An On/Off Link Activation Method for Low-Power Ethernet in PC Clusters

Michihiro Koibuchi[1], Tomohiro Otsuka[2], Hiroki Matsutani[2], and Hideharu Amano[2],[1]

[1]National Institute of Informatics
2-1-2, Hitotsubashi, Chiyoda-ku, Tokyo,
JAPAN 101-8430
koibuchi@nii.ac.jp

[2]Keio University
3-14-1, Hiyoshi, Kohoku-ku, Yokohama,
JAPAN 223-8522
{terry, matutani, hunga}@am.ics.keio.ac.jp

Abstract

The power consumption of interconnects is increased as the link bandwidth is improved in PC clusters. In this paper, we propose an on/off link activation method that uses the static analysis of the traffic in order to reduce the power consumption of Ethernet switches while maintaining the performance of PC clusters. When a link whose utilization is low is deactivated, the proposed method renews the VLAN-based paths that avoid it without creating broadcast storms. Since each host does not need to process VLAN tags, the proposed method has advantages in both simple host configuration and high portability. Evaluation results using NAS Parallel Benchmarks show that the proposed method reduces the power consumption of switches by up to 37% without performance degradation.

1 Introduction

Ethernet has been used for interconnection networks of various PC clusters because of its high performance-per-cost. Unlike the early Beowulf clusters, recent PC clusters with Ethernet employ system software[8] that supports the low-latency zero- or one-copy communication used in system area networks (SANs), such as InfiniBand[2]. High-throughput (non-blocking) commercial Ethernet switches are now available, and the link bandwidth of Ethernet has rapidly increased, evidenced by standardizations, such as 10-Gigabit Ethernet (10GbE). As of Jun. 2008, GbEs were used as interconnects on 57% of the Top500 supercomputers[15].

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. The low-power techniques of interconnects, such as GbE, have become one of the more important research topics for building PC clusters.

Table 1. Power Consumption of GbE Switches (W)

	All except ports	1-Port	Total (Port Ratio)
PC5324	15.0	1.2	42.9 (65%)
PC6248	56.8	2.1	155.2 (63%)
C3750	84.5	1.8	127.7 (34%)

Ethernet 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. The power consumption of the links can be saved by using the port-shutdown operation in the most commercial switches, and we investigate the impact of the port-shutdown operation on Ethernet switches, and the results are listed in Table 1.

In Table 1, "PC5324", and "PC6248" stand for Dell PowerConnect 5324 (24-port non-blocking switch), and 6248 (48-port non-blocking switch), respectively, while "C3750" is a Cisco Systems Catalyst WS-C3750G-24-TS-S. The "All except ports" represents the power consumption of the switch when all ports are shutdown, while "1-Port" represents the power consumption of a single port. "Total (Port Ratio)" is the power consumption when all ports are activated. In this paper, we use the term "the power consumption of a port" as meaning the reduction power by a port-shutdown operation. In addition, the impact of a port shutdown on the power consumption will increase in the case of 10GBase-T. The port-shutdown operation is not originally intended to reduce the power consumption, and it is normally used to block the injection of unexpected frames from the neighboring switches.

In this study, we propose an on/off link activation method

that uses the static analysis of the traffic in order to reduce the power consumption of Ethernet switches. All links are activated when the traffic load is high, while a large number of links are deactivated when the traffic load is low. Depending on which operation is selected, port shutdown or no-shutdown operation, the available network resources (switch and links) are changed. The path set is thus updated by modifying its MAC address table of Ethernet switches.

L2 Ethernet topology is limited to a tree structure, and the MAC-address tables of switches are updated by self-learning when the topology is updated. To immediately stabilize the MAC-address tables of the updated paths on various topologies, we extend the VLAN routing method[7], which attaches or removes a tagged VLAN of frames at switches, for the on/off link activation method.

The rest of this paper is organized as follows. In Section 2, we briefly introduce related work. In Sections 4 and 5, we propose and illustrate the on/off link activation method, and its link selection algorithm. In Sections 6 and 7, we evaluate the overhead of the on/off operation at switches, and the on/off link selection algorithm on a PC cluster. Our conclusions are in Section 8.

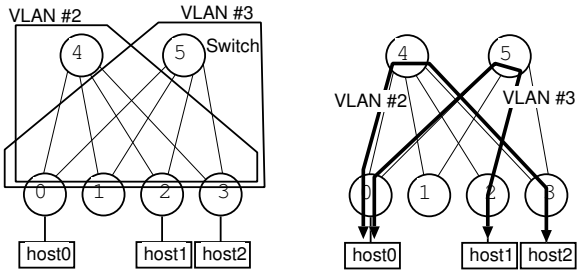
2 Related Work

On/off interconnection networks have been proposed for both off-chip and on-chip communications that have different wakeup times, and required sleep times for break-even point of the power consumption[13][1]. However, Ethernet has unique features of spanning-tree protocol (STP) based management, and the MAC address self-learning that introduces the huge overhead of network reconfiguration, and difficulty in the frequent run-time on/off link activation control and its implementation. Although low-power router architectures, and their performance evaluation using the DVFS (dynamic voltage and frequency scaling) have been discussed[11][14], the architecture of most commercial GbE switches is black-box from users, and operators.

