Virtual Melting Temperature: Managing Server Load to Minimize Cooling Overhead with Phase Change Materials

Matt Skach
University of Michigan
skachm@gmail.com

Manish Arora
Advanced Micro Devices, Inc.
University of California, San Diego
manish.arora@amd.com

Dean Tullsen
University of California, San Diego
tullsen@eng.ucsd.edu

Lingjia Tang
University of Michigan
lingjia@eecs.umich.edu

Jason Mars
University of Michigan
profmars@umich.edu

***Abstract*—As the power density and power consumption of large scale datacenters continue to grow, the challenges of removing heat from these datacenters and keeping them cool is an increasingly urgent and costly. With the largest datacenters now exceeding over 200 MW of power, the cooling systems that prevent overheating cost on the order of tens of millions of dollars. Prior work proposed to deploy phase change materials (PCM) and use Thermal Time Shifting (TTS) to reshape the thermal load of a datacenter by storing heat during peak hours of high utilization and releasing it during off hours when utilization is low, enabling a smaller cooling system to handle the same peak load. The peak cooling load reduction enabled by TTS is greatly beneficial, however TTS is a passive system that cannot handle many workload mixtures or adapt to changing load or environmental characteristics.**

In this work we propose VMT, a thermal aware job placement technique that adds an active, tunable component to enable greater control over datacenter thermal output. We propose two different job placement algorithms for VMT and perform a scale out study of VMT in a simulated server cluster. We provide analysis of the use cases and trade-offs of each algorithm, and show that VMT reduces peak cooling load by up to 12.8% to provide over two million dollars in cost savings when a smaller cooling system is installed, or allows for over 7,000 additional servers to be added in scenarios where TTS is ineffective.

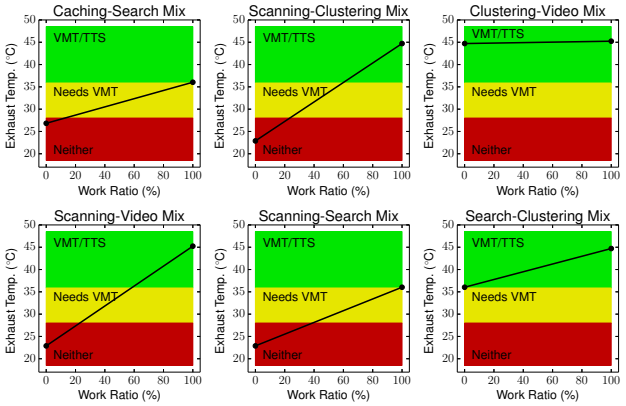


Fig. 1: Thermal time shifting (TTS) can operate in a limited range of temperatures (green), however many mixtures of datacenter workloads lie outside of this range. Virtual Melting Temperature (VMT) manages workload placement to greatly expand the useful range (green+yellow) where deploying phase change materials is beneficial.

I. Introduction

The unprecedented growth of web and cloud services over the last decade spurred an enormous investment in datacenters, also called "warehouse-scale computers" (WSCs) [1]. With the largest datacenter facilities consuming over 200 MW of power each [2], [3] and costing over a billion US dollars to build [4], datacenters represent a huge investment not only in server equipment but also in power, connectivity, facilities, and cooling infrastructure.

In the United States alone, datacenters consumed over 2% of all electrical power generated in 2014 [5], [6]. Extensive prior work investigates how to build more energy efficient processors and remove heat from the processors and servers better [7], [8], [9], [10], [11], [12], [13], but relatively little research has been published on how to maximize utilization and efficiency while minimizing cost to remove this heat from a datacenter facility.

In even a modestly sized datacenter the cooling system cost can exceed hundreds of thousands of dollars per MW of critical power, with large datacenters spending tens of millions in capital costs plus millions more per year in operating expenses to power and maintain the cooling system [14]. Datacenter cooling capital expenses in 2015 alone totaled more than $2.58 billion and are expected to exceed $6 billion by 2023 [15]. Prior work investigating server and datacenter cooling techniques demonstrate efficiency improvements [16], [17], [18], but cannot address the growing cost problem due to a critical assumption: that work and heat are coupled so that all of the heat must be removed at the same time the work is done.

Thermal time shifting (TTS) [19] decouples work and heat by storing excess heat in a phase changing material (PCM) and removing that heat at a later time. TTS with PCM works by placing a quantity of PCM downwind from the CPU sockets in a rack mounted server. During the peak hours (midday through the evening) when users are online and load is high,

2575-713X/18/$31.00 ©2018 IEEE
DOI 10.1109/ISCA.2018.00013

IEEE computer society

the PCM melts absorbing heat to reduce the thermal output of the datacenter. Then during the off hours (late night and early morning) when most users are asleep, load is low, and extra cooling capacity is available, the PCM refreezes and the stored thermal energy is released.

A reduced peak cooling load has two major advantages: the datacenter can employ a smaller cooling system while still meeting the computational demands of peak load, or the same datacenter can run more and/or hotter servers under the same cooling budget. Both benefits can save hundreds of thousands of dollars per year or millions of dollars in capital expenses [19], however it is not a universal solution.

While TTS implemented with commercial grade paraffin wax can be both thermally effective and cost effective (approximately $1,000 per ton [19]), it has key limitations. Most of these limitations stem from the fact that the optimal melting temperature for a datacenter depends on many factors, from ambient temperature, to workload, to power and delivery limits. All of these can change from installation to installation, from season to season, or even from day to day. This is problematic because:

1) Commercial-grade paraffin can only be purchased within a limited range of melting temperatures, typically 40-60 °C, however if a melting temperature outside of this range is needed molecularly pure n-paraffin options cost in excess of $75,000 per ton.
2) Once installed, the wax melting temperature cannot be adjusted. On days when the load does not cause wax to melt, there is no flattening of the diurnal cooling load while on days when all the wax melts too soon, there is no reduction in peak temperature and cooling load.
3) The power and temperature profile of a workload often changes over the multi-year lifetime of a server. As the power profile changes, the ideal (or required) melting temperature can also change to necessitate new wax or leave the range of commercial wax melting temperatures entirely.

In all three cases, deploying wax in the servers provides little to no benefit and TTS is a passive system that cannot adapt.

In this work, we propose a new adaptable technique called Virtual Melting Temperature (VMT) to handle workload power mixtures that TTS alone is unable to cool. VMT does so in such a way that it induces melting of the PCM (and thus heat redistribution) at load and average temperature levels that are (configurably) different than would happen with TTS, thus mimicking the operation of wax with a melting point that is different than the physical melting point of the deployed wax.

