies that are not limited to HPC applications suggests another approach on low power commodity technology, both for the CPU and the network. While an ordinary PC cluster for HPC applications consists of high-performance CPUs consuming tens of Watts and a wide-bandwidth network such as Infiniband[4], we can combine a number of very low-power and medium-speed CPUs designed for laptops and Gigabit Ethernet with a very high performance/cost ratio supported by a non-HPC market.

In our Mega-Scale Computing Project[5], we investigated (i) hardware/software cooperative low-power technology and (ii) workload modeling and model-based management of large scale parallel tasks together with the faults occurring in such a system. Our study covers the processor architecture, compiler, networking, cluster management and programming for a system based on the above concept.

2.2. Conceptual Design of MegaProto

Hereafter, the word "MegaProto" is used to refer to our overall prototyping system used for the feasibility study of our Mega-Scale Computing concept. However, we called our first prototype version by the same name in previous papers[2, 6]. In order to distinguish between the two different versions of the prototype system, we explicitly call the first one "MegaProto/C" and the second one "MegaProto/E". The two suffix letters 'C' and 'E' stand for the code names of the CPUs used, Crusoe and Efficeon, respectively.

In considering the Peta-Flops scale systems based on commodity CPU and network technologies, it is impossible to avoid the issues of power consumption and spacing. We consider the approximate upper limits of such a system as being 10 MW for power consumption and 1,000 racks for space. These specifications are still hard to realize in practice, but are not impossible. With these limitations, our primary goal is to attain 1 TFlops/10 kW/rack as the performance/power/space ratio. MegaProto is a series of prototype systems designed to demonstrate that such systems can be built. Since the system was to be based on commodity system formation, we built it using 19-inch 42U height standard racks. With the exception of the space for the network switch, 32U can be assigned for computation nodes. Thus, the goal was to make a 32 GFlops/310 W/1U building block as the basic unit. Neither the performance or the power consumption criteria can be achieved with today's high-end CPUs such as Intel Xeon, Intel Itanium2 or AMD Opteron, even with multi-way SMP configuration. However, it is possible to satisfy both criteria using state-of-the-art, low power CPUs with DVS and very low voltage drive if we can aggregate 10 to 20 CPUs in a single 1U chassis. Even using blade-style cards, it is impossible to achieve such high densities, but we have finally solved the problem of developing a mother board composed of 16 daughter cards with the size of a one dollar bill.

Another important issue regarding the design of the unit chassis is selecting an interconnection network suitable for such a basic design. If we choose a high-end CPU as the node processor, we will require a high-end network interface with 1 GByte/sec bandwidth or greater for efficient parallel processing. Since we chose a mid-range CPU as the

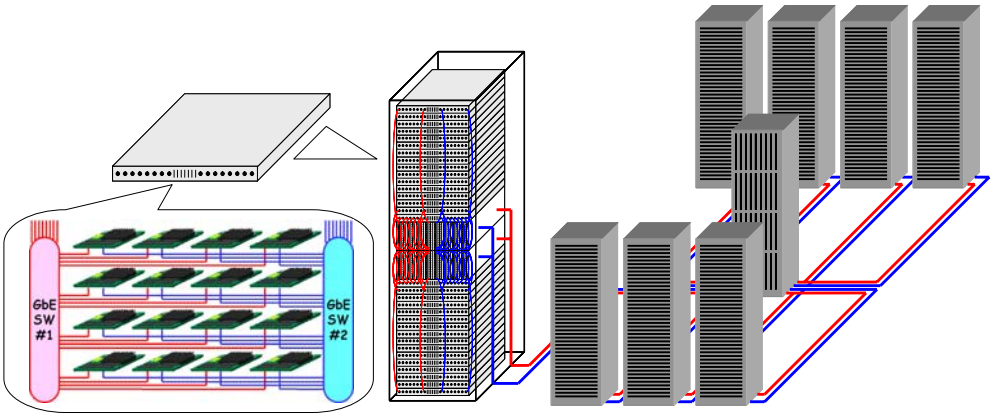

Figure 1. System configuration

node processor, the network bandwidth can be reduced to several hundred MByte/sec. This range can be covered by trunking several channels of a Gigabit Ethernet, and we decided to use dual channels per processing node. It is quite important from the viewpoint of achieving a high performance/cost ratio to introduce a commodity Gigabit Ethernet as the interconnection. Since recent Gigabit Ethernet switching fabric is quite inexpensive and small for 10 to 20 ports of connection, it is possible to implement the basic switch itself on the mother board and to connect all processing nodes which are mounted on the same mother board. Hereafter, we refer to this 1U chassis of the building unit containing multiple CPUs and intra-connection network switches as the "cluster unit".

