Cluster Computing manuscript No.
(will be inserted by the editor)

Imbalance of CPU Temperatures in a Blade System and its Impact for Power Consumption of Fans

Yuetsu Kodama · Satoshi Itoh · Toshiyuki Shimizu · Satoshi Sekiguchi · Hiroshi Nakamura · Naohiko Mori

submitted: 15 April 2011, revised: 25 July 2011

Abstract We are now developing a new metric of data center power efficiency to fairly evaluate the contribution of each improvement for power efficiency. In order to develop it, we built a testbed of a data center and measured power consumption of each components and environmental variables in some detail, including the power consumption and temperature of each node, rack and air conditioning unit, as well as load on the CPU, Disk I/O and the network. In these measurements we found that there was a significant imbalance of CPU temperatures that caused an imbalance in the power consumption of fans. We clarified the relationship between CPU load and fan speed, and showed that scheduling or rearrangement of nodes could reduce the power consumption of fans. We reduced fan power consumption by a maximum of 62 % and total power consumption by a maximum of 12 % by changing the scheduling of five nodes, changing the nodes used from hot nodes to cool nodes.

Keywords Data center · Power consumption · Power efficiency · CPU temperature · Fan speed

Y. Kodama
Graduate School of System and Information Engineering, University of Tsukuba E-mail: kodama@cs.tsukuba.ac.jp

S. Itoh, T. Shimizu, S. Sekiguchi
Information Technology Research Institute, National Institute of Advanced Industrial Science and Technology (AIST)

H. Nakamura
Department of Information Physics and Computing, The University of Tokyo

N. Mori
Applied Network Integration Business Unit, NTT Advanced Technology Corporation
Innovative IP Architecture Center, NTT Communications Corporation

1 Introduction

Power consumption in data centers is increasing year by year and has now become a major issue. Many researchers are grappling with this problem using various approaches, such as development of low power servers, more efficient power supplies / cooling facilities, and more efficient server operation. Because a data center consists of many types of IT equipment and facilities, power consumption and heat transmission are mutually related. Thus it is very difficult to evaluate how much an improvement contributes to overall power reduction. We are modeling data center power consumption and developing a framework which can evaluate individual contributions to total power consumption. This framework will help data center managers to optimize their investment.

In order to establish the evaluation framework, we are now trying to construct a new metric of data center power efficiency by dividing the power consumption not into components such as IT equipment, power facility and cooling facility, but into functions such as processing, power supply and heat removing. In addition to the model, we constructed a small physical environment where the temperature and power consumption of IT equipment and cooling facilities can be measured in some detail. The validity of the model and the evaluation framework are expected to be verified by analyzing actual measurements. We ran several benchmarks in this environment, and measured the power consumption and temperature of several components. In these measurements, we found that there was a significant imbalance of CPU temperatures and this caused an imbalance in the power consumption of fans. We also clarified the relationship between CPU load and fan speed, and pointed out a feasible power reduction scheme by considering the imbalance of CPU temperatures.

The remainder of this paper is organized as follows: In section 2 shows our proposed metric for data center power efficiency. In section 3, the environments, including the facilities involved and the measurement methods used, are described. Section 4 shows the imbalance of CPU temperatures and the speed of fans in our measurement results, and section 5 presents a look at the relationship between CPU load and fan speed. In section 6 we consider two schemes for power reduction, one is based on selecting the nodes to be invoked, and the other is based on rearrangement of node positions in a blade system, and we show the power reductions achieved. We discuss related work in Section 8, and conclude our paper in section 9.

2 A new metric for data center power efficiency

PUE (Power Usage Effectiveness)[1] is a well-known metric for energy efficiency in data center and is useful to understand the current situation of the data center. However, PUE is too macroscopic to evaluate a contribution of an improvement. New metrics, DCeP (Data Center energy Productivity) by TGG (The Green Grid)[2] and DPPE (Data center Performance Per Energy) by GIPC (Green IT Promotion Council)[3], which contain the concept of productivity versus energy, have the same problem with PUE. Server performance-power metric is also useful metric for greenness of server and computer system. However, there are pitfalls in these metrics.

For example, there are two data centers A and B. Data center A has servers which have large fans and PSU (Power Supply Unit) with UPS. Because the large fans generate strong airflow, CRAC (Computer Room Air Conditioners) is enough to generate week airflow with cold air. Small size of UPS is attached for each server. Thus power loss by power distribution as a facility is not so large. On the other hand, Data center B has servers which do not have fans, AC/DC converter and UPS inside servers. CRAC or other cooling facility generates strong airflow which can remove heat from servers. Power facility converts AC to DC, storages electricity and distributes DC to servers directly. Thus power consumption and power loss in the facilities are large. Both data centers have the same productivity and total power consumption. However, PUE indicates that data center A is better than B and sever performance-power metric indicates that servers in data center B is better than A.

PUE requests to know total power consumption of IT equipment and facilities. Server performance-power metric requests to know power consumption of individual server. However we propose to decompose Data Center by function in order to evaluate energy efficiency more accurately and fairly in Fig. 1. In this case, power consumption/loss of fans and PSU in a server has to be measured. Power consumption of fans and CRAC are added up as power consumption for heat removing function in a data center. Power loss of PSU and power facilities, such as UPS and AC/DC converter, are added up as power loss for power supply function in a data center. We propose to decompose Data Center by function in order to evaluate energy efficiency more accurately and fairly. By using power consumption for functions, we propose modified metrics as follows.

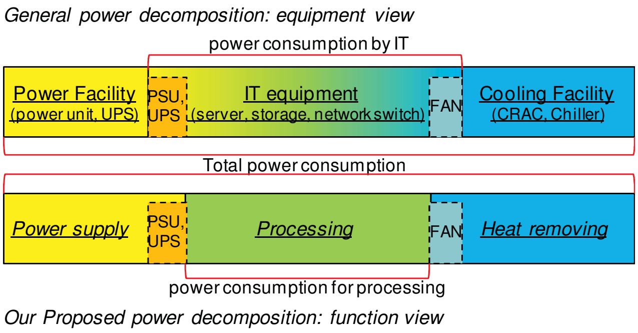

Fig. 1 Data center power model

Modified PUE = total power consumption

/ power consumption for processing

Modified server performance-power metric =

performance / power consumption for processing

It is expected that these metrics give the same value for both Data center A and B.