There are several studies on implementing deadlock-free routing algorithms on Ethernet by statically registering the MAC addresses of hosts without VLAN technology, and deterministic routing algorithms that break cyclic channel dependencies have been studied for lossless interconnection networks that can include Ethernet with the link-level flow control[9] [4][10]. However, it is difficult to stabilize the management of frames using such a configured Ethernet when a broadcast storm occurs.

Routing implementation techniques using VLAN technology for topologies that include loops have been developed so as not to create broadcast storms[5][12][6]. VLAN technology was originally not developed for increasing network throughput, but for partitioning hosts into multiple groups. It has been used in intranets or on Internet backbones for the QoS control[16]. Multiple paths between hosts are obtained using VLANs in the following way: multiple VLANs, each having a different tree of the physical network, are assigned to a physical network that includes loops. Each host is configured as a member of all the VLANs; i.e. it has virtual network interfaces corresponding to all the VLANs. In this way, 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.

Since each path is assigned to a single VLAN, each source host selects a path by specifying the virtual interface that corresponds to the appropriate VLAN. 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-based structures.



(a) VLAN assignment (b) Host-tagged routing

Figure 1. VLAN-based Routing on a Fat Tree

The example of the VLAN-based routing is shown in Figures 1(a) and (b). There are two VLANs, VLAN #2 and #3, each of which is a spanning tree of the physical network including loops. Each host has two virtual interfaces and can send frames to all destination hosts via either of two VLANs as shown in Figure 1(b).

Since communication library used in current PC clusters does not always support VLANs, the design of the VLAN-based routing method cannot be applied for such PC clusters.

Most Ethernet switches support IEEE 802.1D STP (Spanning Tree Protocol) or IEEE 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. Therefore, STP and RSTP must be disabled when a topology that includes loops is used. IEEE 802.1Q-2003 MSTP (Multiple STP) and Cisco Systems' PVST (Per VLAN Spanning Tree) are STPs which support VLANs. They are useful for the VLAN-based routing implementation. However, there are currently only a few cost-effective Ethernet switches that support these protocols.

Although there are a large number of routing algorithms and their implementation on Ethernet, there has been little work on a low-power technique using commercial Ethernet switches in PC clusters.

3 Preliminary

L2 Ethernet topology is limited to a tree structure, and the MAC-address tables of switches are updated by self-learning when the topology is updated. To frequently perform the on/off link activation operations, the MAC-address tables of the updated paths should be immediately stabilized on various topologies. In this section, we extend the VLAN routing method[7], which attaches or removes a tagged VLAN of frames at switches, for the on/off link activation method.

3.1 Frame Tagging in Switches

The switch behavior in VLAN tagging operation is as follows; when an untagged Ethernet frame enters a port, it is tagged with the default VLAN ID tag number (port VLAN ID, PVID), while a tagged frame enters a port with no effect on the tag.

On the other hand, 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 untagged.

3.2 Switch-tagged Routing Method

In the switch-tagged routing method[7], all the paths from a host belong to a single VLAN regardless of the destination host. Both VLAN tagging and untagging operations are performed at each switch port connected to a host, by using the following configuration.

- Set PVID of each port to the ID of the VLAN that is used by the connected host when sending frames.
- Register each port as an "untagged" member of all VLANs.

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 a frame from a host enters a port of a switch, it is tagged with PVID of the port and is regarded as a frame that belongs to a VLAN indicated by the ID tag number. The frame is transferred by the layer-2 Ethernet mechanism as well as the normal VLAN-based routing. Finally, the frame is untagged when it leaves a port

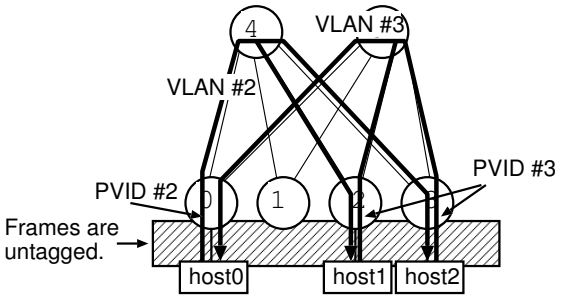


Figure 2. Switch-tagged Routing Method

that is connected to the destination host, because the port is an "untagged" member of the VLAN. The destination host thus receives an untagged frame, and the communication library can easily handle it.

In this way, all the hosts can communicate with each other on various topologies, even if a VLAN tagging operation is not supported by communication library of the hosts.

The example of the switch-tagged routing is shown in Figure 2 corresponding to Figure 1. The port of switch 0 connected with host 0 is configured as PVID #2, and the ports of switches 2 and 3 connected with hosts 2 and 3 respectively both have PVID #3. Thus, frames transmitted by host 0 are routed in VLAN #2 for all destinations, while those by hosts 1 and 2 use VLAN #3.

A host does not process the VLAN tags, and the communication between all the hosts is achieved in various topologies that include loops.

