Vol.2009-ARC-184 No.22
2009/8/5

An On/Off Link Regulation for Low-Power InfiniBand

Jose Miguel Montañana ,[†1] Michihiro Koibuchi ,[†1]
Takafumi Watanabe ,[†2] Tomoyuki Hiroyasu ,[†3]
Hiroki Matsutani [†4] and Hideharu Amano [†5]

Recently, the optimization of the power consumption has become one of the crucial factors to design PC clusters. In this paper, we propose an on/off link regulation method that dynamically deactivates and reactivates the links for power regulations with maintaining the performance in InfiniBand which is a typical system area network used in PC clusters. The proposed method only requires the use of the APM (Automatic Path Migration) and the SMA (Subnet Manager Agent) for monitoring the traffic load. Thus, no modification is required on the hardware of switches, host-channel adapters, nor drivers. Since the available network resources (switch and links) are changed by link activation, the path set is updated by modifying forwarding tables of InfiniBand switches. Evaluation results show that the proposed method reduces the power consumption of InfiniBand by 13% with no performance degradation.

1. Introduction

InfiniBand architecture (IBA) has been used for interconnection networks of high-performance PC clusters because of its high performance-per-cost. High-throughput (non-blocking) commercial InfiniBand switches are now available, and its link bandwidth has rapidly increased, such as SDR (2.5Gbps)/DDR(5.0Gbps) multiplied by 4, 8, and 12. Such rapid improvement of the link bandwidth also increases the power consumption, and the ratio for interconnects in the total power consumption of PC clusters has been grown up. Thus, the power saving techniques of interconnects have become one of the most important research topics for building PC clusters.

The 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. In the case of InfiniBand, the power consumption of the links can be saved by using the port-shutdown operation provided in some commercial switches, or management state change commands stated in IBA specifications. For example, the link deactivation decreases approximately 0.95W per port in the SFS7000D-SK9, which is a 24-port InfiniBand switch, when the link speed is DDR multiplied by 4. It can reduce the total power to 43W when all ports are shutdown. The port-shutdown operation and the management state change commands are not originally intended to reduce the power consumption, and they are normally used to block the injection of unexpected packets from the neighboring switches.

In this paper, we propose a link regulation method by such on/off operations in order to optimize the power consumption of InfiniBand switches according to the traffic load, while maintaining the performance. 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 independently selected, link activation or deactivation, the available network resources (switch and links) are changed. The path set is thus updated by modifying forwarding tables of InfiniBand switches.

The proposed method only requires the use of the APM (Automatic Path Migration) and the SMA (Subnet Manager Agent) for monitoring the traffic load. Thus, no modification is required on the hardware of switches, host-channel adapters, nor drivers. Since the available network resources (switch and links) are changed by link activation, the path set is updated by modifying forwarding tables of InfiniBand switches.

The rest of this paper is organized as follows. In Section 2, we briefly introduce related work. The on/off link regulation method is described in Section 3. In Section 4, we evaluate it using a flit-level simulator. Our conclusions are shown in Section 5.

2. Related Work

On/off interconnection networks have been studied for both off-chip and on-chip

†1 National Institute of Informatics
†2 Internet Initiative Japan Inc.
†3 Graduate School of Engineering, Doshisha University
†4 Research center for advanced science and technology (RCAST), The University of Tokyo
†5 Graduate School of Science and Technology, Keio University

communications, each of which has different wake-up time and break-even time to gain the link deactivation[6] They describe the theory of on/off link regulations and routing strategy that avoids deactivated links. Although there are some Internet-backbone switches supporting on/off link operation, such as AX6000s that require the reboot for changing the low-power mode[1], the details of the procedure to employ them in commercial system area networks (SANs), such as InfiniBand, are not shown.

After links are deactivated, paths are updated in order to avoid them. Thus, routing algorithms should be applied to not only regular topologies but also irregular topologies by deactivating links. SAN is a loss-less network that requires deadlock-free routing algorithms. The interconnection adaptivity by the link activation makes routing algorithm difficult to establish deadlock-free routing paths. There are two existing routing approaches for bypassing deactivated links; fault-tolerant routing and routing algorithms for irregular topologies. The former sometimes requires additional network resources, such as virtual channels. On the other hand, the latter can be used to arbitrary topologies with almost no additional network resources, and it is usually based on an embedded spanning tree. It exploits the connectivity and acyclicity of the tree structure, and a typical routing called up*/down* has been widely used to avoid deadlocks in SANs.

