Performance Evaluation of Power-aware Multi-tree Ethernet for HPC Interconnects

Michihiro Koibuchi[1], Takafumi Watanabe[2], Atsushi Minamihata[2], Masahiro Nakao[3],
Tomoyuki Hiroyasu[2], Hiroki Matsutani[4], and Hideharu Amano[4]

1 National Institute of Informatics / JST, E-mail: koibuchi@nii.ac.jp,
2 Doshisha University, E-mail: aminamihata@mikilab.doshisha.ac.jp, tomo@is.doshisha.ac.jp
3 University of Tsukuba, E-mail: mnakao@ccs.tsukuba.ac.jp
4 Keio University, E-mail: {matutani@ny,hunga@am}.ics.keio.ac.jp.

***Abstract*—Ethernet has been used for interconnection networks of high-performance computing (HPC) systems that include PC clusters. Although a layer-2 Ethernet topology is limited to a tree structure in order to avoid broadcast storms and deadlocks of frames, various topologies with deadlock-free routing, that include loops, suitable for parallel processing can be used by the application of IEEE 802.1Q VLAN technology. However, their performance and power evaluations on real PC clusters with Ethernet have been rarely done. Thus, their evaluation on real PC clusters with Ethernet is important to analyze their impact on the total systems and validate their simulation results. In this paper, firstly we measure tree, and fully connected topologies with link aggregation on a 66-node/528-core PC cluster with a number of Ethernet switches. Secondly, we measure the optimization of the power consumption of Ethernet by a power-aware link regulation in order to reduce the power consumption of Ethernet switches on a real PC cluster. Measurement results show that the fully connected topology achieves 3.21 TFlops in High-Performance Linpack Benchmark (HPL), and its Rmax/Rpeak value is 67%. Up to 23% of the power consumption of networks can be reduced by the power-aware link regulation, while the performance is degraded by less than 1%.**

***Index Terms*—Ethernet, performance measurement, topology, interconnection networks, cluster interconnects.**

I. Introduction

Ethernet has been used for interconnection networks of high-performance computing (HPC) systems that include PC clusters, data center and cloud networks, because of its high performance-per-cost. Non-blocking commercial Ethernet switches are now available, and the link bandwidth of Ethernet has rapidly increased, as evidenced by the consideration of 40/100-gigabit Ethernet. Ethernet is thus used for high-performance computing systems. As of June 2011, gigabit Ethernets (GbEs) were employed as interconnects on 44% of the TOP500 supercomputers[1].

There are two ways of constructing Ethernet: one is to use a switch with several hundreds or more ports, and the other is to connect a number of switches, each having dozens of ports. Since large-scale switches with many ports are quite expensive, the latter way is preferable to make the best use of cost-effectiveness of Ethernet.

Unlike high-performance system area networks (SANs), such as InfiniBand, most current PC clusters using Ethernet have employed simple tree topologies. This is mainly because topologies that include loops are not allowed in order to avoid broadcast storms which circulate packets forever in layer-2 Ethernet. To remove this limitation, there are recently studies using IEEE 802.1Q tagged virtual LAN (VLAN) technology that set up multiple paths between a pair of switches on topologies that include loops, such as mesh[2][3][4]. Their evaluation results show that the VLAN operation overhead is trivial in terms of both latency and bandwidth[4].

Layer-3 Ethernet switches have been developed, and they allow a topology that includes loops. However, they consume higher power than simple layer-2 switches in addition to the disadvantage of high costs, and large switch delay which is an important factor in tightly-coupled computer networks, due to the complexity of their switch functions[4]. Thus, in this paper, we focus on simple layer-2 Ethernet switches.

A large number of researches on topologies and their deadlock-free routing have been done for lossless interconnection networks that can include Ethernet with the IEEE 802.3x flow control[5], however, most of them use a computer simulation with probabilistic models, or execution driven models. Network components including hosts, network interfaces and switches are often simply modeled in such simulations for achieving enough simulation speed. Thus, evaluation in real computer systems, such as PC clusters with Ethernet is important to analyze the impact on total systems and validate the simulation results, though there are a few researches on real PC clusters with Ethernet[6][7][4].

In this paper, our contribution is the performance and power evaluation of power-aware multi-tree Ethernet on a real PC cluster that uses multi-core processors with inter- and intra-processor MPI communications, and we clearly demonstrate that power optimization that explores the influences of various number of deactivated links on the performance in the cluster. For the purpose of the power saving, we apply power-aware link regulation (also known as on/off and multi-speed link regulation) to the PC cluster with Ethernet.

Our preliminary evaluation was done on a traditional PC cluster using single-core processors[7], and we extend the on/off link regulation strategy for making various power-optimization scenario considered in the evaluation. We also reported that a case study when the number of deactivated links is automatically fixed to a single certain value in NAS Parallel Benchmarks[4]. However, the important performance measurement is the exploration of the trade-off between the number of deactivated links and parallel application performance, though our previous works had not done it.

