Thermal Modeling and Analysis of Cloud Data Storage Systems

Xunfei Jiang[1], Mohammed I. Alghamdi[2], Maen M. Al Assaf[3], Xiaojun Ruan[4], Ji Zhang[1], Meikang Qiu[5], and Xiao Qin[1]
[1] Department of Computer Science and Software Engineering, Auburn University, Auburn, AL 36849, USA
[2] Department of Computer Science, Al-Baha University, Al-Baha City Kingdom of Saudi Arabia
[3] King Abdullah II School for Information Technology, The University of Jordan, Amman, Jordan
[4]Department of Computer Science, West Chester University of Pennsylvania, West Chester, PA 19383, USA
[5]Computer Engineering, San Jose State University, San Jose, CA 95192, USA
Email: xqin@auburn.edu; mialmushilah@bu.edu.sa; m_alassaf@ju.edu.jo; xruan@wcupa.edu; meikang.qiu@sjsu.edu

Abstract —An explosive increment of data and a variety of data analysis make it indispensable to lower power and cooling costs of cloud datacenters. To address this issue, we investigate the thermal impact of I/O access patterns on data storage systems. Firstly, we conduct some preliminary experiments to study the thermal behavior of a data storage node. The experimental results show that disks have ignorable thermal impacts as processors to outlet temperatures of storage nodes. We raise an approach to model the outlet temperature of a storage node. The thermal models generated by our approach gains a precision error less than 6%. Next, we investigate the thermal impact of data placement strategies on storage systems. We compare the cooling cost of storage systems governed by different data placement schemes. Our study shows that evenly distributing the data leads to highest outlet temperature for the sake of shortest execution time and energy efficiency. According to the energy consumption of various data placement schemes, we propose a thermal-ware energy-efficient data placement strategy. We further show that this work can be extended to analyze the cooling cost of data centers with massive storage capacity.

Index Terms—Thermal, model, storage system, cloud

I. Introduction

Big data, which is composed of a collection of huge and complex data sets, has been positioned as must have commodity and resource in industry, government, and academia. Processing big data requires a large-scale storage system, which increases both power and cooling costs. In this study, we investigate the thermal behavior of real storage systems and their I/O access patterns, which offer a guideline of building energy-efficient cloud storage systems.

The cooling consumption of data centers can be considerably reduced by using an efficient thermal management for storage systems. However, disk is not considered in traditional thermal models for data centers. In this paper, we investigate the thermal impact of hard disks and propose a thermal modeling approach for storage systems. In addition, we estimate the outlet temperature of a storage server by applying the proposed thermal model, in which the activities of both processor and disk are analyzed. We also study the cooling cost of data storage node by applying different data placement schemes, and propose a thermal-aware energy-efficient data placement strategy.

Motivations. Our proposed thermal model for next generation storage systems is motivated by the following five factors:

1) The continuously increasing cooling and energy costs of broad-scale storage systems;

2) The effect that the temperature may have on the cooling costs of data storage system;

3) The increasing importance of thermal monitoring cost reduction,

4) The capability to estimate cooling costs during datacenter planning phase, and

5) The shortage of studies on the impacts of disk activities on outlet temperatures of data nodes.

Due to high energy and cooling consumption in large-scale storage systems, the electricity costs of a data center for four years approach the costs of building a new datacenter [1]. Improving the energy efficiency of storage systems, as well as cooling systems, has attracted much attention for the recent years.

Increasing evidence indicates that cooling cost is a major contributor to the operational cost of data centers [1], [2]. The power and cooling infrastructure that supports IT equipments have consumed up to 50% of total energy cost in a data center [2]. As existing research reports, the energy efficiency of data centers can be effectively improved by reducing energy dissipation [3], [4]. In particular, maintaining lower outlet temperatures for servers or applying effective optimization on air recirculation can decrease energy costs of cooling systems in data centers [5]. Various workload placement strategies that maintain balanced temperature across data centers through workload management are proposed [3] [4].The experimental results presented by Moore *et al.* show that setting a low outlet temperature of data nodes can save up to 40% of energy consumption [3]. Lowering the temperature of hard disks leads to conservation of the energy consumption in cooling systems, but also increase

Manuscript received April 30, 2013; revised April 20, 2014.
Corresponding author email: xqin@auburn.edu.
doi:10.12720/jcm.9.4.299-311

of the disk reliability and extension the lifetime of the storage systems [6], [7].

There have been some researchers focusing on constructing energy models for storage systems in the past decades. A power consumption model of storage nodes running under certain workload was proposed by Allalouf *et al*. [8]. However, few thermal models of storage systems have been presented in literature. The impact of disk temperatures on the energy efficiency of cool systems in data centers has not been fully explored.

A traditional method to monitor system temperature is to use temperature sensors in data nodes. For example, two temperature sensors are needed to collect the inlet and outlet temperatures of a data node. If interior temperatures of the data node are required, additional sensors should be deployed in the chassis. Although it is practical for temperature collection in small-scale storage systems, it becomes a considerably expensive solution when the storage system consists of thousands of nodes. Deploying a bunch of sensors in a data center is time-consuming. Thus, designing thermal models is a promising alternative for monitoring the temperatures of storage systems.

Building a data center is a large investment for companies. The planning and cost estimation of the investment are critical to decision makers. Cooling and power consumption are two primary contributors to maintenance costs that have to be calculated during the planning process. Thus, accurate estimation of cooling and energy consumption is a key guideline during the planning phase of constructing a data center. Thermal models and simulators could be used to make decisions during the design period.

The outlet temperature of storage systems is affected by a number of factors. A research investigates the impact of the inlet temperature and CPU utilization on the outlet temperature of a data node [5]. Another research presented a temperature forecast model by using historical temperatures and air flow measurements [9]. Researchers also investigated the relationship between seek times and disk temperatures and demonstrated that the disk temperature is affected by platters [10]. Since a storage data server can be comprised of more than 100 hard disks [11], the outlet temperatures of data nodes can be significantly affected by these disks. Unfortunately, there is a lack of study on the impacts of disks on outlet temperatures of a data node.

Contributions. In this paper, we introduce a thermal modeling approach that is able to estimate the outlet temperature of a storage server (a.k.a., data node) based on processor and disk activities. We make the following contributions. First, a thermal profile of a storage server that contains multiple hard disks is created. The profiling results are collected by launching I/O intensive workloads, which are generated by Postmark [12]. The disk temperatures, as well as the inlet and outlet temperatures, are recorded while varying the workload. Second, we create a thermal model that I s able to estimate inlet/outlet temperature difference by using inlet temperatures and workloads. The model can derive outlet temperatures from CPU and dis k workload. Finally, we study the impact of different data placement schemes on cooling cost, and propose a data placement strategy.

Organization. The rest of this paper is organized as follows. Section 2 provides related research issues. Section 3 illustrates four groups of preliminary experiments and observations. We generate linear regression model for each workload scenario. In Section 4, we propose a thermal model for storage systems and a framework for estimating the cooling cost of data nodes. In Section 5, we discuss the impact of data placement on cooling costs and propose a thermal-aware data placement strategy. Finally, Section 6 concludes the paper.

II. Related Work

A. *Energy-Efficient Data Centers*

Increasing number of energy efficiency studies in datacenters have been presented [13], [14]. A study in 2000shows that the total energy consumption in data centers goes up to approximately 1.2% of U.S. energy consumption [15]. The rapid growth of computing and storage capacity is one of the reasons behind the striking energy consumption in data centers.

A variety of energy-saving approaches have been proposed for energy cost reduction. Measurement and management technologies (MMT) were designed by Bieswanger *et al*. for energy-efficient data centers [16]. The MMT model relies on real measurements, in which run-time analysis of energy consumption can be provided by deploying sensors in data centers. Based on these analytical data, an optimal schedule in terms of energy consumption is selected for data center operation.

There were 22 data centers benchmarked and observed by Greenberg *et al*. In their study, the annual energy cost per square foot in a typical data center is more than 15 times of an office building [17]. In addition, a set of best practices, such as air management, optimizing the size of data centers and utilizing chilled water for cooling, were examined. The experimental results indicate that energy savings in data centers is potentially achievable. Verma *et al*. proposed a Sample-Replicate-Consolidate Mapping (SRCMap) approach that enables energy proportionality for dynamic I/O workload [18]. In this approach, a minimal number of physical volumes are activated and used for duplicating a selected subset of data in other volumes. When serving I/O requests, these quests are redirected to replicas on active volumes. Other volumes can be kept in sleep mode as long as possible. According to their experiments, significant power consumption reduction of enterprise storage systems can be achieved by using SRCMap.