The power model of InfiniBand switch has been analyzed[8]. In addition, the low-power router architectures based on DVFS (dynamic voltage and frequency scaling) have been discussed in terms of their performance and power consumption[5),7]. However, the architecture of most commercial InfiniBand switches is black-box from both users and operators. Thus, users and operators are difficult to manage the internal modules of switches.

3. On/Off Link Regulation Method

In this section, we propose a method for on/off link activation in InfiniBand.

In the InfiniBand network, each subnet is managed by a centralized way with the *Subnet Manager* (SM), which can be located in any network device. The SM is the entity that discovers all of the network devices in the subnet at startup time to obtain the topology information and updates routing tables. In each network device, there exists a Subnet Manager Agent (SMA), which is responsible for monitoring port's link integrity. It also performs a periodic sweep of the subnet to detect any change in its topology, by sending control packets to the SMAs.

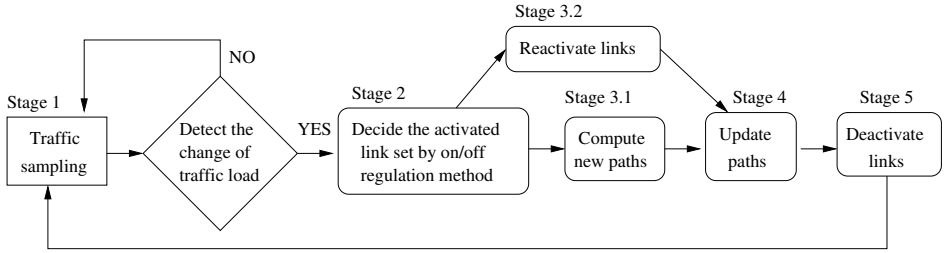

Fig. 1 Outline of on/off link regulation method

3.1 Outline

We show the outline of the on/off link regulation method in Figure 1.

- Stage 1: Monitor and sample the network status.
- Stage 2: When detecting the change of traffic load, decide the activated link set by the on/off link selection algorithm.
- Stage 3: Compute the new paths that avoid deactivated links, and reactivate the links stated in stage 2 in parallel.
- Stage 4: Update the routing paths.
- Stage 5: Deactivate the link set which was not selected in stage 2.

The SM simply completes all stages, and reactivating or deactivating the links. The on/off link selection algorithm, and routing algorithm corresponding to the stages 2 and 3 can be programed to the SM.

3.2 Traffic Sampling

The IBA defines a traffic sampling support for each link using control packets. The sampling counts (1) the amount of data sent and received, (2) the number of packets sent and received, and (3) the transmit queue depth at the start of the interval which can be defined between 4ns and 1us.

When a head-of-line (HoL) blocking occurs in a congested region, the amount of packets sent and received could be small on the congested link. If using only the information of the amount of packets, the SM could regard such a link status as unoccupied. Fortunately, since the transmit queue is fully used in the case of the HoL blocking, SM can correctly judge the network status by the comprehensive

sampling using the above three factors. Subnet Manager Agent (SMA) associated to each switch or host-channel adapter (HCA) manages the sampling results of its port, and the SM can request sampling measures by sending a special control packet to each SMA. The control packets between SMA and SM have a higher priority than the other data packets, since the IBA provides a special mechanism called Directed-Route that transfers the control packets through a reserved virtual channel (VL15). Thus, the control packets will not be delayed so much to get to the destination (SMA or SM), even when a network is crowded. According to the IBA specifications, each control packet is routed by the source routing, and the SM can set the path that avoids any disconnected network component (link or switch).

The SMA cannot monitor the traffic amount for each source-destination pair. On the other hand, communication monitoring of each process for each parallel application at MPI layer is obtained by adding modification of execution environment, and it provides the network 5-tuples, such as, time, source, destination, amount of data, and type. However, to use the traffic monitoring at MPI layer, the target parallel applications are limited, and additional function that sends the information of MPI traces to the SM is needed. Thus, we simply use the existing traffic monitoring tool using the SMA.

3.3 On/Off Link Selection Algorithm

Depending on the operation (link activation or deactivation) which is independently selected, the available network resources (switch and links) are changed. Thus, the paths that avoid deactivated links should be re-computed in this stage.