The rest of this paper is organized as follows. In Section II, we introduce related work. In Section III, we illustrate the implementation of multi-tree Ethernet topology. In Section IV, we explain the implementation of the power-aware link regulation strategy on multi-tree Ethernet topology. In Section V, we measure tree, and fully connected topologies with the power-aware link regulation on the PC cluster with a number of Ethernet switches. Our conclusions are described in Section VI.

II. Related Work

A. Ethernet Topologies and Deadlock-free Routing for High Performance Computing

Ethernet has usually employed a tree topology in order to avoid broadcast storms that circulate broadcast frames or unknown-destination frames forever, and deadlocks of frames that occur when the IEEE 802.3x link-level flow control is enabled[8].

Advanced techniques to use a topology that includes loops on Ethernet rely on the application of commodity VLAN technology[2][3][6][9][7]. Multiple paths between hosts can be obtained by using VLANs as follows: multiple VLANs, each having a different tree of the physical network, are assigned to a physical network with loops. All pairs of hosts can communicate with each other via any VLAN tree topology, and there are multiple paths that consist of different link sets between each pair of hosts. Each tagged frame is transferred by the usual layer-2 Ethernet mechanism within its VLAN topology. Although each VLAN topology is logically a tree, the physical topologies of layer-2 Ethernet are free from tree structures.

High-performance deterministic routing algorithms that break cyclic channel dependencies have been studied on various topologies for lossless interconnection networks that can include Ethernet with the IEEE 802.3x flow control [8][10][11], and some of them can be implemented on Ethernet by statically registered MAC addresses of hosts without VLAN technology. However, it is difficult to stabilize the management of frames with such a configured Ethernet when a broadcast storm occurs.

Most Ethernet switches support IEEE 802.1D STP (Spanning Tree Protocol) or 802.1D-2004 RSTP (Rapid STP) to prevent loops in a network. STP and RSTP are not aware of VLANs. When these protocols are enabled, all links out of a spanning tree are automatically disabled. STP and RSTP are thus disabled when a topology that includes loops is used. The 802.1Q-2003 MSTP (Multiple STP) and Cisco Systems' PVST (Per VLAN Spanning Tree) are STPs which support VLANs. They are quite useful for the VLAN-based routing implementation. In addition, The TRILL (TRansparent Interconnection of Lots of Links) working group has specified a solution for shortest-path frame routing in multi-hop IEEE 802.1-compliant Ethernet networks with arbitrary topologies, and it can work at L2 level[12]. However, there are currently only a few cost-effective Ethernet switches that support these protocols.

Although there are various efficient researches on topology and routing for parallel computing in interconnection networks[5], there are only a few researches on performance evaluation using real PC clusters with Ethernet[6][7]. Thus, their performance evaluation in real PC clusters is important to analyze their impact on total PC clusters and validate the simulation results of interconnection networks.

B. Low-Power Technique of Interconnection Networks

Although the study of power profile and reduction technique has been done in the enterprise networks that include data centers[13][14], our challenge is to save the power of cheap layer-2 switch networks on a PC cluster whose application traffic pattern could be predictable.

When employing topologies that include loops, a number of Ethernet switches and links are used in a PC cluster, which increases the total power consumption of the network compared with a simple tree topology. In addition, the power consumption of interconnects is increased as the link bandwidth is improved in PC clusters. The ratio of interconnects against the power consumption of PC clusters is thus increased in modern PC clusters. Thus, the low-power techniques of interconnects, such as GbE, have become one of the most important research topics for building PC clusters.

In interconnection networks of PC clusters, links consume a large amount of power even if no data is transferred, and its power is almost constant regardless of the traffic injection rates. Thus, power-aware on/off interconnection networks have been studied for both off-chip and on-chip communications that have different wakeup latencies, and minimum sleep duration to compensate for the on/off operations[15][16]. Although Ethernet has unique features of spanning-tree protocol (STP) based management, and the MAC address self-learning that introduces difficulty in the on/off link activation control, we developed a technique to apply the power-aware on/off link regulation to Ethernet[7].

Low-power router architectures, and their performance evaluation using the DVFS (dynamic voltage and frequency scaling) have been discussed[17][18]. Although the architecture of most commercial GbE switches is black-box from operators, we measured that decreasing the link speed reduces the power consumption of commercial GbE switches[7]. Thus, we can apply the existing techniques that vary link speeds and activate/deactivate links to Ethernet so as to reduce the power consumption of the existing commercial switches, and we referred them as power-aware link regulations in this paper.

III. Multi-tree Ethernet Topology

In this section, we describe an implementation of multi-tree Ethernet topologies, such as mesh, by using our previously developed technique called the VLAN routing method[6][9][4]. Unlike the other routing implementation techniques[2][3], since hosts do not need to perform VLAN operations, it has advantages in both simple host configuration and high degree of portability to existing various PC clusters with Ethernet.