3.3 MAC Address Management

A problem introduced by the switch-tagged routing method is the MAC address self-learning at switches. Ethernet switches usually learn unknown MAC addresses when receiving frames. When a path from host A to B and one from B to A use different VLANs, the intermediate switches of both paths cannot learn the destination MAC address. This is because the MAC address self-learning is independently performed on each VLAN. Therefore, We propose a MAC-address self-learning for the switch-tagged routing method that uses the following procedure[17].

1. For each VLAN, make a virtual interface that correspond to the VLAN on each host. In the example of the mesh shown in Figure 4, each host has virtual interfaces for VLANs #101-104. In Linux operating systems, virtual interfaces can be made by the "vconfig" command.
2. Give an IP address to each virtual interface at all the

hosts so that the virtual interface has a unique network address that belongs to a different segment on the physical interface.

3. At each host, broadcast an ICMP or UDP message from each virtual interface so that the switches can learn the MAC address of the host in each VLAN.

In step 2, the IP address is only used for the MAC address registration at each switch.

In step 3, the ping command (ICMP echo req.) at each host can be used in each VLAN segment, and the MAC address of the source host is registered in the MAC address table of each switch. Notice that when the source host uses the ping command, it cannot receive the pong (ICMP echo reply) corresponding to the ping. The VLAN tag of the frame will be removed when the ping is outputted from the switch of the destination host. The learning procedure should be re-taken before the switch aging time expires. Fortunately some commodity Ethernet switches can set the aging time to infinity.

4 Implementation of On/Off Link Activation

4.1 Path Reconfiguration

Before a link is deactivated, the paths that go through it must be changed so that all the paths avoid the set of deactivated links.

When the VLAN-based routing implementation is not used in a topology that includes loops, the MAC address tables should be statically updated after all the links are shutdown, and its operational and time overhead can not be ignored. This is because the network status is unstable while statically updating the MAC address tables in a single LAN without the STP. To stabilize the management of MAC address tables in Ethernet switches, we use the switch-tagged routing method briefly introduced in Section 3.

Figure 3 shows the flow of the on/off link activation method, and its key-point is the implementation on Ethernet. Path reconfiguration by the existing fault-tolerant routing and network reconfiguration is needed when a link status changes to be activated or deactivated on the on/off link selection algorithm. We firstly propose its implementation of Ethernet in this section. Second, the on/off link selection algorithm that deactivates some links so that the performance is not decreased is illustrated in Section 5.

4.1.1 Link Status: On to Off

The following procedure for a path reconfiguration is taken in order to deactivate a target link.

1. Determine the new path set that avoids the target link using a routing algorithm.

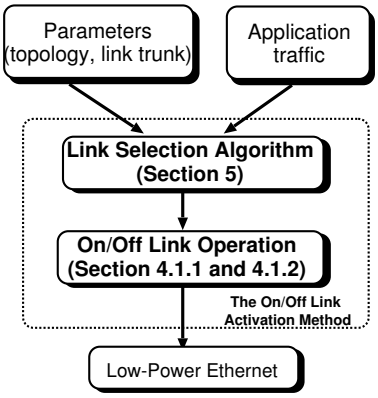


Figure 3. Flow of the On/Off Link Activation Method

2. Create VLANs that include the new path set using the switch-tagged routing method.

3. Make their MAC address tables at each switch using the MAC address self-learning procedure shown in the previous section.

4. Update the PVIDs of the ports for connecting the hosts to the newly-made VLANs in order to use the new path set that avoids the target link.

5. Inactivate the target link.

Until completing step 4, the hosts communicate with each other using the previous paths that may use the target link. For example, in the dimension-order routing in the torus shown in Figure 4, when the link between switches 1 and 5 is deactivated, only switches 1 and 5 update the PVID to #103 in order to avoid using the link between switches 1 and 5.

4.1.2 Link Status:Off to On

The following procedure is taken so that a target deactivated link is re-activated.

1. Determine the path set that uses the target link using a routing algorithm.

2. Create VLANs that include the new path set that uses the target (deactivated) link using the switch-tagged routing method.

3. Activate the target link.

4. Make their MAC address tables in each switch using the MAC address self-learning procedure shown in the previous section.