B. *Green Cloud*

Energy-efficient solutions are proposed in cloud computing as well. Key research challenges were identified when Berl *et al.* extended existing energy-efficient techniques to cloud computing environments [19]. A dynamic resource provisioning and allocation algorithm for energy-efficient management on cloud computing environments was proposed by Buyya *et al.* [20]. In their study, the quality-of-service expectations and characteristics of device power usage were taken into account in resource allocation optimization in terms of energy consumption. They discovered that significant cost savings can be achieved in cloud computing.

Solutions that are able to reduce energy consumption in the virtual machine (VM) environments have been presented [21], [22]. Liu *et al.* proposed a GreenCloud architecture, in which the power consumption can be minimized without penalty of performance degradation [21]. Based on the analysis of data from online monitoring, they proposed a VM placement algorithm that minimized the energy costs via VM migrations. The experimental results show that GreenCloud can save up to 27% of energy.

Ye *et al.* explored energy-efficient techniques on virtual machine environments [22]. Since existing energy-efficient techniques cannot be simply applied in VM environments due to the difference of semantic contexts, on which VM and VM monitor are running, they proposed a mechanism that is able to increase energy efficiency of a hard disk by reducing the number of spin-ups and increasing disk sleep time. Their simulation results indicate a 14.8% decrease on energy consumption with only 0.5% penalty on execution times.

Kaushik and Bhandarkar designed greenHDFS that integrates energy conservation techniques into HDFS [23]. HDFS is classified into hot and cold zones by data popularities. Files are moved back and forth between the two zones based on the time stamps of last accesses. Because the servers in the cold zone are turned down to conserve energy consumption and the ones in the hot zone keep working, the performance of HDFS is not significantly compromised. Simulation results show that $2.4 million can be saved annually if greenHDFS is applied at Yahoo.

C. Thermal-aware Resource Management Strategies

As a major contributor to the power costs of data centers, the power consumption of cooling systems becomes an important issue, on which a number of studies primarily focus. Thermal-aware resource management strategies were proposed to optimize the cooling costs by balancing the temperature distribution among data nodes in datacenters.

A thermal-aware load balancing framework that applies local and regional policies for dynamical workload distribution was proposed by Moore *et al.* [3]. According to their simulations, keeping a uniform temperature distribution, promoted by an asymmetric workload placement, is able to reduce energy consumption and improve equipment reliability. Tang *et al.* studied recirculation process and proposed a task scheduling algorithm, XInt, for homogeneous data centers [5]. Under the help of XInt, there circulation costs can be minimized by balancing the workload within the data center. Motivated by an observation that cooling costs significantly depend on peak inlet temperatures, Tang *et al.* designed a task assignment policy, MPIT-TA, that is able to lower the peak inlet temperature, thereby minimizing the cooling costs [24]. In their experiments that simulate a small-scale data center, MPIT-TA offers at least 20% of cooling energy savings.

Other temperature-aware load-balancing strategies were demonstrated in [25] [26]. In these studies, CPU temperatures are limited to a customized threshold by the strategies for energy conservation. CPU voltage and frequency would be dynamically adjusted with execution time penalty if CPU temperatures exceed a specific threshold.

Thermal-aware resource management techniques which focus on the thermal impact of processors have been widely studied. However, the impact of disks on thermal management has not been fully explored. To provide extensive data capacity, each data node may deploy multiple disks. Under I/O-intensive workloads, disk utilization will be pushed to extremely high. Appropriately managing I/O workload may potentially reduce the cooling costs of data centers.

D. Disk Energy Consumption and Temperature Models

Models that analyze disk energy consumption were presented in [8], [27], [28]. Zedlewski *et al.* built a simulation environment, Dempsey, for modeling disk power consumption [28]. In this environment, performance and power consumption parameters for a disk can be derived without the disk specifications from the manufacturer.

There have been some researches working on disk temperature modeling. For instance, research on modeling the temperature of disk drives has been conducted in the late 1980s [29], and Eibeck *et al.* proposed a thermal model that predicts transient temperatures of IBM 5-1/4-in fixed disk drives [30]. Gurumurthi *et al.* constructed an integrated disk drive model to investigate the thermal behavior of a hard disk [31]. The model estimated the heat generated by internal drive air, spindle motor, the base and cover of disk, the voice-coil motor, and disk arms. Kim *et al.* studied the relationship between seek times and disk temperatures, and explored the thermal behaviors of disks by varying platter types and number of platters [10]. Later, Tan *et al.* evaluate transient temperatures during frequent seeking by using a three-dimensional model [32].

Nevertheless, the impact of workloads on disk temperatures is missing. And it is worth noting that the above studies model disk temperatures in a fine-grained level. In order to estimate the temperature of a new disk, one has to acquire detailed structure or device specification of the disk. In addition, previous energy consumption models ignore the impact of disk temperatures on cooling systems. In this paper, we pay attention to the modeling of disk temperatures from a coarse-grained perspective. We comprehensively evaluate the impacts of CPU and disk temperatures on outlet temperature of a data node.

III. Thermal Impacts of Disk I/O

Various components contribute to outlet temperatures of data nodes (e.g., CPU, disk, motherboard, and ambient temperatures). The thermal impact of CPU has been widely investigated in prior studies ([24]-[26]). However, the thermal impact of disk activities on data nodes remains an open issue. In this section, we conduct four experiments to characterize the impacts of CPU and disks on the outlet temperatures.

A. *Testbed*

The testbed is equipped with four Intel(R) Xeon 2.4 GHz CPU, 2.0 GBytes RAM, and three 160 GBytes SATA disks deployed in a disk array. The detailed information are summarized in Table I. We use the inner sensors to monitor the CPU temperatures and acquired the temperature data by lm-sensors [33]. Disk temperatures are monitored by its interior temperature sensor and could be acquired by *hddtemp* [34]. The inlet and outlet temperatures are collected by using an infrared thermometer.

TABLE I: Testbed Configurations

Hardware	Software
4 * Intel(R) Xeon 2.4 GHz CPU X3430	Ubuntu10.04 Linux kernel 2.6.32
1 * 2.0 GBytes of RAM	lm-sensors [33]
3 * WD 160 GBytesSata disk (WD1600AAJS-75M0A0 [35])	hddtemp [34]

B. *Impact of CPU and Disks on Inlet/Outlet Temperatures*

To investigate the relationship between the workload of CPU/disks and the inlet/outlet temperatures, we conduct preliminary experiments, in which a combination of full utilization and idle state of CPU and disks are considered. CPU utilization is calculated as a percentage of time that CPU is processing instructions; a disk's utilization is measured as a percentage of time during which I/O requests are processed by the disk. Both CPU and disk utilizations are monitor by *iostat* system command.

TABLE II: Configuration of the Preliminary Experiments

Experiments	Workload		Power (W)
	CPU	Disk	
1	idle	idle	73
2	high	idle	135
3	low	high	85
4	high	high	142

The environment temperature (i.e., CRAC temperature) is set to 23.2 °C. The configurations of the experiments are shown in Table II. *stress* [36] is used to generate high workload on CPU, which drives the CPU running under extreme high utilization. Postmark [12] is launched to generate I/O-intensive tasks. We use a power meter to measure the power consumption (or computing cost) of the data node.

1) *Low CPU and low disk utilizations*

In the first experiment, both CPU and disks are staying in idle and the utilizations of CPU and disk are very low. The experimental results are shown in Fig. 1.