Firstly, we explain VLAN tagging operation at a switch for using various topologies. Secondly, we describe the implementation method of the multi-tree Ethernet topologies.

A. Frame Tagging at Switch

A switch behavior of the VLAN tagging operation is stated as follows: when an untagged frame enters a port, it is tagged with a default VLAN ID tag number (port VLAN ID or PVID). Frames leaving the switch are either tagged or untagged depending on the port's VLAN configuration. If the port is a "tagged" member of a VLAN, the output frame is tagged with the respective VLAN ID. If the port is an "untagged" member of a VLAN, the output frame is left untagged. The VLAN untagged operation is originally intended to connect to older equipments that do not support tagged VLAN.

Notice that commodity GbE switches cost from under one hundred dollars to ten thousand dollars. The cheapest switches do not always support VLAN technology, or support a few functions of VLANs, and hence, they cannot operate the frame tagging described in this section; we can use commercial switches that support the operations of the IEEE 802.1Q standard in our VLAN routing method for multi-tree Ethernet topology.

B. VLAN Assignment

In general, there are three functions for representing routing algorithms[19]. The simplest routing relation is based on the $N(source) \times N(destination) \mapsto P$ routing relation (all-at-once)[19], where N is the node set and P the path set. The other routing functions are the $N \times N \mapsto C$ routing relation, which only takes into account the current and destination nodes[19], and the $C \times N \mapsto C$ routing relation, where C is the channel set.

A path set expressed by the $N \times N \mapsto P$ routing relation where all paths from a host are contained by a single tree can be implemented in multi-tree Ethernet topology for an n-switch network as follows[6].

1) Let t_i be the tree topology that contains all paths from switch i. Let v_i be the VLAN that corresponds to t_i, and it is initialized to null. Let i be $zero$.
2) If an existing VLAN v includes t_i, let $v_i := v$; otherwise create a new VLAN that includes t_i, and let v_i be the new VLAN.
3) Set the PVID of the ports of switch i, that connect to hosts, to v_i.
4) Add each port for connecting a switch in t_i to v_i, and make this port "tagged" members of v_i.
5) If $i < n - 1$, let $i := i + 1$ and go to step 2.
6) Register each port connected to a host in all VLANs as an "untagged" member.

C. Behaviour of Frame Transfer

By using the VLAN assignment, a frame is forwarded as follows: a source host transmits a normal (untagged) frame in the usual way by specifying the IP address or MAC address of a destination host. When an untagged frame from a host enters a port of a switch, it is tagged with the PVID of the port stated by the above procedure, and it is regarded as a frame which belongs to the VLAN. Finally, the frame is untagged when it leaves a port connected to the destination host, because such a port is an "untagged" member of the VLAN. The destination host thus receives the usual untagged frame.

Notice that the MAC-address management at switches can be stabilized by using the static MAC address registration or the sophisticated learning procedure[4].

IV. Implementation of Power-aware Link Regulation on Multi-tree Ethernet Topology

In this section, we describe a link regulation on multi-tree Ethernet topology in order to optimize the power consumption of Ethernet switches by using our previously developed technique[7][4].

A. Power Saving

In the multi-tree Ethernet topology, a network connectivity can be maintained if a link is deactivated. Links of recent off-chip interconnection networks consume power even if no data is transferred, and their power consumption is almost constant regardless of the traffic injection rates[15][7]. The power consumption of links can be reduced by using the port-shutdown operation available in most commercial Ethernet switches. Their operation was not originally intended to reduce power consumption; it is normally used to block the injection of unexpected frames from neighboring switches and hosts. In addition to port-shutdown operation, power consumption is further reduced down when the link speed is reduced down to 100 or 10 Mbps[7].

We have proposed to use the port-shutdown and link-speed operation for the implementation of power saving of Ethernet switches[7]. The port-shutdown operation completely reduces the power of the port, even if a physical link is connected to the port in switches we measured. Although additional power could be needed to deactivate or activate links, increasing power is not observed just after links turn off or on in all commercial switches we measured under the condition that the power consumption is captured at intervals of a second.

To monitor and manage ports of switches, most of Ethernet switches support the standard management information base (MIB). The standard MIB gives IP, UDP and TCP traffic information, including the number of input and output frames. A host can obtain them via the simple network management protocol (SNMP).

Ethernet has a unique MAC address management feature for switches that suits a tree topology. The feature makes it difficult to implement power-aware link regulation algorithms on various Ethernet topologies. In our implementation, the power-aware link regulation technique efficiently stabilizes the path update as follows.

B. Link Status: On to Off

The following procedure for a path reconfiguration deactivates a target link.

1) Use an existing routing algorithm to calculate the path set so that it avoids the target link.
2) Implement the path set by the VLAN assignment in Section III.B.
3) Deactivate the target link.

C. Link Status: Off to On

The following procedure re-activates a target deactivated link.

