Optimum control method of workload placement and air conditioners in a GPU server environment

Hikotoshi Nakazato
NTT Network Inovation Center
NTT Corporation
Tokyo,Japan
hikotoshi.nakazato.zd@hco.ntt.co.jp

Seisuke Arai
NTT Network Inovation Center
NTT Corporation
Tokyo,Japan
seisuke.arai.xh@hco.ntt.co.jp

Masashi Kaneko
NTT Network Inovation Center
NTT Corporation
Tokyo,Japan
masashi.kaneko.dr@hco.ntt.co.jp

***Abstract*— High-performance computing datacenters have been rapidly growing, both in number and size. In addition to the conventional CPU servers, more GPU servers are being placed in datacenters to achieve high-speed processing such as image recognition. From our experiments, the power consumption during GPU operation is greatly affected by changes in the temperature of the server intake port. Therefore, to reduce the total power consumption of datacenters, thermal management of datacenters can address dominant problems associated with cooling such as the recirculation of hot air from the server outlets to their inlets and the appearance of hot spots. In this paper, we propose a workload placement method for environments where CPU servers and GPU servers coexist, and an optimum air-conditioning control method that cooperates with the workload placement method to reduce the total power consumption of the servers and air conditioners in datacenters. Experiment results in an actual machine environment showed that our proposed method has valid power-saving effects by adjusting cooling capacity tradeoffs between GPU servers and air conditioners.**

Keywords—workload placement, intake port temperature of GPU server, optimum control of air conditioner, power-saving effects

I. INTRODUCTION

Cloud computing provides dynamic and scalable virtual resources through the Internet to users on demand and is furthering the development of distributed computing, parallel computing, and grid computing [1]. Its main advantage is that it can quickly reduce hardware costs by offering on-demand access to shared computer resources and data; users can access high-quality services at low cost. Because power and energy are first-order concerns in cloud computing, cloud providers require a low operation overhead at datacenters. Optimizing cooling presents the single largest area of opportunity for datacenters to save energy. Due to the increasing power density and heat generation of newer equipment, cooling and air-conditioning energy costs now surpass the cost of powering servers [2]. Moreover, graphics processing unit (GPU) servers have been frequently used in recent years for high-speed, large-volume data processing such as artificial intelligence (AI) inference such as image recognition. Users can obtain calculation results more quickly by multi-parallel processing. Generally, GPU servers consume more power than central processing unit (CPU) servers, and the amount of heat released to the datacenter space increases in proportion to the power consumption of the server, so the power consumption of the air conditioners also increases accordingly [3]. On the other hand, our basic experimental results make it clear that the power consumption of the GPU server fluctuates more in the same temperature range than that of the CPU server due to the changes in the server intake port temperature. Figs. 1-4 show the changes in the GPU parameters when the server intake port temperature is 20 or 33 °C. We use HPE Apollo6500 Gen10 with in 4 QuardroRTX8000 GPU cards inserted and FAST N-Body simulation with CUDA [4] for running applications on the GPU server in our experiment. Fig. 1 shows the temperature change in the GPU card during GPU operation when the intake port temperature is 20 or 33 °C. When the intake port temperature is 20 °C, the temperature of all GPU cards increases up to and stays around 55 °C while the application is running. On the other hand, when intake port temperature is 33 °C, some GPU card temperatures exceed 70 °C, and the average temperature of each one is about 10° C higher than when the intake port temperature is 20 °C. Fig. 2 shows the power consumption of the GPU cards. The average power consumption was 15W per GPU card higher when the intake port temperature was 33 °C than when it was 20 °C. The difference in fan rotation rate can be considered as one cause of the increase in the power consumption of the GPU cards, shown in Fig. 3. The fan rotation rate during application running was constant when the intake port temperature was 20 °C, whereas it increased as the GPU card temperature increased when the intake port temperature was 33 °C. Fig. 4 shows the changes in GPU sever power consumption due to changes in intake port temperature. In the GPU server as a whole, the power consumption is about 9% higher at the inlet temperature of 33° C than at the inlet temperature of 20 °C. In the same temperature range of inlet temperature, such a large fluctuation in power consumption has not been confirmed in conventional CPU servers so far. Under a high inlet temperature such as 30 °C or more, temperature is more likely to rise inside the GPU server than the CPU server, thereby increasing the amount of power consumption due to fan rotation for cooling the heat inside the server.