Generally it is not easy to measure power consumption of fans and PSU in a server. However, recent blade system mounts many sensors to monitor power consumption of fans and PSU. Split of power consumption to components is possible. Rack-mount servers are majorly used in a data center. Recent servers have sensors for fan speed, but may not have sensors for power consumption of fans. By changing load for server and conditions of environment, it is possible to guess power consumption of fans in a rack-mount server.

3 Environment

In order to verify our model of the power consumption of a data center, we constructed a small environment where the temperature and power consumption of IT equipment and cooling facilities can be measured in some detail[4]. This environment is located in a part of our cluster room, where our large scale clusters were located until this April. We covered a booth with aluminum frames 2000mm in height and plastic curtains, as shown in Fig. 2. Although the room has sixteen CRACs (Computer Room Air Conditioners), we used 4 CRACs for our environment and divided the space under the floor by walls so that the cold air from the CRACs blew only for our booth. We controlled each CRAC individually, allowing us to control the number of active CRACs, the volume of airflow and the temperature of the airflow.

Six racks for IT equipment with 3 x 2 columns each were located in the booth. One rack was used for two blade systems, and five racks were used for 1U servers. Each rack for the 1U servers contained twenty 1U servers assigned as Web servers, two 1U servers as DB servers, one storage unit, two network switches and a KVM switch. Racks were placed back to back, and spaces without equipment in a rack were filled by fillers in order to prevent mixing of cold and hot air, but we did not use aisle capping outside of the racks. We measured the voltage, current and power consumption of each node individually every second. We also measured the total power consumption of the servers on a rack simultaneously. We measured the temperature. We placed temperature sensors at 87 points in and out of racks with three levels of height position in the booth. We monitored the temperature of servers and CPUs using IPMI or the sensors command, and also monitored the load of each server by means of the sar command.

One rack included two blade enclosures. They were HP BladeSystem c7000 units, and each of them included sixteen PROLIANT BL460c blade nodes. One of the blade systems had an Intel Xeon 5160 (3.00GHz, 2core) on a blade node, and the other had dual CPUs on a blade. We used dual CPUs nodes in the following evaluation. The HP Blade enclosure includes an OA (Onboard Administrator) that provides us with various items of information, such as the power consumption of the enclosure, fans and switches, as well as the temperature of the enclosure, each blade, switches, etc. The blade enclosure has six power supply units (PSUs). Their status, such as active or standby, is dynamically controlled and the OA provides us with the PSU status, the AC power input for the PSUs and DC power output from each PSU. The blade enclosure has ten fan units, and the OA also provides us with the speed and power of the fans. The power consumption and value of the virtual fan of each node, that is described later, are monitored by IPMI command.

4 Measurements

We ran the LINPACK benchmark on each node of the blade system repeatedly for about two hours. The LINPACK is a dense N by N system of linear equations solved by Gaussian elimination with partial pivoting. It is optimized for floating point computing, and the load of the FPU is a major part of total power consumption, so the power consumption under LINPACK will be almost the maximum possible in the system. We used the Intel LINPACK package, linpack_10.2.5[5], as a program, and ran it on a single node with 4 cores. Since the memory size of each node was 2 GBytes, the array size was set to 15000. The average performance of LINPACK was 36.4 GFLOPS per node. The performance is 76 % of theoretical peak performance. The average elapsed time of a run was 61.8 seconds. Fig. 3 shows the LINPACK performance of each node. We ran LINPACK more than 100 times in two hours, and the graph shows the maximum, minimum and average of the performance. The variation of LINPACK performance repeated on a node was very small. The standard deviation was 0.17. The variation