1) Activate the target link.
2) Use an existing routing algorithm to calculate the path set so that it uses the target link.
3) Implement the path set by the VLAN assignment in Section III.B.

D. Changing the Link Speed

Slowing down the link speed further reduces the power consumption of switches. When operating to change the link speed, the communication is interrupted in the link for a few seconds.

Table 1. Specifications of each host

CPU	Quad-Core AMD Opteron 2.3GHz
Memory	DDR2 667 MHz 8GB
NIC & driver	Broadcom BCM95721, Tigon3
OS	CentOS 4.6
Kernel	2.6.9-67.0.15.ELsmp

The following procedure thus changes the speed of the target link in order to hide the communication interruption of the target link.

1) Implement the temporal path set in order to avoid the target link by the link on-to-off procedure in Section IV.B.
2) Change the speed of the target link.
3) Implement the path set in order to include the target link by the link off-to-on procedure in Section IV.C.

V. Performance Measurement

We show the performance measurement of tree, and fully connected topologies with link aggregation, in which each switch is directly connected by each other, on a PC cluster.

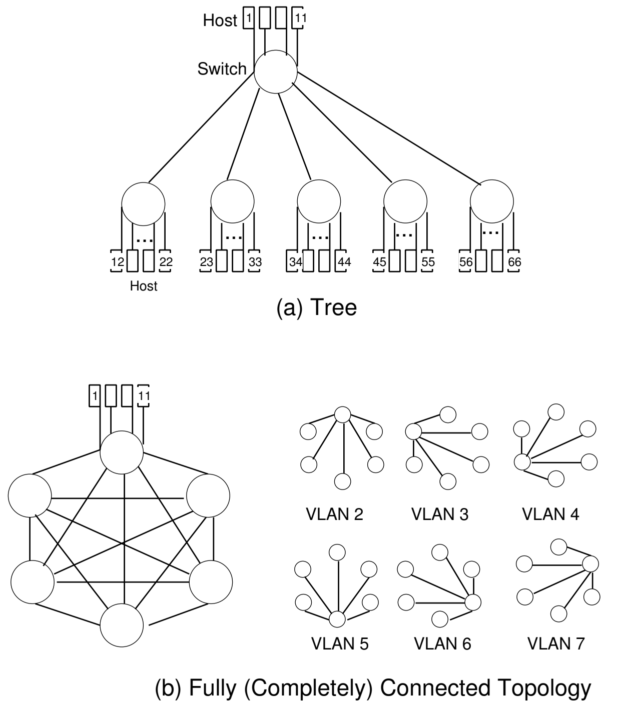
A. PC Cluster

We used a 66-host/528-core PC cluster using six GbE switches (Dell PowerConnect 6248, 48 ports). Its specifications are listed in Table 1. Each switch in the PC cluster connects to 11 hosts, and each host has two Opteron processors. The PC cluster provides TCP/IP with MPICH 1.2.7p1 or Open MPI 1.3, and MAC addresses of hosts are registered by self-learning before the measurements were made. The IEEE 802.3x link-level flow control was enabled at every port. The Dell PowerConnect 6248 switches have 48 ports whose speed can be automatically determined to 10, 100, and 1,000Gbps. It is a layer-3 switch, however we only used the layer-2 function in the measurement. It supports tagging and port-based 1,024 VLANs.

B. Throughput for Synthetic Traffic Patterns

1) Measurement Environment: We evaluated the network throughput of each topology for typical synthetic traffic patterns. Throughput was measured for two synthetic traffic patterns, bit-reversal and matrix transpose. In bit-reversal traffic, a host with the identifier $(a_0, a_1, \cdots, a_{n-1})$ sends a packet to the host whose identifier is the bit reversal $(a_{n-1}, \cdots, a_1, a_0)$ of the source host. In matrix transpose traffic, a host (x, y) sends a packet to the host $(k-y-1, k-x-1)$ or $(k-x-1, k-y-1)$ when $x+y = k-1$, where k is the number of hosts in each dimension.

The transfer of Tperf-1.5[20] was used for measuring the throughput of each transfer pair, and the sender and receiver processes were running on each host. The frame size was set to the maximum UDP datagram size, 1,470 Bytes.



(a) Tree

(b) Fully (Completely) Connected Topology

Figure 1. Topologies used in measurements

We compared three topologies (tree, and fully connected topologies) in Figure 1. The fully connected topology requires six VLANs, while the other topologies do not use VLANs. We measured the overhead of VLAN operation at the Ethernet switch[7], and it was reported that its throughput overhead and latency overhead are trivial (almost $zero$) in all the evaluated commercial switches.

Although there are a large number of topologies, such as mesh, fat tree, or torus, the fully connected topology achieves the highest performance in the case of the 225-host PC cluster with Ethernet whose position was 93rd in the Top500 ranking as of Nov. 2003[1][9][7]. We used the same switch as that of the 225-host PC cluster, and we mainly focus on the comparison of tree, and the fully connected topologies.