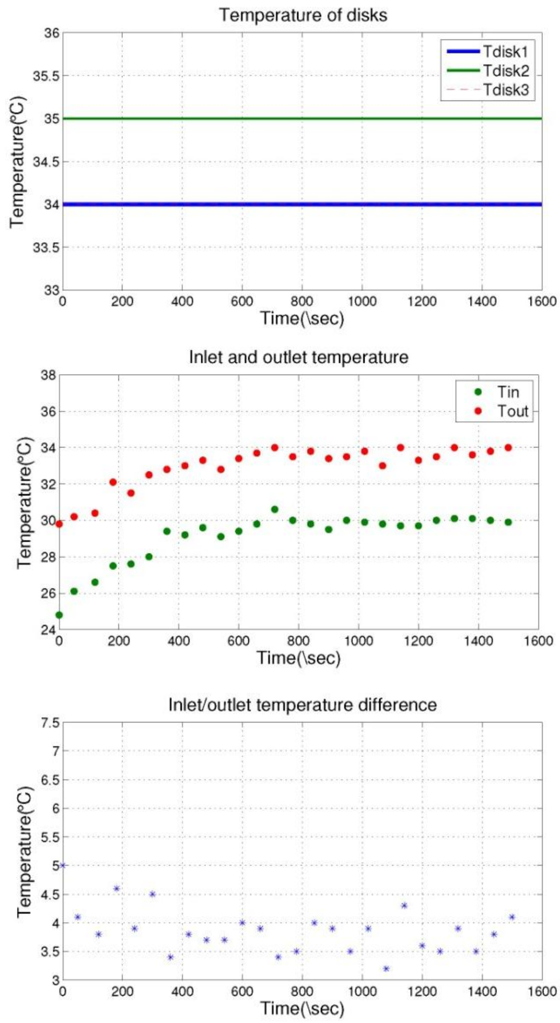


Fig. 1. Temperature evaluation under the low CPU and low disk utilizations.

We observe that the inlet temperature of the data node changes between 24.8 °C and 30.6 °C, which lead the outlet temperature to vary accordingly. When the outlet temperature increases, the inlet temperature also raises due to heat recirculation around the data node.

On average, we find a discrepancy of 3.87 °C between the inlet and outlet temperatures, ranging anywhere between 3.2 °C and 5.0 °C. In this case, the discrepancy between inlet and outlet temperatures is expressed as a constant. Thus, we have:

$$T_{diff1}(t) = 3.87 \quad (1)$$

Compared with real measurements, we demonstrate that this model has a low precision error of 1.00%.

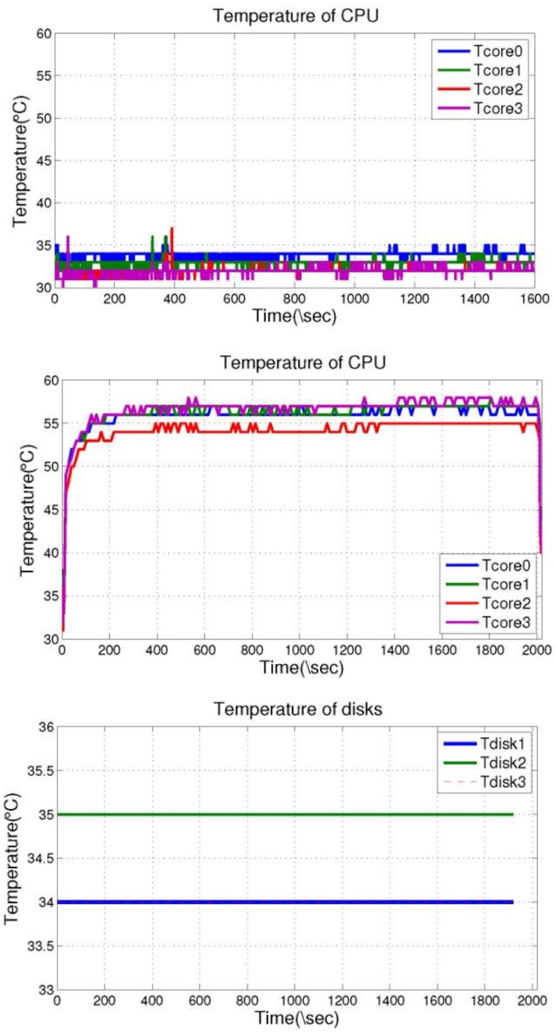
2) High CPU and low disk utilizations

In the second experiment, we keep CPU extremely busy (i.e., CPU utilization approaches 100%) while making disks remain in the idle state.

Fig. 2 shows that the CPU temperatures go up very quickly with an average increment of 20°C in 4 minutes. After that, the temperatures of the four CPU cores are steadier, with their temperatures between 53 °C and 57 °C. Since disks are idle, there are no big changes for the disk temperatures. Discrepancy between the inlet and outlet temperatures increases about 2°C gradually (i.e., from 4.6°C to 6.6°C) in the first 600 seconds, and then maintains steady in the following 1200 seconds. We denote the inlet and outlet temperature difference as T_{diff2}, where t is the time at which the data node has run under extremely high CPU utilization and disk is sitting idle. We generate a linear-regression model to fit the difference between inlet and outlet temperature:

$$T_{diff2}(t) = \begin{cases} 0.0023 * t + 4.88; & if\ t \leq 600 \\ 6.27; & if\ t > 600 \end{cases} \quad (2)$$

The precision error of this model is as low as 4.30%.



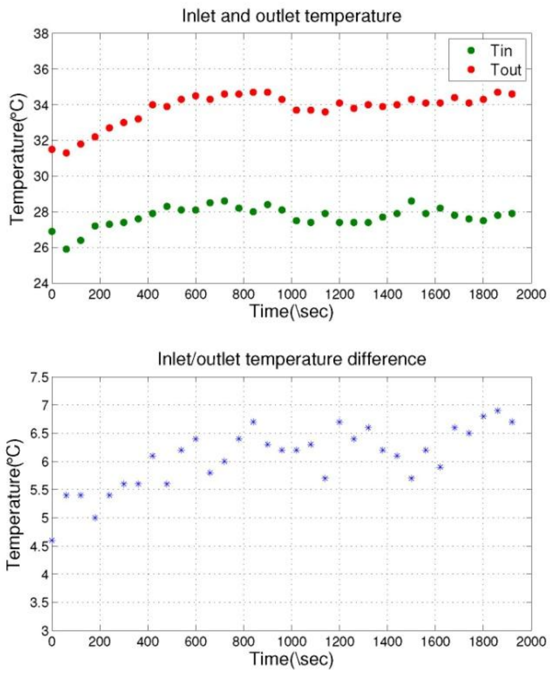
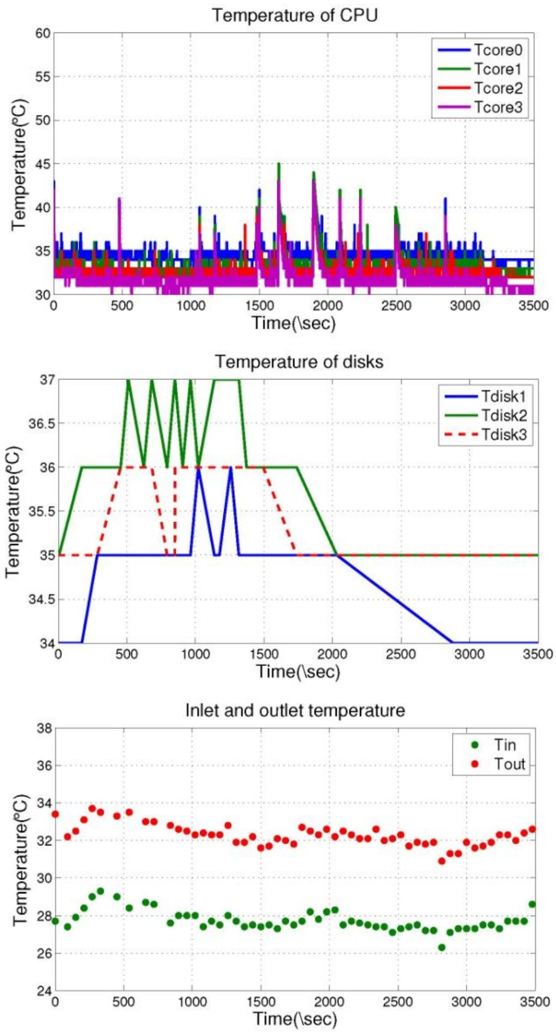


Fig. 2. Temperature evaluation under the high CPU and low disk utilizations.



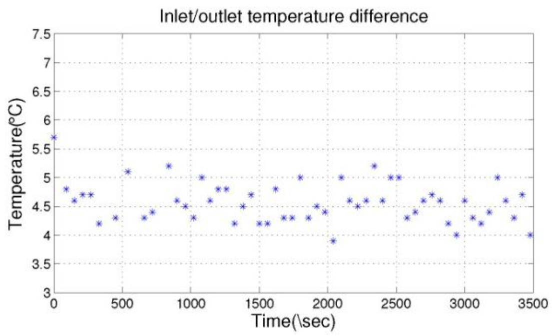


Fig. 3. Temperature evaluation under the low CPU and high disk utilizations.