Fig. 2 Photograph of our Booth

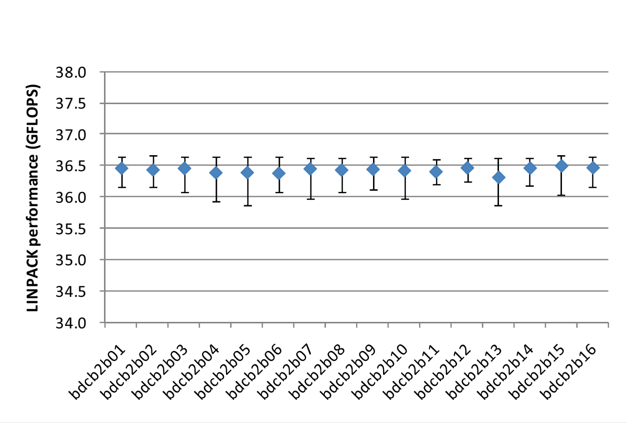

Fig. 3 LINPACK performance of each node and runs

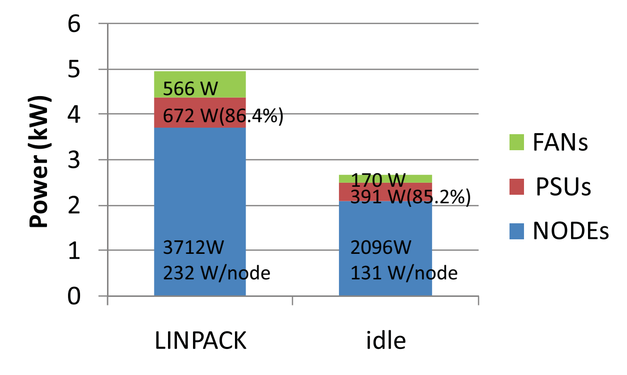

Fig. 4 Total Power Consumption using LINPACK

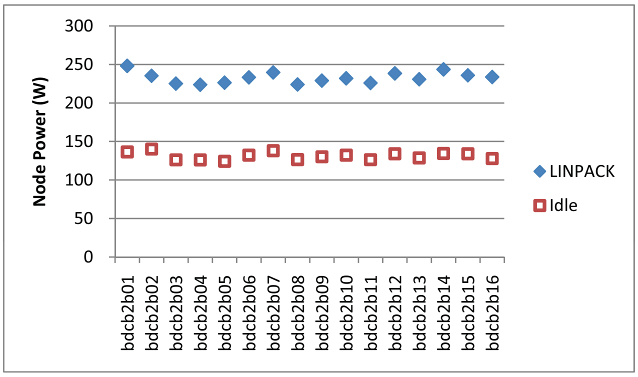

Fig. 5 Power consumption of each node using LINPACK

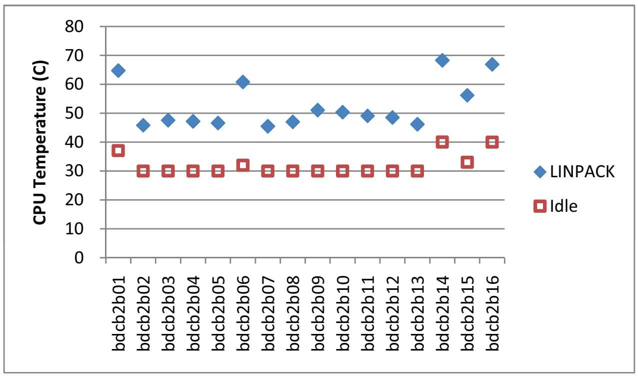

Fig. 7 CPU temperature of each node using LINPACK

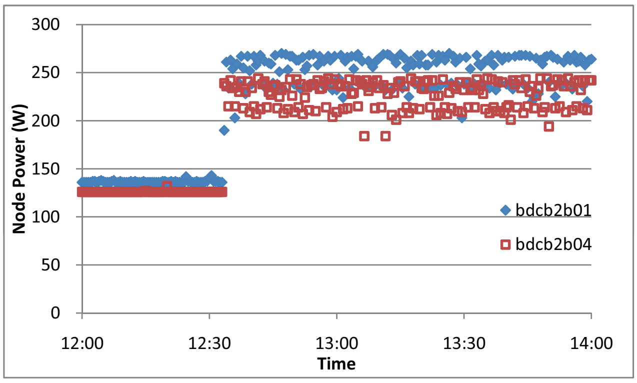

Fig. 6 Power consumption of node1 and node4 using LINPACK

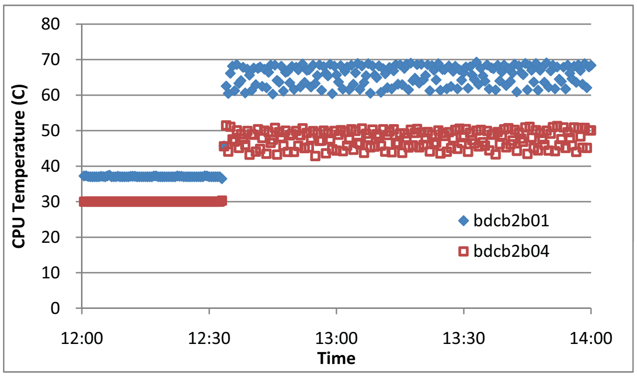

Fig. 8 CPU temperature of node1 and node4 using LINPACK

of average performance between nodes was also small, and the standard deviation was 0.045.

We measured the power consumption of nodes, CPU temperature, fan speed, etc., while running LINPACK, and compared with the idle state. Fig. 4 shows the power consumption both while running LINPACK and idle. In the figure, total power is divided to power of fans, loss of PSUs and power of nodes. The power of fans is the sum of power consumption of the 10 fans in the blade enclosure. The loss of PSUs is the sum of difference of Input AC power and Output DC power of the 6 PSUs. The value in the parentheses is the efficiency of PSUs, which is the ratio of output power to input power. The power of nodes is the sum of power consumption of the 16 nodes. The average power consumption of a node is also shown. This information was monitored every 2 seconds through the OA (Onboard Administrator).

While the total power consumption is 2649 W at idle, it increased to 4944 W during a LINPACK run. The loss of PSU increased from 391W to 672W, but the efficiency of a PSU was high and almost the same as at idle. This is because PSUs were controlled with dynamic load balance. During the idle time, only 4 PSUs were active, and the remaining two PSUs were on standby. Six PSUs were active while running LINPACK. The increase of total fan power was 396 W, and the average increase of Node Power is 101 W per node. The increase of total fan power was large, and was almost same as the increase of all four nodes.

Fig. 5 shows the average power consumption of each node. The average power consumption among nodes during idle was 131 W, the difference between the maximum and minimum was 15.9 W, and the standard deviation was 4.8, 3.7 % of the average. On the other hand, while running LINPACK, the average among nodes was 232 W, the difference between the maximum and minimum was 24.4 W, and standard deviation was 7.2, 3.1 % of the average. The power of each node was measured every 15 seconds by IPMI in order not to affect LINPACK performance. Fig. 6 shows the measured power with the measured time, where bdcb2b01 is the node name where power was the maximum and bdcb2b04 is the node name where power was the minimum during a LINPACK run. The data is an average of 30 seconds. The power consumption was minimized periodically during a LINPACK run. This is because the benchmark repeats a LINPACK and it takes about 60 seconds with the parameter used in this evaluation. Just when a LINPACK was finished, the load on the nodes became low. But power consumption during a LINPACK run was almost constant, and the difference between two nodes was also almost constant. The difference was small during idle and was large during a LINPACK run.

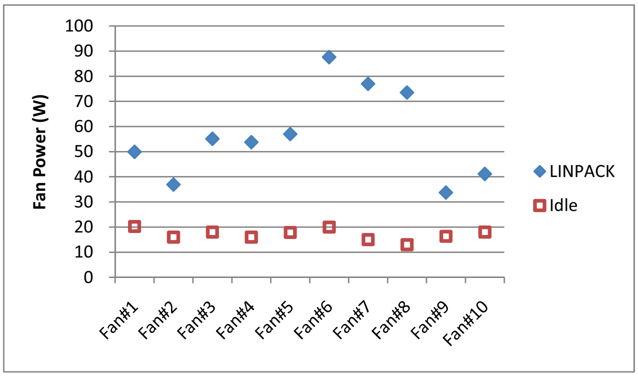

Fig. 9 Power consumption of fans using LINPACK

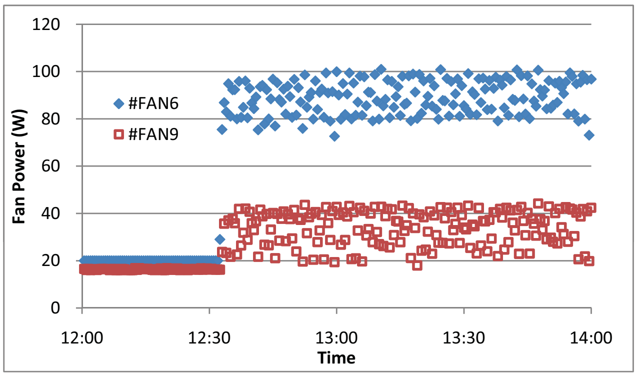

Fig. 10 Power consumption of FAN#6 and FAN#9 using LINPACK

The above results show that the performance of each node was almost the same but the power of each node had some variance. To check the reason and effect, the temperature of each CPU was compared. Fig. 7 shows the average temperature of CPUs on each node. The variance of temperature was larger than that of power. The overall average was 32.0 degrees during idle and 52.6 degrees during a LINPACK run. The difference between the maximum and minimum was 10.0 degrees during idle and 22.8 degrees during a LINPACK run. The difference became larger while LINPACK was running. CPU temperature was measured by the OA every 2 seconds. Fig. 8 shows the time variation of two nodes, where bdcb2b01 was a high temperature node and bdcb2b04 was a low temperature node. The data is an average of 30 seconds. Sometimes there were low values during a LINPACK run. This is for the same reason as that of the node power and the repeat timing of LINPACK. Otherwise, the value was relatively constant, and the difference between two nodes was also constant.