Therefore, in a datacenter containing GPU servers in addition to the conventional CPU servers, a new power-saving control method is required that adjusts the tradeoff between cooling capacity of GPU servers and air conditioners. In this paper, we propose an optimum air-conditioning control method that cooperates with workload placement of Kubernetes-based [5] CPU and GPU workloads. The proposed method minimizes total power consumption of air conditioners and servers by considering changes in power consumption of GPU servers due to fluctuations in inlet temperature. The power consumption of air conditioners is minimized by our original environmental self-adaptation method that categorizes learning history with the conditions composed of exhaust heat distribution and temperature information in the room as an environmental classification standard.

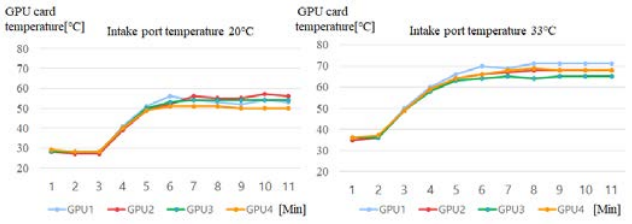

Fig. 1 Temperature change of the GPU card

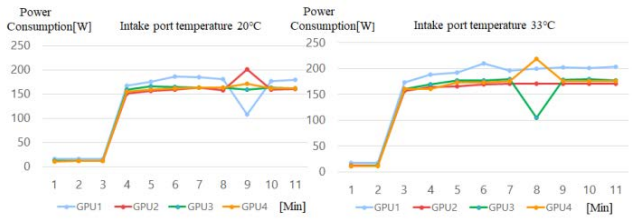

Fig. 2 Power consumption change of the GPU card

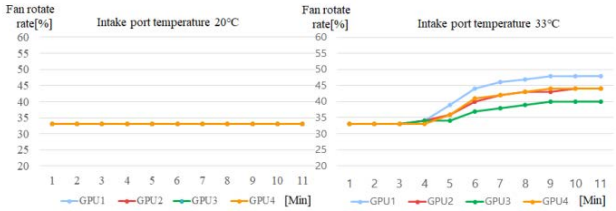

Fig. 3 Fan rotate rate change of the GPU card

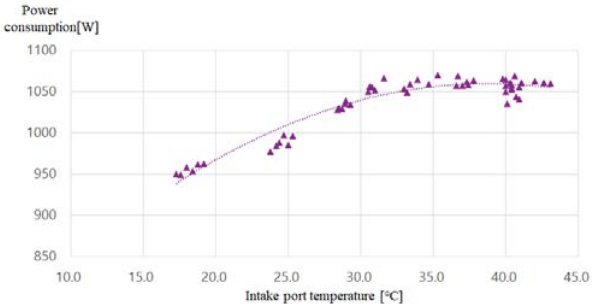

Fig. 4 Changes in GPU server power consumption

II. RELATED WORK

A. Workload Placment

Workload placement in datacenters has long been investigated and is currently still a hot topic in various research domains. Deelman et al. [6] have done considerable work on planning, mapping, and data-reuse in workflow scheduling. They proposed Pegasus [6], which is a framework that maps complex scientific workflows onto distributed resources such as the Grid. DAG Man, together with Pegasus, schedules tasks to the Condor system. Kim et al. [7] implemented a guest-aware priority-based scheduling scheme, which is specifically designed to support latency-sensitive workloads. The proposed scheduling scheme prioritizes the virtual machines (VMs) to be allocated by using the information about priorities and status of guest-level tasks in each VM. It preferentially selects the VMs that run latency-sensitive applications to be scheduled, and in this way, reduces the response time to the I/O events of latency-sensitive workloads. Wang et al. [8] proposed a novel VM scheduling algorithm for virtualized heterogonous multicore architectures. The algorithm exploits the core performance heterogeneity to optimize the overall system energy efficiency. Takouna et al. [9] also addressed the VM scheduling of heterogeneous multicore machines. A scheduling policy is designed to schedule each virtual machine to an appropriate processing core on the basis of the performance sensitivity to the CPU clock frequency and the performance dependency on the host. Balouch and Bejarzahi [10] targeted a scheduling algorithm aiming at allocating VMs to physical hosts of data centers in such a way that the target host will not be overloaded or over-heated by scheduling VMs with respect to the temperature and CPU utilization of processors.