IBA supports deterministic routing using virtual lanes (VL). Up to 15 VLs can be implemented, and the VL selection is specified by a service-level (SL) label in each packet header. However, fixing a path with a unique SL could lead to a mapping conflict. It occurs when two packets labeled with the same SL enter a switch through the same input port. and they required to be routed through the same output port but along different VLs.

To simply avoid the VL and SL mapping conflicts, we apply existing deadlock-free routing algorithms for irregular topologies without the VL transition for computing the paths, such as up*/down* routing with multiple spanning trees[3].

3.4 Updating the Paths

The IBA states that the SM entity is responsible for computing the forwarding tables. IBA allows each destination node to be identified with up to 128 different virtual addresses with different destination label identifiers (DLIDs). It is achieved by adding up to 7 bits to the ID (LID) which are masked at destination. The forwarding tables on switches are able to store the required entries for all the virtual addresses. Thus, old and new paths when available network resources are changed can be identified by using the different virtual addresses. Both paths can be easily managed by the Alternative Path Migration mechanism (APM)[4] which allows to migrate to an alternative ID for each destination independently.

Figure 2 shows an example of APM on a host which has two alternative DLIDs, each one for routing in a different path. The APM simply switches paths by specifying the DLID, and two DLIDs are used to each destination in the on/off link regulation method.

The APM, which is implemented on each IBA device, provides a fast mechanism for the migration of the initial path to the alternative one. Once the path migration is completed, the APM mechanism can be loaded in advance with an alternative DLID before the next migration could be required. Notice that DLID is stored in the packet header and cannot be changed.

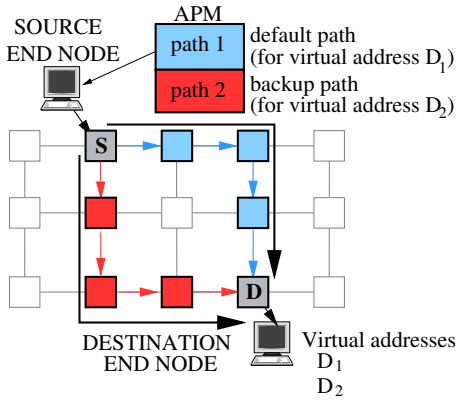

Fig. 2 Example of selecting alternative paths at APM.

Here, we apply to use the APM mechanism for managing old and new routing paths when the topology is updated by activating and deactivating links.

When the SM has completed the sending of the new routing information, it will wait for the reception of a confirmation from all the switches whose tables have been modified. This is required in order to ensure that packets can be appropriately routed when using the new LIDs.

3.5 Applicable Commercial Switches

There are various commodity InfiniBand switches whose costs range from a thousand dollars to hundreds thousand dollars. Their number of ports ranges from four to thousands, and their link speed is SDR/DDR multiplied by 1, 2, 4, or 12. Although InfiniBand switches follow the specifications of IBA, all the functions are not always supported especially in low cost switches. To apply the on/off link regulation method, the following functions are needed in target switches; (1) on/off link operation can be performed by operators or users, and (2) forwarding tables can be managed by them.

4. Evaluation

In this section, we firstly show the power consumption of an existing InfiniBand switch. We also show the overhead of port-shutdown at the switch. Finally, we evaluate the network throughput with the on/off link activation method using a flit-level network simulator.

4.1 Power Consumption of a Switch

We measured a power consumption of an InfiniBand switch using a watthour meter, namely the System Artware's SHW3A. Table 1 shows the power consumption of the Cisco SFS7000D-SK9, which is a 24-port InfiniBand switch.

The "Single port (DDR, ×4)" is the power consumption of a single port whose speed is DDR multiplied by 4. This power can be cut by the port-shutdown operation. The "All except ports" is the power consumption of the switch when all the ports are shutdown. The results show that the power consumption of the InfiniBand switch can be reduced by the port-shutdown operation, by up to 34%.

Its power is almost constant regardless of the traffic injection rates. Although there exist various InfiniBand switches, their maximum power consumption can be taken from their specifications. Using these parameters, we can estimate their power consumption when the on/off link activation method is applied.