Figure 1 shows the conceptual view of the 1U cluster unit and the overall system. In the figure, 16 CPUs are equipped with two channels of Gigabit Ethernet NICs and individual two switches are mounted on a cluster unit. A single system rack contains 32 cluster units and interconnection switches, and finally hundreds of system racks makes up a Peta-Flops system.

3. Design and Implementation of MegaProto/E

In this section, we describe the detailed design and implementation of MegaProto/E compared with the previous version, MegaProto/C[2, 6].

3.1. Implementation of MegaProto/C

Before introducing the detailed design and implementation of MegaProto/E, we first describe the implementation of the previous model. We planned to develop the MegaProto as the first and second versions of the prototype according to the availability of parts and modules. When we started the design and implementation of MegaProto, Transmeta Crusoe was the best candidate for the CPU to be used, and IBM Japan had already provided a processor card with CPU, memory and I/O bus extension as a commercial product for embedded controlling systems. In MegaProto/C, Transmeta Crusoe (TM5800) with 933 MHz clock frequency was employed. It has the peak performance of 933 MFlops because it can only issue a single floating point operation per clock. Thus, the peak performance with 16 CPUs on a cluster unit is limited to 14.9 GFlops, and it could not achieve the goal described in the previous section.

However, we considered it as a good start for the development because we just had to develop a mother board to contain 17 of these processor cards as daughter cards. Therefore, we decided to develop the mother board at the first stage of the plan and to develop a new daughter card later when the Efficeon processor became available. Thus the production schedule of the Efficeon processor suited our two-staged plan. Actually, MegaProto/C (with Crusoe) was a kind of "prototype of a prototype" for software development, including Linux kernel tuning, drivers for NICs and switches, compiler and MPI library settings, as well as determining the environment of power consumption measurements[7].

On a cluster unit, there are two categories of interconnection networks, the data network and the management network. Hereafter, the 16 processor cards for computation are called "computation nodes" while the processor card for system management is called a "management node".

Data Network: It consists of two individual Gigabit Ethernet with switches. Each computation node is equipped with dual Gigabit Ethernet ports, and each port is connected to a 24-port Gigabit Ethernet switch (Broadcom BCM5692). Since only the computation nodes are connected to this network, there are 8 unconnected ports on each switch. These 8 links are connected to external RJ-45 ports for inter-unit connection outside the cluster unit (See Section 5). A computation node can drive

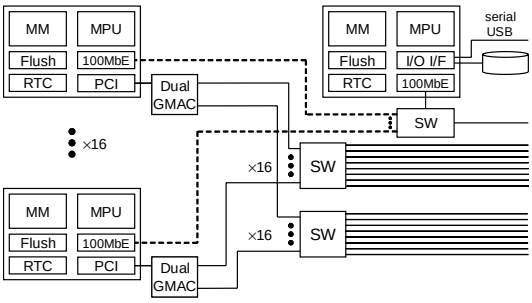

Figure 2. Block diagram of MegaProto/C cluster unit

both network links simultaneously for trunking (double bandwidth) or duplicated communication (fault tolerance) by software. This network is mainly used for data exchanging during parallel processing.

Management Network: It consists of a Fast Ethernet with a switch. All computation nodes and management node are equipped with a Fast Ethernet port, and these links are bound to the Fast Ethernet switch with dual upper-level Gigabit Ethernet links (Broadcom BCM5646). These upper-level links are connected to external RJ-45 ports for inter-unit connection. This network is mainly used for network management on the operating system (Linux) for NIS, NFS, remote login, remote shell, etc.

Figure 2 shows the block diagram of MegaProto/C cluster unit. Only the management node is equipped with a 2.5 inch hard disk drive with 60 GByte capacity to contain all system files for the 17 nodes in the cluster unit. At the system boot time, all disk-less computation nodes are booted via the Management Network sharing the binary images on the HDD of the management node. User's home directories are built on the outside file server to be shared by multiple cluster units through the external links of the Management Network.

