VMAP: Proactive Thermal-aware Virtual Machine Allocation in HPC Cloud Datacenters

Eun Kyung Lee, Hariharasudhan Viswanathan, and Dario Pompili
NSF Cloud and Autonomic Computing Center
Department of Electrical and Computer Engineering, Rutgers University, New Brunswick
e-mail: {eunkyung_lee, hari_viswanathan, pompili} @cac.rutgers.edu

***Abstract*—*Clouds* provide the abstraction of nearly-unlimited computing resources through the elastic use of federated resource pools (virtualized datacenters). They are being increasingly considered for HPC applications, which have traditionally targeted grids and supercomputing clusters. However, maximizing energy efficiency and utilization of cloud datacenter resources, avoiding undesired thermal hotspots (due to overheating of over-utilized computing equipment), and ensuring quality of service guarantees for HPC applications are all conflicting objectives, which require joint consideration of multiple pairwise tradeoffs. The novel concept of *heat imbalance,* which captures the unevenness in heat generation and extraction, at different regions inside a HPC cloud datacenter is introduced. This *thermal awareness* enables *proactive* datacenter management through prediction of future temperature trends as opposed to the state-of-the-art reactive management based on current temperature measurements. VMAP, an innovative proactive thermal-aware virtual machine consolidation technique is proposed to maximize computing resource utilization, to minimize datacenter energy consumption for computing, and to improve the efficiency of heat extraction. The effectiveness of the proposed technique is verified through experimental evaluations with HPC workload traces under single- as well as federated-datacenter scenarios (in the machine rooms at Rutgers University and University of Florida).**

***Index Terms*—Virtualized datacenters, thermal awareness, heat imbalance, consolidation.**

I. Introduction

Datacenters are a growing component of society's IT infrastructure and their energy consumption surpassed 237 billion kWh/year worldwide and 76 billion kWh/year in the US in 2010 [1]. Even though these numbers are lower than what the US Environmental Protection Agency predicted in 2007 [2], they correspond to 6% and 2% of the total electricity usage in the US. The impact of this proliferation of datacenters on the environment and society includes increase in CO_2 emissions, overload of the electricity supply grid, and rise in water usage for cooling leading to water scarcity [3]. The scale and complexity of datacenters are growing at an alarming rate and their management is rapidly exceeding human ability, making *autonomic* (self-configuration, self-optimization, self-healing, and self-protection) management approaches essential.

High-Performance Computing (HPC) applications are resource-intensive scientific workflow (in terms of data, computation, and communication) that have typically targeted Grids and conventional HPC platforms like super-computing clusters. *Clouds* – composed of one or more virtualized datacenters providing the abstraction of nearly-unlimited computing resources through the elastic use of federated resource pools – are being increasingly considered to enable traditional HPC applications. However, maximizing energy efficiency and utilization of cloud datacenter resources, avoiding undesired thermal hotspots (due to overheating of over-utilized computing equipment), and ensuring Quality of Service (QoS) guarantees for HPC applications are all conflicting objectives, which require joint consideration of multiple pairwise tradeoffs.

From our feasibility study and proof-of-concept experiments conducted at our machine room in the NSF Cloud and Autonomic Computing Center (CAC), Rutgers University, we have inferred that one of the fundamental problems in HPC-cloud datacenters is the *local unevenness in heat-generation* and *heat-extraction rates*: the former can be attributed to the non-uniform distribution of workloads (of different types and intensities) among servers and to the heterogeneity of computing hardware; the latter can be attributed to the non-ideal air circulation, which depends on the layout of server racks inside the datacenter and on the placement of Computer Room Air Conditioning (CRAC) unit fans and air vents. The heat-generation and -extraction rates may differ, which over time causes *heat imbalance*. This heat imbalance will be large if the rates are significantly different from each other or if their difference prolongs over extended time periods.

A large negative heat imbalance at a particular region inside a datacenter will result in energy-inefficient overcooling and, hence, in a significant decrease in temperature. Conversely, a large positive heat imbalance will lead to a significant temperature increase, which may result in undesired thermal hotspots and server operation in the unsafe temperature range. Thus, *thermal awareness,* which is the knowledge of heat imbalance at different regions inside a datacenter, is essential to maximize energy and cooling efficiency as well as to minimize server system failure rate. Our novel concept of heat imbalance enables *proactive* datacenter management decisions (such as resource provisioning, cooling system optimization) through prediction of future temperature trends as opposed to the state-of-the-art *reactive* management decisions based on current temperature measurements.

In virtualized HPC datacenters, one or more Virtual Machines (VMs) are created for every application request (with one or more workloads) and each VM is provisioned with

This work was supported by the NSF Award No. CSR-1117263.

resources that satisfy the application QoS requirements, which are based on Service Level Agreements (SLAs). Once VMs are provisioned, they have to be allocated to servers. We propose a novel thermal-aware proactive VM consolidation solution referred to as VMAP. The benefit of employing VMAP is three-fold: i) energy spent on computation can be saved by turning off the unused servers after workload (or VM) consolidation; ii) the utilization of servers that are in the "better cooled" areas of the datacenters (with high heat extraction) can be maximized; iii) heat can be extracted more efficiently (by doing a lower amount of work) by the CRAC system from the consolidated server aisles, which are hotter than non-consolidated server aisles. Note that (iii) is possible due to the fact that the efficiency of heat extraction increases with increase in return-air temperature.

VMAP also exploits the heterogeneity in the cloud infrastructure (federated datacenters) – in terms of electricity cost, hardware capabilities (CPU, memory, disk I/O, and network subsystems), tunable parameters of the CRAC system, and local regulations (governing CO_2 emission and water usage) – to maximize energy efficiency. VMAP is aimed at increasing the energy and cooling efficiency and at decreasing equipment failure rates so to minimize both the impact on the environment and the Total Cost of Ownership (TCO) of datacenters. VMAP can significantly contribute to energy efficiency (9%, 9%, and 35% average reduction in energy consumption compared to the traditional temperature-based reactive thermal management schemes: first-fit-decreasing, best-fit-decreasing, and "cool-job" [4] allocation, respectively) while not violating recommended operating temperature range. The following are the main contributions of our work.

- We introduce the novel notion of heat imbalance and validate a simple yet robust heat-imbalance model, which helps predict future temperature trends and make proactive resource provisioning decisions;
- We propose a proactive thermal-aware VM consolidation solution, VMAP (for self-optimization of computing resources), which minimizes energy consumption for computation, increases resource utilization, and improves efficiency of cooling;
- We validate our proposed approach through extensive experiments – in a single-datacenter as well as in federated-datacenters (at different sites of NSF CAC, Rutgers University and University of Florida).

The remainder of this paper is organized as follows: in Sect. II, we outline our broader vision for thermal-aware autonomic datacenter management; in Sect. III, we present details on the design and validation of our heat-imbalance model; in Sect. IV, we describe our proposed thermal-aware VM allocation scheme (VMAP); in Sect. V, we study the performance of VMAP using experiments and simulations; in Sect. VI, we present an overview of the state of the art in autonomic thermal-aware management of datacenters; and finally, in Sect. VII, we briefly discuss future work and conclude.