B. Kubernetes Scheduler

Kubernetes (K8s), an open-source container orchestration tool, has become valuable for managing complex container-based applications [5] and has been gaining a lot of attention. Kube-scheduler is the default task scheduler in K8s that uses Pod as the smallest deployable unit and has been one of the most active research topics. The scheduler determines which Nodes are the proper placement for each Pod in the scheduling queue on the basis of constraints and available resources. Chang et al. [11] proposed a platform that dynamically manages the number of Pods deployed on a K8s cluster in accordance with Node resource usage. In their platform, Nodes are monitored using multiple monitoring tools, and the number of Pods is increased or decreased when the overall CPU usage is above or below a certain threshold. Townend et al. [12] clarified the importance of considering the characteristics of hardware and software when scheduling Pods. In their scheduler, Nodes are monitored and modeled using specialized machines. In an ideal environment, this scheduler facilitates a reduction in overall power consumption. Douhara et al. [13] proposed both a Workload Allocation Optimizer (WAO)-scheduler and WAO Load Balancer architecture, which optimize Pod allocation and task allocation, respectively, as an AI-based power consumption reduction function for K8s.

C. Optimum control of air conditioners at datacenters

Asa et al. [3] solved a combinatorial optimization problem by minimizing the total power consumption of air conditioning and IT equipment in a specific IT load placement pattern. Nakamura et al. [14] proposed a datacenter energy management system that reduces power consumption of cooling and a novel method for semi-optimal workload placement and cooling. An algorithm for a semi-optimal solution is introduced, and optimal cooling is calculated using the resultant workload placement and linear programming, which is faster than mixed integer programming.

However, in an environment where GPU servers and CPU servers are placed together, an optimal control method has yet to be proposed that minimizes the total power consumption of servers and air conditioners in consideration of the power consumption characteristics of GPU servers due to changes in the inlet temperature. Moreover, no K8s scheduler has been proposed that implements a placement method that appropriately adjusts the tradeoff of cooling efficiency between air conditioners and GPU servers by simultaneously

controlling GPU load placement and Pod placement in cooperation with suitable air-conditioning settings.

III. PROPOSAL

First, Fig. 5 shows the definition settings of our proposed method. In the server room of a datacenter, GPU servers and CPU servers are placed in several Server Placement Areas, which indicate which server is located in which area of the room. Air conditioners are also placed at regular intervals. The Air-Conditioning Control Area is an area to measure the room temperature effect of air-conditioning control, and it faces either the intake port side or the discharge port side of the server. The air blown from the air conditioner is blown out from the Air-Conditioning Control Area at the inlet side (Areas 3 and 4 in Fig. 5) via pipes provided under the floor. Then, in the Air-Conditioning Control Area at the outlet side (Areas 1, 2, 5, and 6 in Fig. 5), whose temperature has risen due to the heat of each server exhaust, and an air flow returning to the air conditioner is generated. Many temperature sensors are installed in both the Air-Conditioning Control Area and Server Placement Area, and a temperature sensor is also installed outside the datacenter. It is assumed that the correspondence of which temperature sensor indicates the intake port temperature of each GPU server is also set in advance, and our power management control system can obtain this temperature information in real time. We assume fixed amounts of CPU and GPU workloads are generated at regular time intervals we call "control turns," and the problem to solve is how to schedule these tasks to CPU and GPU servers with suitable air-conditioning settings that minimize the total power consumption of servers and air conditioners.

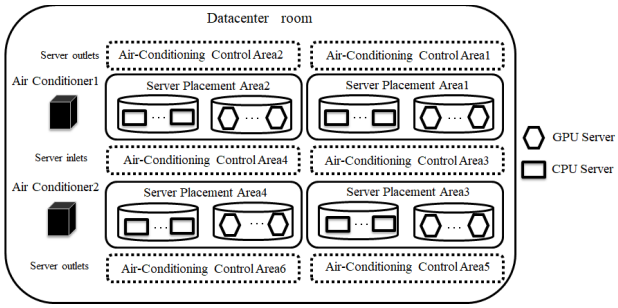

Fig. 5 Definition settings of proposal

Formula (1) calculates total power consumption.

$$P=\sum_{t=ts}^{te}\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{o} a\_\{i\}(t)+c\_\{j\}(t)+g\_\{k\}(t) \quad (1)$$