3.2. Design of MegaProto/E

As mentioned in the previous subsection, we designed and implemented the second version of MegaProto while simultaneously constructing the software environment of MegaProto/C and evaluating it. The new version of cluster unit was named "MegaProto/E" (with Efficeon). On this version, we developed a new processor card (daughter card) equipped with an enhanced processor, memory and I/O bus compared with those on MegaProto/C. Several minor changes were also made to the mother board to improve the system stability.

The processor card was designed to fit the connection socket of the mother board of MegaProto/C, however, the density of each processor card was higher than that of MegaProto/C. The processor card consists of two small PCBs which are vertically stacked. The I/O bus to connect the processor card to the Data Network was improved from 32 bit/33 MHz PCI to 64 bit/66 MHz PCI-X, which provides four times the bandwidth of the old system. It provides 533 MByte/sec of theoretical peak bandwidth to support dual bidirectional Gigabit Ethernet links which have a peak bandwidth of 500 MByte/sec. As a result of using a commercial embedded controller module as the computation node on MegaProto/C, there was a severe bandwidth bottleneck on it. This was improved by using an upgraded I/O bus as reflected by several benchmark performances (See Section 4). The memory throughput was also improved from SDR-133 to DDR-266 in addition to the doubled capacity on MegaProto/E.

Figure 3. Photograph of MegaProto/E cluster unit

Since the computation performance is not directly limited by the performance of the management node, the processor card for the management node was kept the same as that used in MegaProto/C, that is, we did not use an Efficeon processor here. Therefore, the management node and computation nodes had a heterogeneous CPU configuration on MegaProto/E. There was no actual problem encountered at this point.

3.3. Implementation of MegaProto/E

As described above, the main work done in implementing MegaProto/E was on the processor card for computation nodes. The improvements to the processor card from that of MegaProto/C are summarized in Table 1. In particular, the enhancements to memory throughput and PCI bus are expected to be reflected in performance improvements of both the single CPU and parallel processing derived by

	MegaProto/C	MegaProto/E
MPU	TM5800 (0.93 GHz)	TM8820 (1.0 GHz)
TDP	7.5 W	3 W
Peak Perf./Power	124.0 MFlops/W	666.7 MFlops/W
Caches	L1=64KB(I)+64KB(D)	L1=128KB(I)+64KB(D)
	L2=512KB(D)	L2=1MB(D)
Memory	256 MB SDR (133 MHz)	512 MB DDR (266 MHz)
Flush	512 KB	1 MB
I/O Bus	PCI (32 bit, 33 MHz)	PCI-X (64 bit, 66 MHz)

Table 1. Processor card specification

the high network bandwidth.

Although the TDP of each CPU is less than the half of that of Crusoe CPUs, the power consumption on the memory module and the bus and the PCI-X bridge are greater. As a result, the power consumption of each processor card is slightly increased, and the maximum total power consumption of the MegaProto/E cluster unit is 320 W while that of MegaProto/C is 300 W. However, this small power increase is acceptable considering the greatly enhanced memory and I/O bus performance as well as more than twice the floating point performance of the CPU. A photograph of the cluster unit is shown in Figure 3.

The most important performance improvement on MegaProto/E is that we are able to obtain an excellent performance/power/space ratio of 1.024 TFlops/10.24 kW/rack, which satisfies our goal.

3.4. Air Cooling

MegaProto/E is equipped with ordinary small sized multi-fan to fit to its 1U cluster unit chassis. We have taken much care of the air flow issue because 17 of node processors are implemented only with small heat sinks without individual cooling fans. We designed the mother board of the unit to make the air flow over all heat sinks. However, it is impossible to control the surface temperature of all processors evenly since the air flows from the bottom to the top in Figure 3 where four processors on the top of three horizontal blocks.

To confirm the actual temperature on the different locations, we measured the approximate temperature on the heat sinks of several points by a thermography camera. Figure 4 shows the result image of thermograph and the thermal transition when running NAS Parallel Benchmark kernel CG (class-B) with 16 computation nodes. On the thermograph, each small rectangle displays a node processor. As shown here, even the highest temperature on the surface is lower than 40 ℃ and the difference between points A and C is within 8 ℃. As a result, the air cooling on MegaProto/E works well and actually we have no problem on thermal condition so far.