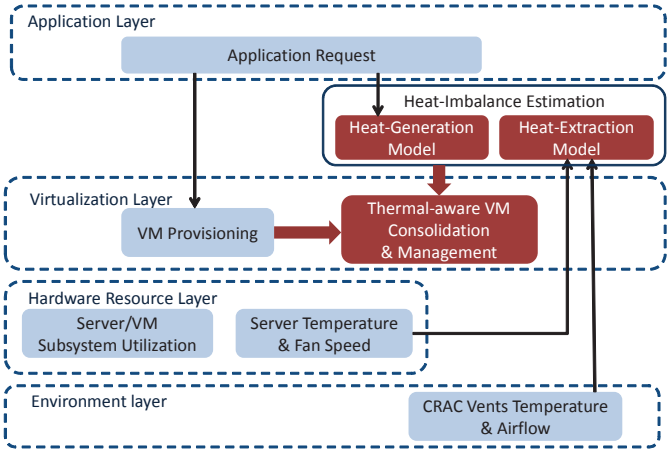

Fig. 1. Envisioned cross-layer approach to autonomic management of virtualized datacenters. The main focus of this paper is indicated in red boxes.

II. Proposed Approach

We propose a proactive *cross-layer* approach to autonomic datacenter management, which is *information centric* and requires continuous processing and analysis of real-time feedback from multiple layers of abstraction (as depicted in Fig. 1). The *application layer* provides information regarding the applications' (and, hence, the workloads') characteristics such as their computing resource requirements, energy consumption, and performance on different hardware platforms. Modern blade servers (*hardware resource layer*) are equipped with a number of internal sensors that provide information about server fan speed and subsystem operating temperatures as well as utilization. However, information extracted from the application and hardware resource layers alone cannot capture the complex thermodynamic phenomena of heat and air circulation inside a datacenter.

Information from the *environment layer,* comprising of an heterogeneous sensing infrastructure (with scalar temperature and humidity sensors, thermal cameras, and airflow meters) is key to characterize the thermal behavior of a datacenter under a given load (information from the *application layer*) [5]. As mentioned earlier, the estimation of heat imbalance requires estimation of the heat-generation and heat-extraction rates. The heat-generation model exploits the information provided by the application layer while the heat-extraction model leverages information provided by the environment as well as hardware resource layers. The *virtualization layer* – which provisions, allocates, and manages VMs (created based on application requests) – exploits the knowledge of heat imbalance to predict future temperature trends for optimal resource allocation in datacenters. In this paper, we focus on the design and validation of the heat-imbalance model and on how the knowledge of heat imbalance can be exploited to perform energy-efficient proactive VM consolidation in datacenters (shown in red boxes in Fig. 1).

While proactive VM consolidation has several clear advantages, namely, reduced energy cost for computation (through high utilization of fewer computing resources) as well as for

cooling (through better heat extraction at higher operating temperatures), it has certain drawbacks. Increased utilization of servers results in continuous operation of computing hardware at temperatures close to the upper bound of the recommended operating temperature range. This, however, is not a major concern due to the following reasons: i) manufacturers usually provide a conservative upper bound for the recommended operating temperature range; ii) our consolidation solution is thermal-aware and does not let the operating temperatures go beyond the recommended range (referred to as *thermal violation*) unlike other temperature-agnostic solutions; iii) the frequency of equipment upgrades (due to tremendous rate of innovation in computing hardware) is much higher than the rate of replacement due to failures. Last, but not least, our heat-imbalance based predictive approach allows us to create a *fingerprint* of expected hotspots inside the datacenter. This fingerprint can then be used for detecting anomalies such as mis-configuration of servers (mismatch between VM allocation decision and the actual allocation) and/or attacks (unidentified VMs running illegitimate workloads on servers) in the thermal domain.

Another drawback of traditional server consolidation is violation in SLAs (in terms of application runtime) due to greater resource contention at higher utilization levels. However, this is not a concern for virtualized HPC clouds as i) users are guaranteed the resources they specifically ask for, ii) VMs are isolated from each other, and iii) we do not multiplex resources, i.e., the total subsystem utilization of all VMs in a server will not exceed the total subsystem capacity of that server. In our prior work [6], we have shown through simulations that heat-imbalance-based proactive datacenter management (cooling system optimization) is superior in terms of energy efficiency and minimization of risk of equipment failures compared to its conventional temperature-measurement-based reactive counterpart. Our envisioned approach represents a transformative shift towards cross-layer autonomics for datacenter management problems, which have so far been considered mostly in terms of individual layers. In this paper, we first focus on our novel heat-imbalance model, which incorporates information from the application, hardware resource, and environment layers. We then present our heat-imbalance-based proactive VM allocation solution, which resides in the virtualization layer.

III. Heat-Imbalance Model

A VM is created for every application request and is provisioned with resources (CPUs, memory, disk, and network capacity) that satisfy the application's QoS (usually deadline) requirements. Without any loss of generality, we assume that this provisioning has already been performed using techniques such as the ones described in [7]. The provisioned VMs now have to be allocated to physical servers housed within racks in datacenters. Let $\mathcal{M}$ be the set of VMs to be allocated and $\mathcal{N}$ be the set of servers. An associativity binary matrix $\mathbf{A} = \{a_{mn}\}$ (with $a_{mn} \in \{0, 1\}$) specifies whether VM m is hosted at server n or not. A VM m is specified as a vector $\mathbf{\Gamma}_m = \{\gamma_m^s\}$, where $s \in \mathcal{S} = \{CPU, MEM, IO, NET\}$ refers to the server subsystems and γ_m^s's are the VM subsystem requirements (e.g., CPU cores, amount of volatile memory [MB], disk storage space [MB], network capacity [Mbps]).

Representation (or *mapping*) of a VM's subsystem requirement (γ_m^s) as a factor of physical server subsystem capacity is straightforward if all the servers of the datacenter are assumed to be homogeneous. For example, a VM *m* requiring 4 virtual CPUs, 2GB of RAM, 64GB of hard-disk space, and 100Mbps network capacity can be represented as $\mathbf{\Gamma_m} = \{0.25, 0.125, 0.125, 0.1\}$ if all the servers in a datacenter have 16 CPU cores, 16GB of RAM, 512GB of local hard-disk space, and a gigabyte ethernet interface. However, homogeneity is rarely the case as datacenters usually have a few different generations of each subsystem, for example, CPUs with different clock rates and number of cores (1.6/2.0/2.4 GHz and 4/8/16/32 cores), different generations or sizes of RAMs (SDR/DDR SDRAM or sizes ranging from 4 to 32 GB), network switches of varying capacities (0.1, 1, or 10 Gbps), etc. The mapping problem becomes non trivial in an heterogeneous environment. However, assuming that only a small finite number of generations of each subsystem are present in the datacenter, we create such a mapping for each generation of every subsystem.

Estimation of heat-generation rate: The total power dissipation of a server is estimated based on power dissipation as heat at the CPU and other subsystems. All the subsystems are composed of semiconductor devices, hence we can calculate the leakage power dissipated as heat P_{leak} as given in [8]; P_{leak} provides us with the direct relation between the subsystem utilization and heat dissipation. The heat dissipation factor of a server subsystem is given by $\alpha^s = \frac{P^s_{leak}}{P^s}$, where P^s [W] is the average power utilized and P^s_{leak} [W] is the leakage power for subsystem s. When Advanced Configuration and Power Interface (ACPI) [9] is enabled, a subsystem can potentially transition between multiple 'on' and 'idle' states (apart from the 'off' state). As power management is an operating system functionality, we abstract the details and use the following power utilization model. The average power utilized by the subsystems is given by $P^s = P^{s,on} \cdot u^s + P^{s,idle} \cdot (1 - u^s)$, where $P^{s,on}$ [W] is the average power utilization when subsystem s is in 'on' state, $P^{s,idle}$ [W] is the average power utilization when the subsystem is in idle state, and u^s is the subsystem utilization factor.