3) Low CPU and high disk utilizations

In the third experiment, we keep a low CPU utilization while increasing disk utilization up to approximately100%. We use Postmark to launch three I/O-intensive tasks, each of which drive a single disk to be nearly fully utilized. Fig. 3 shows the temperature of the four CPU cores frequently fluctuates between 31 °C and 35 °C because the CPU issues I/O requests. Nevertheless, the CPU utilization remains fairly low. In this case, the thermal impact of CPU is negligible. In contrast, we observe a slowly increase of disk temperatures at the rate of 2 °C per 1000 seconds. The discrepancy between inlet and outlet temperatures can be modeled as follow:

$$T_{diff3}(t) = \begin{cases} 0.0001 * t + 4.61; & if\ t \leq 1000 \\ 4.71; & if\ t > 1000 \end{cases} \quad (3)$$

This model exhibits a small precision error of 5.46%.

4) High CPU and high disk utilizations

In the final experiment, we launch CPU-intensive task and disk I/O-intensive tasks at the same time. Both CPU and disks utilizations are pushed to be fully utilized, as shown in Fig. 4. We observe an increment of 20 °C for CPU temperature at the beginning, and it goes back to its initial temperature after the CPU-intensive task is completed.

Thus, we investigate the thermal behaviour of the data node before the CPU-intensive task is finished. We observe that the inlet and outlet temperature difference ranges between 4.3°C and 7.5°C. In the first 660seconds, the inlet/outlet temperature discrepancy rises rapidly and then does not fluctuate much. We could conclude from the experiment that CPU and disks have significant impact on the outlet temperature, and the discrepancy between inlet and outlet temperatures can be expressed as follow with a precision error of 5.77%.

$$T_{diff4}(t) = \begin{cases} 0.0023 * t + 5.37; & if\ t \leq 660 \\ 6.89; & if\ t > 660 \end{cases} \quad (4)$$

Cold-start time is a time period in which a component is heating up from an initial temperature to a steady temperature in the active state. From Fig. 4, we also observe that the average cold-start time for the three disks is more than 1200 seconds, which is much larger than the cold-start time of CPU (i.e., 100 seconds).

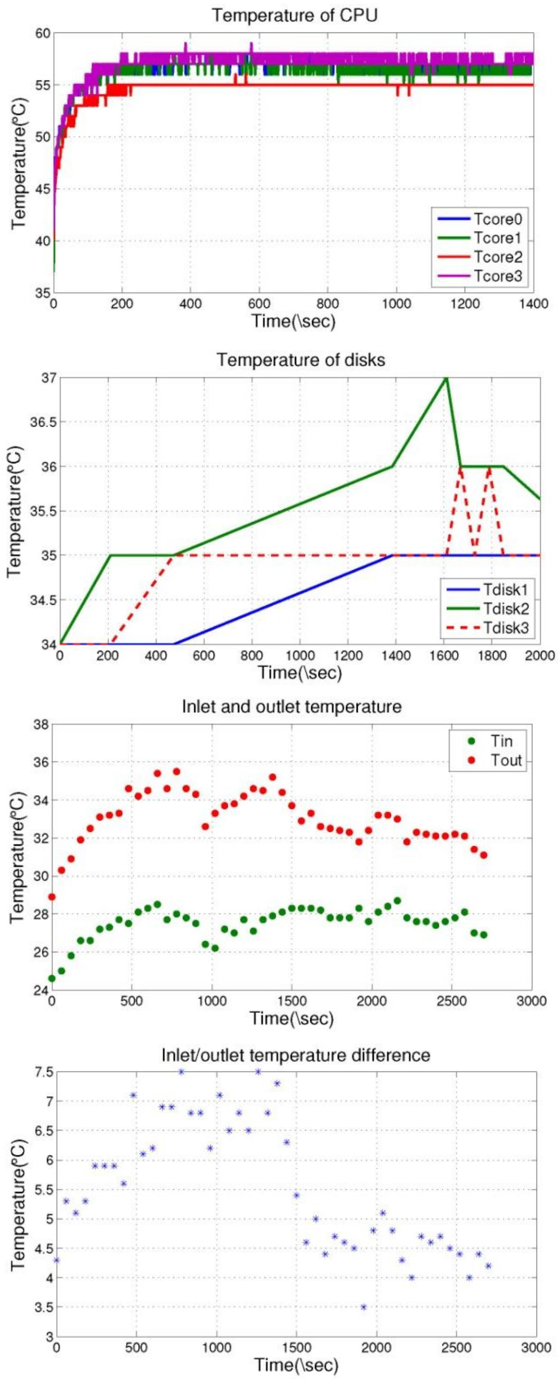


Fig. 4. Temperature evaluation under the high CPU and high disk utilizations.

IV. Thermal Models

Accurately modeling the energy consumption relationship between computing and cooling systems is extremely challenging. Cooling costs of data storage systems is determined not only by the cooling settings (e.g., inlet temperatures and cooling equipment placement), but also by heat dissipated by computing facilities. CPU and disks are major components and heat contributors in data nodes. In this section, we develop a thermal model which aims at estimating outlet temperatures by considering the impacts of both CPU and disks. In addition, our thermal model could be used to predict the impact of CPUs and disks on cooling cost by combining a coefficient of performance (COP, for short) model that estimating cooling costs by CRAC supply temperature and computing cost of IT facilities [3].

A. Framework

Fig. 5 displays our thermal-modeling framework, which consists of three components, namely, inlet/outlet-temperature model, computing cost model and COP.

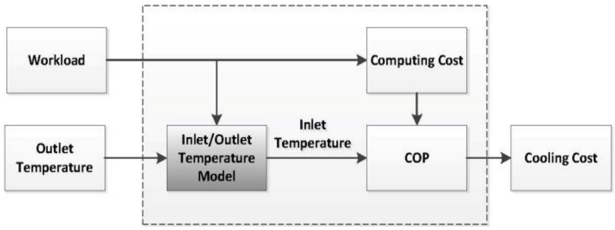


Fig. 5. Framework of estimating cooling cost.

The inlet/outlet-temperature model builds up the relationship between inlet and outlet temperatures by profiling analysis. In addition, given an outlet temperature, our model estimates inlet temperatures under certain workloads. The computing cost model estimate the power consumption of data nodes according to the utilization of the components. The COP model estimates cooling costs by considering inlet temperatures offered by the inlet/outlet-temperature model.

The main contributions of this framework are: (1) a thermal model that characterizes the relationship between inlet/outlet temperature difference and the workload of a data node and (2) cooling costs estimation of data storage nodes.

B. An Inlet/Outlet Temperature Model

Considering CPU and disk utilizations, we classify workload of a node into several basic types (i.e., see Section III.B for a combination of high and low CPU and disk load). The previous section shows a new modeling approach, in which the inlet/outlet temperature difference is derived from CPU and disk load.

When the CPU, disk, inlet temperatures are changing, we apply the proposed modeling approach to refine the thermal model. When CRAC supply temperature is changed, our approach can be used as a guideline to build thermal models. For any application, the workload of a node can be decomposed into a number of sub-periods, in which the node runs under various combinations of environment temperature and CPU/disk load.

We use the following linear regression model to fit the trend of the inlet/outlet temperature difference:

$$T_{diff}(t) = \begin{cases} a * t + b \\ c \end{cases} \quad (5)$$

, where t is the time period measured in second, a, b, c are the parameters generated by our modeling approach.

Given workloads and a number of sub-periods T = $\{t_1 \ldots t_n\}$, we derive the outlet temperature as follow:

$$T_{diff}(T) = \frac{\sum_{i=1}^{n} T_{diff}(t_i)}{|T|} \quad (6)$$

C. Computing Cost Model

The electrical power consumption of IT facilities has been demonstrated to be proportional to their utilizations. Thus the computing cost of a data node could calculated as following: by taking the workload as a input, the utilization of CPU and disks could be determined; then we generate the computing cost by considering the utilization of each component of the data node. Thus, the computing cost of a data node which has n components could be expressed as follow:

$$P_C = \sum_{i=1}^{n} \left(P_{Comp}^{Min}(i) + U_{Comp} * \left(P_{Comp}^{Max}(i) - P_{Comp}^{Min}(i) \right) \right) \quad (7)$$

, where P_C is the total computing cost of a data node, $U_{Comp}(i)$ represents the utilization of the ith component, $P_{Comp}^{Max}(i)$ is the power of i^{th} component when it is fully utilized and $P_{Comp}^{Min}(i)$ is the power of the i^{th} component when it maintains in idle state.