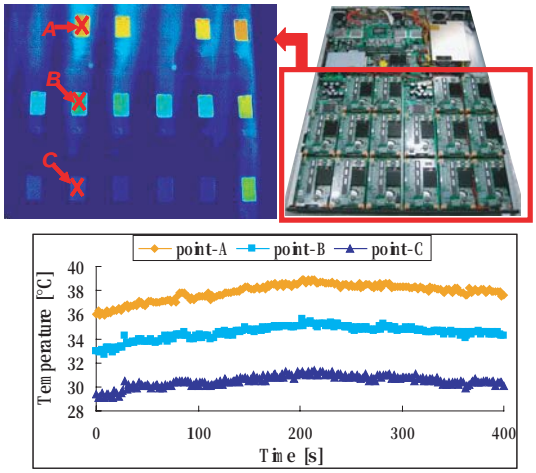

Figure 4. Thermograph and temperature transition on NPB kernel CG

4. Performance Evaluation

In this section, we evaluate the basic performance of a single cluster unit of MegaProto/E compared with that of MegaProto/C and of an ordinary high-end PC server with dual Xeon in SMP configuration. To assist in comparing network performance, we also refer to the performance of a two-node system with the same configuration of dual-Xeon servers.

The benchmark programs referred to in this evaluation are the commonly used HPL (High Performance Linpack)[8] and NPB (NAS Parallel Benchmarks)[9] kernels. For NPB kernels, the problem size is class-A. For HPL, the performance with $N = 10,000$ is shown. All sources were compiled with gcc/g77 version 3.2.2, linked with LAM-MPI version 7.1.1, and executed using a Linux kernel version 2.4.22mmpu. The environment of the dual-Xeon 1U server had similar software environments but slightly different versions; gcc/g77 version 3.4.3, LAM-MPI version 6.5.6 and Linux kernel 2.4.20-20.7smp. For all the benchmarks, only a single channel Gigabit Ethernet was used in order to avoid software overhead of trunking and to keep the fairness of the two-node dual-Xeon servers.

4.1. Performance Improvements from MegaProto/C

Figure 5 shows the overall performance comparison between MegaProto/C and MegaProto/E. In all graphs, the speed-up ratios compared with the performance of MegaProto/C with 4 CPUs are shown.

As shown in these results, the performance of MegaProto/E is between 1.06 to 2.38 times that of MegaProto/C. We analyzed these results as follows:

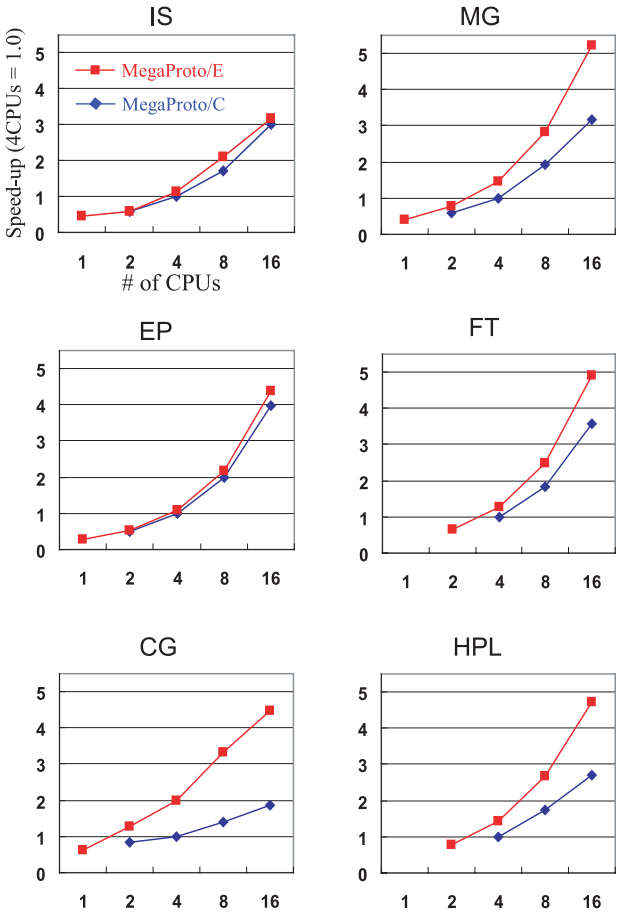

Figure 5. Speed-up comparison between two versions of MegaProto

FT, MG: 1.37 $\sim$ 1.65 times performance gain is due basically to the improvement in floating point operation speed, which is more than twice that of MegaProto/C.