Table 2. Topological properties

Topology	Ave. distance of hosts	# of links between switches
Tree (1link)	2.39	5
Tree (5link)	2.39	25
Compl (1link)	1.83	25
Compl (5link)	1.83	125

2) Topology and Link Aggregation: Table 2 shows that the topological properties that enable to estimate the switches and link costs of these topologies. In the table, "Ave. distance" denotes the average path hops between hosts. Figure 2 shows the average throughput of all transfer pairs on each topology. In the table, "Tree" represents the tree topology, and "Compl" represents fully connected topology. The parenthesized number indicates the number of the 1Gbps links between switches.

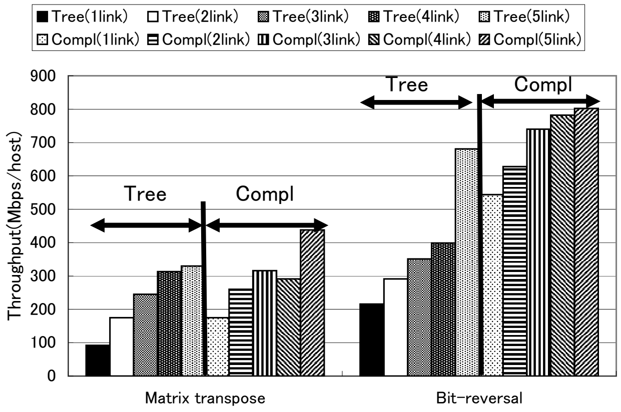


Figure 2. Throughput of topologies in synthetic traffic patterns (TCP, 64 hosts)

As shown in Figure 2, the fully connected topologies with five aggregation links outperform the tree topology with link aggregation by up to 25 %. Although the fully connected topologies use a larger number of links than that of the tree, the cable and port costs of cheap Ethernet switches are trivial against the system's initial costs. Since the number of aggregation links is limited in most of commercial switches, the fully connected topology has the advantage of employing the larger number of links than that of tree topology in a PC cluster. Thus, the fully connected topology can improve the throughput.

3) Power-aware Link Regulation: We used a simple power-aware link selection algorithm in order to show its impact on the total power consumption of the network in the PC cluster, although we could have used more sophisticated method of the power-aware link regulation [17][18][16][7]. In this paper, since we mainly focus on the power reduction of the network in the PC cluster, we omit the behaviour of the dynamic network reconfiguration that usually requires a few seconds latency in the PC cluster with Ethernet.

The simple power-aware link regulation algorithm selects the links to be deactivated or slowed down in the fully connected topology with five aggregation links as follows[7].

1) Set the acceptable maximum amount of traffic on a single channel as $peak$.
2) Estimate the amount of the traffic in each channel by the pre-execution of the target parallel application when all the links are activated. Notice that a link consists of two uni-directional channels.
    a) Associate a counter to each channel, and initialize it to $zero$.

   b) Select a path that has not been traced from all the source-destination pairs.
   c) Increment the counters of all channels that compose the path by the amount of its traffic.
   d) Repeat the second and third steps until all the communication pairs are traced.
3) Reduce the number of links between switches under the condition that the amount of the application traffic on every channel is less than $peak$.

By varying the parameter $peak$, the number of deactivated links can be varied; thus the ratio of the power reduction can be optimized to the traffic pattern and required throughput.

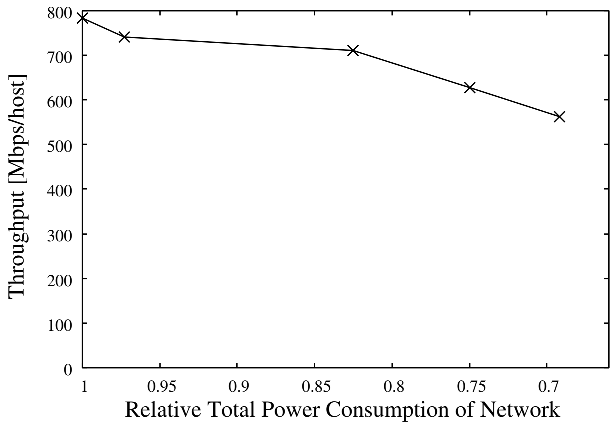


Figure 3. Throughput and relative power consumption of the power-aware link regulation in bit reversal traffic (TCP, 64 hosts)

Figure 3 shows the measurement results of the power-aware link regulation, and the baseline topology is the fully connected graph with five aggregation links. The x-axis represents the reduction rate of the total network power consumption by the power-aware link regulation. Its value is calculated by the power analysis results of the PowerConnect 6248 switch obtained from [7]. The higher value of y-axis is better. As the number of deactivated links increases, the total network power consumption decreases.

Figure 3 illustrates that the power-aware link regulation reduces not only the total power consumption of the network but also the performance in the synthetic traffic pattern, since each host injects frames as much as possible, resulting in a heavy link utilization. Thus, the power-aware link regulation is not efficient for the congested network by the synthetic traffic pattern.