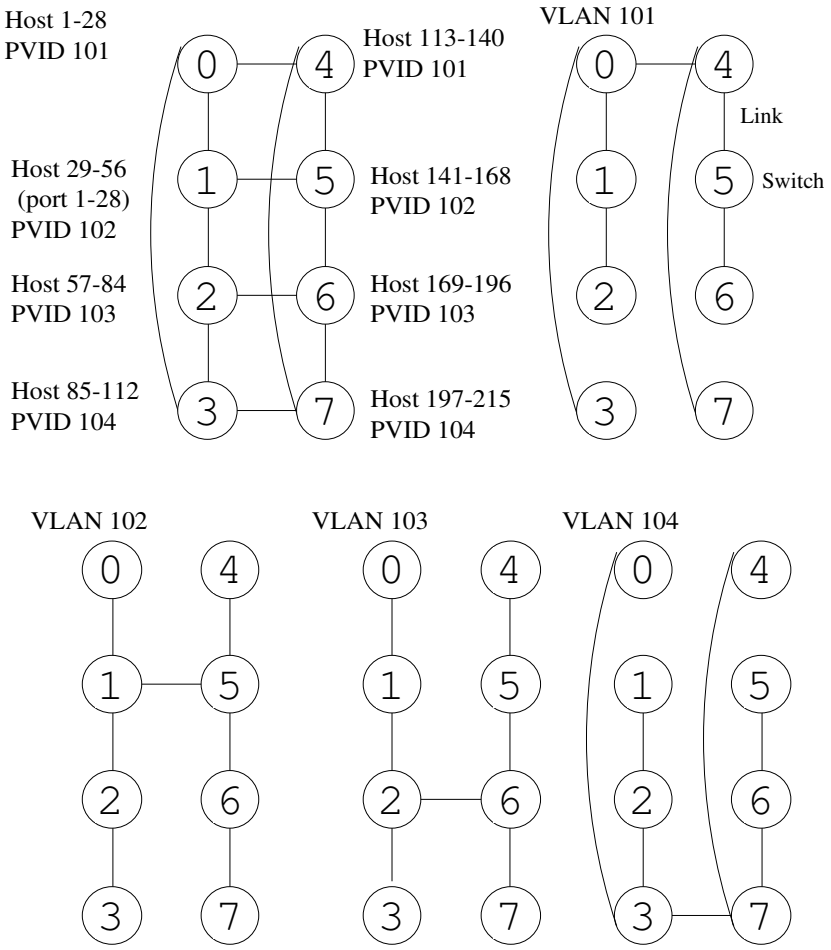


Figure 4. Example of Switch-tagged Routing Method on a PC Cluster

5. Update the PVIDs of the ports for connecting hosts to the newly-added VLANs in order to use the new path set.

Since hosts do not always insert the traffic at the maximum injection rate in parallel applications, all the network resources are not always fully used. Since the delay caused by the link wakeup and negotiation tends to be several seconds (details are shown in Section 7), the on/off link activation would be done by a per-application operation or a per-day operation based on the traffic prediction.

In the case of most cheap commercial switches, the on/off procedures could be manually performed. However, if the traffic is monitored via SNMP and the activation of the ports is executed by the remote operation, the on/off procedures will be automatically performed by a server host.

4.2 Limitations on Existing Commodity Switches

4.2.1 Applicable Commercial Ethernet Switches

There are various commodity GbE switches whose costs are from dozens of dollars to ten thousands dollars. However, the cheapest switches do not support VLAN technology, or only a few functions of VLANs and cannot employ the proposed method. The on/off link activation method can be applied to commercial switches that support the operations of the IEEE 802.1Q standard described in Section 3.

4.2.2 Upper Limit of Number of Hosts

The number of hosts is limited to the sizes of the MAC address tables in Ethernet switches. Each entry of a MAC address table consists of the destination MAC address, VLAN ID, and port. The maximum number of hosts, H, is thus

estimated as follows:

$$H = \frac{T}{V} \quad (1)$$

where T is the number of table entries at a switch, and V the number of used VLANs. T is usually around ten thousands such as $8k$ or $12k$ in commodity cost-effective switches, such as Dell PowerConnect 5324.

5 An On/Off Link Selection Algorithm

In this section, we propose an on/off link selection algorithm that uses the information of the traffic on Ethernet. The proposed algorithm classifies links into on/off links, and ever-on links, as well as an algorithm by Soteriou and Peh [13]. To guarantee the connectivity of the network even when the traffic load is close to $zero$, a spanning tree that consists of ever-on links is embedded in the network.

The on/off link selection algorithm deactivates some links so that the performance is not decreased. Its implementation is based on the switch-tagged routing method, as proposed in Section 4.

In the on/off link selection algorithm, the traffic patterns are pre-analyzed by the pilot execution of parallel programs, traffic history in a certain term using monitoring tools, such as IPTraf[3] at hosts, or its sampling. We assume that the existing fault-tolerant routing, or network reconfiguration are applied to the path computation when a link is deactivated. The traffic amount of each link is computed at a server host that manages Ethernet switches.

The on/off link selection algorithm deactivates links that transfer traffic that is smaller than *trf* in the following way.

1. Estimate the amount of the traffic in each channel 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 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.
2. Build the (ever-on) spanning tree.
   (a) Select the switch whose channel counter has the highest value as the root of the spanning tree.
   (b) Add links and their switches to the spanning tree by the order that firstly selects the link with the highest value in the counters.
3. Set the $threshold$ to the highest value among trf and the counters, which is used to judge whether a channel is activated, or not in the step 6.
4. Select the channel whose counter has the lowest value among the on/off channels that have not been checked.
5. Determine the path set that avoids the channel selected in the step 4 using a routing algorithm. The route is not changed when the channel is a member of the link aggregation.
6. Compare the $threshold$ and the highest value of counters under the updated path set in order to judge whether the channel selected in step 4 is deactivated or not.
   - If the $threshold$ is equal or higher, deactivate the channel, update the path set avoiding it, and update its counter value.
   - If the $threshold$ is lower, discard the updated paths.
7. Repeat steps 4-6 until all the channels outside the (ever-on) spanning tree have been selected in the step 4.
8. Deactivate links whose two channels are deactivated.