We assume the start and end times of a control turn as *ts* and *te*, there are m air conditioners, n CPU servers, and o GPU servers respectively at the evaluation environment. The power consumption of each air conditioner, CPU server, and GPU server is shown as $a\_\{i\}(t)$, $c\_\{j\}(t)$, $g\_\{k\}(t)$ respectively. $a\_\{i\}(t)$ is affected by the amount of exhaust heat generated by the workload in each Air-Conditioning Control Area, and the amount of heat exhausted from the server increases in proportion to the amount of power consumption [3]. We estimate $a\_\{i\}(t)$ by categorizing the workload placement pattern from learning history in the specific room environment, on the basis of the standard composed of power consumption of server in each Server Placement Area and some temperature information of the room. In the workload pattern with GPU servers, $a\_\{i\}(t)$ and $g\_\{k\}(t)$ are in a tradeoff relationship in terms of the adjustment of cooling capacity of GPU servers and air conditioners. Therefore, the total power consumption needs to be predicted on the basis of the change in the intake port temperature of the GPU servers for each workload distribution and air-conditioning setting pattern, and the pattern needs to be selected that minimizes the total power consumption among them.

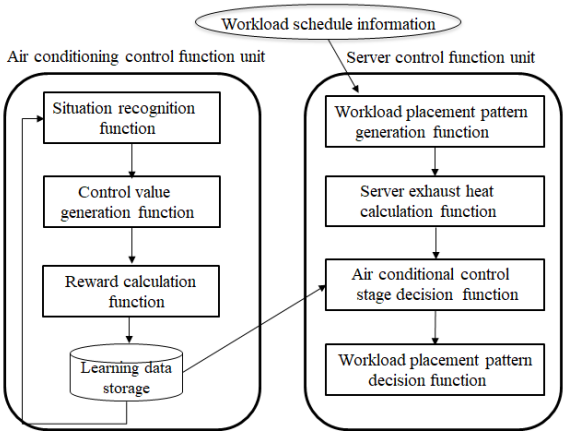

Fig. 6 Functions of power management control systems

Fig. 6 shows all functions of our proposed power management control system. There are two main function units in our system: the server control function unit and air-conditioning control function unit. In the air-conditioning control function unit, suitable air-conditioning settings need to be found to meet the overall power consumption requirement by self-adaptive control.

We define environment standardization called the "Situation" to express the current server heat environment in the room of a datacenter. When the Situation differs, the power consumption of cooling capacity also differs. Fig. 7 shows the parameter factors of the Situation in our proposed method. These factors are expected to have a significant effect on the power consumption of air conditioners in the room. The average temperature of the total room, outside temperature, and server calorific value in each Server Placement Area are parameter elements of the Situation. Each parameter is divided into multiple ranges, and the combination of the areas of each range is defined as one situation. At each start time of a control turn, the situation recognition function obtains the latest parameters of the Situation and decides the current Situation by judging each range of the parameter. In Fig.7 for example, when the average temperature of the total room (factor 1) divides 0-48 degrees evenly with 8 degrees into 6 ranges (0 degrees or more and less than 8 degrees as factor 1-1; 8 degrees or more and less than 16 degrees as factor 1-2) and identifies which range the current value belongs to. By identifying the range of all factors, the Situation is determined. Then the control value generation function generates air-conditioning settings with several levels of strength. For example, air volume (Hz) and target temperature (°C) can be adjusted as air-conditioning control values. In our proposed method, the range of the maximum and minimum values of each control value is evenly divided, and multiple control stages are provided. At the end of each control turn, pass/fail judgement about room temperature reward is calculated at the reward calculation function. The temperature reward condition is threshold-based, and the threshold is set to an appropriate temperature that does not exceed the predetermined GPU power consumption on the basis of the relationship between the server intake port temperature and

power consumption of GPU server. Moreover, from the viewpoint of secure operation of a datacenter, sensor temperature at each Air-Conditioning Control Area also needs to be under the fixed threshold value at both inlet and outlet sides of servers. Among several control stages of air conditioning in each Situation, the control stages that exceeded the reward threshold are assumed as candidates of control solutions. Additionally, the learning data storage saves learning history of air-conditioning control in each Situation.

Situation example

Factor1	Factor2	Factor3	Factor4	Factor5	Factor6
15	28	31	31	42	42

Situation Category
Factor1-2_Factor2-4_Factor3-4_Factor4-4_Factor5-5_Factor6-5

Factor	Parameter	Classification definition	Range	Factor range identification
Factor1	Average room temperature	Divide 0-48℃ into 6 divisions	8-16	Factor1-2
Factor2	Outside temperature	Divide 0-48℃ into 6 divisions	24-32	Factor2-4
Factor3	Exhaust heat in Server Placement Area1	Divide 0-200KW into 20 divisions	30-40	Factor3-4
Factor4	Exhaust heat in Server Placement Area1	Divide 0-200KW into 20 divisions	30-40	Factor4-4
Factor5	Exhaust heat in Server Placement Area1	Divide 0-200KW into 20 divisions	40-50	Factor5-5
Factor6	Exhaust heat in Server Placement Area1	Divide 0-200KW into 20 divisions	40-50	Factor6-5