One possible reason for the imbalance of temperatures among nodes was the node position in a blade enclosure. To determine the effect of this, we replaced adjacent nodes, for example we exchanged node 1 and node 2 and node 15 and node 16, and ran LINPACK again. However, the results were almost the same, independent of position. We concluded this imbalance was in the nodes themselves, such as the parts lot, etc. First, we thought the imbalance itself had less impact on other items. However, we found it had significant impact on the power consumed by fans.

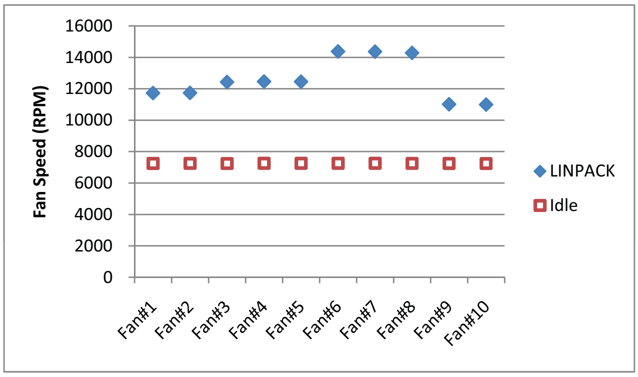

Fig. 11 Fan speed using LINPACK

Fig. 9 shows the power of ten fans in an enclosure. The variations in power consumption during idle was small, but the variations in power during a LINPACK run was quite large. The maximum power was 88 W for FAN#6, and the minimum power was 34 W for FAN#9. The power consumption of fans was monitored every 2 seconds by the OA. Fig. 10 shows the time variation of FAN#6 and FAN#9. The data is an average of 30 seconds. Although the values changed rapidly, the power consumption of FAN#6 was always larger than the power consumption of FAN#9.

The amount of power consumption of a fan depends on the speed of the fan. The speed of the fan may be the control parameter. Fig. 11 shows that average speed of each fan. The speed of fans was monitored every two seconds by the OA. The speeds of the fans were constant during idle. The speeds of the fans during a LINPACK run can be classified into several groups. FAN#6-#8 are the highest group, and FAN#3-#5 are the second highest group. The reason why there was an imbalance in the power consumption of fans while they were the same speed is based on differences in the voltage of fans, the efficiency of fans, etc.

5 Relation between the loads on nodes and fan speed

In this blade system, each node has no fans, and enclosure has fans. The number of nodes is 16, and the number of fans is 10, so nodes and fans are not corresponding one by one. First, we should clear the relationship between temperature of CPU and power of fans. We tried to get the relationship from the manuals of blade system, but we cannot find it. We found a parameter called Virtual FAN in a list of IPMI results from blade nodes. The values may be a good indicator for the requests made on fans in an enclosure.

Fig. 12 shows the average value of Virtual FAN for each node during a LINPACK run and idle. The values of Virtual FAN during idle are same among all nodes. The value of

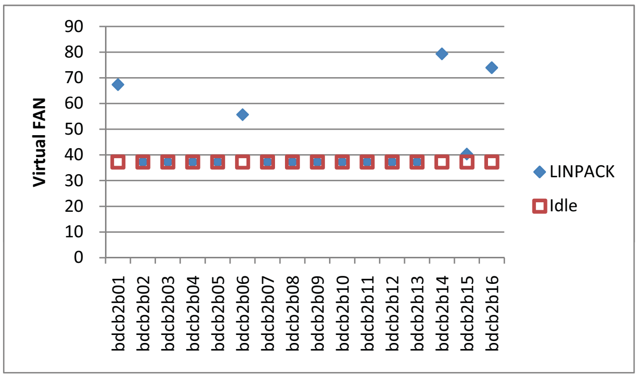

Fig. 12 Virtual FANs Values for nodes using LINPACK

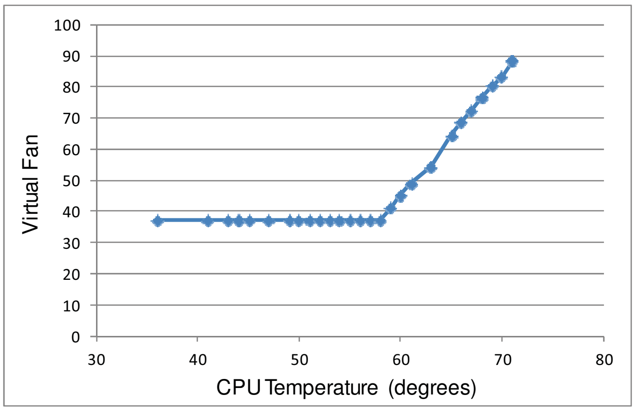

Fig. 13 Relationship between Virtual FAN value and CPU temperature

Virtual FAN did not change for any nodes except node 1, 6, 14, 15 and 16 during a LINPACK run.

Fig. 13 shows the relationship between CPU temperature and Virtual FAN. In the figure, Virtual FAN was a fixed value of 37 if the CPU temperature was less than 59 degrees Celsius. If the CPU temperature was larger than the threshold, Virtual FAN increased in proportion to the CPU temperature. If the value of Virtual FAN was increased, the speed of fans related to the node increased and thus the power consumption of the fans increased. If the CPU temperature was less than the threshold, the value of Virtual FAN was not changed, and thus the power consumption of the fans may also not increase.

While it was appeared that the CPU temperature controls the Virtual FAN with a threshold, it is not yet appeared that how the value of Virtual FAN controls each physical fan on an enclosure.

We ran LINPACK on nodes not simultaneously, but one by one, and checked the value of Virtual FAN. Fig. 14 shows the results. The x axis refers to time. We ran LINPACK from node 1 to node 16, one by one. The sub-grid labeled 1 specified the time when only node 1 ran LINPACK. The y axis refers to Virtual FAN of each node. Each value is an average in a minute. The value of a Virtual FAN increased only when the node ran LINPACK and it did not changed when other node ran LINPACK. The maximum value of Virtual FAN on each node was quite different, and the high temperature node had large value of Virtual FAN, such as node 1, 6, 14, 15 and 16.