This is accomplished by rebalancing the load to raise temperatures in some of the servers above the PCM's melting temperature and storing energy in select servers, with the benefits of reduced cooling load and reduced power. Strategically employing VMT enables fine-grained control of wax melting and cooling, allowing VMT to reduce the peak cooling load when TTS cannot.

In this paper, we make the following contributions:

1) We introduce VMT, a method to manage the thermal properties of a PCM-enabled datacenter by controlling workload placement. We introduce and discuss two workload placement algorithms to enable VMT.
2) We perform a scale out study of VMT with both algorithms, using a previously verified simulation methodology to execute a design space exploration of VMT on a cluster of 1,000 PCM-enabled servers. We examine two VMT algorithms in a cluster running five different workloads, each with unique thermal characteristics.
3) We quantify the impact of VMT at the cluster and datacenter levels, providing useful discussion of how best to use VMT in a datacenter and quantifying the potential benefits of VMT-enabled cooling oversubscription policies.

At the cluster level, we find that VMT can reduce the peak cooling load by 12.8% even when the average thermal output of the the cluster is too low for TTS. At the datacenter level VMT reduces the peak cooling load by up to 3.2 MW, allowing for up to 7,339 more servers under the same cooling budget or for the datacenter to operate at full capacity with a smaller cooling system saving $2.6 million in scenarios where TTS provides no measurable benefit.

II. Background – TTS

Prior work [19] showed that TTS can greatly reduce the peak cooling load of a datacenter, providing significant cost savings by reducing the size of the cooling system needed or providing cooling for thousands more servers under the same cooling budget.

This is particularly important as datacenters continue to grow because, while the cooling system is an integral and critical part in the design of every datacenter, the cooling system itself does not directly contribute towards revenue generation. With cooling infrastructure costing millions of dollars for even modestly sized datacenters [14] and consuming millions of MWh annually [5], working towards more efficient and affordable cooling systems is of critical importance.

TTS proposes to place a small amount of PCM in each server in a datacenter running primarily user-facing workloads. These types of workloads typically see a diurnal load pattern with a high peak during the afternoon/evening and a large trough during the late night [1], [20], however a diurnal cycle is particularly problematic for the cooling system because the system size must be provisioned for peak load even though it spends most of the day running at levels considerably below the peak (Figure 2). TTS addresses this by raising the minimum load and lowering the maximum load, increasing average utilization in an appropriately resized cooling system. In the right configuration, TTS can accommodate the same load without overheating or thermal downclocking.

To enable TTS, the PCM must have a melting temperature that lies between the peak and trough such that during the peak hours wax melts and stores thermal energy, and then during the off hours when the load is low the PCM solidifies and releases the stored energy. The total amount of energy stored is proportional to the latent energy of the PCM (the amount of energy absorbed during the phase transition), and how much PCM melts. The sensible heat (energy required to raise the temperature of the PCM without a phase transition) also stores energy, but typically stores several times less energy than

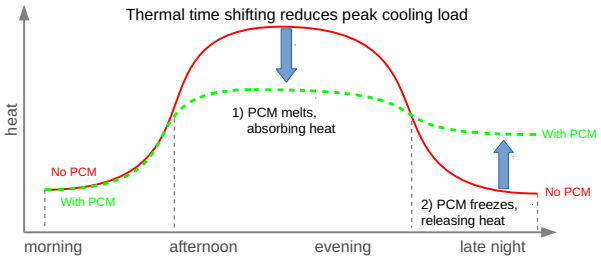


Fig. 2: Thermal time shifting with PCM.

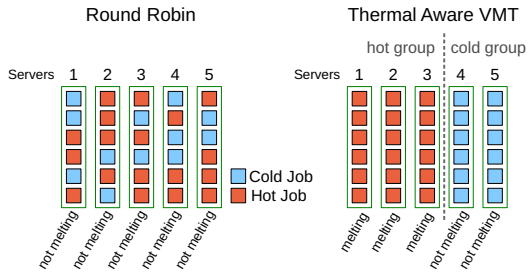


Fig. 3: Thermal Aware VMT scheduling.

the phase transition [19], [21], [22]. TTS does not inherently remove heat or reduce the amount of heat that must removed from the datacenter.

TTS proposes to alter this paradigm, storing thermal energy at the peak and releasing it during the off hours to flatten the cooling load. This enables two major opportunities for cost savings. First, the cooling system in a datacenter may now be sized for a reduced peak load, saving hundreds of thousands of dollars per year in amortized TCO, or alternatively second: more servers with a higher peak power load may be added to the same datacenter without increasing the peak cooling load and saving millions of dollars over a new cooling system in a retrofit scenario [19].

However, a passive management system for TTS that only melts or cools wax at a specific and set threshold [19] (the physical melting temperature of the wax) cannot handle many mixtures of different workloads, especially as the types, prevalence and power characteristics of these workloads change over the lifetime of the datacenter and may change as frequently as day to day or hour to hour.

PCM Selection - Thermal energy storage can be accomplished with any PCM, however not all PCMs are appropriate for deployment in a datacenter. Commercial paraffin wax is particularly advantageous, not only because it is non-corrosive and non-conductive in case of a leak, but also because it is cheap and available with a range of melting temperatures typically between 40 and 60 °C. Molecular n-paraffins can have lower melting temperatures, but are cost prohibitive to deploy in a datacenter [19], [21], [22], [23].

Wax Placement - TTS proposes to place the wax directly inside of each server, behind the CPU heat sinks occupying empty air space left available for expansion card slots and other configuration options. Prior work demonstrated the benefits of TTS in a variety of servers including low power and high throughput commodity servers as well as high density Microsoft Open Compute servers for workloads with heterogeneous thermal profiles [19].

III. Virtual Melting Temperature

VMT actively manages workload placement to control the distribution of temperatures within the datacenter, raising the temperature in a subset of servers to melt wax (and thus store heat) while lowering the temperature in other servers to reduce the peak cooling load for the whole datacenter. This creates a "virtual" melting temperature where, although the average temperature is unable to melt wax, we initiate melting in a subset of servers to benefit from heat storage. VMT gives the system or operator active control over the melting and cooling cycles of wax in the datacenter.

Without VMT, the datacenter's ability to target the best physical melting temperature (the point at which temperature of the server is held stable while the material melts) is relatively limited and, most importantly, remains constant for the life of the server unless the wax is swapped out and replaced (labor intensive). VMT is a technique that allows a datacenter to vary the apparent melting temperature in the datacenter to melt wax even if it would not normally melt. With a diverse workload, we can create thermal imbalance via job placement. With a homogeneous workload we can do the same through load imbalance; for this work we assume the former.