Notice that we will show that it works well in the real parallel benchmarks, Linpack and a number of applications of NAS Parallel Benchmarks in the following sections.

C. Linpack

We evaluated topologies and their link aggregation using High-Performance Linpack Benchmark (HPL) HPL2.0[21] in the PC cluster, and it requires the accuracy of a numerical solution. We used Open MPI 1.3 and GotoBLAS 1.26 which were compiled using gcc 3.4.6/g77 3.4.6 for the operation library. In HPL, it is possible to tune the parameter suited to the feature of the system. The main parameters of HPL used for this measurement are shown in Table 3.

Table 3. Main parameters of HPL

N	234960
NB	220
(P, Q)	(6, 88)
BCAST	1ring

Each host has multi-core processors that use NUMA architecture, in which the number of MPI processes and threads strongly affects the HPL performance. Since in GotoBLAS library, matrix calculation can be executed by the number of threads in parallel on a node, we preliminary evaluated the influences of the number of threads in the matrix calculation on the performance. As a result, the following parameters achieved the highest performance; the number of processes per host is 8 in MPI, while the number of threads is 1. We thus used 8 MPI processes (1 MPI process per core), and 528 processes in total.

To minimize the overhead to access remote memory by MPI processes, we used the affinity function that associates a series of four MPI processes with four cores in a processor. We confirmed that using the affinity function of OpenMPI improves the HPL performance by 300 GFlops.

HPL is one of the implementation of Highly Parallel Computing of Linpack Benchmark, and this class requires the accuracy of a numerical solution.

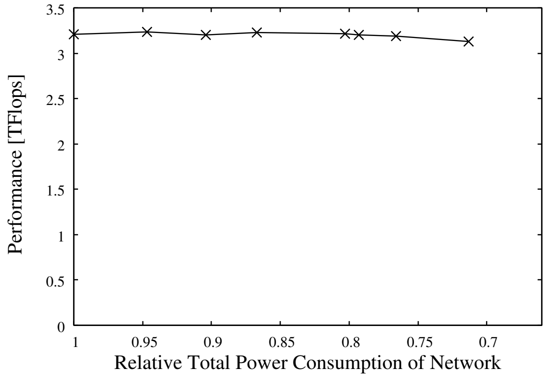


Figure 4. HPL results of the power-aware link regulation, 2 NICs

Figure 4 shows the relationship between the performance, and network power consumption results of the power-aware link regulation in the PC cluster. The baseline network topology is the fully connected topology with five aggregation links used in the previous subsection. To select a link to be deactivated or slowed down, we use the simple algorithm described in the previous subsection. The Rmax/Rpeak value is 67% in the case of the fully connected topology with five aggregation links. Since the value of Rmax/Rpeak is less than 60% in the most of the top500 supercomputers with Ethernet in recent five years whose performance is TFlops order, it can be said that the multi-tree topology is efficient for increasing the HPL performance of PC clusters.

The x-axis represents the reduction rate of the total network power consumption. Its value is calculated by the analysis results of the PowerConnect 6248 switch obtained from [7]. The y-axis represents the HPL performance whose unit is TFlops. As shown in Figure 4, the performance is gracefully reduced, as the network power is decreased. Up to 23% of the power consumption of networks can be reduced by the power-aware link regulation, while the performance is degraded by less than 1%. Thus, it can be said that in the case of executing the HPL, we can reduce the network power consumption, while the performance is almost maintained in the PC cluster.

D. NAS Parallel Benchmarks

We evaluated the performance of the topologies on NAS parallel benchmarks (NPB) 3.2[22], and we set the problem size to class C in all the measurements. The number of processes for parallel execution was fixed to 64 or 128. While BT, CG, FT, IS, LU, and MG applications can run under 64 processes, CG, FT, IS, LU, and MG benchmarks can run under 128 processes because of their specifications.

In this measurement, we used the trace files of the applications with the smallest class (W or A) for the traffic estimation in the procedure in Section V.B.3 of the power-aware link regulation, and we used the unit, byte, in MPI level communication.

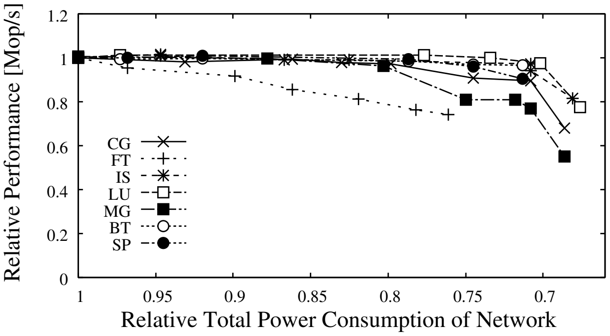


Figure 5. NAS Parallel Benchmarks results of power-aware link regulation, 64 processes

Figures 5 and 6 show that the power consumption of all switches in the PC cluster. The value of the power consumption is calculated by the analysis results of the PowerConnect 6248 switch obtained from [7]. The baseline topology is the fully connected topology with five aggregation links. the performance of tree topology. The unit of performance is the mega operations per second (Mop/sec), and the y-axis is its relative value. A higher value is thus better.