The average power utilization of a subsystem on a server n running a set of VMs is determined by the subsystem 'on' time $t^{s,on} = f(\sum_{m \in M} a_{mn} \cdot \gamma_m^s, n, s)$. Here, the γ_m^s used takes into account the appropriate generation of subsystem in use as specified earlier. The utilization factor for a given $\delta = t^{s,on} + t^{s,idle}$ is given by $u^s = \frac{t^{s,on}}{\delta}$ [s]. The heat-generation rate h_n [W] at a server n hosting a set of VMs is given by,

$$h_n = \sum_{s \in S} [P_n^{s,on} \cdot u_n^s + P_n^{s,idle} \cdot (1 - u_n^s)] \cdot \alpha_n^s. \quad (1)$$

Estimation of heat-extraction rate: Heat is extracted by the fans in the server enclosure and by the fan in the CRAC

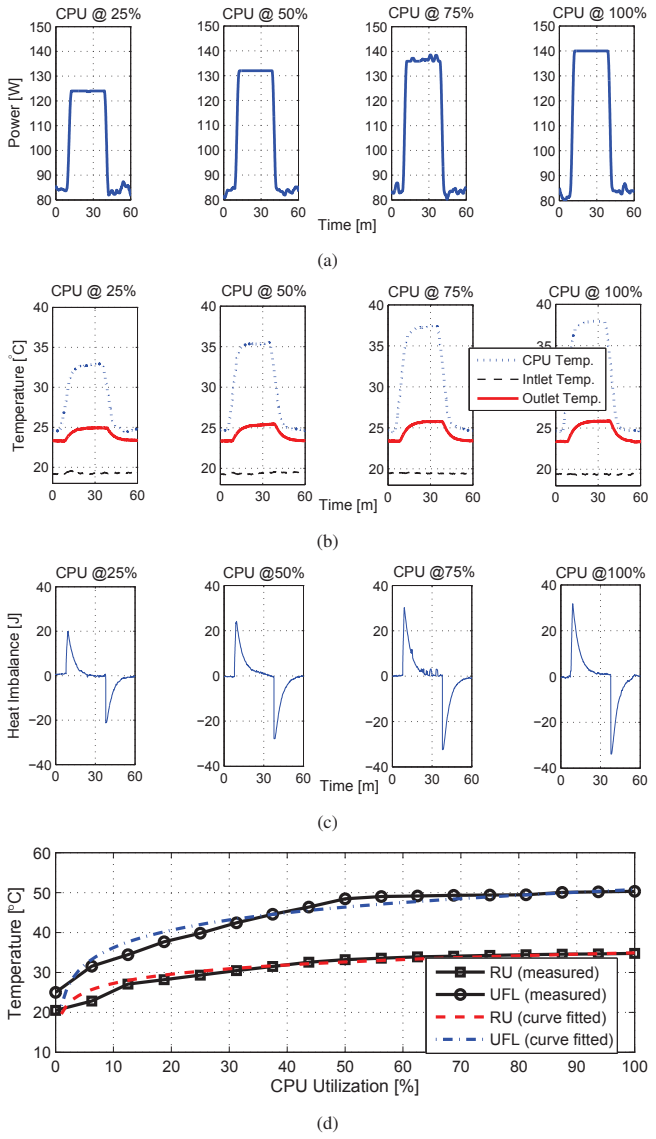

Fig. 2. Empirical data collected from servers at RU – (a) power consumption, (b) CPU and external (inlet and outlet) temperatures, and (c) calculated heat imbalance – when a representative CPU-intensive workload is run at different CPU utilization levels. (d) Relationship between ΔT and CPU utilization using data from both RU and UFL servers.

unit. Most datacenters use chilled-water air conditioning system. The efficiency of cooling can be determined by factors such as airflow and chilled water temperature, and can be quantified by the Coefficient Of Performance (COP)[1]. As the COP is inversely proportional to W, a higher COP means that more heat Q can be removed by doing less work W [4]. As the CRAC supply temperature increases, the COP also increases (in compliance with the second law of thermodynamics). The rate of heat extraction q_n [W] at a server n given by,

$$q_n = m_n^{in} \cdot c_p \cdot (T_n^{out} - T_n^{in}), \tag{2}$$

depends on the mass air flow rate (m_n^{in}) at the cold air inlet of the server and on the temperatures at the cold-air inlet (T_n^{in}) and hot-air outlet (T_n^{out}). Here, c_p is the specific heat capacity of air. In our solution, we use the current measurements for m_n^{in} and T_n^{in} obtained from air flow meters and external temperature sensors, respectively, of our multi-tier sensing infrastructure (environment layer).

Estimation of heat imbalance: We formulate the heat-imbalance model in a datacenter based on heat-generation and heat-extraction rates as follows,

$$\Delta I_n = \int_{t_0}^{t_0+\delta} (h_n - q_n)dt = M_n \cdot C \cdot \Delta T^n_{[t_0,t_0+\delta]}, \tag{3}$$

where ΔI_n [J] denotes the heat imbalance of CPU inside server n during the time between t_0 and $t_0 + \delta$, and M_n and C denote the mass and specific heat capacity, respectively, of the CPU. Note that if ΔI_n is positive (i.e., $h_n > q_n$), the temperature of the CPU at server n increases in the time interval $[t_0, t_0 + \delta]$ (hence, $\Delta T^n > 0$); conversely, if ΔI_n is negative (i.e., $h_n < q_n$), the temperature of the CPU at server n decreases (hence, $\Delta T^n < 0$).

This estimated heat imbalance helps us predict the increase or decrease in temperature, given by ΔT^n, to take management decisions such as VM placement, VM migration, and cooling system optimization.

Validation of the proposed models: Certain parameters in the proposed heat-imbalance model are determined empirically as they cannot be obtained directly (e.g., from server specification documents). The heat dissipation factor α in (1) is one of the key parameters that is determined empirically. Similarly, the server outlet temperature T_{out} in (2) varies with time and is a function of CPU temperature, which is what the heat-imbalance model is designed to estimate. Hence, the relationship between T^{out} and ΔT is determined empirically (assuming T^{in} is known and is constant in the time interval $[t_0, t_0 + \delta]$) and is substituted in the heat-imbalance model so to eliminate an extra unknown. We performed simple experiments (measurements shown in Fig. 2) to obtain α, to derive the relationship between T^{out} and ΔT, and to validate the resulting heat-imbalance model by comparing its output (predicted increase in the CPU temperature ΔT at a server) with actual observation (shown in Fig. 3).

We started from an initial idle condition, with 0% CPU utilization and a corresponding zero heat imbalance, and increased the CPU utilization from 0% to 25%, 50%, 75%, and 100% progressively as shown in Fig. 2. The CPU was subject to each of the aforementioned load levels for around 60 minutes so to allow the CPU temperature to reach steady state. To increase the CPU utilization we used *Lookbusy* (a synthetic load generator for Linux systems), which keep the CPU(s) at the chosen utilization level by adjusting its own load up or down to compensate for other loads on the system. We measured the corresponding increase in power consumption (Fig. 2(a)) as well as CPU and server outlet temperatures (Fig. 2(b)), and also calculated the variation in heat imbalance over time (Fig. 2(c)). Obtaining the value of α using (3) is now straightforward as the heat imbalance, heat extraction,

[1] $COP = \frac{Q}{W}$ is the ratio of amount of work done by the CRAC unit (W [kWh]) to extract a unit quantity of heat (Q [kWh]).