D. The COP Model

The energy consumption of a data node is contributed by the electrical energy cost of the node and the cooling cost. We use COP (i.e., the Coefficient of Performance model), described in [3], to compute the cooling cost.

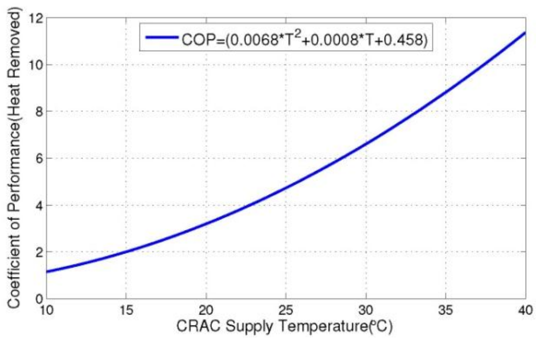


Fig. 6. Coefficient of the performance curve for the chilled-water CRAC units at the HP labs utility data center [3]

Fig. 6 plots COP values that increase with the supply temperatures of CRAC. Larger COP value indicates higher energy efficiency.

$$COP(T) = 0.0068 * T^2 + 0.0008 * T + 0.458 \quad (8)$$

In 8, COP represents the ratio of heat removed to the energy cost of the cooling system for heat removal. T is the supply temperature of CRAC. If we use P_C to represent the energy cost of IT facilities in data center, and P_{AC} to represent the energy cost of the air-conditioner, then cooling cost is inversely proportional to the COP value.

$$P_{AC} = \frac{P_C}{COP(T)} \quad (9)$$

Given the power of a data node, we can calculate the energy dissipation. The total energy cost of the data node could be expressed as follow:

$$P_{Total} = \sum_{j=1}^{m}\left(P_C(j) + \frac{P_C(j)}{COP(T_{Inlet}(j))}\right) \qquad (10)$$

, where j represents small time period in the execution process, $T_{Inlet}(j)$ is the inlet temperature in the j^{th} time period. We conduct some case studies and show how to apply our model to estimate the total energy cost under different workload scenarios in a previous work [37]. Experimental results show our thermal model could accurately estimate the inlet/outlet temperature difference.

V. Data Placement Strategies

A. Thermal Impacts of Data Placement

Our evidence shows that disks have non-negligible thermal impacts on data nodes (see Section III). In this section, we demonstrate that data placement strategies can significantly affect thermal performance of data nodes. In this data placement study, we use the same testbed described in Section III. The number of disks in this group of experiments is set to two and three, respectively.

1) The two-disk case

In the first group of experiments, two disks are configured in the HP server. It is noteworthy that both disks are placed inside the node's chassis rather than an external disk array. These two disks are of the same type. Compared with disk 2, disk 1 is kept closer to the fan. The initial disk temperature of disk 1 is 36 °C, and the initial disk temperature of disk 2 is 38 °C. Two I/O-intensive tasks are running on the two disks. We leverage Postmark to create 100 files, the size of which ranges anywhere between 1 to 100 MBytes.

Each of the two tasks issues a total of 2,000 I/O requests to access the files stored on the disks. We set up three scenarios summarized in Table III. In scenario 1, the two tasks are keeping both disks busy. In scenarios 2 and 3, the two tasks are accessing on one disk while keeping another disk idle.

TABLE III: The Two-Disk Scenarios

Scenarios	Disk 1	Disk 2
1	Task 1	Task 2
2	Task 1& 2	
3		Task 1& 2

Fig. 7 shows the disk temperatures in the three tested scenarios. In scenario 1, the temperature of disk 1 increases by 4°C, and disk 2 increases by 3°C. In scenario2, after running for a few minutes, the temperature of the disk1 increases by 3 °C , and the temperature of disk2 increases by 1 °C. In scenario 3, the temperature of disk 2 increases by 4 °C, and the temperature of disk1 increases by 2 °C as well. Table IV compares the execution times and peak average temperatures of the two disks we tested in the three scenarios. Task execution time is the sum of the two tasks' execution times; application execution time is the maximum execution time of the two tasks involved in the application. We observe that scenario 3 results in the shortest accumulative active disk time (i.e., 3,981 seconds) compared with scenario 1 (i.e., 4,136 seconds) and scenario 2 (i.e., 5,323 seconds), concluding that disks tested in scenario 3 may consume the least energy. Evenly distributing requests issued by the application to the two disks (see scenario 1) produces a high average disk temperature. However, scenario 1 exhibits smaller application execution time than those of scenarios 2 and 3. More interestingly, issuing requests to disk 1 that is closer to the fan in the chassis (see scenario 2) gives rise to the lowest average disk temperature. This result reveals that scenario 2 is more thermal friendly than the other two scenarios.

TABLE IV: Peak Average Disk Temperatures and Total Task/Application Execution Times.

Scenarios	Peak Average Temperature (°C)	Execution Time(s) Task	Application
1	40.5	4,136	2,250
2	39.0	10,632	5,323
3	40.0	7,948	3,981

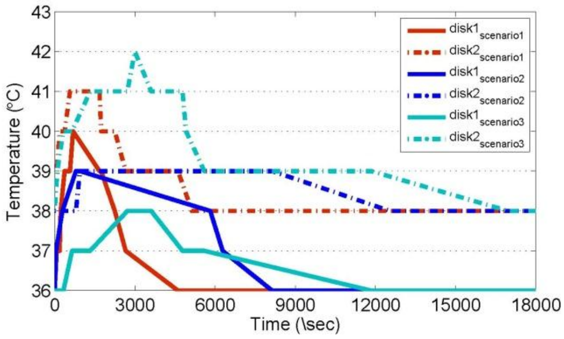


Fig. 7. Thermal impacts of data placement in the two-disk case.

TABLE V: The Three-Disk Scenarios

Scenarios	Disk 1	Disk 2	Disk 3
1	Task 1	Task 2	Task 3
2	Task 1 & 2 & 3		
3		Task 1 & 2 & 3	
4			Task 1 & 2 & 3
5	Task 1 & 2	Task 3	
6	Task 1 & 2		Task 3
7	Task 1	Task 2 & 3	
8	Task 1		Task 2 & 3
9		Task 1	Task 2 & 3
10		Task 3	Task 1 & 2

2) The three-disk case

We deploy three disks inside a disk-array chassis connecting to the data server. The disk-array chassis has a fan to cool down disks. We use postmark to initially create 100 files, the size of which ranges from 1 to 100MBytes. Three postmark tasks issue 1,000 requests to the disks. Ten scenarios (see Table V) are investigated in this group of experiments. In the first scenario, the three tasks are accessing the three disks. In the next three scenarios, the three tasks are sharing a single disk. And for the other scenarios, different task assignments are examined.

Fig. 8 plots the disk utilization and temperature of the ten scenarios examined in the three-disk case. The peak average temperatures of three disks, the task/application executing times and the estimated cooling cost of each scenario are summarized in Table VI, where task execution time is the sum of the three tasks' execution times; application execution time is the maximum execution time of the three tasks within the application.

TABLE VI: PEAK AVERAGE DISK TEMPERATURES, EXECUTION TIMES AND ESTIMATED COOLING COSTS.

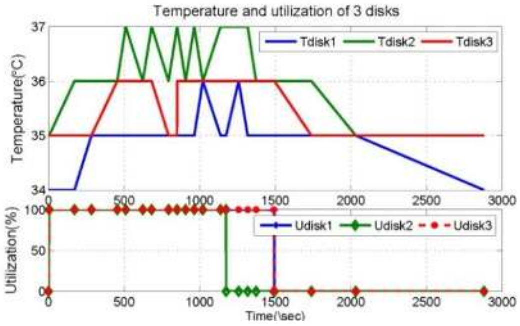
Scenarios	Peak Average Temperature (°C)	Execution Time(s) Task	Application	Cooling Cost (J)
1	36.35	4144	1500	23,655
2	35.33	3010	3010	48,527
3	35.00	3024	3024	48,671
4	35.00	3126	3126	50,065
5	35.34	2616	1768	28,469
6	34.67	4271	2551	41,169
7	35.34	3032	2134	34,340
8	35.00	4466	2751	44,370
9	35.33	2717	1846	29,723
10	35.35	3227	2063	33,244

We observe that evenly distributing tasks to the disks(i.e., Scenario 1) leads to higher temperatures on average than forcing all the tasks to share a single disk, however, it takes 1,500 seconds (the shortest time) to complete all the I/O requests. Fig. 8(a) shows that the temperatures of disk 1 and 2 increase by 2 °C; the temperature of disk 3 increases by 1 °C. When the three tasks are sharing one disk, the disk temperature increases by 2 °C, whereas temperatures of the other two disks remain unchanged. We conclude that sharing a disk among multiple tasks can maintain low disk temperatures at the cost of increased I/O processing time (e.g., from 1,500 to3,000 seconds).