Figure 5 shows an example of the on/off link selection algorithm (*trf* parameter is $zero$), and it omits hosts. In this example, three links out of the spanning tree are deactivated.

6 Fundamental Evaluations

In this section, we show the power consumption of existing GbE switches, and the overhead of port-shutdown and VLAN operations at switches.

6.1 Power Consumption of GbE Switches

We measured the power consumption of GbE switches using System Artware, a watthour meter (SHW3A), and Table 2 lists their results. "GbE/port" is the power consumption of a single port. The switches can enforce port speeds to 1000 Mbps, and 100 Mbps. The "All except ports" is the power consumption of switches when all the ports are shutdown. Its power is almost constant regardless of the traffic injection rates. When a link is deactivated, two ports are shutdown.

In Table 2, "PC5324", "PC6224" and "PC6248" stand for Dell PowerConnect 5324, 6224 (24-port non-blocking switches), and 6248 (48-port non-blocking switch), respectively, while "SF-420" is PLANEX Communications INC, SF-0420G (24-port switch).

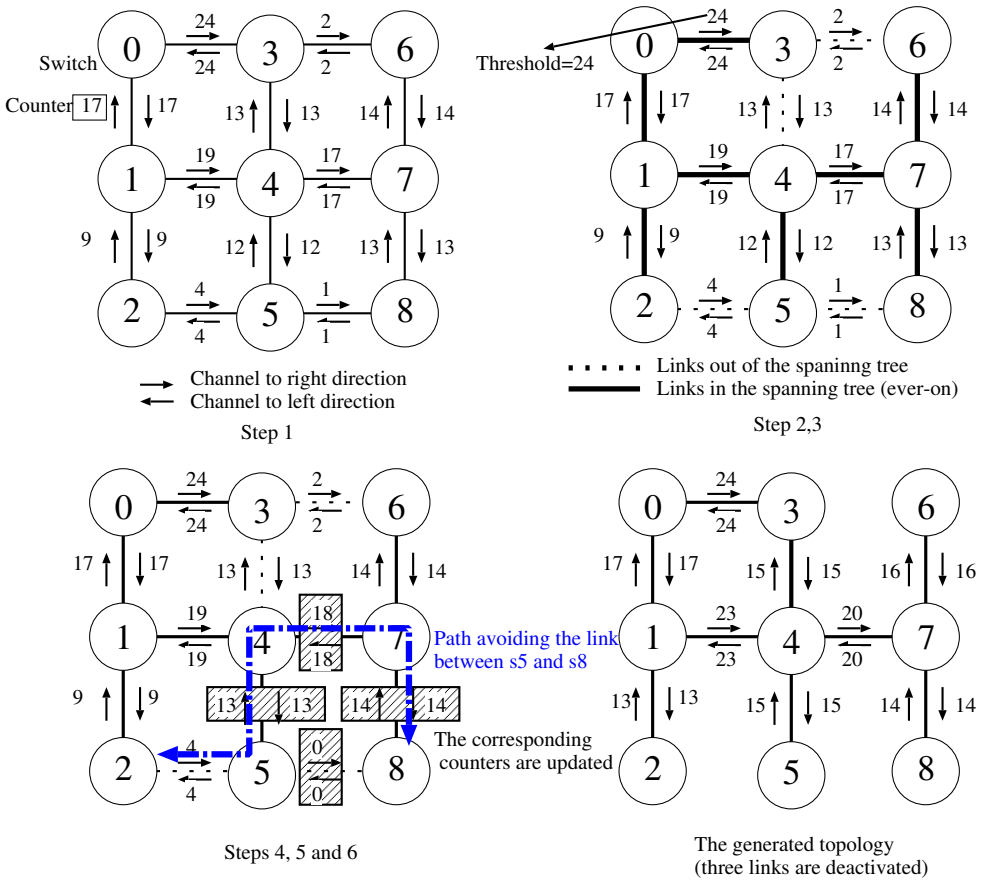


Figure 5. Example of The On/Off Link Selection Algorithm

Table 2. Power Consumption of GbE Switches (W)

	All except ports	GbE/port
PC5324	15.0	1.2
PC6224	42.5	2.0
PC6248	56.8	2.1
SF-420	32.6	1.0

Since the PC6224 and PC6248 switches are L3 switches, they provide a larger number of services than that of other L2 switches. Thus, the power consumption of the PC6224 and PC6248 switches is larger than that of PC5324 and SF-0420G switches.

These results show that the port-shutdown operation strongly affects the reduction of the power consumption of the GbE switches.

6.2 Overhead of On/Off Operations

We measured the overhead of the on/off link operation at a switch. The proposed on/off link activation implementation can be decomposed into a port-shutdown operation, and the VLAN operation at the switches in order to update the paths.

In the measurement of the overhead of the port-shutdown operation, while the ping command (ICMP message: 64 bytes) between two hosts is executed at intervals of 0.1 second, a switch continuously operates the shutdown, and no-shutdown (resume) of the port. While the link is deactivated, the communication (ping frame) is interrupted, and we state it as the overhead of the on/off link operation.