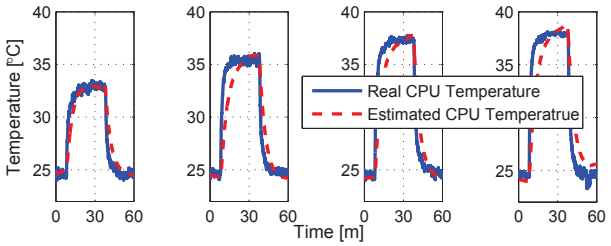

Fig. 3. CPU temperature – measured and estimated (using the heat-imbalance model) – when a representative CPU-intensive workload is run at different CPU utilization levels.

and power consumption are known. On the contrary, deriving the relationship between T^{out} and ΔT is non-trivial.

First, we use logarithmic regression equations to model the relationship between CPU utilization ($u_n\%$) and the increase in CPU temperature (ΔT_n $^\circ C$, shown in Fig. 2(d)), i.e., $\Delta T_n = \alpha \ln(u_n) + \beta$. Then, based on this knowledge and our observation from Fig. 2(c), we derive a simple linear regression model that represents the relationship between ΔT_n and T_n^{out} (for a fixed server inlet temperature and airflow rate) for use in our heat-imbalance model. We verify the accuracy of the logarithmic regression equations with the empirically determined coefficients (α and β) as well as the linear regression model by repeating the aforementioned experiment again and comparing the predicted CPU temperatures over time with the actual CPU operating temperature as shown in Fig. 3. Prediction of future CPU operating temperatures using our heat-imbalance model is sensitive to the variable heat and air circulation patterns (thermodynamic phenomena) at different regions inside a datacenter.

IV. Thermal-aware VM Consolidation

For a given set of VMs, minimizing the number of servers that are in operation (consolidation) will help reduce the energy overhead and, hence, the total energy consumption. In addition to saving the energy spent on computation, thermal-aware VM consolidation also helps achieve a higher COP of cooling. In this section, we first formulate the VM allocation problem as an optimization problem, which employs our heat-imbalance model. As this optimization is NP-hard, we then present our heuristic solution, VMAP (thermal-aware proactive VM mapping solution). The motivation for formulating the optimization problem is to gain insight and make key design decisions for our heuristic solution.

Optimization Problem: The total energy consumption in a datacenter can be split into energy consumption for *computing* (E^{comp} [kWh]), i.e., for running the workloads (or VMs) on servers, and energy consumption for *cooling* (E^{cool} [kWh]). We assume that the cooling system parameters (fan speed and compressor duty cycle of the CRAC) are fixed, i.e., the energy spent on cooling is fixed ($E^{cool} = const$) for the duration δ. Note that E^{cool} can be optimized independently at a periodicity $\Delta \gg \delta$. The goal is to find an *optimal* mapping of VMs to physical servers (represented by the binary associativity matrix $\mathbf{A}$) so to minimize E^{comp} while simultaneously increasing COP of cooling. The known (given as well as measured) parameters and optimization variables of the optimization problem can be summarized as,

$$\begin{aligned}
&\textbf{Given (offline)} && : \mathcal{N}, T^{reco}, \delta, M_n, C_p;\\
&\textbf{Given (online)} && : \mathcal{M}, \mathbf{\Gamma}_m \ \forall m \in \mathcal{M};\\
&\textbf{Measured (online)} && : T_n^{t_0}, m_n^{in}, T_n^{in}, \mathbf{\Lambda}_n \ \forall n \in \mathcal{N};\\
&\textbf{Find} && : \mathbf{A} = \{a_{mn}\},\ m \in \mathcal{M}, n \in \mathcal{N}. \quad (4)
\end{aligned}$$

Here, $T_n^{t_0}$ and $\mathbf{\Lambda}_n = \{\lambda_n^s\}$ represent the current CPU temperature and the maximum residual capacity of each subsystem s at server n, respectively. The objective of the optimization problem is,

$$\textbf{Minimize}: \ E^{comp} = \sum_{n \in \mathcal{N}} E_n^{comp}, \quad (5)$$

$$E_n^{comp} = \sum_{s \in \mathcal{S}} \left(P_n^{s,on} \cdot t_n^{s,on} + P_n^{s,idle} \cdot t_n^{s,idle} \right) \cdot \alpha_n^s; \quad (6)$$

Subject to : **C1, C2, C3.**

The first constraint (**C1**) ensures that a VM is allocated to *one* and *only one* server, i.e.,

$$\textbf{C1:} \sum_{n \in \mathcal{N}} a_{mn} = 1, \forall m \in \mathcal{M}. \quad (7)$$

The second constraint (**C2**) ensures that the resource requirements of all VMs allocated to one server do not exceed the maximum capacity of a server subsystem and is given by,

$$\textbf{C2:} \sum_{m \in \mathcal{M}} a_{mn} \cdot \gamma_m^s \leq \lambda_n^s, \forall n \in \mathcal{N}, \forall s \in \mathcal{S}. \quad (8)$$

The third constraint (**C3**) ensures that the predicted CPU temperature – sum of the current CPU temperature $T_n^{t_0}$ and the predicted temperature increase $\Delta T^n_{[t_0, t_0+\delta]}$ calculated using (3) – is always below the recommended maximum operating temperature (T^{reco}) and is represented as,

$$\textbf{C3:} \ T_n^{t_0} + \Delta T^n_{[t_0, t_0+\delta]} \leq T^{reco}, \forall n \in \mathcal{N}. \quad (9)$$

The optimization problem presented here naturally forces VM consolidation. As heat generation increases logarithmically with increase in CPU utilization (shown in Fig. 4), the optimization problem prefers already loaded active servers for VM allocation when all the constraints (**C1**, **C2**, and **C3**) are met. This is because the additional cost of placing a VM in an already loaded server (in terms of increase in temperature) decreases as the load increases. Also, constraint **C3** ensures that more VMs are allocated to servers in better-cooled areas of the datacenter. Such thermal-aware VM consolidation leads to better utilization of computing resources. In addition, consolidation increases the return air temperature in the consolidated server aisles thus increasing the efficiency of cooling. This can be attributed to the fact that higher the CRAC return air temperature the higher the COP of cooling.

VMAP - Thermal-aware Proactive VM Consolidation: We characterize the aforementioned optimization problem as a *variable-size multi-dimensional bin-packing problem* [10],

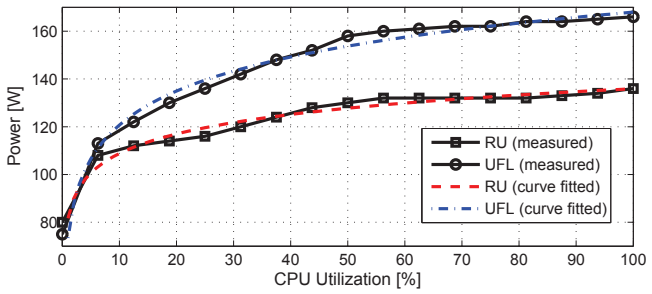

Fig. 4. Relationship between power consumption and CPU utilization in multi-core multi-threaded systems at RU and UFL servers.