Fig. 7 Parameter factors of Situation

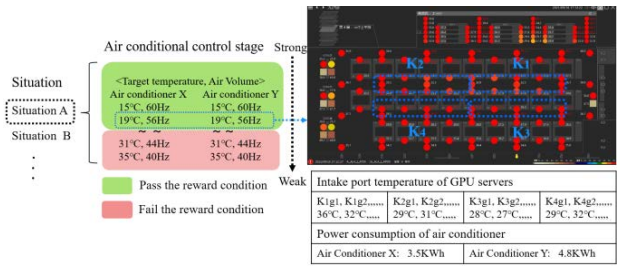

Fig. 8 Data types saved in learning data storage

In order to calculate the total power consumption of air conditioners and servers, it is necessary to save the power consumption of air conditioners of every control stages in each situation. Fig.8 shows the image of data types saved here. Data is classified by every control stage of air conditioning in each situation, and only the data exceeding the temperature reward is saved as a candidate control solution. In each data, predicted temperature transition information of the sensors and power consumption of operating air conditioners during the turn are saved. As for temperature information, we assume that for each GPU placed in each Server Placement Area *Ki* for *Kig1*,*Kig2*…*Kigm*, our system holds the temperature of each sensor closest to the intake port of *Kigj*(j=1,,,,m) and obtains this information from every Server Placement Area to record the temperature distribution around the intake port of the GPU servers in the entire room at every snapshot of the control turn. Handing these leaning history data to the server control function enables the total power consumption of servers and air conditioners of each workload placement pattern to be calculated by predicting the power consumption of placing servers on the basis of the temperature transition of the intake port during the turn and then adding power consumption of air conditioners at the specific control stages to it. In this way, the tradeoff between cooling capacity of servers and air conditioners can be adjusted by controlling the temperature distribution in each Server Placement Area at the turn.

The server control function mainly calculates the total power consumption at each workload placement pattern and determines the best pattern to most reduce it by selecting an appropriate workload distribution to each Server Placement Area and air-conditioning control stage in that Situation. First, it obtains the workload schedule information, which contains how many CPU and GPU workload processing requests will come to the datacenter room at the start of the control turn. On the basis of this information, the server control function generates workload placement patterns to distribute these CPU and GPU workloads at each Server Placement Area. It can divide the entire load of CPU and GPU into *n* pieces respectively, and each Server Placement Area can be assigned a minimum of 0 workloads and a maximum of *n* workloads of CPU and GPU. After generating workload placement patterns, the server control function determines the server calorific value in each Server Placement Areas of each workload placement pattern. Server calorific value is defined as Formula (2).

$$W(Si) = Pc(Si) \times kc + Pg(Si) \times kg \qquad (2)$$

Pc(Si) and *Pg(Si)* are basic power consumptions of a CPU server and GPU server in the Server Placement Area *Si*. These values are the basic power consumption amounts before an increase in power consumption due to an increase in the temperature of the server intake port. *Kc* and *kg* are thermal resistance coefficients of CPU server and GPU server, which show the proportional relationship between server calorific value and power consumption of the CPU server and GPU server, respectively. Since these coefficient values vary depending on the server model used and the operating room environment, the basic value needs to be calculated on the basis of the calorific value and the power consumption amount in advance in the room environment. For example, if there are 12 Pods and 5 GPU workload requests at Serve Placement Area 1, 0.025KW per Pod workload and 1KW per GPU workload are required as power consumption of server, and both *kc* and *kg* are 1 in the operating room. Then the server calorific value in Server Placement Area 1 is calculated as 5.3KW.

On the basis of the average temperature of the total room at the start of the control turn, outside temperature, and server calorific value in each Server Placement Area, the appropriate Situation is determined for each workload placement pattern. Next, the suitable air-conditioning control stage in that Situation is determined. There are several air-conditioning control stages in each Situation, and we assume the learning history data of each control stage, which includes predicted temperature transition information of the sensors near the intake port of GPU servers and power consumption of operating air conditioners in that turn, is saved in advance. The control stage that most reduces the total power consumption of servers and air conditioners will be selected in that Situation.