VMT can also raise the melting temperature by locating hot jobs in a subset of servers with already melted wax, preserving wax in anticipation of a very hot peak still to come. However, the focus of this work is on reducing the melting point rather than increasing it.

In this section we present two scheduling algorithms to enable virtual melting temperature: a thermal aware algorithm that sorts and places jobs based upon their thermal properties, and a wax aware algorithm that additionally reallocates jobs away from fully melted servers.

A. VMT with Thermal Aware Job Placement

VMT with thermal aware job placement (VMT-TA) proposes to divide the cluster into a hot group of servers and cold group, then schedule jobs with a hot thermal profile in the hot group while jobs with a cold thermal profile are placed in the cold group (Figure 3). (Note that hot group and cold group servers do not need to physically clustered: they can be distributed throughout datacenter to maintain the same cluster or DC-level temperature distributions.)

Jobs are placed into either the hot group or the cold group based on the thermal profile of the workload they belong to: if a server filled with only a single workload can melt significant wax over a peak load cycle, regardless of whether the jobs can be colocated with itself enough times to do so as long as they could be collocated with other hot jobs, the workload is considered hot and VMT will attempt to located these jobs together in the hot group. Otherwise, the workload is labeled cold and VMT attempts to place jobs in the cold group.

In such a configuration, the hot group can melt wax even if the mean temperature within all of the servers or mean

temperature with round robin/non-thermal-aware scheduling is not high enough to melt wax.

To calculate the number of servers placed in the hot group, VMT-TA uses a ratio of the user-set Grouping Value (GV) divided by the Physical Melting Temperature (PMT) of the wax in the following formula:

$$hot_group_size = \frac{GV}{PMT} \times num_servers \quad (1)$$

Where num_servers is the number of servers in the cluster and hot_group_size is the number of servers in the hot group.

There is not a general solution that maps the GV to a Virtual Melting Temperature (VMT) because such a mapping depends on the PMT as well as the workload power profile and workload mixture, however a mapping can be experimentally derived for a given combination. In Section V-A we show a GV to VMT mapping for our test datacenter.

After calculating the hot group size, the cold group is simply composed of the remaining servers:

$$cold_group_size = num_servers - hot_group \quad (2)$$

To implement VMT-TA, workload types are first classified as hot jobs or cold jobs based upon thermal characteristics. This can be done using on-package thermal sensors and/or power sensors or models (e.g. Intel RAPL). Once deployed, hot jobs are placed in the hot group of servers while cold jobs are placed in the cold group .

Within each group, jobs are distributed evenly among the servers. Here care must be taken to ensure each group is large enough to support the peak load for its respective subset of workloads else individual queries must be dropped or queued causing QoS degredation. This can be handled by dynamically adjusting the VMT to modify the group sizes or by allowing jobs to be scheduled to the other group if one group fills up.

B. VMT with Wax Aware Job Placement

Last, we propose VMT with wax aware job placement (VMT-WA). Where VMT-TA has no mechanism to handle all of the wax in the hot group melting early, VMT-WA monitors the melted state of the wax and automatically increases the size of the hot group if all of the wax melts before the end of the load peak.

At its simplest, VMT-WA schedules just like VMT-TA until wax on a server in the hot group is fully melted. Unlike VMT-TA, once the wax is fully melted in a server VMT-WA moves a server from the cold group to the hot group, maintains just enough load on the melted servers to keep the wax melted, and moves the additional load to the newly added server to continue melting wax (Figure 4). A detailed description of the algorithm follows.

VMT-WA begins by calculating the size of the hot and cold groups using Equation 1, the same as VMT-TA, however the group sizes are dynamically updated as wax melts and cools.

Periodically, the cluster scheduler updates the size of the hot and cold groups by scanning the amount of wax melted on each server. The scheduler compares each server against the Wax Threshold, the fraction of the wax melted above which the server is considered completely melted, and adds each server above the threshold to a list of fully melted servers. After counting the number of servers in this list, servers are removed

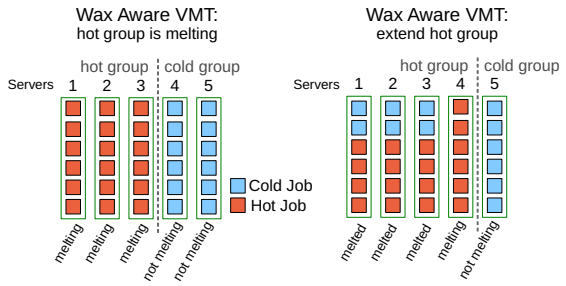


Fig. 4: Wax Aware VMT scheduling.

from the cold group and added to the hot group based upon current load trends. During each update, the scheduler restarts from the minimum hot group size (Equation 1) and adds servers in order. To the extent possible, we do not transition servers from the hot group to the cold group during the peak because cooling a melted server releases heat.

When placing individual jobs, the scheduler considers the job's thermal classification (the same as VMT-TA) but does not strictly place the job in the corresponding server group. For hot jobs, the scheduler first attempts to schedule the job in the hot group by considering a subset of servers in the hot group that are currently below a certain amount of wax melted (the wax threshold) or are below the wax melting temperature. Placing a hot job on either such server will attempt to melt more wax or keep already molten wax melted (both advantageous for reducing cooling load).

If there are no hot group servers meeting these characteristics (possible with sudden spikes in load), then servers are added to the hot group from the cold group sequentially until the hot group includes a server that is below the wax threshold or melting temperature. In the event that no such servers exist (a corner case where all servers are added to the hot group) then the job is scheduled on any server below the melted threshold or, barring that, any remaining servers.

To place a cold job, the scheduler first attempts to place the job in the cold group. If the job cannot be placed in the cold group (as may occur when the hot group grows), the scheduler attempts to place the job on a server in the hot group that is already above the melted threshold and melting temperature to minimize thermal impact. If the job cannot be placed in these servers either, then the job is placed into one of the remaining hot group servers.

This ordering of scheduling policies will only fail to schedule a job in the case where a thermally unconstrained datacenter would also run out of computational space, so we do not model that case.

Tracking Wax State - VMT-WA requires knowledge of the current melted state of wax in servers in the cluster to adjust the size of the hot group properly. A single temperature sensor on the exterior of the wax container can tell us when the wax starts melting or freezing, then we add a light weight model of current wax state running on each server. The model uses the CPU power consumption and temperature sensors already in the server to estimate the current state of the wax based upon a lookup table [24].