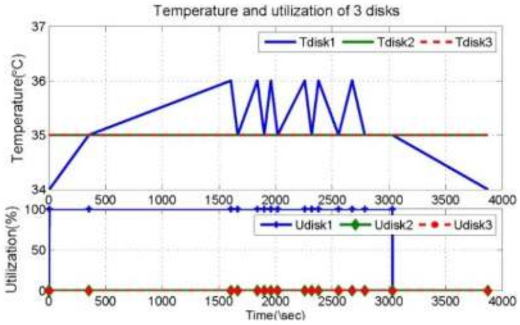
In both Scenarios 5 and 6, two tasks are issuing I/O requests to disk 1 and the third task is sending I/O requests to another disk. The task execution times in these two scenarios are 2,616 and 4,271 seconds, respectively. The long execution time of Scenario 6 keeps the three disks in a higher temperature than the initial state. Fig. 8(e) shows that the temperature of disk 1 increases by 3°C, and the temperature of disk 2 increases by 1°C. Fig. 8(f) indicates that the temperatures of disks 1 and 3 both increase by only 1°C.

In Scenarios 7 and 9, two tasks are assigned to disk 2 and the third one is allocated to the third disk. The execution times of these three tasks are very close. Fig. 8(g) and Fig. 8(i) show that the temperature of disk 2 increases by 3°C. The temperature of disk 1 in Scenario 7 rises by only 1°C; however, the temperature of disk 3 in Scenario 9 goes up by 2°C. The disks lead to higher energy consumption in Scenario 7 than in Scenario 9.

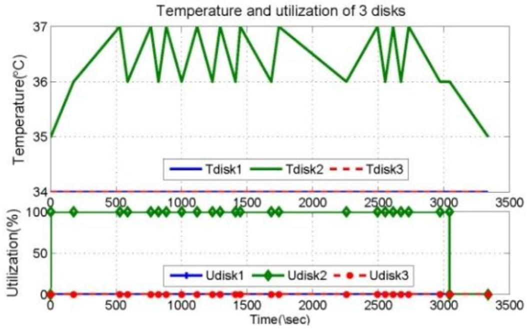
When it comes to Scenarios 8 and 10, disk 3 handles requests from two tasks, and another disk deals with the requests from the third task. The task execution time of the Scenario 8 is much longer than that of Scenario 10. Let us consider the first 4,000 seconds during the testing process. Fig. 8(h) and Fig. 8(j) illustrate that the average temperature of the three disks in Scenario 10 is higher than that in Scenario 8. These results confirm that assigning tasks to a disk sitting in the middle can give rise to high disk temperatures and low energy efficiency.



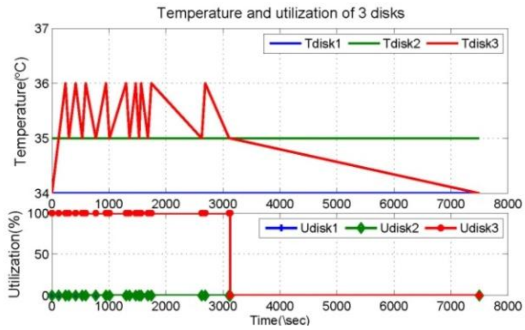
(a) Scenario 1



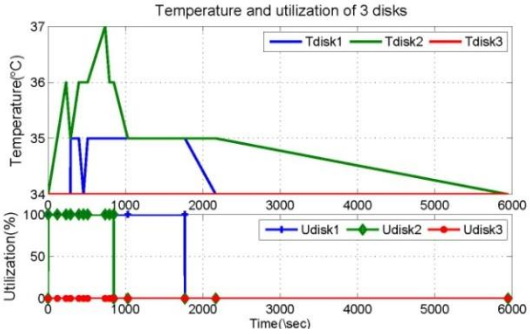
(b) Scenario 2



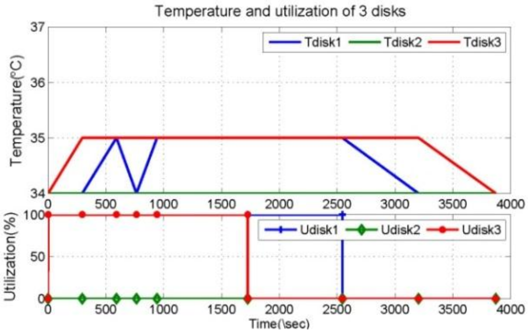
(c) Scenario 3



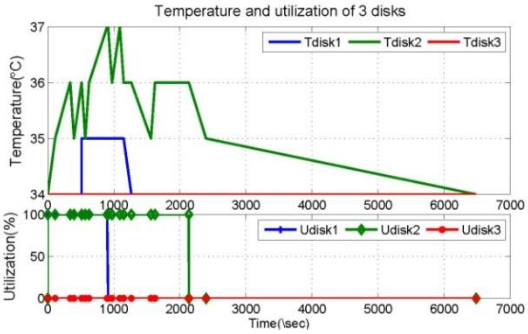
(d) Scenario 4



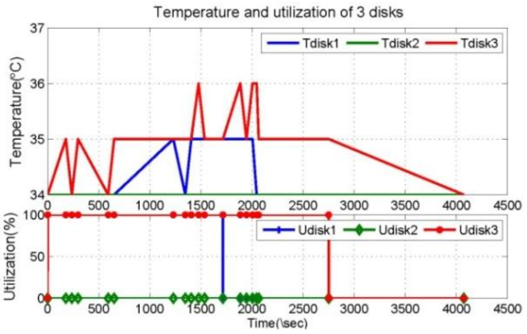
(e) Scenario 5



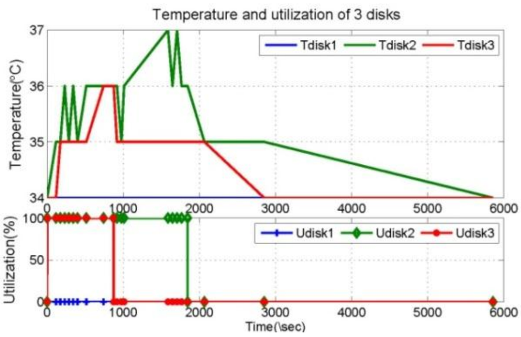
(f) Scenario 6



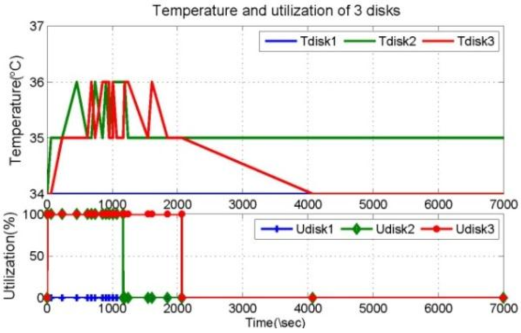
(g) Scenario 7



(h) Scenario 8



(i) Scenario 9



(j) Scenario 10

Fig. 8. Thermal impacts of data placement in the three-disk case.

From Table VI, we observe that the cooling cost of Scenario 1 is the least and cooling cost of Scenario 4 is the most. From the above experiments, we conclude that though evenly distribute tasks have the highest peak average temperature because a load balancing strategy which makes disks stay in high temperatures for less time offers better overall performance, and it is more energy-efficient.

B. Thermal-Aware Data Placement

The previous subsection shows evidence that outlet temperatures affected by disks vary greatly among the tested cases. In the three-disk case, we chose to evaluate ten scenarios out of many other possibilities. For example, one possible scenario might be that the workload is composed of tasks that are of different disk utilizations or of different execution times. And to provide large storage capacity, one may increase the number of disks in each data node. Manually measuring all possible scenarios is a time-consuming and impractical process. A promising solution is to use real measurements collected in simple disk configurations, and to model the thermal characteristics of other complicated scenarios. Our results suggest that disk temperatures significantly affect the outlet temperatures of a node. Disk temperatures in turn depends on data placement and I/O activities. These observations motivate us to study thermal-aware data placement strategies, which aim to migrate data among disks in order to minimize the cooling costs.