Figure 6 shows the topology needed to evaluate the overhead of VLAN operations. While the ICMP messages are transferred between two hosts at 0.1 sec intervals, the PVID of the port to the sender host is updated at a switch. While the port is deactivated, the communication is interrupted, and we say this is the overhead of the VLAN operations. Table 3 also shows the results of the overhead of VLAN op-

Table 3. Overhead of On/Off Operation (Sec)

GbE SW	On/Off Operation	VLAN Modification
PC6224	3.4	0
PC6248	2.2	0
PC5324	4.0	0
SF-420G	12.0	0

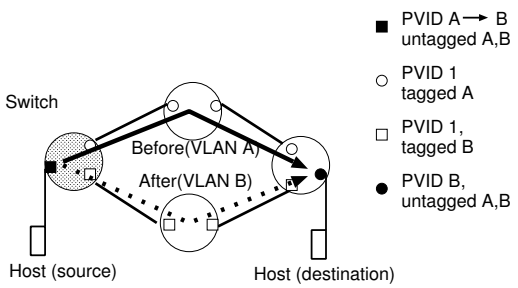


Figure 6. Topology Considered in Evaluations

erations, and the overhead is quite smaller than that of the port-shutdown operation.

7 Evaluations using NAS Parallel Benchmarks

7.1 PC Cluster

We used part of a 225-host PC cluster in Doshisha University whose interconnect consists of eight GbE switches (Dell PowerConnect 6248 (48 ports))[17]. It uses a TCP/IP with MPICH 1.2.7.p1, and the MAC addresses of hosts are registered by the MAC address self-learning procedure in Section 3 before the measurements. The IEEE 802.3x link-level flow control is enabled at every link.

Table 4. Specifications of Each Host (SuperNova Cluster)

CPU	AMD Opteron 1.8 GHz $\times$ 2
Chipset	AMD 8131+8111
Memory	PC2700 Registered ECC 2 GB
OS	Debian GNU/Linux 4.0
Kernel	2.6.18-4-amd64
MPICH	1.2.7p1

The topology is 4×2 2-D torus using six links between switches (link aggregation) with the switch-tagged routing method. As shown in Figure 4, four VLANs with dimension-order routing are employed. The circle in the figure represents a switch. In Figure 4, frames from host 1-28 are tagged with the VLAN ID #101 in the port of switch 0, and they are forwarded along the VLAN #101 for their destinations. The VLAN ID tag #101 is removed when the frames leave the switch that connects to the destination hosts.

In this evaluation, we used a turn-model based routing called L-Turn[4]. The L-Turn (deadlock-free) routing can be used for arbitrary topologies, and it includes the path set of the dimension-order routing on 2-D mesh. Since there are six links between switches, only a few path modifications occur in the on/off link selection algorithm; it mainly adjusts the number of links between switches according to the traffic of the NPB in the PC cluster.

We used the trace file of the NPB with the smallest class (W or A) in the on/off link selection algorithm. We used a unit, bytes in MPI level communication between each source-destination pair, as "the amount of the traffic" in the on/off link selection algorithm proposed in Section 5.

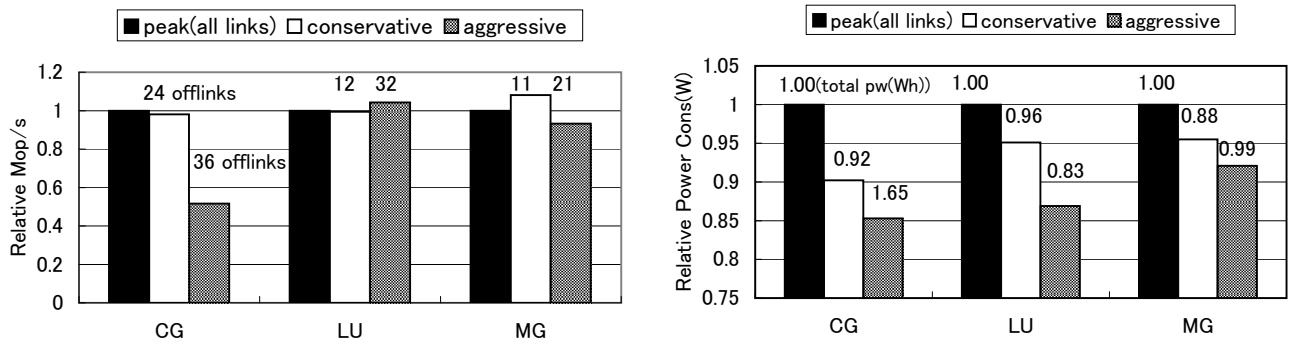
In the case of 128 processes, each switch connects to 16 hosts, while it connects to 8 hosts in the case of 64 processes. Each host executes a single process.