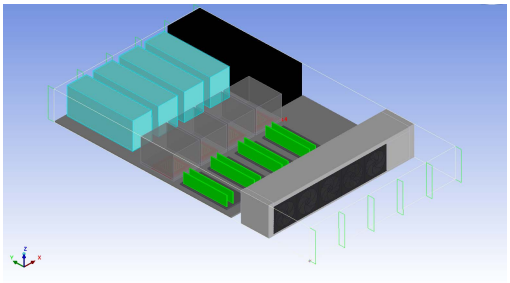


Fig. 5: Test server with 4.0 liters of wax (light blue) behind the CPUs.

IV. Methodology

A. Datacenter Architecture

We consider a datacenter running a Google-style suite of workloads: all are user facing with different latency requirements that, with modern contention reduction techniques [25], [26], [27], allow for collocation on the same servers. Within the datacenter, servers are divided into homogeneous clusters and job scheduling is performed at the cluster level. In Section V we perform a number of cluster-level scale out studies on a cluster of 1,000 servers (with some parameter sweeps performed with 100 servers to reduce total compute time). To perform a datacenter-level TCO analysis, we consider many clusters summing to a critical power of 25 MW, just shy of the 27.25 MW median critical power for large scale datacenter reported by Ghiasi et al [3].

We provision the datacenter with 2U high throughput servers (Figure 5), based upon the internal layout of a Sun Fire X4470 server but populated with 4x Xeon E7-4809 v4 CPUs. In this form factor, this corresponds to approximately 20 servers per rack and 50 racks per cluster. Each server has a peak power consumption of 500 W, and an idle power consumption of 100 W. Per core power consumption is approximated using a linear model [14].

Based upon computational fluid dynamics (CFD) design space exploration, this server can hold 4.0 liters of wax without exceeding CPU thermal limits [19]. The wax in each server is commercial paraffin wax with a melting temperature of 35.7 °C, the lowest commercially available temperature [28].

The paraffin wax is divided between four aluminum containers that contain the wax when molten and provide surface area contact for heat transfer from the air to the wax. Even though the paraffin wax has a melting temperature of 35.7 °C, the lowest of a commercially available paraffin wax, for many workload compositions this is not low enough to melt wax even at peak load due to the thermal characteristics of the datacenter and the workloads.

Each server in the cluster maintains its own model of the state of the wax inside of it [24]. We update the model once per minute based upon load and temperatures in the last minute, and report the wax state to the cluster level scheduler when it is updated. Running only once per minute, the update process provides a negligible impact on server and network performance.

TABLE I: Workloads considered for scaleout study (power is normalized to a single 8 core Xeon E7-4809 v4 CPU; each server contains four CPUs).

Workload	CPU Power	VMT Class
WebSearch	37.2 W	hot
DataCaching	13.5 W	cold
VideoEncoding	60.9 W	hot
VirusScan	3.4 W	cold
Clustering	59.5 W	hot

B. Workloads

We consider a cluster of servers inside of datacenter running 5 different workloads (Table I). All of the workloads can be co-located within the same server, however they are assigned separate physical cores and never share simultaneous multithreading (SMT) contexts to reduce the complexity of contention mitigation techniques.

Of the five workloads two are user-facing, latency critical workloads that demand immediate responses back from the server: Web Search and Data Caching. These workloads have strict QoS requirements on the order of milliseconds or microseconds.

The other three workloads perform user-facing functions that demands a degree of QoS (that is, they are not batch jobs that can be scheduled hours later) but are not as strict as web search and data caching. Video Encoding, e.g. for Youtube, Virus Scanning files, e.g. for uploading files to Google Drive, and Clustering, e.g. for web advertisements, demand computation be near when the action is initiated to ensure benefit (responsive file downloads, relevant ads, etc.) but a runtime difference on the order of seconds will not greatly reduce the user experience. On these workloads, we consider a datacenter running contention mitigation techniques [25], [26], [27] that allow a small performance penalty to ensure that the latency critical workloads meet their QoS requirements.

To enable sorting and placement using VMT, jobs are classified as either a 'hot' or 'cold' based upon whether their power and temperature profile would enable them to melt significant wax if run in isolation.

Web Search - We consider the CloudSuite 2.0 Web Search benchmark [29]. Web Search shards queries to multiple servers, each holding a portion of the index, and returns the results based upon the users query. Using power profiling of Web Search [30], we classify it as a hot job for VMT.

Data Caching - For Data Caching, we consider the CloudSuite 2.0 implementation using the Memcached server framework to meet the demands of a social media service [29]. The Memcached server must respond in real-time to user requests, performing a number of memory operations on large sets of data. With relatively low CPU power consumption [30], VMT classifies data caching as a cold job.

Video Encoding (h264) - We consider the SPEC 2006 implementation of h264 video encoding [31]. Video media uploaded to video sharing sites such as Youtube are re-encoded to several different file sizes [32] before users can share or view the video. As such, although this is not a batch job where the host provider can leave the user waiting potentially several hours until the encoding can be scheduled during off hours [20],

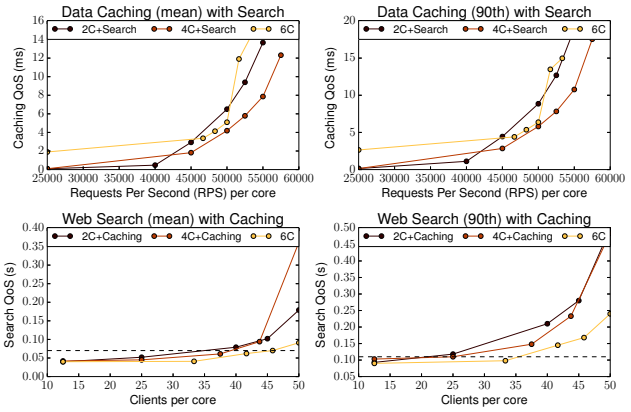


Fig. 6: Latency scaling with load and cores for Web Search and Data Caching colocated on a Xeon server, without contention reduction techniques. Caching contention is within an acceptable range when colocated versus not, and Web Search exhibits behavior that can be managed using previously studied contention mitigation techniques.

a small delay of seconds or even minutes for longer videos is tolerable. Based upon power measurements of h264 Video Encoding [19], we classify video encoding as a hot job for VMT.

Virus Scanning - Files uploaded to a file host like Google Drive are scanned for viruses before they are shared, converted or downloaded [33]. We consider a virus scanner [30] running on freshly uploaded files. Similarly, these are not latency critical but cannot be delayed for batch job scheduling. Based upon power profiling of VirusScan [30], virus scanning is classified as a cold job for VMT.