[11]. This is a generalized version of the traditional fixed-size one-dimensional bin-packing problem as the *bins (servers)* and *objects (VMs)* are represented as "hypercuboids" with multiple dimensions d (5 in our problem) and all the bins need not have the same capacity along each dimension. The size of each VM along the five different dimensions are its four normalized subsystem utilization requirements and the heat-generation rate. The size of a server along the five different dimensions are the normalized residual capacity (or availability) of each of the four subsystems and the heat extraction rate. The first four dimensions corresponding to VM subsystem requirements (in the object definition) and server subsystem residual capacities (in the bin definition) are straightforward to interpret and incorporated into a bin-packing problem. However, the relationship between the heat-generation (in the object definition) and heat-extraction (in the bin definition) rates is more involved. The bin capacity along the fifth dimension is actually the difference between current CPU temperature (T^{t_0}) and the upper bound of the recommended temperature range T^{reco}.

We use a multi-dimensional best-fit-like algorithm [12] to allocate a set of VMs ($\mathcal{M}$) that have arrived in a time window to a set of physical servers ($\mathcal{N}$). First, the VMs are sorted in decreasing order of their deadlines (or running time). Note that this is a shift from the traditional method of sorting based on one of the dimensions. This is because, in HPC clouds, the subsystem requirements of VMs are comparable and, hence, their durations play a pivotal role in determining energy consumption. It is desirable to pack longer duration VMs together so that server that host smaller duration VMs can be switched off at the completion of workload tasks so to save energy. Once the VMs are sorted according to their deadline, each VM $m \in \mathcal{M}$ is allocated a server $n \in \mathcal{N}$ whose residual volume (of the hypercuboid) is the lowest of all servers' after assignment. The time complexity of the aforementioned heuristic is $\mathcal{O}(|\mathcal{M}| \cdot \log |\mathcal{M}| + d \cdot |\mathcal{M}| \cdot |\mathcal{N}|)$, where the first and second components correspond to the sorting step and the assignment steps, respectively.

The objective of bin packing (minimize the number of bins used) is in line with the objective of the optimization problem, i.e., the fewer the active physical servers, the lower the energy consumption. This is also made possible due to the logarithmic behavior (as shown in Fig. 4) of CPU temperature as well as energy consumption with respect to CPU utilization in multi-core multi-threaded systems (which are the most common computing equipment configuration in cloud datacenters). In addition, bin-packing heuristics require that the objects are not further manipulated (i.e., divided or rotated) and do not overlap inside the bins (similar to constraint **C1**), the total volume of all the object inside a bin cannot exceed the bin's volume (similar to constraints **C2** and **C3**).

VMAP has the ability to optimize resource allocation across a network of heterogeneous yet federated datacenters. Heterogeneity here refers to the difference in characteristics and capabilities of computing (e.g., heat-generation rate of servers, processing power, network capacity, etc.) and cooling (e.g., COP of air-chilled vs. water-chilled cooling) equipment, sources of energy for operation and cooling (e.g., renewable or non-renewable), and environmental regulations in the respective geographical region (e.g., cap on CO_2 footprint or cap on water temperature increase caused by cooling systems). We follow a two-step approach in which the problem of deciding which datacenter should handle the VM and which physical server should host the VM are determined sequentially. For example, if reducing the CO_2 footprint and the aggregate TCO are the goals, the solution will load datacenters that rely on renewable sources of energy as long as the following conditions are met: high COP of cooling, compliance with requirements of VMs/workloads and with environmental regulations such as cap on water consumption and cap on water temperature increase caused by the cooling system. As mentioned earlier, we have a testbed of geographically separated yet federated datacenters to validate our solutions.

V. Performance Evaluation

We evaluated the performance of VMAP via experiments on a small-scale testbed and via trace-driven simulations. The system model used in our simulations has the same characteristics of our real testbed. First, we provide details on our testbed and experiment methodology (workload traces, performance metrics, and competing approaches). Then, we elaborate on the experiment and simulation scenarios aimed at highlighting the benefits of thermal-aware VM consolidation.

A. Testbed and Experiment Methodology

Testbed: We have fully equipped machine rooms at two sites of NSF CAC – Rutgers University (RU) and University of Florida (UFL) – with state-of-the-art computing equipment (modern blade servers in enclosures) and fully controllable CRAC systems. The blade servers at both sites are equipped with a host of internal sensors that provide information about server subsystem operating temperatures and utilization. In addition, the machine room at RU is instrumented with an external heterogeneous sensing infrastructure [5] to capture the complex thermodynamic phenomena of heat generation and extraction at various regions inside the machine room. The sensing infrastructure comprises of scalar temperature and humidity sensors placed at the server inlet (cold aisle)

and outlet (hot aisle), airflow meters at the server outlet, and thermal cameras in the hot aisle.

The computing equipment configuration at RU is two Dell M1000E modular blade enclosures. Each enclosure is maximally configured with sixteen blades, each blade having two Intel Xeon E5504 Nehalem family quad-core processors at 2.0 GHz, forming an eight core node. Each blade has 6 GB RAM and 80 GB of local disk storage. The cluster system consists of 32 nodes, 256 cores, 80 GB memory and 2.5 TB disk capacity. The cooling equipment at RU is a fully controllable Liebert 22-Ton Upflow CRAC system. The computing equipment configuration at UFL is two IBM Blade Center with sixteen blades in each, each blade having two Intel Xeon E5504 Nehalem family quad-core processors at 2.0 GHz, forming an eight core node. Each blade has 24 GB RAM and 80 GB of local disk storage. The cluster system consists of 32 nodes, 256 cores, 768 GB memory and 2.5 TB disk capacity. The cooling equipment at UFL consists of two fully controllable Liebert 14- and 9-Ton CRAC system (Model FH302C-CA00 and FH147C-CAEI) with humidifier and reheating capacity.

Workloads: We used real HPC production workload traces from the RIKEN Integrated Cluster of Clusters (RICC) [13]. The trace included data from a massively parallel cluster, which has 1024 nodes each with 12 GB of memory and two 4-core CPUs. As the RICC is a large-scale distributed system composed of a large number of nodes, we scaled and adapted the job requests to the characteristics of our system model. First, we converted the input traces to the Standard Workload Format (SWF) [14]. Then, we eliminated failed and canceled jobs as well as anomalies. As the traces did not provide all the information needed for our analysis, we needed to complete them using a model based on [15].

The entire trace consists of 400,000 requests spread over 6 months. We extracted three versions out of this long trace, one for use in the small-scale experiments (with tens of servers) and two for use in medium-scale (hundreds of servers) simulations. The trace used in our experiments have 100 requests over the course of one day. The two other traces used in our simulations, however, have 5,200 requests spread over 2 days and 10,000 requests spread over 3 days. We assigned one of four benchmark profiles (based on *Sysbench* for CPU-intensive and *TauBench* for CPU-plus-memory-intensive workloads) to each request in the input trace, following a uniform distribution by bursts. The bursts of job requests were sized (randomly) from 1 to 5 requests.

Competing Strategies: We compared the performance of VMAP against six strategies, namely, Round Robin (RR), First-Fit-Decreasing (FFD), Best-Fit-Decreasing (BFD), First-Fit-Decreasing Reactive (FFD_R), Best-Fit-Decreasing Reactive (BFD_R), and Cool-Job (CJ) [4] allocation. Of these six strategies, RR, FFD, and BFD are thermal-unaware while FFD_R, BFD_R, and CJ make reactive allocation decisions based on current temperature measurements.