CG: The communication data amount in CG is larger than other benchmarks. The performance improvement on PCI-X bus gives rise to a performance gain of 2.38 times.

HPL: 1.74 times performance gain is obtained as a result of the upgraded floating point performance similar to that of FT and MG. However, the efficiency of Linpack performance to the theoretical peak is only 30.1% due to the small capacity of the memory. MegaProto/C achieved 38% of peak performance with 16 processors, and the memory capacity problem is more serious on MegaProto/E in reference to its powerful floating point performance.

IS: Almost no performance gain because of the lack of floating point operations. The gain of CPU clock frequency is only 7%, which is reasonable.

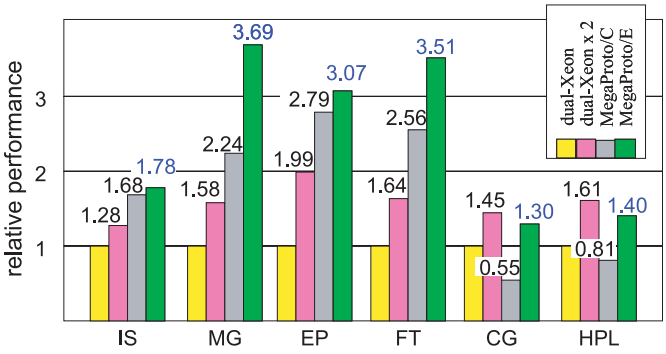

Figure 6. Comparison with Xeon-base servers

EP: This performance result seems to be anomalous because EP is basically a floating point bound benchmark. We conjecture that this benchmark involves a large number of fundamental numerical functions such as *log* or *sqrt* which may not be well-tuned in an ordinary gcc math library. If this is the case, then the large overhead of function calls will inhibit any gain in performance.

Overall, we can see a good performance gain, especially for the memory and the network throughput bound benchmarks. The memory capacity problem on HPL is serious, and it shows that our solution is not suitable for non fine-grained applications.

4.2. Comparison with Xeon-base Systems

Figure 6 shows the relative performance among dual-Xeon, two-node dual-Xeon, MegaProto/C and MegaProto/E. A Dual-Xeon system is a single node PC server in SMP configuration, and two-node dual-Xeon (labeled as "dual-Xeon x 2") is a small cluster to connect two dual-Xeon nodes with a single Gigabit Ethernet. All performances are given relative to that of a dual-Xeon. The PC server used here was Appro 1124Xi with 3.06 GHz Intel Xeon and 1 GByte of DDR memory. The total TDP and peak performance of processors were 170 W and 12.2 Gflops, respectively[10], and the maximum AC power rating of the entire 1U system was 400 W. Since these specifications are all comparable to our cluster unit with 16 processors, the dual-Xeon server is a good reference for performance comparison.

First, we can see the performance of MegaProto/E always exceeds that of dual-Xeon, ranging from 1.30 to 3.69 times improvement. In MG, EP and FT, it achieves a remarkable score. Especially for MG and FT, which are CPU performance and memory throughput bound benchmarks, MegaProto/E achieves excellent performance. Although MegaProto/C showed a markedly inferior perfor-

mance to that of the dual-Xeon for CG and HPL, MegaProto/E overcomes this with its improved network bandwidth and increased memory capacity.

With the exception of CG and HPL, MegaProto/E achieved higher scores than the two-node dual-Xeon system, even though it runs with less than half the power consumption of the two-node dual-Xeon. Since both systems are based on a single channel Gigabit Ethernet, this means that MegaProto/E is equipped with an interconnection network having a better performance balance between CPU performance and network bandwidth than the Xeon-base system. Since this is the case, the scalability of MegaProto solution could be much better than the Xeon-base HPC cluster if it were possible to adopt commodity Ethernet as the interconnection network. Such a performance balance is quite important for large scale parallel processing systems, and it has been shown that our solution based on commodity technology imported from non-HPC world works well in this arena.

5. Multi-Unit System

5.1. How to Utilize Multiple Upper-Links

After having confirmed the excellent performance of MegaProto/E, we are currently building a multi-unit system with more than 16 processors. At the first stage, we will build a cluster with 128 computation nodes connecting 8 MegaProto/E cluster units. Since a MegaProto/E cluster unit is equipped with 8 ports of external links for each Data Network on a channel, the total potential bandwidth from the cluster unit is 16 Gbps or 2 GByte/sec with the dual channel of the Data Network.