Clustering - Clustering is commonly used to deliver ads targeted ads based upon user actions on the web [34]. This is a computationally intensive task with some leeway for contention mitigation, but the sooner it can finish then the sooner relevant ads can be delivered to the end user [34]. This makes batch job scheduling possible but not ideal in many cases. Based upon power profiling [35], we classify clustering as a hot job for VMT.

1) Workload Migration: If a job cannot be migrated at all and load cannot be redirected to a different host then VMT cannot be used, however this is a relatively rare case.

Of the diverse workloads we consider, all can be migrated or reallocated but some are more portable than others. Virus Scanning and Video Encoding, for example, are very portable with data requirements dominated by the incoming files to be scanned or encoded. Web Search on the other hand requires a large amount of data that is not very portable, however multiple copies of the data are already distributed throughout the datacenter to enhance query speed and redundancy [36]. This allows a degree of flexibility in job placement without requiring data migration that VMT exploits.

C. Workload Colocation and Interference

In Figure 6, we consider latency scaling with load and cores for mixtures of Web Search and Data Caching servers from the Cloudsuite 3.0 benchmark suite [29] running on a 6 core E5-2420 CPU with Turbo Boost turned off. All jobs are scheduled on separate cores and have sufficient main memory to prevent swapping to disk but still may interfere in the last level cache and memory bandwidth. No contention reduction techniques are applied during this test. Data Caching RPS per server core was fixed at was 45k when collocated with Web Search, while Web Search clients per server core was fixed at 37.5 when collocated with Data Caching.

For Data Caching, we observe that that at very low loads, when QoS targets are most often met, 6 cores running together provides the best latency. Similarly at high latencies when QoS targets are violated 6 cores once again provides slightly better average QoS, however in the middle range for Data Caching a mixture provides similar or better performance than homogeneous workloads as the memory resources are split between memory intensive data caching and more compute intensive web search. Given that the total load must meet QoS with all cores allocated to one workload at load, (thus dividing peak resource utilization approximately even by the number of cores), we assert that the high latency sensitive workloads will be able to coexist in general with other high latency sensitive workloads. Corner cases that may arise (e.sg. specific cache thrashing access patterns) can be mitigated by dynamic management and recompilation techniques [25], [26], [27] or allocated to non-VMT-enabled servers.

For Web Search, we observe decreased performance across the whole range of clients per core. Here, it is important to observe that even with 6 cores running only Web Search the clients per core are limited by QoS targets to return data to the user. As these are compute heavy workloads and sufficient memory bandwidth was available with 6 cores, the slowdown is likely caused by cache interference which can be mitigated by BubbleUp and Protean Code [25], [26].

D. Server Reliability

The impact of thermal wear on computer and server components has been extensively studied [37], [38], [39]. Using VMT, servers in the hot group experience higher average utilization and the temperature of many components increases relative to a round robin or coolest first scheduler (thus exhibiting a higher failure rate) while servers in the cold group experience the opposite.

To ensure even wear leveling across components, servers should therefore be rotated between the hot group and cold group regularly [40], [41], [42], [43]. In Figure 7, we plot the 6 month and 36 month (3 year) cumulative failure rates using a RR scheduler and a VMT-WA scheduler.

To model the failure rate, we first begin with a 70,000 mean time before failure (MTBF) at 30 °C based on numbers from Intel [44]. We scale this reliability using the rule of thumb that a 10 °C increase in temperature doubles the failure rate of components [45], [39] to adjust the rate of failure to the temperatures in our test datacenter.

We then assume a 20% rotation per month, where each server spends two months in the cold group and three months in the hot group based upon our breakdown of workloads. After 3 years [1], the cumulative failure rate of all servers for VMT-WA is only 0.6% higher than for round robin.

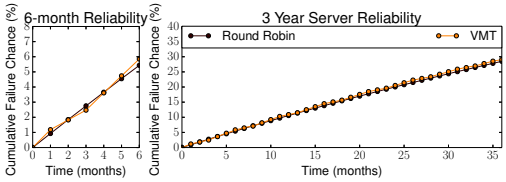


Fig. 7: Server reliability for round robin versus VMT-WA when 20% of servers are rotated each month (3 months in the hot group, 2 months in the cold group). After 3 years, the cumulative failure rate for VMT-WA is 0.4% higher than for RR.

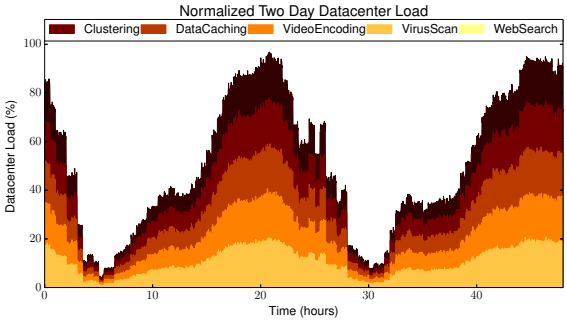


Fig. 8: Two day trace (cumulative for 100 servers).

E. Simulation Infrastructure

We perform our scale out simulation using DCsim [14], an event-driven simulator to model a cluster of 1,000 servers. The wax model for the server is based upon real hardware measurements to validate a CFD model of a test server [19], and then the CFD result is used to derive model parameters for DCsim. (CFD is more accurate, but computationally infeasible to solve at the granularity needed to evaluate VMT in a cluster-level scale out study.) The cluster results from DCsim are then multiplied linearly to calculate the effects of VMT workload placement policies on the datacenter level.

We use a two day trace of datacenter load from Google [46], normalized using a similar procedure to Kontorinis et al. [14]. The total load is divided between our five workloads, providing a roughly 60-40 split between hot jobs and cold jobs (Figure 8). The load pattern on these two days, up to 95% server utilization, represent the worst case days for the the cooling system. Servers are usually provisioned such that peak daily load is much lower than the total capacity [1], resulting in server and cooling capacity that is underutilized. We consider atypically high day-to-day utilization over two days to realistically stress the cooling system and VMT algorithms in our evaluation.

F. TCO Model

To quantify the cost-saving benefits of VMT, we consider the TCO of the cooling system in a datacenter. When constructing a datacenter, the age of non-IT infrastructure (facilities, cooling, power distribution, etc.) is typically expected to outlast the IT infrastructure (servers, networking equipment, etc.).

To estimate lifetimes, costs and benefits we adapt the TCO calculations from Kontorinis et al. [14] to our datacenter. They use a 10 year linear depreciation for non-IT infrastructure including the cooling system, and a 4 year depreciation for servers.