7.2 Evaluation Results

7.2.1 128 Hosts

Figure 7(a) shows the relative performance(Mop/s) of topologies that include deactivated links selected by the on/off link selection algorithm. It is normalized by the "peak (all links)" case when all the links are activated. Figure 7(b) shows their power consumption of all switches. It is calculated using the results in Table 2. The "conservative" stands for the topology generated by the on/off link selection algorithm whose *trf* parameter is $zero$, while "aggressive" is one whose *trf* is varied for the further power reduction. The numerical values in Figure 7(a) show the number of deactivated links on torus topology, and those in Figure 7(b) shows the relative total power consumption of the target application that is calculated by the execution time and the power consumption of all switches. We can estimate the total power consumption of all switches in the target application, since its power is almost constant regardless of the traffic injection rates.

Both figures illustrate that the power consumption of all switches is reduced by up to 13% with almost no performance degradation in the case of the LU application. The crucial factor for improving the total power consumption is to maintain the performance (Mop/s) close to that when all links are activated, since the execution time of the application strongly affects the total power consumption. In the



(a) Performance Evaluation

(b) Power Consumption Evaluation (PC6248)

Figure 7. NPB Benchmarks (128 Process, Class C)

case of the on/off link selection algorithm whose *trf* parameter is $zero$, all applications maintain their performance. The case where all links are activated is expected to achieve the highest performance (Mop/sec) in theory. Although in practice, the "peak" cases are sometimes just slightly inferior to "conservative" and "aggressive" cases, we consider that it is an acceptable error range when they can be regarded as the same performance on the evaluation using the large PC cluster. Notice that, when the value of *trf* is higher than that of "aggressive" cases in MG and LU, the performance is drastically reduced.

7.2.2 64 Hosts

Figure 8 shows the performance of the on/off activation method, and the power-consumption results of switches under the condition that there are 64 hosts using up to 5 links between switches. The other parameters are the same as the cases of 128 hosts. The performance of the IS application is drastically decreased when the power consumption of switches is reduced by 12%, because it frequently performs MPI_All to all function and thus requires a large bisection bandwidth that makes it difficult to greatly save power of Ethernet without experiencing performance degradation. The EP application maintains its performance even if the power consumption of switches is reduced by 26%, because it generates a small number of frames. It can be said that the tuning of the *trf* parameter strongly affects the performance of the application. However, along with the results from using 64 and 128 hosts, for an on/off link selection algorithm whose *trf* parameter is $zero$, all applications maintain their performance.

7.2.3 Power Estimation of Other Switches

We estimated the power consumption of the switch, Dell PowerConnect 6248 which is non-blocking in the previous subsection. The other 24-port switches, Dell PowerConnect 5324 and Planex SF-0420G, are also non-blocking, and their performance is close to that of Dell PowerConnect 6248.

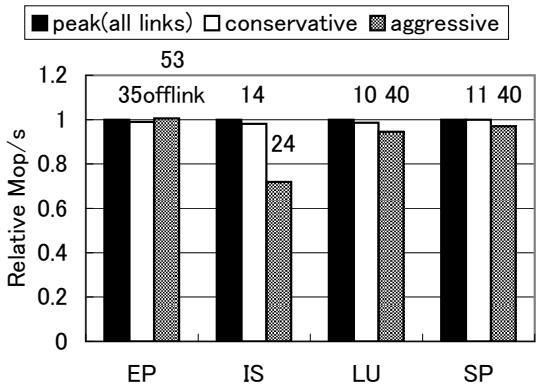
In the case of the evaluations using 64 hosts in the previous subsection, each switch used 23 ports: 8 ports for connecting hosts, and the remaining ports for connecting the neighboring switches. Thus, we can use the other 24-port switches, Planex SF-0420Gs or Dell PowerConnect 5324s, instead of PowerConnect 6248s in the 64-host evaluations on the cluster. Figure 9 estimates their total power consumptions using the results in Tables 2.

As shown in Figure 9, the power consumption can be reduced by up to 37% in the other switches, and these results show that the on/off link activation method strongly improves the power consumption of commercial Ethernet switches in the PC cluster.

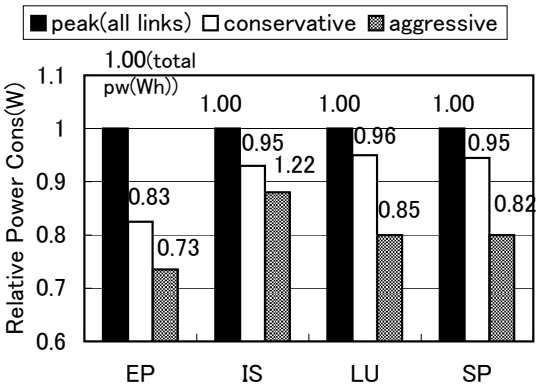
Figures 8(b) and 9 show the proposed technique efficiently reduces the power consumption of the simple L2 switch (such as PowerConnect5324) compared with that of the sophisticated L3 switch (such as PowerConnect6248), since various services that consume the power are always running on sophisticated switches regardless of link status.