However, there is a problem in utilizing the 8 uplink ports for external inter-unit connection. Since the on-board Gigabit Ethernet switches on MegaProto/E is in Layer-2, a so-called “broadcast storm” occurs if we simply connect two or more uplink ports of multiple cluster units. The broadcast storm occurs when multiple links make one or more loops including the node PC and any of the intermediate Layer-2 switches.

There are two ways to solve this problem:

1. Using Layer-3 switchs which have the IP-base routing function, and then connecting all intermediate links through these switches. All looped connections are logically cut and all loops disappear.
2. Using tagged-VLAN (Virtual LAN) for pseudo static routing to separate multiple links in an isolated domain, and cutting the loops as shown in Figure 7[11].

The first method is simple but requires expensive Layer-3 switches to support a large number of Gigabit Ethernet ports. Such switches cost more than US$5,000 and this

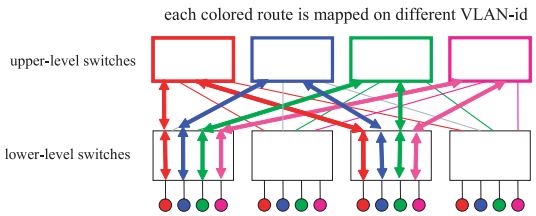

Figure 7. Fat Tree Network with VLAN-based routing

method is not acceptable in ordinary situations. The second method is reasonable because most of today’s medium class Layer-2 switches are equipped with a tagged-VLAN (IEEE802.1q)[12] feature. However, this requires multiple IP addresses on each node, and very complicated and tricky IP route setting is necessary for the whole network with a standard VLAN driver on Linux[11].

To solve the problem in a sophisticated way, we developed a special device driver to handle the source-destination IP routing with attaching/detaching VLAN-tag on Ethernet frames. With this technique, we can build a flat IP-space network with a single IP address per node to exploit the multiplied bandwidth on multiple upper-level links and switches. This network system is called VFREC-NET (VLAN-based Flexible, Reliable and Expandable Commodity Network)[13]. For example, we need 8 sets of Layer-2 8-port Gigabit Ethernet switches with the IEEE802.1q feature as the upper-level switches to combine 8 MegaProto/E cluster units. Then we configure the whole network so that each uplink from a cluster node is connected to one of these 8 switches. On the upper-level switches, from any source to any destination cluster unit, there exists a unique link which can be tagged as one of 4095 tags which is the physical limit of the IEEE802.1q protocol.

The on-board Gigabit Ethernet switch on MegaProto can handle this mechanism, and we can scale up the entire system to utilize all the allowed tags. It seems expensive to introduce eight 8-port switches, but actually a switch can be virtualized into multiple logical switches with VLAN. If we carefully select the assigned VLAN-tag without any conflict in the system, it is possible to configure the system network with a minimum set of Layer-2 switches.

5.2. MegaProto/E Multi-Unit System at SC2005 StorCloud Challenge

In some special cases, we can utilize the first method described above. For a StorCloud demonstration at SC2005 in Seattle, we brought four sets of MegaProto/E cluster units and operated them with an iSCSI server with 64 processors through 64 Gigabit Ethernet links. The operation of StorCloud challenge was performed by AIST (Advanced

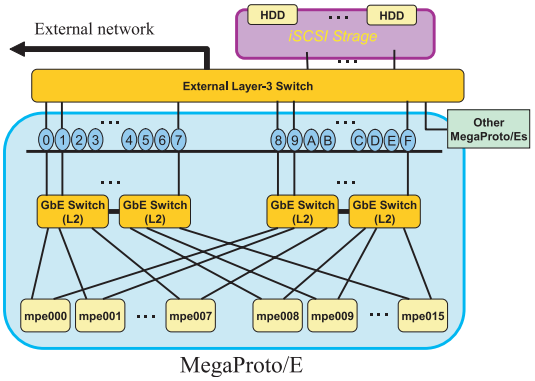

Figure 8. Block diagram of StorCloud with MegaProto/E at SC2005

Institute of Science and Technology, Japan)[14], and these MegaProto/E units ran through the event with just a couple of hardware failures on Gigabit Ethernet ports of computation nodes. However, these failures could be tolerated using the other ports of Data Network, and the total bandwidth was not lost.