To calculate the depreciation cost of the cooling system alone, they report a depreciation cost of $7.00 per kilowatt of critical power per month. With a cooling system expected to depreciate over 10 years, this adds up to $84,000 per MW of critical power per year, or $21 million total for 25 MW of critical power.

We evaluate only the cost savings in the cooling system for VMT. The cost to add wax to each server is very small (less than 0.5% of the purchase cost per server at a wax price of $1000/ton), as is the cost savings from utilizing lower electricity prices during the off-peak hours [19].

V. Evaluation

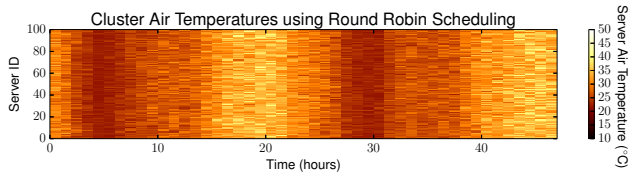
In our experiments, we consider two baselines. The first is a round robin scheduler, the same used in prior work on TTS [19]. The second is a more advanced coolest-first scheduler that presumes the coolest servers have the greatest thermal headroom available and schedules on them first.

In Figures 9 and 10, we plot a heat map of the temperature inside of 100 servers and the portion of wax melted in those servers when jobs are placed according to the round robin and coolest first schedulers, respectively. Both schedulers receive the same workload. Temperature peaks (around 20 hours and 46 hours) and troughs (around 5 and 29 hours) correspond with the peaks and troughs seen in the workload pattern (Figure 8). Compared to round robin, coolest first maintains a much tighter temperature distribution between servers as expected of a thermal aware load balancing scheduler, however due to the diverse thermal profiles of these workloads the average temperature in either cluster and the temperatures in each server never reach levels high enough to melt a significant amount of wax.

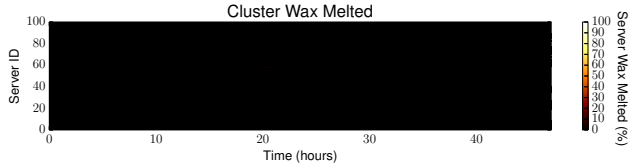
A. Thermal Aware VMT

First, we consider VMT-TA in a cluster of 1,000 servers. As noted in Section III, the GV used to calculate the size of the hot and cold groups does not directly correlate to a temperature but is used to control the ratio of servers in the hot group to servers in the cold group.

As noted in Section III-A, the grouping value is used to determine the size of the hot group. We empirically derive a mapping from GV to the VMT in Table II by running multiple experiments where the wax heat of fusion is modified to match the available thermal energy storage in the hot group and the PMT is swept above and below the starting melting temperature, 37.5 °C. (VMT temperatures above GV=20 are indistinguishable because the datacenter no longer melts wax.) Note that the relationship is non-linear, and is specific to this configuration; the GV to VMT relationship can vary with different mixtures of the PMT and workload composition.

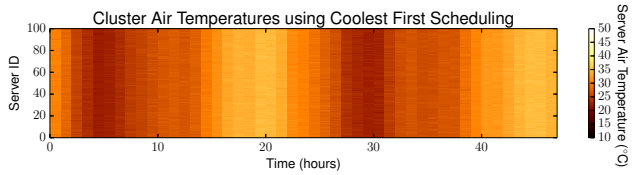


(a) Air temperatures at the wax using round robin placement.

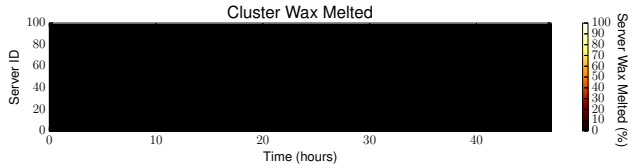


(b) No wax melts using round robin placement.

Fig. 9: Air temperatures and wax melted for 100 servers using round robin placement. The cluster does not benefit from TTS because both the average temperature and individual server temperatures are not hot enough to melt wax.



(a) Air temperatures at the wax using coolest first placement.

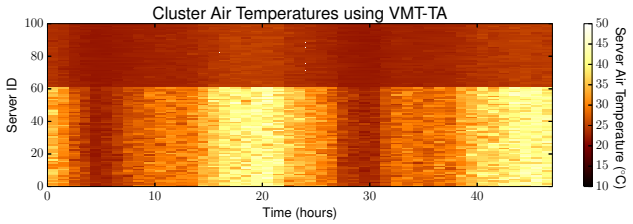


(b) No wax melts using coolest first placement either.

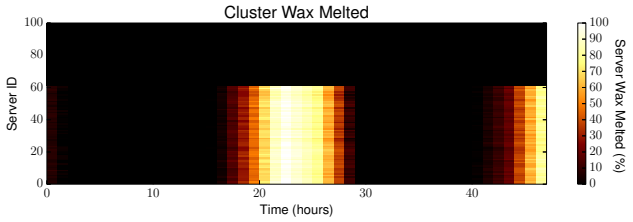
Fig. 10: Air temperatures and wax melted for 100 servers using coolest first placement. Coolest first scheduling produces a much lower temperature deviation between the servers versus round robin, but similarly does not melt significant wax.

TABLE II: Experimentally derrived mapping between the Grouping Value (GV) and Virtual Melting Temperature (VMT) for the test datacenter.

GV	VMT (°C)	ΔPMT (°C)
20.03	37.7	+2.0
20.14	36.7	+1.0
20.23	35.7	+0.0
20.83	34.7	-1.0
21.25	33.7	-2.0
21.55	32.7	-3.0
21.69	31.7	-4.0
21.84	30.7	-5.0
23.99	29.7	-6.0
30.75	28.7	-7.0



(a) Air temperatures at the wax using VMT-TA.



(b) Wax melted using VMT-TA.

Fig. 11: Air temperatures and wax melting for 100 servers using VMT-TA with GV=22.

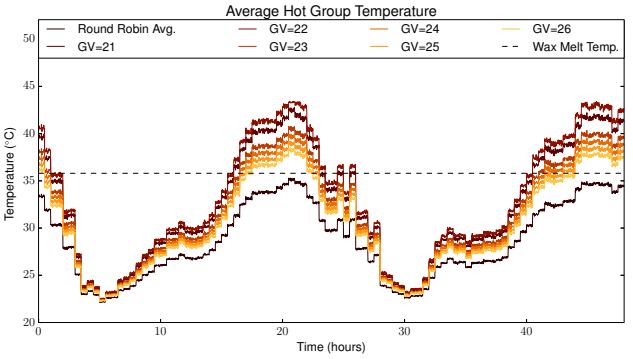


Fig. 12: Average temperature in the hot group using VMT-TA as the GV is adjusted for a cluster of 1000 servers.