- In RR, the VMs are allocated sequentially to servers.
- In FFD, the VMs corresponding to the requests that have arrived in the previous time window (of duration δ [s]) are first sorted in the decreasing order of volumes of the hypercuboids representing the VMs. Then, each VM is allocated to the first server (w.r.t. server ID) that satisfies all the four subsystem utilization requirements.
- In BFD, the VMs are again sorted according to volume as in FFD. Then, each VM is allocated to the first physical server (w.r.t. server ID), which not only satisfies all the four subsystem utilization requirements but also has the least residual volume after packing that VM.
- In FFD_R, VMs are first placed following the FFD policy. Then, VMs in overheated servers are relocated to cooler servers again based on the FFD principle.
- In BFD_R, VMs are first placed following the BFD policy. Then, VMs in overheated servers are relocated to cooler servers again based on the BFD principle.
- In CJ, each VM (that has arrived in the previous δ [s]) is allocated to the first "coolest" physical server, which satisfies all the four subsystem utilization requirements. Similar to FFD and BFD, the VMs are sorted in the decreasing order of their normalized volume. Note that CJ does not predict future temperatures like VMAP does.

Metrics: We evaluate the impact of our approach in terms of the following metrics: *energy consumption* (in kilo- Watt-hour [kWh]), and *thermal violation* (duration in second per day[s/day]). The thermal violation was calculated by monitoring the average time the servers were operating in the unsafe temperature region in a day (24 hours). Unsafe temperature region here refers to temperatures greater than the upper bound of the recommended range specified by equipment manufacturers. A higher percentage of thermal violation results in greater risk of equipment failure and/or drop in performance.

B. Energy savings

Non-consolidation vs consolidation: We performed trace-driven simulations to quantify the energy savings achieved by VMAP in a large-scale setting (180 servers and 10,000 VM requests spread over 3 days). Figure 5(a) shows VMAP's energy savings in comparison to each competing algorithm. RR and CJ are the least energy efficient in comparison to VMAP as they spread the workload (VMs) over the entire datacenter (to balance the load in the case of RR and in search for the coolest server in the case of CJ). The other four schemes consolidate VMs like VMAP does, however, they consume more energy than VMAP. In Fig. 5(b), we analyzed different components (and their percentage of the total) of VMAP's energy savings. The main reasons for VMAP's superior energy performance are savings due to 1) increased server utilization, 2) efficient cooling because of the higher COP, and 3) turning off idle servers. Even though the actual amount of energy savings ranges from 17 (in comparison to BFD) to 148kWh (in comparison to CJ), the ratio of the three components of savings does not fluctuate significantly. It can be clearly observed that increased server utilization is the largest contributor to energy efficiency followed by shutdown of idle servers.

Non-thermal-aware vs thermal-aware: Figure 6 shows thermal violation of the same simulation performed above.

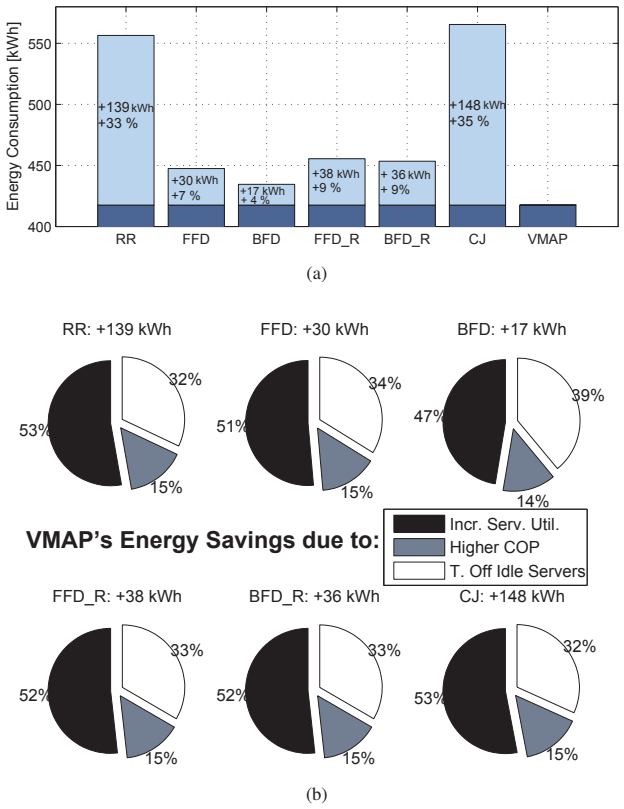

Fig. 5. (a) VMAP's overall energy savings [kWh] in comparison to the competing algorithms; (b) Components (and their percentage) of energy savings due to: 1) increased server utilization rate, 2) efficient cooling because of the higher COP, and 3) turning off idle servers.

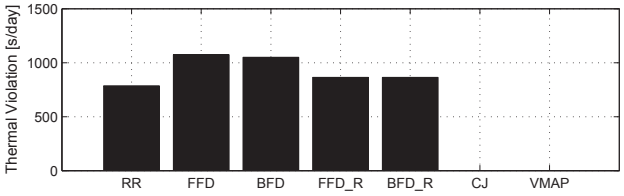

Fig. 6. Thermally violated duration of CPU temperature per server in a day.

Thermal-aware algorithms (FFD_R, BFD_R, CJ, VMAP) exhibit a smaller degree of violation in comparison with non-thermal-aware algorithms (FFD, BFD). FFD_R and BFD_R perform better in comparison to FFD and BFD because VMs from overheated servers are reallocated in reaction to thermal violation alarms. However, due to the reactive nature of these techniques, undesired equipment overheating is still an issue. VMAP and CJ avoid thermal violations. However, CJ's performance in terms of this metric is similar to VMAP's, it comes at a very high energy cost as shown in Fig. 5(a).

C. Consolidation in "better-cooled" areas

We performed trace-driven simulations to show how VMAP can exploit unevenness in heat imbalance inside a datacenter (with homogeneous computing equipment) caused by unevenness heat-extraction rates due to difference in server inlet temperatures. Evaluation was carried out in a small-scale setting (180 servers and 5,200 VM requests spread over 2 days). We studied the performance of the four thermal-aware techniques (FFD_R, BFD_R, CJ, and VMAP) under different degrees of Gaussian variation in the server inlet temperature; $\mathcal{N}(25,1)°\mathbf{C}$, $\mathcal{N}(25,5^2)°\mathbf{C}$, and $\mathcal{N}(25,9^2)°\mathbf{C}$. Unevenness of inlet temperature of each server can be attributed to the non-ideal air circulation, which depends on the layout of server racks inside the datacenter and on the placement of CRAC unit fans and air vents. Figure 7 shows that the total energy consumption (for computation as well as cooling) of VMAP decreases as the degree of unevenness increases. This is because VMAP consolidates VMs in better-cooled areas where a higher heat-extraction rate leads to a lower increase in CPU temperature (with the same heat-generation rate).

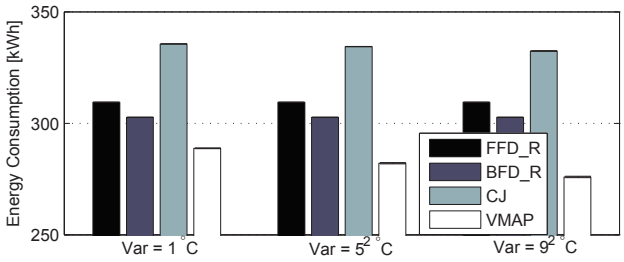

Fig. 7. Energy consumption [kWh] under different degrees of unevenness in heat extraction (induced by difference in server inlet temperatures).