Power consumption in each Server Placement Area is calculated by predicting the power consumption of working CPU servers and GPU servers in that area to process certain types of workload. As for CPU servers, the server control function uses server resource utilization such as CPU utilization as input information and predicts power consumption of the server on which the Pods are running. As for GPU servers, the server control function considers the change in temperature of the server intake port and uses this parameter, the number of GPU processing cards, and sensor temperature inside GPU enclosure as input parameters to

predict power consumption. This learning model is prepared in advance on the basis of learning history of running workload in the specific room environment and server model. By adding power consumption of the entire Server Placement Area and air conditioners during the control turn, the best control stages of air conditioning in each Situation can be determined. We assume this total power consumption of servers and air conditioners to be the best solution at the specific workload placement pattern to adjust room temperature distribution to an appropriate level to most reduce total power consumption of cooling functions in the room.

Finally, the workload placement pattern decision function chooses the workload placement pattern that has the lowest power consumption from all generated patterns. In this way, the optimum workload distribution and air-conditioning control settings are determined. Even though only CPU servers exist in the server environment, there is no need to consider change in the intake port temperature. Also, the server environment including GPU servers shows a different tradeoff adjustment approach and is also greatly dependent on the number of GPU servers and air conditioners and the level of exhaust heat generation, etc. in the specific room environment. Therefore, our self-adaptive approach to adjusting cooling capacity in specific room environment is effective to reduce the total power consumption in datacenter rooms, especially ones containing GPU.

IV. EVALUATION

A. Evaluation settings

To evaluate the effectiveness of our proposed method, we generate 3 different scales of exhaust heat patterns (30, 60, 120 KW) in a datacenter room and evaluate the total power consumption of servers and air conditioners by changing workload placement patterns and control stages of air conditioners. Exhaust heat is generated by a pseudo heat generator that simulates heat generation of GPU and CPU workloads. By measurement in the datacenter room in advance, thermal resistance coefficients *kc* and *kg* are set to 1, and basic power consumptions *Pc(Si)* and *Pg(Si)* of each CPU and GPU workload are set to 0.025 and 1, respectively, in this experiment. For example, at 30KW exhaust heat patterns, we simulate 25 GPU workloads and 200 CPU workloads by Pods placed in the room at the start of a control turn. The workloads at the 60KW and 120KW patterns are two and four times as much as at the 30KW pattern, respectively. Datacenter rooms used in this experiment have four Server Placement Areas and two air conditioners as shown in Fig. 5. The length of one control turn is set to 30 minutes.

Table 1 shows our workload distribution ratio pattern and air-conditioning setting patterns. We prepared a workload distribution ratio divided by 25% to create patterns, which is very normal at datacenters. Also, control stages of air conditioners are also comprehensively prepared in the possible control values. We change them at three exhaust heat patterns. Table 1 also shows how we applied these setting parameters to each exhaust heat pattern in this experiment. We compared the difference in total power consumption between the best pattern selected by our proposed method and the randomly selected pattern among patterns that can be taken in this experiment. Power consumption of GPU servers in each Server Placement Area during the control turn is estimated on the basis of the intake port temperature obtained from the temperature sensor nearest to each intake port in the room, using learning data shown in Fig. 4.

Table 1. Evaluation settings

Workload distribution ratio pattern		Air conditioning settings pattern		Exhaust heat pattern settings
P1	(0%,50%,50%,0%)	C1	(15°C,60Hz)	30KW: P1～P5 & C1～C6
P2	(50%,50%,0%,0%)	C2	(19°C,56Hz)	60KW: P1～P5 & C1～C4
P3	(0%,0%,50%,50%)	C3	(23°C,52Hz)	120KW: P3,P4 & C1～C4
P4	(25%,25%,25%,25%)	C4	(27°C,48Hz)	
P5	(50%,0%,0%,50%)	C5	(31°C,44Hz)	
		C6	(35°C,40Hz)	

B. Evaluation results

Fig. 9 shows our evaluation results for three exhaust heat patterns. In all patterns, our proposed method achieved the lowest power consumption. It reduced total power consumption by 33.7, 9.4, and 6.8% compared with the other method at 30, 60, 120KW exhaust heat patterns, respectively. The power consumption of air conditioners alone is reduced by 85.9, 35.1, and 44.5%, respectively. This shows that our proposed method significantly reduced the total power consumption while considering the increase in GPU server power consumption due to the intake port temperature change, lowering the air-conditioning control stage as much as possible, and balancing the cooling capacity of air conditioners and server itself in the datacenter room. This feature is maintained even if the exhaust heat generated in the room changes, by categorizing heat distribution in each room area of the current control turn with the Situation we defined, estimating the temperature distribution at the control turn, and adopting the suitable air-conditioning control stage that most reduces the total power consumption in each Situation. Also, in the environment where GPU servers are placed, our proposed method considers the power increase of the GPU server due to the increase in the intake port temperature, and this phenomenon was not observed in CPU servers even in very common temperature ranges such as over 30 °C. Especially in the case where the number of processing GPU servers increases in the room, a solution needs to be derived that suppresses the total power consumption by accurately estimating the power consumption of the servers at the specific control stage of air conditioners.