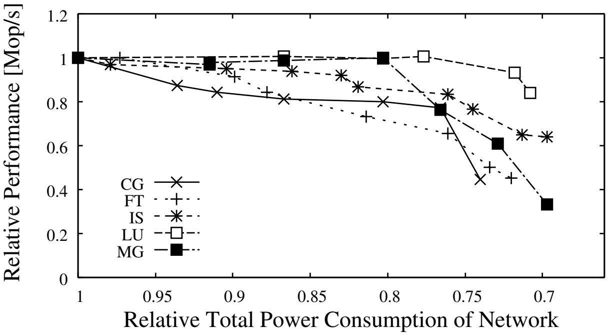


Figure 6. NAS Parallel Benchmarks results of power-aware link regulation, 128 processes

Table 4. Average computation and communication time of CG, BT and FT benchmarks under Compl. topology (5links) (sec)

	# of Processes	Total	Comput	Comm
CG	64	37.270	7.858	29.412
BT	64	82.719	51.705	31.014
FT	128	28.182	5.145	23.037

Table 4 shows the average computation and communication time of processes. These results clearly show that the NPB performance is affected by the communication overhead of topologies.

It is important to maintain the cluster's performance close to that when all links are activated, since the execution time of the application strongly affects the total power consumption of the PC cluster. In the case of 64 processes with the BT, CG, IS, LU, and MG benchmarks, the power consumption of all switches is reduced down by up to 26%, while the performance is maintained. However, in the case of FT benchmarks, the performance is decreased, as the power reduction rate increases. It can be said that the power-aware link regulation is efficient in a number of applications taken from NAS Parallel Benchmarks.

E. Comparison to Simulation Study

Topology issues of interconnection networks have been analyzed by the following techniques: 1) theoretical analysis, 2) probabilistic simulation, 3) execution driven simulation, and 4) execution on real computer systems. Although a full system simulation is recently available for chip multiprocessors [23][24], it is still difficult to simulate a large PC cluster system because of its computation costs.

The theoretical analysis would be difficult to preciously model large complicated network systems. Although the probabilistic simulation has been frequently used for analysis of routing algorithms and topologies, traffic pattern is not usually based on real applications. Although execution driven simulation could take a long time to simulate a minutely modeled host with operating system especially for large systems, performance measurement and analysis of PC clusters with monitoring tools have been done. However, most of these performance evaluation studies focus on fundamental performance of network interfaces and switches. A simulation study using a probabilistic model has reported that the impact of the routing and topology on performance of interconnection networks[5]. Fortunately, the performance tendency of the topology on the PC cluster in this paper is consistent with that of the most performance analysis, and simulation studies. Thus, this study concretes the validation of the existing analysis and simulation results in which the topology is quite crucial to the performance of lossless interconnection networks that include Ethernet with the link-level flow control.

VI. Conclusions

Although a layer-2 Ethernet topology is limited to a tree structure in order to avoid broadcast storms and deadlocks of frames, various topologies with deadlock-free routing, that include loops, suitable for parallel processing can be used by the application of IEEE 802.1Q VLAN technology.

In this study, our main contribution is the performance measurement of power-aware multi-tree Ethernet on a real PC cluster and we clearly demonstrate that power optimization that explores the influences of various number of deactivated links on the performance in the cluster. For the purpose of the power saving, we apply power-aware link regulation (also known as on/off and multi-speed link regulation) to the PC cluster with Ethernet.

Evaluation results showed that the fully connected topology achieved 3.21 TFlops in High-Performance Linpack Benchmark (HPL) whose Rmax/Rpeak value was 67%. Up to 23% of the power consumption of networks can be reduced by the power-aware link regulation, while the performance is degraded by less than 1% in the HPL.

Acknowledgements

This work was partially supported by JST CREST (ULP-HPC: Ultra Low-Power, High-Performance Computing via Modeling and Optimization of Next Generation HPC Technologies), and the NII Joint Work.

References

[1] Top 500 Supercomputer Sites, http://www.top500.org/.

[2] T. Kudoh, H. Tezuka, M. Matsuda, Y. Kodama, O. Tatebe, and S. Sekiguchi, "VLAN-based Routing: Multi-path L2 Ethernet Network for HPC Clusters," in *Proc. of IEEE International Conference on Cluster Computing (Cluster)*, Sep. 2004.

[3] S. Sharma, K. Gopalan, S. Nanda, and T. Chiueh, "Viking: A Multi-Spanning-Tree Ethernet Architecture for Metropolitan Area and Cluster Networks," in *Proc. of 23th Annual Joint Conference of the IEEE Computer and Communications Societies (Infocom)*, Mar. 2004, pp. 2283–2294.