Table 1 Relationship between a single node and fans

	FAN #									
node #	1	2	3	4	5	6	7	8	9	10
1	△	△	√	√	√	△	△	△	△	△
2	–	–	√	√	√	–	–	–	–	–
3	–	–	√	√	√	–	–	–	–	–
4	–	–	√	√	√	–	–	–	–	–
5	√	√	√	–	–	–	–	–	–	–
6	√	√	√	–	–	–	–	–	–	–
7	√	√	√	–	–	–	–	–	–	–
8	√	√	√	–	–	–	–	–	–	–
9	–	–	–	–	–	–	–	√	√	√
10	–	–	–	–	–	–	–	√	√	√
11	–	–	–	–	–	–	–	√	√	√
12	–	–	–	–	–	–	–	√	√	√
13	–	–	–	–	–	√	√	√	–	–
14	△	△	△	△	△	√	√	√	△	△
15	–	–	–	–	–	√	√	√	–	–
16	△	△	△	△	△	√	√	√	△	△

We also checked the speed of the fans. Fig. 15 shows the results. The x axis refers to time. We ran LINPACK from node 1 to node 16, one by one. The y axis refers to the speed of the fans, the top of the graph is FAN#1 and the bottom of the graph is FAN#10. The figure shows that the speed of only some of the fans increased, and the rest did not change when a single node was activated. For example, when node 6 ran LINPACK, only the speed of FAN#1, #2 and #3 increased but the others did not change. Nodes 1, 14 and 16 affected all of the fans, while the remaining nodes affected only three fans. This may be because the former's CPU temperature was very hot and the enclosure decided that the speed of all fans had to be increased. We supposed there may be two kinds of threshold of CPU temperature, the lower threshold affects the three fans near the node, and the higher threshold affects all fans. Table 1 summarizes the relationships. In the table, '√' indicates that the speed of the fan is increased and a '–' indicates that the speed of fan does not change when the node is activated. A '△' indicates that the speed of the fan may not be changed if the node of the CPU is not too hot although it was increased in Fig. 15 because the CPU was too hot.

The table can be summarized by saying the nodes and fans are classified into four groups and a group of nodes affects a group of fans. For example, nodes 5, 6, 7 and 8 make up a group of nodes (NodeGroup2), FAN#1, #2 and #3 make up a group of fans (FanGroup1), and a node in NodeGroup2 affected only the fans in FanGroup1 when the CPU temper-

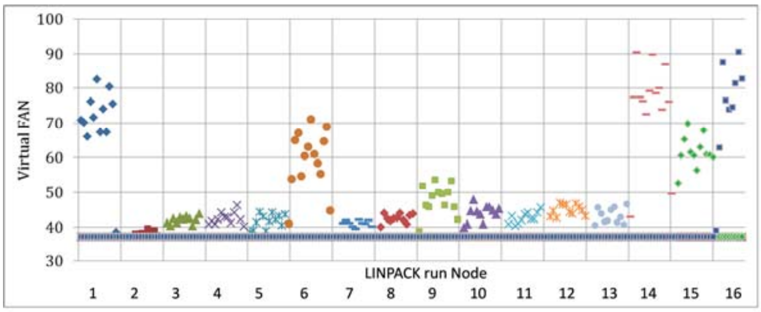

Fig. 14 Virtual FAN during a single node execution of LINPACK

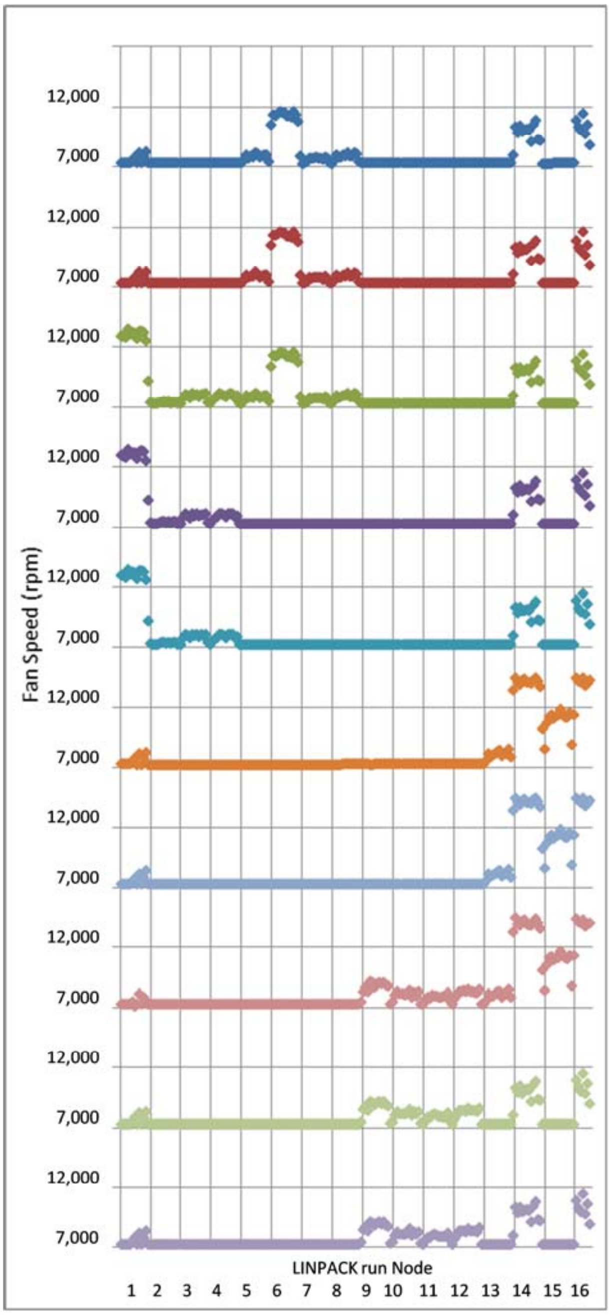

Fig. 15 Fan speed during a single node execution of LINPACK

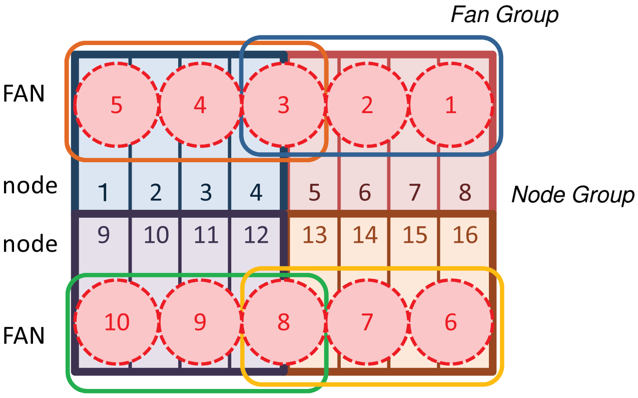

Fig. 16 Relationship between nodes and fans in a blade enclosure

ature was less than the high threshold. FAN#5 and #8 belong to two FanGroups. Fig. 16 shows the relationship. This figure is a front view of an enclosure, and nodes are numbered from left top to right bottom. At the same time, fans are placed in the rear and they are numbered from left top to right bottom in the rear view, so they are numbered from right top to left bottom in the front view.