Let us consider a storage cluster containing a large number of data nodes. Encouraged by our experimental results presented in the previous sections, we propose a thermal-aware data placement strategy that is composed of two stages:

- Initial stage: placing data sets in data nodes in a way that all the nodes have very similar outlet temperatures.
- Redistribution stage: migrating data according to temperature distribution measured by sensors and predicted by our models.

In the initial stage, a large amount of data must be written into data nodes of a storage cluster. A straight-forward strategy is to evenly distribute data across all the data nodes in the system. Data nodes of a storage cluster can be configured in two ways. The first strategy is designed for storage clusters where nodes have the same number of disks deployed. In this strategy, more amounts of data are placed on disks whose temperature in the idle state is higher than other disks. The second strategy is tailored for heterogeneous storage clusters where nodes have different number of disks. In this case, data nodes equipped with more disks should handle a less amount of data in order to reduce heat stress.

After the initial stage of a storage cluster, the data access patterns are likely to change dynamically. For example, some data sets are accessed more frequently than the other data. The storage cluster tends to exhibit unbalanced outlet temperatures of the data nodes. To balance thermal stresses, with the data placement mechanism, hot data sets will be migrated from nodes with high outlet temperatures to those with low outlet temperatures. The data redistribution process is triggered by a threshold of outlet temperatures. For instance, when the maximum outlet temperature is 25% higher than the average temperature of all the nodes, the data redistribution process begins. To maintain high I/O performance, our mechanism delays the redistribution process until the nodes involved in the migration procedure have very large I/O load.

VI. Conclusions and Future Work

Energy efficiency of data storage systems must be urgently addressed because there have been fast increase of energy consumption and cooling costs of large-scale storage systems in data centers for the past decade. Recent studies show that cooling costs contribute to a growing portion of the operational costs of data centers. Thermal management techniques are adopted to reduce the energy consumption of cooling systems for improving the energy efficiency of data centers. Thermal models play a key role in thermal management; however, traditional thermal models of data centers give little thought of disks.

In this study, we proposed a thermal modeling approach to predict the outlet temperature of data nodes, which offers the following two benefits. First, the thermal model approach makes it possible to estimate the outlet temperature without the specification of CPU and disks. Our thermal models were developed at a coarse-grained level, where there is no need to know the details of data nodes. In case of changing environment temperatures, one may adjust a model by conducting profiling experiments to refine parameter values of the model. Our method enables data storage systems to reduce thermal monitoring costs. Second, our thermal model enables data center designers to make intelligent decisions on thermal management during the design phase. Thermal management of storage systems helps to cut cooling costs and boost system reliability. Monitoring temperatures is a key issue in thermal management techniques; however, it is prohibitively expensive to acquire and set up a huge number of sensors in a large-scale data center. Our modeling method is an alternative to monitoring temperatures of storage systems. In addition, we study the impact of data placement on the cooling cost and thermal performance of storage system, and propose a thermal-aware energy-efficient data placement strategy.

There are several research directions we intend to address in the future. First, we plan to design a thermal-aware data placement system. Through the deployment of our proposed strategy, this new system aims to handle the placement of new data and the management of existing data. In addition, the system is able to load balancing I/O load to reduce cooling cost of data centers. Second, we will apply our thermal-aware data placement scheme to address the thermal management issues in a real-world data center. We intend to conduct experiments to verify the performance and energy efficiency of our thermal-aware data placement system in the data center.

Acknowledgment

This research was supported by the U.S. National Science Foundation under Grants CCF-0845257 (CAREER), CNS-0917137 (CSR), CNS-0757778 (CSR), CCF-0742187 (CPA), CNS-0831502 (CyberTrust), CNS-0855251 (CRI), OCI-0753305 (CI-TEAM), DUE-0837341 (CCLI), and DUE-0830831 (SFS). Mohammed I. Alghamdi's research was supported by AL-Baha University and the King Abdulaziz City for Science and Technology (KACST). Meikang Qiu was partially supported by NSF CNS-1359557.

References

[1] Global Data Center Energy Demand Forecasting. (2011). Technical Report, Institution. [Online]. Available: http://www.datacenterdynamics.com/research/energy-demand2011-12

[2] U.S. Environmental Protection Agency, Report to Congress on Server and Data Center Energy Efficiency, Technical Report, August 2007.

[3] J. Moore, J. Chase, P. Ranganathan, and R. Sharma, "Making scheduling "cool": Temperature-aware workload placement in data centers," in *Proc. Annual Conference on USENIX Annual Technical Conference*, USENIX Association, Berkeley, CA, USA, 2005, pp. 5.

[4] N. Vasic, T. Scherer, and W. Schott, "Thermal-aware workload scheduling for energy efficient data centers," in *Proc. 7th International Conference on Autonomic Computing*, New York, NY, USA, 2010, pp. 169–174.

[5] Q. Tang, S. Gupta, and G. Varsamopoulos, "Thermal-aware tasks cheduling for data centers through minimizing heat recirculation," in *Proc. IEEE International Conference on Cluster Computing*, Sept 2007, pp. 129 –138.

[6] E. Pinheiro, W.-D. Weber, and L. A. Barroso, "Failure trends in a large disk drive population," in *Proc. 5th USENIX Conference on File and Storage Technologies*, USENIX Association, Berkeley, CA, USA, 2007, pp. 2.

[7] T. Xie and Y. Sun, "Understanding the relationship between energy conservation and reliability in parallel disk arrays," *J. Parallel Distrib. Comput.*, vol. 71, pp. 198–210, February 2011.

[8] M. Allalouf, Y. Arbitman, M. Factor, R. I. Kat, K. Meth, and D. Naor, "Storage modeling for power estimation," in *Proc. SYSTOR*