Next, Fig.10 shows the difference in power consumption of the best solution in each workload placement pattern from P1 to P5 at 60KW heat cases. The results show that when the workload placement pattern differs, the total power consumption of best solution, which with air control stages achieves the most energy-saving, also differs. At 60KW heat cases, the P3C2 pattern totally requires 36.89KWh during the control turn, so it reduces the power consumption the most. On the other hand, the P1C4 pattern requires 39.77KWh, so there was a difference of 7.2% in total power consumption between these two patterns. This reduction in overall power consumption can be a non-negligible effect as the scale of servers operating at the datacenter increases. Our proposed method generates multiple workload placement patterns of the Server Placement Area for the same amount of processing requests generated in a room, predicts the power consumption of air conditioning and servers, and selects the patterns with the smallest total power consumption. As the room environment differs, the best pattern also differs in accordance with the distance and wind angle between air conditioners and servers, hot spots in the room, number of servers, etc., It can be said that our self-adaptive approach effectively responds to the changes in these environmental factors.

Finally, Table 2 compares of energy-efficiency parameters of the best solution in 30, 60, and 120KW heat patterns, average power consumption per air conditioner, server power consumption ratio to basic power consumption, and total power consumption ratio when converted to 120KW. When the temperature of the intake port rises, power consumption of GPU servers increases, and this leads to a higher ratio of server power consumption to server basic power consumption. Strengthening the air control stages will reduce this ratio but increase the power consumption per air conditioner, so there is in a tradeoff relationship between these two parameters. Table 2 also shows that when the exhaust heat scale differs, the proper value of these two parameters also changes. In the case where the server power consumption ratio rises extremely due to the temperature of server inlets rising and a large number of GPU servers working at the turn, power consumption of servers especially will increase.

With our proposed method, the power consumption ratio can be suppressed to a certain level by strengthening air-conditioning control stages to effectively adjust the tradeoff between cooling capacity of air conditioners and GPU servers. From the comparison of the 3rd parameter, the total power consumption ratio when converted to 120KW, it also can be said that after optimizing control within each single datacenter room, considering workload distribution among multiple datacenter rooms can further improve power efficiency. In the case of our experimental settings, assuming there are enough air-conditioned rooms with entirely the same environment to handle workload requests, the best power effective solution is to process 120KW in 4 rooms with 30KW each, and the worst solution is to process 120KW in 2 rooms with 60KW each. It is worth noting that there is a about 10% power efficiency difference between these two solutions, and this optimal solution depends heavily on the room environment. As the next step of optimizing the workload placement and air-conditioning settings in a single datacenter room, we will next consider how to distribute the total workload generated at a datacenter between the rooms in consideration of the characteristics of each room.

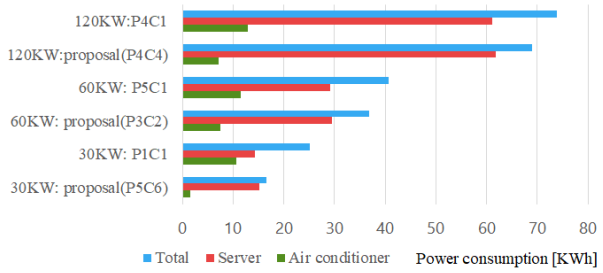

Fig. 9 Evaluation results of 3 exhaust heat patterns

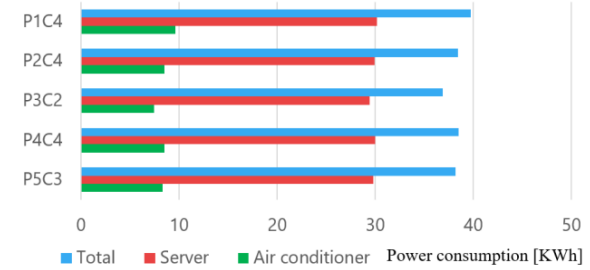

Fig. 10 Comparison of workload placement patterns (60 KW)