By understanding that relationship between fan group and node group, we can understand why power consumption of fans was greatly different. In Fig. 9, the FAN#6 was the largest power consumption because it was affected from node 13 to node 16, and node 14 and node 16 was very high temperature in LINPACK runs. FAN#9 was the lowest power consumption because it was affected from node 9 to node 12, and all of four nodes were not high temperature.

6 Reduction methods for fan power consumption

As we shown in Fig. 4, the power consumption of fans was about 10% of total power consumption when nodes were active with high load. The power consumption of a fan depended on temperature of some of nodes, and there ware imbalance of temperature between nodes. We thought we could reduce the power of fans using the imbalance of node temperature. We propose two schemes, one is selecting nodes to

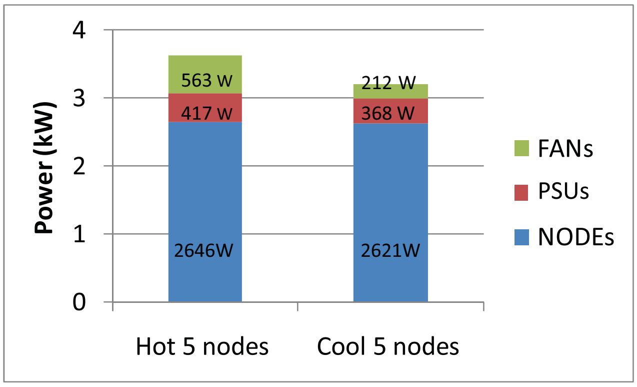

Fig. 17 Power consumption of node groups using LINPACK

be executed, and the other is changing node position in an enclosure to reduce the power of fans.

6.1 power reduction by scheduling nodes

First idea is that the power of fans will be reduced by selecting nodes when we use only some of nodes in a blade system. The power consumption of nodes itself was almost same as shown in Fig. 5, even if the temperature of nodes was different. On the other hand, the power of fans was quite different if the temperature of nodes was different as shown in Fig. 9.

We ran LINPACK only on 5 nodes in order to determine the effect of scheduling nodes for the power consumption of fans. We defined two groups, one is a hot 5 nodes, and the other is a cold 5 nodes. The hot 5 nodes included node 1, 6, 14, 15 and 16. These nodes were high temperature in LINPACK runs as shown in Fig. 7. The cold 5 nodes included node 2, 3, 4, 5 and 7. These nodes were low temperature in LINPACK runs as shown in Fig. 7. The results are shown in Fig. 17. In the figure, total power is divided to power of fans, loss of PSUs and power of nodes. It shows that the power of nodes was almost same between hot nodes and cool nodes, but the power of fans was greatly different. In the cool 5 nodes it was 212W, and this value was almost same as idle case. But in the hot 5 nodes it was 563W, and this value was almost same as all nodes active case. The reduced power consumption in fans was 351W, and this was about 62% of fan power. The loss of PSUs in the cool 5 nodes was also reduced about 10% according to the reduction of total power. This results showed that by selecting nodes to be executed, it reduced 425W of power in total, this was 12% of total power.

We also ran LINPACK using cool 11 nodes, which included all nodes except the hot 5 nodes. The power of fans was almost same as the above cool 5 nodes. Although the number of nodes in the 11 cool nodes was more than twice the number of nodes in the 5 hot nodes, the fan power consumption of the 11 cool nodes was smaller than that of the 5 hot nodes. The fan power consumption in the 11 cool nodes increased slightly in comparison with the idle case. This is because the value of Virtual FAN on the executed node increased from 37 to 41. In addition, as shown in Fig. 15, the fan speed increased when a single node was activated. In that case, the value of Virtual FAN for the node also increased. However, as shown in Fig. 12, Virtual FAN of the nodes except hot 5 nodes remained at the minimum value of 37 when all nodes ran LINPACK. This is because when all nodes were running, the 5 hot nodes increased the speed of all fans, so the CPU temperature of the 11 cool nodes remained under the threshold of Virtual FAN. When only the 11 cool nodes ran, there were no requests to increase the speed of all fans, so the CPU temperature of the 11 cool nodes increased and exceeded the threshold of Virtual FAN. But the increase of Virtual FAN was quite small and the power consumption of the fans also remained small.

From these results, we find that if we do not have to run on all nodes, we can reduce the power consumption portion of total power by scheduling jobs on the nodes where CPU temperature does not become too hot. As in the example above, if we select 5 nodes from the 11 cool nodes, the power consumption will be reduced by 350 W from the case where the 5 hot nodes are selected. This is about 10 % of total power consumption.

6.2 power reduction by rearranging node position

Above idea is only applicable when we use not all nodes, but some of nodes. Second idea is applicable even if we use all nodes. It is based on consideration that the increase of Virtual FAN for a node impacts only a group of fans.

We checked two types of rearrangement of node position. One is a concentrated positioning that gathers the hot nodes in a node group, and the other is distributed positioning that distributes the hot nodes to different node groups. In the concentrated scheme, since the effect from hot nodes is mainly limited to a fan group, the total power consumption of fans will be reduced. On the other hand, in the distributed scheme, since each hot node affects each group of fans, all the fans are affected, and the total power consumption of fans will increase.