Table 1 Power consumption of the SFS7000D-SK9 switch [W]

All except ports	43.4
Single port (SDR, ×1)	0.21
Single port (DDR, ×1)	0.77
Single port (SDR, ×4)	0.26
Single port (DDR, ×4)	0.95
Maximum power consumption	66.1 (The ratio of ports is 34%)

Table 2 Overhead of on/off operation [sec]

	Ping delay
SDR, x1	2.8
SDR, x4	2.9
DDR, x1	4.9
DDR, x4	4.5

Although some commercial switches can change their link speed by specific commands, the IBA specifications do not define such commands. They only provide the management state change commands that performs the on/off link operations. Thus, the link speed (SDR/DDR multiplied 1, 2, or 4) is usually determined by the speed of the connecting devices.

4.2 Overhead of On/Off Operations

We measured the overhead of the on/off link operation at the switch. In the measurement of the overhead of the port-shutdown operation, while the ping command (ICMP message of 64 bytes) between two hosts is executed at intervals of 0.1 second on IP-over-InfiniBand environment, a switch continuously operates the shutdown and no-shutdown (resume) of a port. While the link is deactivated, the communication (ping frame) is interrupted and delayed. Here we regard the delay as the overhead of the on/off link operation.

As shown in Table 2, the overhead is several seconds which will fit with per-application or per-hour on/off operations.

4.3 Network Simulation

4.3.1 Simulation Environment

To evaluate the proposed mechanism, we have developed a detailed simulator that allows us to model the network at the register transfer level. The simulator models an IBA network, following the IBA specifications[4].

Packets are routed at each switch by accessing the forwarding table. This table contains the output port to be used at the switch for each possible destination. The routing time at each switch will be set to 100 ns. This time includes the time to access the forwarding tables, the crossbar arbiter time, and the time to

set up the crossbar connections.

For each simulation run, we assume that the packet generation rate is constant and the same for all the end-nodes. Except when specified, the number of simulated packets is 80,000 and results will be collected from the last 40,000.

We use two-dimensional torus as a baseline topology when all links are activated. The destination of a packet is determined by the traffic patterns, i.e., *uniform* or *bit-reversal*. In addition to the synthetic traffic, we use the trace of NAS Parallel Benchmarks (NPB). The class of problem was set to "W", and the numbers of tasks was 16 or 64.

The number of ports is assumed to 24, since we measured and modeled the power consumption of the 24-port InfiniBand switch. Two or more links are used between switches, while a single link is used between host and switch. Both link speed is set to DDR multiplied by 4, according to the specification of the target InfiniBand switch.

The crucial factor of the on/off link activation method is the routing algorithm, and on/off link selection algorithm. In this evaluation, we evaluated two patterns; the dimension-order routing and up*/down* routing that commonly used in high-performance computing systems with the simple approach to select deactivated links proposed in[2] that only regulate the number of links between switches.

4.3.2 Power Consumption

Figure 4.3.2 estimates the power consumption of networks using values of Table 1. When the number of activated links is decreased, the power consumption is linearly reduced by up to 20%. The power consumption of the InfiniBand switches only depend on the number of the activated links. As the network size increases, the reduction rate of the power consumption is almost constant.

4.3.3 Performance Evaluation

We evaluate the throughput of the on/off link activation method when the traffic injection rate is varied. We have evaluated also the synthetic uniform and bit-reversal traffic patterns on 8×8 torus in Figures 5.a and 5.b, respectively. The "bt"and "cg" are applications taken from NPB in Figure 4.3.3.

In each figure, X-axis is the number of deactivated links from the original mesh or torus topology, while Y-axis is the average reception traffic rate at a destination node that is referred as throughput in this paper.

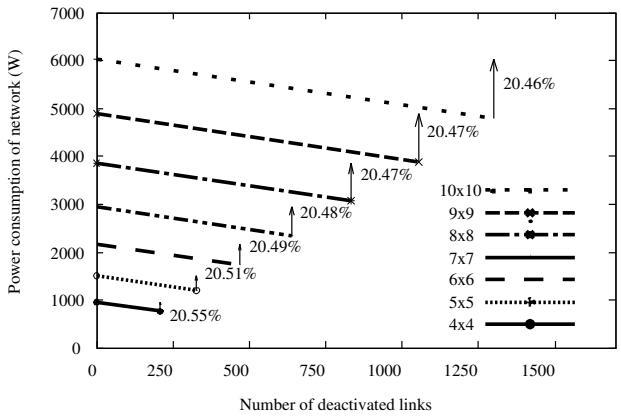

Fig. 3 Power consumption of networks 2-D torus

As observed in all the figures (4.3.3, 5.a, and 5.b), little decrease of network performance allows to deactivate a larger number of links. Thus, a large number of links can be deactivated without decreasing the network performance, and the total power consumption of IBA switches can be reduced down to 13%.