Figure 8 shows the block diagram of MegaProto connection to the iSCSI server through a Layer-3 switch at StorCloud at SC2005. Throughout the demonstration, we could confirm the stability of the MegaProto/E cluster unit and its applicability for a variety of applications besides computational cluster computing.

6. Conclusions and Future Works

We have developed a building block for a large-scale power-aware PC cluster for high performance computing based only on commodity processor and network technologies, named MegaProto. The latest version, named the MegaProto/E cluster unit with Transmeta Efficeon processor, achieves 32 GFlops/320 W/1U of performance/power/space density and is suitable for mounting on a standard 19-inch rack. Including the space required for the inter-unit network switches, we can construct 1TFlops/10kW/rack Linux ready cluster based on dual-link Gigabit Ethernet.

The benchmark results show that our MegaProto/E demonstrates much better performance on NPB kernels and HPL than an ordinary high-end PC server with dual Intel Xeon in SMP configuration, having the same space occupancy and less power consumption. It has also been shown that typical applications with a certain degree of parallelism can be effectively solved on our platform, which promises to make very high density scalable cluster systems viable.

Besides building the hardware platform, we are also developing the software tools to connect these cluster units to thousands of processors in a system based on commodity Gigabit Ethernet routing with VLAN technology.

Future work includes the construction of a medium size system with hundreds of processors, performance evaluation of the system including network performance equipped with our VLAN solution, the verification of fault tolerance software not only for the processing node but also for the interconnection network.

By doing research on a Mega-scale computing based on low-power technology and workload modeling, we will continue to seek an effective way of achieving Peta-Flops computing.

Acknowledgments The authors would like to express their appreciation to the technical staff of IBM Japan for their contributions and support. This research work is supported by Japan Science and Technology Agency as a CREST research program entitled “Mega-Scale Computing Based on Low-Power Technology and Workload Modeling.”

References

[1] M. Warren, et al., “High-density computing: A 240-node Beowulf in one cubic meter”, in Proc. Supercomputing 2002, Nov. 2002.

[2] H. Nakashima, et al., “MegaProto: A Low-Power and Compact Cluster for High-Performance Computing”, in Proc. HP-PAC05 (with IPDPS2005), Apr. 2005.

[3] N. R. Adiga, et al., “An overview of the BlueGene/L supercomputer”, in Proc. Supercomputing 2002, Nov. 2002.

[4] http://www.infiniband.org/

[5] Mega-Scale research team, “Mega-Scale computing based on low-power technology and workload modeling”, http://www.para.tutics.tut.ac.jp/megascale/, 2005.

[6] H. Nakashima, et al., “MegaProto: 1 TFlops/10kW Rack Is Feasible Even with Only Commodity Technology”, in Proc. Supercomputing 2005, Nov. 2005.

[7] Y. Hotta, et al., “Measurement and characterization of power consumption of microprocessors for power-aware cluster”, in Proc. COOL Chips VII, Apr. 2004.

[8] A. Petitet, et al., “HPL - a portable implementation of the high-performance Linpack benchmark for distributed-memory computers”, http://www.netlib.org/benchmark/hpl/, Jan. 2004.

[9] D. H. Bailey, et al., “The NAS parallel benchmarks”, in Proc. Intl. J. Supercomputer Applications, 5(3):63-73, 1991.

[10] Intel Corp. Datasheets of the following Intel processors on 90nm process: Xeon (302355-001), Pentium 4 (303128-004), Mobile Pentium 4 (302424-002), Celeron M (300302-003) and Pentium M (302189-004), 2004.

[11] T. Kudoh, et al., “VLAN-based Routing: Multi-path L2 Ethernet network for HPC Clusters”, in Proc. of CLUSTER2004, Sep. 2004.

[12] http://www.ieee802.org/1/pages/802.1Q.html

[13] S. Miura, et al., “Low-cost high-bandwidth tree network for PC clusters based on tagged-VLAN technology”, in Proc. I-SPAN2005, Dec. 2005.

[14] O. Tatebe, et al., “High-performance KEKB/Belle data analysis using Gfarm Grid file system”, StorCloud Challenge, SC2005, Nov. 2005.