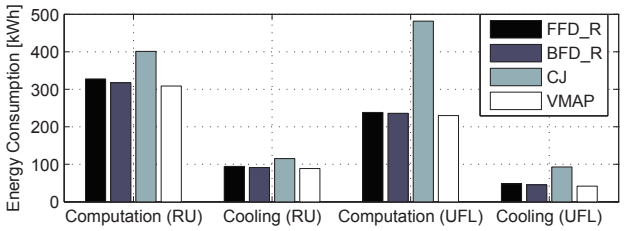

Fig. 8. Energy consumption for computation and cooling at different datacenters – RU and UFL

D. Performance Under High COP

We performed trace-driven simulations to study VMAP's performance under varying COP. First, based on the system model of the infrastructure at RU and UFL, we carried out evaluations in a large-scale setting (180 servers and 10,000 VM requests spread over 3 days at each site). The CRAC outlet temperature in the UFL system model was set to a higher value ($30\ °C$) compared to the $25\ °C$ in RU system model. It can be seen in Fig. 8 that the energy consumption for cooling at UFL is lower than the one at RU because the COP of the CRAC system model at UFL is higher than the one at RU. COP of a CRAC unit increases with increase in the outlet temperature [4] as the work that needs to be done to reduce the hot-air temperature to $30\ °C$ is lower than the work that needs to be done to reduce it to $25\ °C$. We then studied the performance of VMAP and the other thermal-aware techniques using one system model (RU's) with different CRAC COPs. In Fig. 9, we also show that VMAP does not incur thermal violation while others do even for the servers in higher temperature.

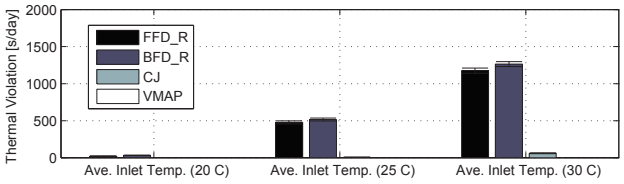

Fig. 9. Thermal violations under different average server inlet temperatures.

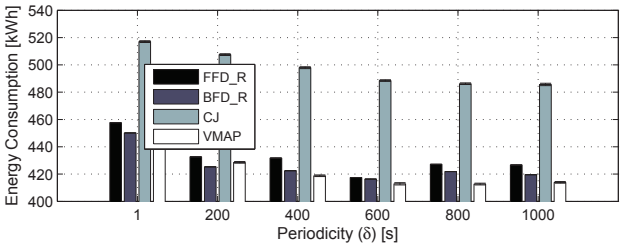

Fig. 10. Energy consumption of servers based on different periodicity

E. Impact of Decision Window (δ)

We studied the impact of the periodicity (δ) on the performance of VMAP. Evaluation was carried out in a large-scale setting (180 servers and 10,000 VM requests spread over 3 days). If δ is big, the complexity increases because the number of VM requests ($|\mathcal{M}|$) increases. If δ is small, the complexity decreases because $|\mathcal{M}|$ decreases, but it is less efficient as only fewer VM requests can be optimized. Generally, VMAP can do better packing and save energy when δ is big but δ cannot exceed certain time bound because the extra delay incurred may violate SLA. Figure 10 shows high energy consumption for small δ but lower energy consumption for large δ. VMAP outperforms the other thermal-aware strategies for any δ. The best choice of δ is, however, dependent on the workload pattern and its statistics.

VI. Related Work

Prior research efforts on thermal management of datacenters [16] have focused exclusively on only one of the two fundamental approaches: management of heat extraction [17] or management of heat generation inside a datacenter [18], [19]. The first approach aims at improving cooling system efficiency by effectively distributing cold air inside the datacenter (cooling system optimization), while the second approach focuses on how to balance or migrate workloads in such a way as to avoid overheating of computing equipment. In contrast, we focus on a joint approach so to minimize the risk of overheating of servers while simultaneously maximizing the cooling efficiency.

In [20], the authors profile and benchmark the energy usage of 22 datacenters. They perform energy benchmarking using a metric that compares energy used for IT equipment to the energy used for the CRAC system and conclude that the key to energy efficiency is air circulation management (for effective and efficient cooling). As many datacenters employ raised floors with perforated tiles to distribute the chilled air to racks, researchers have tried to gain valuable insights into efficient airflow distribution strategies in such datacenter layouts [17]. Other research efforts were aimed at improving the efficiency of cooling systems through thermal profiling (knowledge of air and heat circulation) of datacenters. Basic mathematical modeling and parameters for profiling datacenter are proposed in [21]. However, capturing complex thermodynamic phenomena using complex Computational Fluid Dynamic (CFD) models [22] is prohibitive in terms of computational overhead. Measurements from scalar sensors alone [23] cannot capture the complex thermodynamic phenomena inside a datacenter. Hence, we used a heterogeneous sensing infrastructure [5] – composed of temperature and humidity scalar sensors, thermal cameras, and air flow meters – to thermally profile datacenters in space and time so to exploit that information for resource provisioning and cooling system optimization.

Several solutions that employ temperature-aware job distribution and migration have been proposed for alleviating undesired thermal behavior (higher operating temperatures) inside datacenters. Moore et al. [18] proposed thermal management solutions that focus on scheduling workloads considering temperature measurements. They designed a machine-learning-based method to infer a model of thermal behavior of the datacenter online and to reconfigure automatically the thermal load management systems for improving cooling efficiency and energy consumption. Bash and Forman [24] developed a policy to place the workload in areas of a datacenter that are easier to cool, which results in cooling power savings. They used scalar temperature sensor measurements alone to derive two metrics that help decide whether to place workload on a server or not: the first metric, Thermal Correlation Index (TCI), gives the efficiency with which any given CRAC can provide cooling resources to any given server; while the second is Local Workload Placement Index (LWPI). Tang et al. [19] investigated the mechanism to distribute incoming tasks among the servers in order to maximize cooling efficiency while still operating within safe temperature regions. They developed a linear, low-complexity process model to predict the equipment inlet temperatures in a datacenter given a server utilization vector; they mathematically formalize the problem of minimizing the datacenter cooling cost as the problem of minimizing the maximal (peak) inlet temperature through task assignment. However, the work was validated only through simulations. In [25], the authors explore a spatio-temporal thermal-aware job scheduling as an extension to spatial thermal-aware solutions like [18], [19], [26].

Heath et al. [26] propose emulation tools ('Mercury' and 'Freon') for investigating the thermal implications of power management. In [27], the authors present 'C-Oracle', a software infrastructure that dynamically predicts the temperature and performance impact of different thermal management reactions (such as load redistribution and dynamic voltage and frequency scaling) into the future, allowing the thermal management policy to select the best reaction. However, neither of the aforementioned thermal-aware workload placement solutions explicitly take into account the direct impact of workload distribution on cooling system efficiency and vice-versa. Thermal-aware management of datacenters should

strive to minimize the TCO of datacenters, i.e., to minimize the cost of running servers through energy-aware workload distribution as well as to minimize the energy spent on cooling, by thoroughly understanding the effect of one on the other. That is why we combined thermodynamic models and real-time measurements (from temperature and humidity scalar sensors as well as air flow meters) to capture the complex thermodynamic phenomena of heat generation (due to specific workload distribution) and heat extraction (due to cooling system parameters and characteristics), in order to predict the future temperature map of the datacenter for enabling proactive thermal-aware datacenter management decisions.