[4] M. Koibuchi, T. Otsuka, T. Kudoh, and H. Amano, "A Switch-Tagged Routing Methodology for PC Clusters with VLAN Ethernet," *IEEE Transaction on Parallel and Distributed Systems*, vol. 22, no. 2, pp. 217–230, Feb. 2011.

[5] J.Duato, S.Yalamanchili, and L.Ni, *Interconnection Networks: an engineering approach*. Morgan Kaufmann, 2002.

[6] T. Otsuka, M. Koibuchi, T. Kudoh, and H. Amano, "A Switch-tagged VLAN Routing Methodology for PC Clusters with Ethernet," in *Proc. of the International Conference on Parallel Processing (ICPP)*, Aug. 2006, pp. 479–486.

[7] M. Koibuchi, T. Otsuka, H. Matsutani, and H. Amano, "An On/Off Link Activation Method for Low-Power Ethernet in PC Clusters," in *IEEE International Symposium on Parallel and Distributed Processing (IPDPS)*, May 2009.

[8] F. D. Pellegrini, D. Starobinski, M. G. Karpovsky, and L. B. Levitin, "Scalable Cycle-Breaking Algorithms for Gigabit Ethernet Backbones," in *Proc. of 23th Annual Joint Conference of the IEEE Computer and Communications Societies (Infocom)*, Mar. 2004, pp. 2175–2184.

[9] T. Watanabe, M. Nakao, T. Hiroyasu, T. Otsuka, and M. Koibuchi, "The impact of topology and link aggregation on pc cluster with ethernet," in *Poster(Work-in-progress presentation), IEEE International Conference on Cluster Computing (Cluster2008)*, 2008.

[10] A. Jouraku, M. Koibuchi, and H. Amano, "An Effective Design of Deadlock-Free Routing Algorithms Based on 2-D Turn Model for Irregular Networks," *IEEE Transaction on Parallel and Distributed Systems*, vol. 18, no. 3, pp. 320–333, Mar. 2007.

[11] S.-A. Reinemo and T. Skeie, "Effective Shortest Path Routing for Gigabit Ethernet," in *IEEE International Conference on Communications (ICC)*, Jun. 2007, pp. 6419–6424.

[12] TRILL Working Group Charter,
http://datatracker.ietf.org/wg/trill/charter/ .

[13] P. Heller, S. Seetharaman, P. Mahadevan, Y. Yiakoumis, P. Sharma, S. Banerjee, and N. McKeown, "ElasticTree: Saving Energy in Data Center Networks," in *NSDI*, 2010, pp. 249–264.

[14] P. Mahadevan, S. Banerjee, and P. Sharma, "Energy proportionality of an enterprise network," in *Green Networking*, 2010, pp. 53–60.

[15] V. Soteriou and L.-S. Peh, "Exploring the Design Space of Self-Regulating Power-Aware On/Off Interconnection Networks," *IEEE Transactions on Parallel and Distributed Systems*, vol. 18, no. 3, pp. 393–408, Mar. 2007.

[16] M. Alonso, J. M. Martinez, V. Santonja, P. Lopez, and J. Duato, "Power Saving in Regular Interconnection Networks Built with High-Degree Switches," in *IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, Apr. 2005.

[17] L. Shang, L.-S. Peh, and N. K. Jha, "Dynamic Voltage Scaling with Links for Power Optimization of Interconnection Networks," in *Proceedings of the International Symposium on High-Performance Computer Architecture*, Jan. 2003, pp. 91–102.

[18] J. M. Stine and N. P. Carter, "Comparing Adaptive Routing and Dynamic Voltage Scaling for Link Power Reduction," *IEEE Computer Architecture Letters*, vol. 3, no. 1, pp. 14–17, Jan. 2004.

[19] W. D. Dally and B. Towles, *Principles and Practices of Interconnection Networks*. Morgan Kaufmann, 2003.

[20] Tperf, http://www.am.ics.keio.ac.jp/~terry/tperf/.

[21] HPL - A Portable Implementation of the High-Performance Linpack Benchmark for Distributed -Memory Computers,
http://www.netlib.org/benchmark/hpl/.

[22] The NAS Parallel Benchmarks, http://www.nas.nasa.gov/Software/NPB/.

[23] P. S. Magnusson *et al.*, "Simics: A Full System Simulation Platform," *IEEE Computer*, vol. 35, no. 2, pp. 50–58, Feb. 2002.

[24] M. M. K. Martin, D. J. Sorin, B. M. Beckmann, M. R. Marty, M. Xu, A. R. Alameldeen, K. E. Moore, M. D. Hill, and D. A. Wood, "Multifacet General Execution-driven Multiprocessor Simulator (GEMS) Toolset," *ACM SIGARCH Computer Architecture News (CAN'05)*, vol. 33, no. 4, pp. 92–99, Nov. 2005.