2009: The Israeli Experimental Systems Conference, New York, NY, USA, 2009, pp. 3:1–3:10.
[9] L. Li, C.-J. M. Liang, J. Liu, S. Nath, A. Terzis, and C. Faloutsos, "Thermocast: A cyber-physical forecasting model for datacenters," in *Proc. 17th ACMSIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, 2011, pp. 1370–1378.
[10] Y. Kim, S. Gurumurthi, and A. Sivasubramaniam, "Understanding the performance-temperature interactions in disk i/o of server workloads," in *Proc. Twelfth International Symposium on High-Performance Computer Architecture*, Feb 2006, pp. 176–186.
[11] A. Pavlo, E. Paulson, A. Rasin, D. J. Abadi, D. J. DeWitt, S. Madden, and M. Stonebraker, "A comparison of approaches to large-scale data analysis," in *Proc. ACM SIGMOD International Conference on Management of Data*, New York, NY, USA, 2009, pp. 165–178.
[12] J. Katcher, "Postmark: A new file system benchmark," *System*, pp. 1–8, 1997.
[13] D. J. Brown and C. Reams, "Toward energy-efficient computing," *Commun. ACM*, vol. 53, no. 3, pp. 50–58, March 2010.
[14] L. A. Barroso and U. Hölzle, "The case for energy proportional computing," *Computer*, vol. 40, no. 12, pp. 33–37, December 2007.
[15] J. G. Koomey, "Estimating total power consumption by servers in the U.S. and the world," Technical Report, Lawrence Derkley National Laboratory, February 2007.
[16] A. Bieswanger, H. F. Hamann, and H.-D. Wehle, "Energy efficient datacenter," *It - Information Technology*, vol. 54, pp. 17–23, 2012.
[17] S. Greenberg, E. Mills, B. Tschudi, P. Rumsey, and B. Myatt, "Best practices for data centers: Lessons learned from benchmarking 22 data centers," *ACEEE Summer Study on Energy Efficiency in Buildings*, 2006.
[18] A. Verma, R. Koller, L. Useche, and R. Rangaswami, "Srcmap: Energy proportional storage using dynamic consolidation," in *Proc. 8th USENIX Conference on File and Storage Technologies*, USENIX Association, Berkeley, CA, USA, 2010, pp. 20–20.
[19] A. Berl, E. Gelenbe, M. D. Girolamo, G. Giuliani, *et al.*, "Energy-efficient cloud computing," *The Computer Journal*, vol. 53, no. 7, pp. 1045–1051, September 2010.
[20] R. Buyya, A. Beloglazov, and J. Abawajy, "Energy-efficient management of data center resources for cloud computing: A vision, architectural elements, and open challenges," in *Proc. International Conference on Parallel and Distributed Processing Techniques and Applications*, Las Vegas, USA, July 12-15, 2010.
[21] L. Liu, H. Wang, X. Liu, X. Jin, W. B. He, *et al.*, "Greencloud: a new architecture for green data center," in *Proc. 6th International Conference Industry Session on Autonomic Computing and Communications Industry Session*, New York, NY, USA, 2009, pp. 29–38.
[22] L. Ye, G. Lu, S. Kumar, C. Gniady, and J. H. Hartman, "Energy-efficient storage in virtual machine environments," *SIGPLAN Not.*, vol. 45, no. 7, pp. 75–84, March 2010.
[23] R. T. Kaushik and M. Bhandarkar, "Greenhdfs: Towards an energy-conserving, storage-efficient, hybrid hadoop compute cluster," in *Proc. International Conference on Power-Aware Computing and Systems*, USENIX Association, Berkeley, CA, USA, 2010, pp. 1–9.
[24] Q. Tang, S. K. S. Gupta, and G. Varsamopoulos, "Energy-efficient thermal-aware task scheduling for homogeneous high-performance computing data centers: A cyber-physical approach," *IEEE Trans. Parallel Distrib. Syst.*, vol. 19, no. 11, pp. 1458–1472, November 2008.
[25] O. Sarood, A. Gupta, and L.V. Kale, "Temperature aware load balancing for parallel applications: Preliminary work," in *Proc. IEEE International Symposium on Parallel and Distributed Processing Workshops and Phd Forum*, May 2011, pp. 796 –803, May 2011.
[26] O. Sarood and L. V. Kale, "A 'cool' load balancer for parallel applications," in *Proc. International Conference for High Performance Computing, Networking, Storage and Analysis*, New York, NY, USA, 2011, pp. 21:1–21:11.
[27] A. Hylick, R. Sohan, A. Rice, and B. Jones, "An analysis of hard drive energy consumption," in *Proc. IEEE International Symposium on Modeling, Analysis and Simulation of Computers and Telecommunication Systems*, Sept 2008, pp. 1–10.
[28] J. Zedlewski, S. Sobti, N. Garg, F. Zheng, A. Krishnamurthy, and R. Wang, "Modeling hard-disk power consumption," in *Proc. 2nd USENIX Conference on File and Storage Technologies*, USENIX Association, Berkeley, CA, USA, 2003, pp. 217–230.
[29] N. S. Clauss, "A computational model of the thermal expansion within a fixed disk drive storage system," Master's thesis, University of California, Berkeley, 1988.
[30] P. A. Eibeck and D. J. Cohen, "Modeling thermal characteristics of a fixed disk drive," *IEEE Transactions on Components, Hybrids, and Manufacturing Technology*, vol. 11, no. 4, pp. 566–570, Dec 1988.
[31] S. Gurumurthi, A. Sivasubramaniam, and V. K. Natarajan, "Disk drive roadmap from the thermal perspective: A case for dynamic thermal management," *SIGARCH Comput. Archit. News*, vol. 33, no. 2, pp. 38–49, May 2005.
[32] C. P. H. Tan, J. P. Yang, J. Q. Mou, and E. H. Ong, "Three dimensional finite element model for transient temperature prediction in hard disk drive," in *Proc. Magnetic Recording Conference*, Asia-Pacific, Jan 2009, pp. 1–2.
[33] Lm-sensors. [Online]. Available: http://www.lm-sensors.org/
[34] Hddtemp. [Online]. Available: http://manpages.ubuntu.com/manpages/natty/man8/hddtemp.8.html
[35] Wd1600aajs Specification. [Online]. Available: http://www.wdc.com/wdproducts/library/SpecSheet/ENG/2879-701277.pdf
[36] Stress. [Online]. Available: http://www.unixref.com/manPages/stress.html
[37] X.-F. Jiang, M. I. Alghamdi, J. Zhang, M. Al Assaf, *et al.*, "Thermal modeling and analysis of storage systems," in *Proc. 31st IEEE Int'l Performance Computing and Communications Conf*, 2012.

Xunfei Jiang is a PhD student in Computer Science and Software Engineering, Auburn University. She received the B.S. and M.S. degrees in Computer Science from Huazhong University of Science and Technology (HUST), China, in2004 and 2007. Then she joined Digital Video Networks Co., Ltd. in 2007 and Cisco Systems (Shanghai) Video Technology Co., Ltd in2010. Her research interests include storage systems, big data, parallel and distributed computing, data-intensive computing, cluster computing, thermal modeling and management, task scheduling and resource management. She is a student member of IEEE.

Mohammed I. Alghamdi received the B.S. degree in Computer Science from King Saud University, Riyadh, Saudi Arabia in 1997. He received the M.S. degrees in Software Engineering and Information Technology Management from Colorado Technical University, Denver, Colorado, 2003. He received the Ph.D. degree in Computer Science from New Mexico Institute of

Mining and Technology. Currently, he is an Assistant Professor with the Department of Computer Science, Al-Baha University, Kingdom of Saudi Arabia.

Maen M. Al Assaf received BS degree in computer science from Applied Science University, Amman, Jordan in 2006 and MA degree in Computer and Network Security from DePaul University (Jordan Campus) in 2008 and the Ph.D. degree in Computer Science from Auburn University, AL in 2011. Currently, he is an assistant professor of Computer Science in King Abdullah II School for Information Technology -The University of Jordan, Amman, Jordan. His research focus is distributed operating systems. He is a member of IEEE, ACM, and ASIST.

Xiaojun Ruan is an Assistant Professor in the Department of Computer Science, West Chester University of Pennsylvania. He received B.S. degree in Computer Science and Technology from Shandong University, Jinan, China, in 2005and his Ph.D. degree in the Department of Computer Science and Software Engineering, at Auburn University 2011 advised by Dr. Xiao Qin. His research interests include Energy Efficient Storage Systems, parallel and distributed systems, Computer Architecture, Operating Systems, HDD and SSD technologies, and Computer Security.

Ji Zhang is a Ph.D student in the Department of Computer Science and Software Engineering at Auburn University. He received the B.S. and M.S.in Computer Science from HUST, China in 2004and 2007. He worked as a software engineer in Huawei Technologies from 2007 to 2010. His research interests include I/O-intensive computation, parallel and distributed file systems and geographic information systems.

Meikang Qiu received the BE and ME degrees from Shanghai Jiao Tong University, China. He received the M.S. and Ph.D. degree of Computer Science from University of Texas at Dallas in2003 and 2007, respectively. Currently, he is an associate professor of Computer Engineering at San Jose State University. He has worked at Chinese Helicopter R & D Institute, IBM, etc. Currently, he is an IEEE Senior member and ACM Senior member. His research interests include cyber security, embedded systems, cloud computing, smart grid, microprocessor, data analytics, etc. A lot of novel results have been produced and most of them have already been reported to research community through high-quality journal (such as IEEE Trans. on Computer, ACM Trans. on Design Automation, IEEE Trans. on VLSI, and JPDC) and conference papers (ACM/IEEE DATE, ISSS+CODES and DAC). He has published 3 books, 200 peer-reviewed journal and conference papers (including 90 journal articles, 110 conference papers), and 3 patents. He has won ACM Transactions on Design Automation of Electrical Systems (TODAES) 2011 Best Paper Award. His paper about cloud computing has been published in JPDC (Journal of Parallel and Distributed Computing, Elsevier) and ranked No 1 in 2012Top Hottest 25 Papers of JPDC. He has won another 4 Conference Best Paper Award (IEEE/ACM ICESS'12, IEEE GreenCom'10, IEEEEUC'10, and IEEE CSE'09) in recent four years. His research is supported by US government such as NSF, Air Force, Navy and companies such as Nokia, TCL, and Cavium.

Xiao Qin received the B.S. and M.S. degrees in computer science from the HUST, China, and the Ph.D. degree in computer science from the University of Nebraska-Lincoln, Lincoln, in 1992, 1999, and 2004, respectively. Currently, he is an Associate Professor with the Department of Computer Science and Software Engineering, Auburn University. His research interests include parallel and distributed systems, storage systems, fault tolerance, real-time systems, and performance evaluation. He received the U.S.NSF Computing Processes and Artifacts Award and the NSF Computer System Research Award in 2007 and the NSF CAREER Award in 2009.He is a senior member of the IEEE.