VII. Conclusions and future work

We first introduced and validated the novel concept of *heat imbalance*, which captures the unevenness in heat generation and extraction, at different regions inside a HPC cloud datacenter. We then proposed thermal-aware (knowledge of heat imbalance) proactive Virtual Machine (VM) mapping (consolidation) solution, VMAP. Our solution maximizes computing resource utilization, minimizes datacenter energy consumption for computing, and improves the efficiency of heat extraction, while not violating recommended maximum operating temperature. We verified the effectiveness of VMAP through experimental evaluations with HPC workload traces at Rutgers University and University of Florida machine rooms. We observed that the VMAP is 9%, and 35% more energy efficient than best-fit and "cool job", respectively, two state-of-the-art reactive thermal-aware solutions. Currently, we are investigating the joint optimization of duty cycle for the VMAP allocation (δ) and cooling (Δ). As the heat generation is mainly due to the running workload (related to δ) and heat extraction is mainly due to the cooling (related to Δ), both δ and Δ should be jointly considered to reduce energy consumption.

VIII. Acknowledgments

We thank Prof. Manish Parashar, Rutgers University, for his insightful comments and suggestions. We thank Profs. Jose Fortes and Renato Figueiredo of the University of Florida for their help with the multi-site experiments. We also thank RU undergraduate students Christopher Camastra and Kalyan Yalamanchi for developing a VM provisioning-plus-management tool and a server performance monitoring tool, respectively, which were used extensively in our experiments.

References

[1] Growth in Data center electricity use 2005 to 2010. [Online]. Available: http://www.analyticspress.com/datacenters.html

[2] "Report to congress on server and data center energy efficiency," U.S. Environmental Protection Agency, Tech. Rep., August 2007.

[3] Water Efficiency Management in Datacenters (Part I). [Online]. Available: http://www.hpl.hp.com/techreports/2008/HPL-2008-206.pdf

[4] J. Moore, J. Chase, P. Ranganathan, and R. Sharma, "Making Scheduling "cool": Temperature-aware Workload Placement in Data Centers," in *Proc. of USENIX Annual Technical Conference (ATEC)*, Apr. 2005.

[5] H. Viswanathan, E. K. Lee, and D. Pompili, "Self-organizing Sensing Infrastructure for Autonomic Management of Green Datacenters," *IEEE Network*, vol. 25, no. 4, pp. 34–40, Jun. 2011.

[6] E. K. Lee, I. Kulkarni, D. Pompili, and M. Parashar, "Proactive Thermal Management in Green Datacenter," *The Journal of Supercomputing*, vol. 60, no. 2, pp. 165–195, 2012.

[7] I. Rodero, J. Jaramillo, A. Quiroz, M. Parashar, F. Guim, and S. Poole, "Energy-efficient Application-aware Online Provisioning for Virtualized Clouds and Data Centers," in *Proc. of Intl. Green Computing Conference (GREENCOMP)*, Chicago,IL, Aug. 2010.

[8] G. Chandra, P. Kapur, and K. Saraswat, "Scaling Trends For The On Chip Power Dissipation," in *Proc. of IEEE Interconnect Technology Conference (IITC)*, Jun. 2002.

[9] Advanced Configuration & Power Interface (ACPI). [Online]. Available: http://acpi.info/DOWNLOADS/ACPIspec50.pdf

[10] M. R. Garey, R. L. Graham, and D. S. Johnson, "Resource Constrained Scheduling as Generalized Bin Packing," *Journal of Combinatorial Theory*, vol. 21, no. 3, pp. 257–298, 1976.

[11] R. M. Karp, M. Luby, and A. Marchetti-Spaccamela, "A Probabilistic Analysis of Multidimensional Bin Packing Problems," in *Proc. of ACM Symposium on Theory of computing (STOC)*, Apr. 1984.

[12] K. Maruyama, S. K. Chang, and D. T. Tang, "A General Packing Algorithm for Multidimensional Resource Requirements," *International Journal of Parallel Programming*, vol. 6, pp. 131–149, 1977.

[13] The RIKEN Integrated Cluster of Clusters (RICC) Log. [Online]. Available: http://www.cs.huji.ac.il/labs/parallel/workload/l_ricc/index.html

[14] D. Feitelson, "Parallel Workload Archive," 2010. [Online]. Available: http://www.cs.huji.ac.il/labs/parallel/workload/

[15] U. Lublin and D. G. Feitelson, "The Workload on Parallel Supercomputers: Modeling the Characteristics of Rigid Jobs," *Journal of Parallel Distributed Computing*, vol. 63, no. 11, pp. 1105–1122, Nov. 2003.

[16] J. Rambo and Y. Joshi, "Modeling of data center airflow and heat transfer: State of the art and future trends," *Distributed and Parallel Databases*, vol. 21, no. 2-3, pp. 193–225, Jun. 2007.

[17] R. R. Schmidt, E. E. Cruz, and M. K. Iyengar, "Challenges of data center thermal management," *IBM Journal of Research and Development*, vol. 49, no. 4/5, pp. 709–723, 2005.

[18] J. D. Moore, J. S. Chase, and P. Ranganathan, "Weatherman: Automated, Online and Predictive Thermal Mapping and Management for Data Centers," in *Proc. of Intl. Conf. on Autonomic Comp. (ICAC)*, 2006.

[19] 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, 2008.

[20] S. Greenberg, E. Mills, and B. Tschudi, "Best Practices for Data Centers: Lessons Learned from Benchmarking 22 Data Centers," in *Proc. of American Council for an Energy-Efficient Economy (ACEEE)*, Pacific Grove, CA, Aug. 2006.

[21] R. Sharma, C. Bash, and R. Patel, "Dimensionless Parameters For Evaluation of Thermal Design and Performance of Large-Scale Data Centers," in *Proc. of ASME/AIAA Joint Thermophysics and Heat Transfer Conference*, Jun. 2002.

[22] A. Beitelmal and C. Patel, "Computational Fluid Dynamics Modeling of High Compute Density Data Centers to Assure System Inlet Air Specifications," *Distributed and Parallel Databases*, vol. 21, no. 2-3, pp. 227–238, 2007.

[23] J. Liu, B. Priyantha, F. Zhao, C. Liang, Q. Wang, and S. James, "Towards Discovering Data Center Genome Using Sensor Networks," in *Proc. of the Embedded Networked Sensors (HotEmNets)*, Jun. 2008.

[24] C. Bash and G. Forman, "Cool job allocation: Measuring the power savings of placing jobs at cooling-efficient locations in the data center," in *Proc, of USENIX Annual Technical Conf. (ATEC)*, 2007.

[25] T. Mukherjee, A. Banerjee, G. Varsamopoulos, S. Gupta, and S. Rungta, "Spatio-temporal thermal-aware job scheduling to minimize energy consumption in virtualized heterogeneous data centers," *Computer Networks*, vol. 53, no. 17, pp. 2888–2904, Dec. 2009.

[26] T. Heath, A. P. Centeno, P. George, L. Ramos, Y. Jaluria, and R. Bianchini, "Mercury and freon: temperature emulation and management for server systems," in *Intl. Conf. on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*, 2006.

[27] L. Ramos and R. Bianchini, "C-oracle: Predictive thermal management for data centers," in *Proc. of Symp. on High-Performance Computer Architecture (HPCA)*, 2008.