Because of this fluid behavior, we plot the GV rather than VMT in our evaluation.

In Figure 11, we plot air temperatures at the wax for VMT-TA for GV=22 on 100 servers (one tenth of the cluster). In Figure 11a, the separation between the hot group and cold group is immediately apparent. Although the temperature trends with the load pattern in both groups, placing the hot jobs together in the hot group allows the hot group to exceed the melting temperature of the wax (thus storing energy as the wax melts), even though the average temperature remains unchanged. This effect is visible in Figure 11b, where only wax in the hot group melts.

In Figure 12, we plot the average temperature in the hot group of the 1,000 server cluster versus the GV from GV=21 to GV=26. The round robin job placement algorithm almost but does not quite reach the melting temperature. With VMT-TA, the average cluster temperature remains the same as round

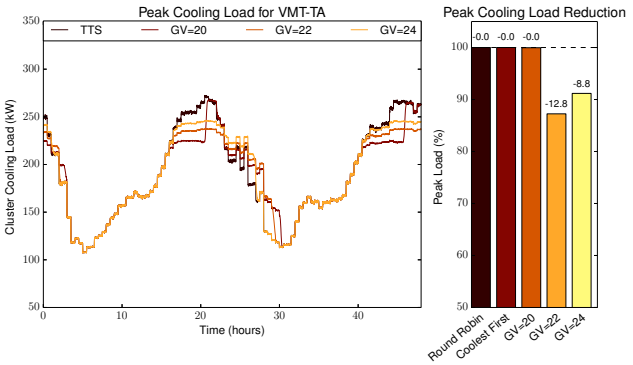


Fig. 13: Cooling load reduction with VMT-TA at 3 different GVs for a cluster of 1000 servers. GV=20 begins melting too soon and runs out of wax capacity before the end of the peak. GV=24 begins melting too late, and still has a significant amount of unmelted wax at the end of the peak.

robin but temperatures in the hot group exceed the melting temperature of the wax.

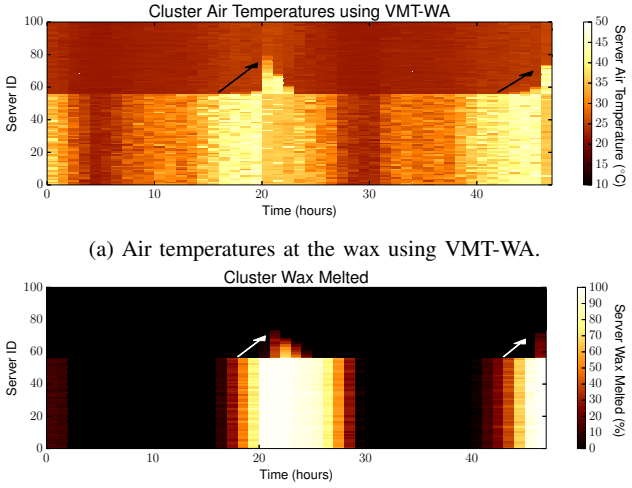
The degree to which the hot group temperature exceeds the average temperature is inversely proportional to the GV value. As the GV setting is decreased the temperature of the hot group increases because there are fewer servers to spread the hot jobs out across, but there is also less thermal energy storage capacity in the hot group because wax is allocated per server.

In Figure 13, we plot the cooling load for three GV values: GV=20, GV=22 and GV=24. GV=22 provides the best peak cooling load reduction of 12.8%. GV=24 works about two-thirds as well (8.8%) because the hot group is still hot enough to melt the wax, but not hot enough for long enough to melt all of the wax so some thermal energy storage capacity goes unused. GV=20 on the other hand is even hotter than GV=22, but melts too fast: just over halfway through the peak all of the wax is melted and the thermal storage capacity is exhausted. At this time, the cooling load increases to provide no benefit for the rest of the peak.

B. Wax Aware VMT

In Figure 14 we plot heat maps of server temperature and wax melted on 100 servers using the VMT-WA job placement algorithm with GV=20. At this setting, which does not provide a significant cooling load reduction using VMT-TA because all of the wax melts prematurely, VMT-WA instead extends the group of hot servers once wax in the hot group servers is melted and continues to melt wax to store energy in the newly added servers.

The temperature impact of this extension is first observable in Figure 14a at the 19th hour, then more clearly after hour 20 where the hot group is expanded by around 20 servers as more servers in the hot group reach the wax melting threshold. As the hot group is expanded, hot jobs are still scheduled on the servers originally in the hot group to maintain a temperature above the melting temperature. This prevents the premature freezing and release of stored thermal energy; however, additional load that



(a) Air temperatures at the wax using VMT-WA.

(b) Wax melted using VMT-WA.

Fig. 14: Heat map of Air temperature at the wax, and wax melted, for a cluster of 100 servers using VMT-WA scheduling (GV=20). The hot group servers (bottom) have a consistently higher temperature than the cold group servers (top). Note the expansion of the hot group around 20 and 45 hours correspond with peak load and wax in the hot group reaching the wax threshold.

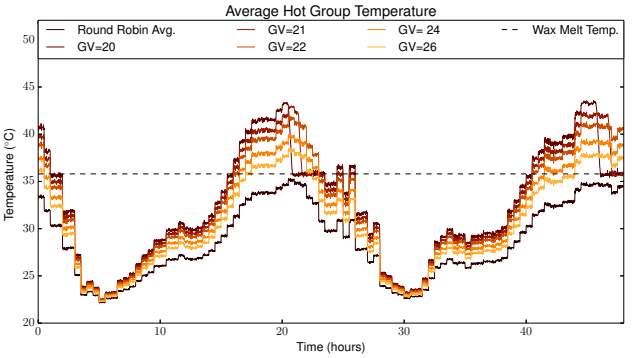


Fig. 15: Average temperature in the hot group using VMT-WA as the GV is adjusted for a cluster of 1000 servers. The hot group is extended when the average temperatures for GV=20 and 21 drop.

would have gone to those servers now goes to newly added servers with unmelted wax. This has the double advantage of moderating the temperature of the melted servers (at the melting point) and moving new jobs to unmelted servers where more thermal storage capacity is available. As a result, we can see a quick drop in temperatures in the hot group in Figure 14a and a capping of the cooling load in Figure 16 at the same time. This quick drop is a result of the granularity in which VMT-WA adds servers to the hot group.