8 Conclusions

The power consumption of the interconnects is increased as the link bandwidth is improved in PC clusters. In this paper, we proposed an on/off link activation method that uses a static analysis of the traffic in order to reduce the power consumption of Ethernet switches while maintaining the performance of a PC cluster. When a link whose utiliza-

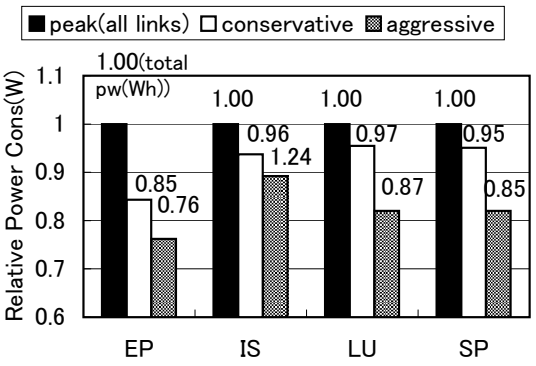


(a) Performance Evaluation

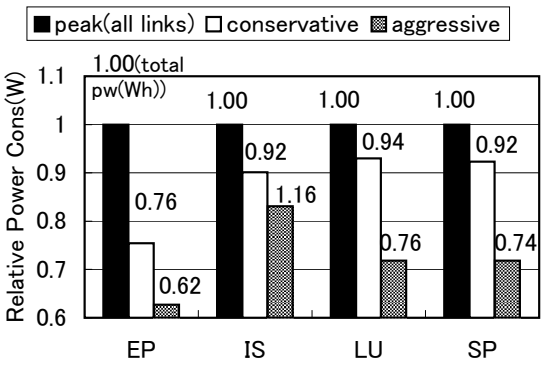


(b) Power Consumption Evaluation (PC6248)

Figure 8. NPB Benchmarks (64 Process, Class C)



(a) SF-0420G



(b) PowerConnect 5324

Figure 9. Power Consumption (64 Processes, NPB Class C, Other Switches)

tion is low is deactivated, the proposed method renews the VLAN-based paths that avoid it without creating broadcast storms. Since each host does not need to process VLAN tags, the proposed method has advantages in both simple host configuration and high portability.

The evaluation results using NAS Parallel Benchmarks show that the proposed method reduces the power consumption of switches by up to 37% with no performance degradation.

We are planning to extend and implement the proposed power-saving technique that will vary the port speed from 100Mb to 10Gb for the further power reduction, and its tool for PC clusters with Ethernet.

Acknowledgments

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). The authors would like to thank Prof. Tomo Hiroyasu, Dosisha University, for allowing us to use the SuperNova PC cluster.

References

[1] M. Alonso, J. M. Martinez, V. Santonja, P. Lopez, and J. Duato. Power Saving in Regular Interconnection Networks Built with High-Degree Switches. In *International Parallel and Distributed Processing Symposium*, 2005.

[2] InfiniBand Trade Association. `http://www.infinibandta.org/`.

[3] IPTraf: IP Network Monitoring Software. `http://iptraf.seul.org/`.

[4] 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, 18(3):320–333, Mar. 2007.

[5] 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 2004 IEEE International Conference on Cluster Computing (Cluster2004)*, Sept. 2004.

[6] T. Otsuka, M. Koibuchi, A. Jouraku, and H. Amano. VLAN-based Minimal Paths in PC Cluster with Ethernet on Mesh and Torus. In *Proc. of the 2005 International Conference on Parallel Processing (ICPP-05)*, pages 567–576, June 2005.

[7] T. Otsuka, M. Koibuchi, T. Kudoh, and H. Amano. A Switch-tagged VLAN Routing Methodology for PC Clusters with Ethernet. In *Proc. of the 2006 International Conference on Parallel Processing (ICPP-06)*, pages 479–486, Aug. 2006.

[8] PC Cluster Consortium. `http://www.pccluster.org/`.

[9] 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 2004)*, pages 2175–2184, Mar. 2004.

[10] S.-A. Reinemo and T. Skeie. Effective Shortest Path Routing for Gigabit Ethernet. In *IEEE International Conference on Communications (ICC)*, pages 6419–6424, June 2007.

[11] 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*, pages 79–90, Jan. 2003.

[12] S. Sharma, K. Gopalan, S. Nanda, and T. Chiueh. Viking: A Multi-Spanning-Tree Ethernet Architecture for Metropolitan Area and Cluster Networks. In *Infocom*, pages 2283–2294, Mar. 2004.

[13] 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*, 18(3):393–408, Mar. 2007.

[14] J. M. Stine and N. P. Carter. Comparing Adaptive Routing and Dynamic Voltage Scaling for Link Power Reduction. *IEEE Computer Architecture Letters*, 3(1):14–17, Jan. 2004.

[15] Top 500 Supercomputer Sites. `http://www.top500.org/`.

[16] S. Urushidani, S. Abe, K. Fukuda, J. Matsukata, Y. Ji, M. Koibuchi, and S. Yamada. Architectural Design of Next-generation Science Information Network. *IEICE Transaction*, E90-B(5):1061–1070, May 2007.

[17] T. Watanabe, M. Nakao, T. Hiroyasu, T. Otsuka, and M. Koibuchi. The Impact of Topoloy and Link Aggregation on PC Cluster with Ethernet. In *Proc. of IEEE International Conference on Cluster Computing (Cluster2008)*, pages 80–85, Sept. 2008.