Table 2. Comparison of energy-efficiency parameters

Exhaust heat patterns	Average power consumption /air conditioner	Server power consumption ratio to server basic power consumption	Total power consumption ratio (converting120KW)
30KW	0.75KWh	100.80%	96.50% (4 rooms)
60KW	3.73KWh	98.10%	107.10% (2 rooms)
120KW	3.55KWh	103.00%	100% (1 room)

V. Conclusion

We proposed an optimum control method for workload placement and air conditioners in datacenter rooms where CPU servers and GPU servers coexist. Experiment results in an actual machine environment showed that our proposed method has valid power-saving effects by adjusting cooling capacity tradeoffs between GPU servers and air conditioners.

REFERENCES

[1] Rimal, B.P. and Choi, E. Lumb, "A Taxonomy and Survey of Cloud Computing Systems," In: Fifth International Joint Conference on INC, IMS and IDC, vol. 218, pp. 44–51, 2009.

[2] Info-Tech, "Top 10 energy-saving tips for a greener data center," Info-Tech Research Group, London, ON, Canada, Apr. 2010.

[3] Yasuhiro Asa, Tadakatsu Nakajima, Jun Okitsu, Takeshi Kato, Tatsuya Saito, and Yasuhiro Kashirajima, "Air Conditioning Optimum Control Cooperative with IT System for Environment-Conscious Data Center" Fit, vol.9, Issue 01, pp97-102, 2010.

[4] Nbody: https://developer.nvidia.com/gpugems/gpugems3/part-v-physics-simulation/chapter-31-fast-n-body-simulation-cuda

[5] D. Bernstein, "Containers and cloud: From LXC to docker to kubernetes," IEEE Cloud Computing, vol. 1, pp. 81–84, Sep. 2014.

[6] E. Deelman, G. Singh, M.-H. Su, J. Blythe, Y. Gil, C. Kesselman, G. Mehta, K. Vahi, G. B. Berriman, J. Good, A. Laity, J. C. Jacob, and D. S. Katz. Pegasus, "A framework for mapping complex scientific workflows onto distributed systems," Sci. Program., 13(3):219–237, 2005.

[7] Kim, D., Kim, H., Jeon, M., Seo, E., and Lee, J., "Guest-Aware Priority based Virtual Machine Scheduling for Highly Consolidated Server," 14th International Conference on Parallel and Distributed Computing (Euro-Par 2008), pages 285–294, 2008.

[8] Kolodziej, J., Khan, S., Wang, L., Kisiel-Dorohinicki, M., and Madani, "Security, Energy, and Performance-aware Resource Allocation Mechanisms for Computational Grids," Future Generation Computer Systems, 2012.

[9] Takouna, I., Dawoud, W., and Meinel, C., "Efficient Virtual Machine Scheduling-policy for Virtualized heterogeneous Multicore Systems," In Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications, 2011.

[10] Abdolkhalegh Balouch and Abdolvahed Bejarzahi, "Thermal and Power-Aware VM Scheduling on Cloud Computing in Data Center," International Journal of Engineering Research & Technology (IJERT), Vol. 8. Issue 06, 2019.

[11] C. Chang, S. Yang, E. Yeh, P. Lin, and J. Jeng, "A kubernetes-based monitoring platform for dynamic cloud resource provisioning," in Proc. IEEE Global Communications Conference '17, Dec. 2017, pp. 1–6.

[12] P. Townend, S. Clement, D. Burdett, R. Yang, J. Shaw, B. Slater, and J. Xu, "Invited paper: Improving data center efficiency through holistic scheduling in kubernetes," in Proc. IEEE International Conference on Service-Oriented System Engineering '19, Apr. 2019, pp. 156–15 610.

[13] Ryuki Douhara, Ying-Feng Hsu, Tomoki Yoshihisa, Kazuhiro Matsuda, and Morito Matsuoka, "Kubernetes-based Workload Allocation Optimizer for Minimizing Power Consumption of Computing System with Neural Network," International Conference on Computational Science and Computational Intelligence (CSCI), 2020.

[14] Masayuki Nakamura, Hideaki Hashimoto, Ryta Nakamura, and Joji Urata, "Optimization of Cooling and Workload Placement for Power Saving of Data Centers," Transactions of the Society of Instrument and Control Engineers 2014 Vol.50, Issue7, pp 518-527