In the case of uniform traffic, the throughput is linearly reduced, as the number of deactivated links increases. However, in the other traffic patterns, the throughput is maintained, until the number of deactivated links is quite large. Thus, when the traffic pattern includes the access locality, the on/off link activation method can reduce the power consumption of networks with maintaining the throughput.

5. Conclusions

The optimization of the power consumption is one of the crucial factors to design PC clusters, as well as the performance improvement. In this paper, we proposed an on/off link regulation method that dynamically deactivates and reactivates links for power regulations with maintaining the performance in IBA.

The proposed method only modifies the configurations of the APM (Automatic Path Migration) that manages forwarding tables of switches and the SMA (Sub-

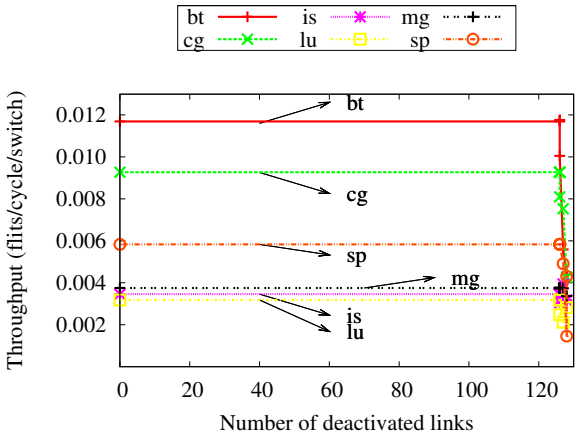

Fig. 4 Throughput of networks 4×4 torus, NPB traces

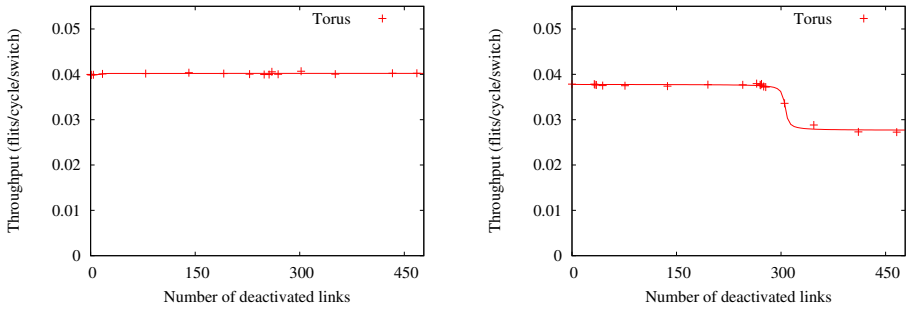

Fig. 5 Throughput of networks in 8×8 torus (a) uniform, and (b) bit-reversal traffic

net Manager Agent) for monitoring the traffic load. Thus, no modification is required on the hardware of switches, host-channel adapters, nor drivers. Since the available network resources (switch and links) are changed by link activation, the path set is updated by modifying forwarding tables of InfiniBand switches through Subnet Manager. Evaluation results show that the proposed method reduces the power consumption of IBA by 13% with no performance degradation.

We are planning to implement the proposed on/off link regulation method in a small InfiniBand network.

Acknowledgment

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).

References

1) ALAXALA Networks. http://www.alaxala.com/en/index.html.
2) 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*, page5b, 2005.
3) J.Flich, P.Lopez, J.C. Sancho, A.Robles, and J.Duato. Improving infiniband routing through multiple virtual networks. In *Proceedings of the International Symposium on High Performance Computing*, pages 49–63, 2002.
4) InfiniBand Trade Association™. *InfiniBand Architecture specification release 1.2*, October 2004.
5) 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.
6) 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.
7) 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.
8) H.-S. Wang, L.-S. Peh, and S.Malik. A Power Model for Routers: Modeling Alpha 21364 and InfiniBand Routers. *IEEE Micro*, 23(1):26–35, Jan. 2003.