Fig. 18 shows the results with different rearrangements of node position. A comparison of two node position arrangements revealed that the concentrated arrangement reduced the power consumption of fans by 310 W, the power loss of PSUs by 41 W and the total power consumption by 351 W from that of the distributed arrangement, while both power consumptions of nodes were the same. Fig. 19 shows the speed of each fan for each arrangement. In the distributed arrangement, the average speed of all fans was over 14000 RPM. On the other hand, in the concentrated arrangement,

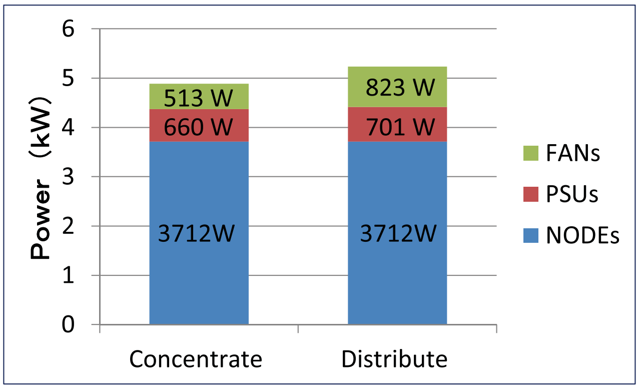

Fig. 18 Power consumption with position rearrangement

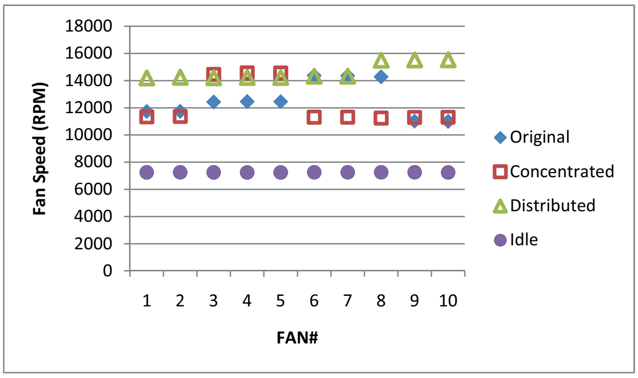

Fig. 19 Fan speed with node rearrangement

the speed of only three fans was over 14000 RPM, and all the others were less than 12000 RPM. In the original position, three fans were also over 14000 RPM, but the other three fans were over 12000 RPM, and they were the reasons of the difference in power consumption. Since the results shown by the original position were near those of the concentrated arrangement, the reduction of total power was only 65 W, which was 1.3 % of the total power used. But if the original position was near the distributed arrangement, the reduction of total power consumption was 351 W, which was 6.7 % of total power.

We can combine these two schemes. First select the optimal positions of nodes, and then schedule tasks from cool nodes. When a task with heavy load is requested, the task on the cool node may be migrated to the hot node and the task with heavy load may be scheduled on the cool node if the migration cost is lower than the effects of the schedule.

7 Discussion

We use LINPACK in previous evaluation. This is because LINPACK has about 80 % of peak performance of floating point arithmetic and it increases CPU temperature and power consumption of nodes to almost the peak. Since our scheme uses the difference of hot nodes and cool nodes, the effect is larger when the CPU temperature of hot nodes is larger. So LINPACK produces the maximum effect for our scheme.

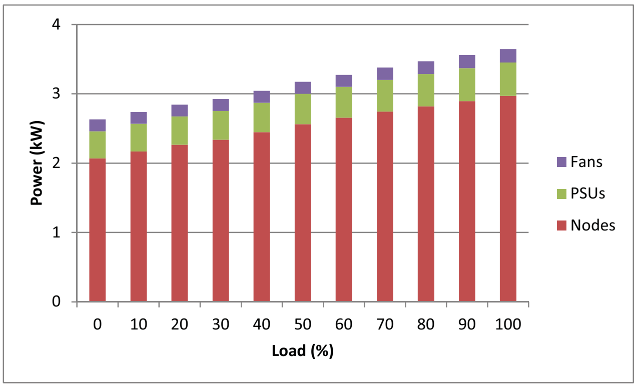

Fig. 20 Power consumption on SPECPower

We evaluated the SPECPower benchmark. SPECPower is a benchmark for measuring server power efficiency simulating web server load. It is available from the Standard Performance Evaluation Corporation (SPEC) [6]. Fig. 20 is the results of SPECPower on the same platform of perivious section. The load was changed from 0 % to 100 % by 10%. The power consumption of nodes increased in almost proportional to the load. But the increase of power of nodes with 100 % of load was about 54% of LINPACK case. The loss of PSUs was almost proportional to the total power consumption. The power of fans did not increased when the load was less than 60 %. This is because the CPU temperature was less than the threshold increasing Virtual FAN of each node. The power of fans was increased when the load was more than 60 %. But the maximum increase of power of fans with 100 % of load was only 23W, it was only 13 % of the power of fans in an idle case. This result shows that our scheme can be reduce power of fans at most 23W. This is very small with comparison of 350W in LINPACK case. This is because the CPU temperature was not so high even when the load was 100 % in SPECPower. The computation intensive application will increase the CPU temperature and our scheme will be effective. Or from the other viewpoint, the fan power during idle will be reduced more by controlling the speed of fans or the number of active fans.

We tried to evaluate our scheme on another platform, DELL PowerEdge E1000e blade system. But the fan speed was not changed even when all nodes ran LINPACK. This may be a difference policy for controlling speed of fans in the blade enclosure. But since it is a fact that there is a room for reducing power consumption of fans, our scheme will be applied when the speed of fans are controlled according to the temperature of nodes.

8 Related Works

Many works assume that the homogeneous nodes have the same power consumption [7, 8], but in [9], the authors revisit and focus on such “truths” commonly assumed concerning the energy usage of servers. The authors measured the real consumption of homogeneous nodes with typical applications and shown there are some difference. They assume that the consumption depends on its position on the rack and its temperature, but they cannot propose a model of the link between those factors and the variations in energy consumption.

In data center level, the pioneering work in [10, 11] proposes to reduce server idle power by concentrating the data center loading on a subset of the servers and powering-off the rest of the servers. However, in [12], the authors pointed out that these approach significantly reduced idle power, but they also increased cooling power due to hot spot created by concentrating the data center loading. The authors proposed a new approach, which trades-off idle power and cooling power for each other to reduce the total power.

9 Conclusion

This paper describes a new metric of data center power efficiency to fairly evaluate the contribution of each improvement for power efficiency. The proposed metric is based on the power of each function such as processing, power supply and heat removing, while other existing metrics are based on the power of each component such as IT equipment, power facility and cooling facility.

This paper also shows that an imbalance of CPU temperature can exist even with the same specifications and the same load in a blade system, and that the imbalance affected the power consumption of fans. It also shows that the power consumption of fans can be reduced by scheduling nodes or by rearrangement of node positions taking into account the imbalance of CPU temperatures. We achieved a reduction in fan power consumption of a maximum of 62 % and in total power consumption of a maximum of 12 % by changing the five nodes to be used in the schedule from hot nodes to cool nodes.

These results were measured on HP’s blade system, and in the future we plan to determine whether this scheme can be applied to other systems or not, and if it can, what is the effect. The blade system that we used in this study was a part of a system used to evaluate power consumption in a data center in 2007. 2007 was the planning year of our project. We had to rent a system built in 2007. It was difficult to rent such a system, and many blade nodes were gathered from different enclosures. The blade nodes had the same specifications, such as processor speed, memory size, etc., but their production lots might be different. This may have caused an exaggeration of the imbalance of CPU temperatures. However, when we scale up the blade system, the same imbalance may occur, so it is useful to show the possibility of power reduction by scheduling or rearrangement of nodes, taking the imbalance of CPU temperatures into consideration.

Acknowledgment

This research was partially supported by a New Energy and Industrial Technology Development Organization (NEDO) research project entitled “research and development project for green network/system technology (Green IT Project).”

References

1. Belady, C.,: Green grid data center power efficiency metrics: PUE and DCIE. White paper: Metrics & Measurements, http://www.thegreengrid.org (2007)
2. Anderson, D., etc.: A Framework for Data Center Energy Productivity. White paper: Metrics & Measurements, http://www.thegreengrid.org (2008)
3. Green IT Promotion Council: Concept of New Metrics for Data Center Energy Efficiency. http://www.greenit-pc.jp/e/topics/release/100316_e.html (2010)
4. Itoh, S., Kodama, Y., Shimizu, S., Sekiguchi, S., Nakamura, H., Mori, N.: Power consumption and efficiency of cooling in a Data Center. Energy Efficient Grids, Clouds and Clusters Workshop (E2GC2), in conjunction with the 11th ACM/IEEE Int. Conf. on Grid Computing (Grid 2010), 305–312 (2010)
5. Intel Math Kernel Library, http://software.intel.com/en-us/intel-mkl/
6. SPEC, http://www.spec.org/
7. Kim, K.H., Buyya, R., Kim, J.: Power Aware Scheduling of Bag-of-Tasks Applications with Deadline Constraints on DVS-enabled Clusters. Proc. of the IEEE Int. Symp. on Cluster Computing and the Grid (CCGRID 2007), 541–548 (2007)
8. Steinder, M., Whalley, I., Hanson, J.E., Kephart, J.O.: Coordinated management of power usage and runtime performance. Proc. of the IEEE Network Operations and Management Symposium (NOMS 2008), 387–394 (2008)
9. Orgerie, A.C., Lefevre, L., Gelas, J.P.: Demystifying Energy Consumption in Grids and Clouds. Work in Progress in Green Computing (WIPGC) Workshop, in conjunction with the first Int. Green Computing Conf. (IGCC 2010), 335–342 (2010)
10. Chase, J.S., Anderson, D.C., Thakar, P.N., Vahdat, A.M., Doyle, R.P.: Managing energy and server resources in hosting centers. Proc. of the 18th ACM symp. on Operating systems principles (SOSP 2001), 103–116 (2001)
11. Pinheiro, E., Bianchini, R., Carrera, E.V., Heath, T.: Load balancing and unbalancing for power and performance in cluster-based systems. Workshop on Compilers and Operating Systems for Low Power (COLP 01), in conjunction with the 10th Int. Conf. on Parallel Architectures and Compilation Techniques (PACT’01), (2001)
12. Ahmad, F., Vijaykumar, T.N.: Joint Optimization of Idle and Cooling Power in Data Centers While Maintaining Response Time. Proc. of the 15th Int. Conf. on Architectural Support for Programming Languages and Operating Systems (ASPLOS 2010), 243–259 (2010)

13. NEDO (New Energy and Industrial Technology Development Organization): Outline of NEDO 2008-2009. http://www.nedo.go.jp/kankobutsu/pamphlets/kouhou/2008gaiyo_e/ , 124–125 (2009).

Yuetsu Kodama received his B.E, M.E. and Ph.D degree in engineering from the University of Tokyo in 1986, 1988 and 2003. He has been engaged in the research of parallel computer architecture since he joined Electrotechnical Laboratory in 1988. He is currently a professor at Graduate School of Systems and Information Engineering, University of Tsukuba since Feb. 2011. He is a member of IEEE CS, IEICE and IPSJ.

Satoshi Itoh obtained the Ph. D degree in physics from University of Tsukuba, Japan, in 1987. From 1987 to 2002 he worked for high performance and parallel computing in the both area of material science and business application at Central Research Laboratory, Hitachi, Ltd. In 2002, he moved to AIST (National Institute of Advanced Industrial Science and Technology), Japan and has researched on Grid computing, Cloud computing, and Green IT. He is currently the Deputy Director of Information Technology Research Institute, AIST.

Satoshi Sekiguchi received BS from The University of Tokyo, ME from University of Tsukuba, and Ph.D. in Information Science and Technology from The University of Tokyo, respectively. He joined Electrotechnical Laboratory (ETL), Japan in 1984 to engage research in high performance and parallel computing widely from the computer architecture, compiler, numerical algorithm, and performance evaluation methods. He served as the director of Grid Technology Research Center (GTRC), National Institute of Advanced Industrial Science and Technology (AIST) in 2002-2008., and is currently the Director of Information Technology Research Institute, AIST. He has been contributing to the Open Grid Forum as a member of board of directors, is a member of IEEE, SIAM, IPSJ.

Toshiyuki Shimizu is a director of development division at SynergeTech, Inc. in Japan. He graduated from Osaka Electro-Communication University. He had been working for developing LSI test equipments in advantest corporation from 1981 to 1989. In 1989, he participated in establishment of SynergeTech, Inc. and has been working for developing a high-speed LSI test equipment, and managing a large scale computer cluster. He is a member of IEICE and IPSJ.

Hiroshi Nakamura received the BE, ME, and Ph.D. degrees in Electrical Engineering from the University of Tokyo in 1985, 1987, and 1990, respectively. He was a visiting associate professor at the University of California, Irvine from 1996 to 1997. He is currently a Professor of Department of Information Physics and Computing at the University of Tokyo. His research interests include low-power processor, VLSI design, power-aware computing, high-performance computer systems, and dependable computing. He is a senior member of IEEE and ACM.

Naohiko Mori received his B.E. and M.E. from Kyoto University in 1982 and 1984. He was engaged in the creation of commercial services based on a distributed file system and a secret sharing scheme while he was with NTT Communications Corporation. He is currently with NTT Advanced Technology Corporation since April, 2011.