In Figure 14b, the effects of extending the hot group can

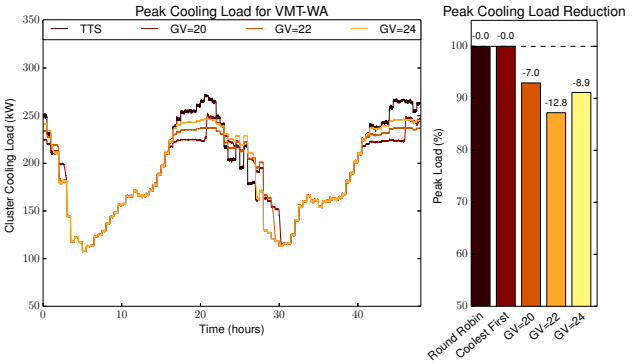


Fig. 16: Cooling load reduction with VMT-WA at 3 different GV levels for a cluster of 1000 servers. For GV=20 when the hot group becomes fully melted, VMT-WA adds more servers to the hot group to and rebalanced load to continue melting wax.

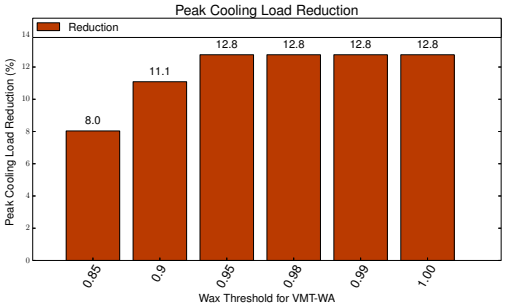


Fig. 17: Peak cooling load reduction as the Wax Threshold is adjusted for VMT-WA (GV=22) for 100 servers. Maximum reduction is achieved above 0.95.

be seen in the distribution of wax melted. None of the newly added hot group servers reach a fully melted state, but because the thermal energy storage happens during the melting process and they do melt more wax than otherwise is melted and more thermal energy is stored.

The effect is further visible in Figure 15, where we plot the average temperature in the hot group servers on GV values from GV=20 to GV=26. The average temperature drops abruptly (at roughly 20 hours for GV=20 and 21 hours for GV=21) when the wax in the original group of servers for GV=20 and GV=21 melts to the wax threshold. Although the average temperature is now lower, the VMT-WA carefully places jobs to maintain the already melted wax and schedules the newly added servers to exceed the melting temperature and melt as much additional wax as possible. For larger GV values, where the wax never becomes fully melted, the temperature and peak cooling load reduction of VMT-WA closely match that provided by VMT-TA.

In Figure 16, we plot the cooling load for GV=20, GV=22 and GV=24. As with VMT-TA, GV=22 provides the greatest peak cooling load reduction (12.8%). This is expected because the overall workload distribution is very close (approximately

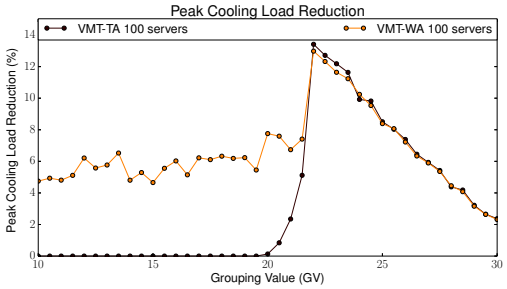


Fig. 18: Peak cooling load reduction as the GV is adjusted for VMT-TA and VMT-WA for 100 servers. Both achieve peak cooling load reduction at GV=22.

60% hot jobs) to the ratio of GV/PMT used to size the group when GV=22. GV=24 also provides results similar to VMT-TA (8.9%), but GV=20 provides significantly better results.

Unlike VMT-TA, where once the hot group is fully melted the cooling load immediately returns to the level without wax, the cooling load with VMT-WA increase once the wax in the initial hot group is melted but levels off as new servers are added to the hot group and hot jobs are placed on these servers to melt more wax. GV=20 does not provide quite as much benefit as the GV=22 or 24, but still manages a 7.0% reduction in peak cooling load.

Figure 17 plots the result of varying the wax threshold, above which VMT-WA considers the wax in a server to be ”fully melted,” from 0.85 to 1.00. (We fix the wax threshold at 0.98, or 98% melted, in all other experiments.) A wax threshold of 1.00 means that all of the wax is melted, however in practice this can be hard to maintain because mild temperature fluctuations can cause small portions of wax to freeze again prematurely. A lower threshold means we are less likely to overshoot the desired temperature, but also risks leaving more wax unmelted and sacrifices some thermal storage. We see from these results that the threshold can be set as low as 0.95 without a noticeable loss in capacity.

C. VMT-TA vs. VMT-WA

In Figure 18, we plot the results of sweeping the GV=10 to GV=30 on 100 servers using VMT-TA and VMT-WA. Both provide peak reduction at GV=22, and as the GV is increased both trend downwards together closely. This is the best GV for this specific combination of workloads and PMT, and will vary from datacenter to datacenter. However because VMT gives the ability to control GV, it provides a necessary degree of flexibility and adaptability that TTS does not.

To evaluate VMT-TA versus VMT-WA, the advantage of VMT-WA is most apparent below 22: while the peak cooling load reduction using VMT-TA quickly drops to zero when the hot group melts too quickly and cannot adjust, the reduction using VMT-WA drops to around 6% immediately then continues to decrease much more slowly afterwards.

First, both perform similarly well at GV=22 and above. This is because there is a fixed amount of energy that can be absorbed from the air before the temperature in the hot group will drop below the melting temperature and no more heat can

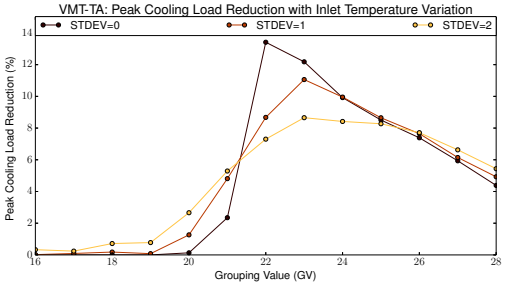


Fig. 19: Peak cooling load reduction using VMT-TA with normally distributed inlet temperature variation (average of 5 runs of 100 servers each).

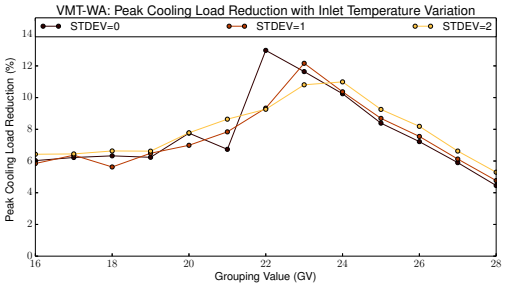


Fig. 20: Peak cooling load reduction using VMT-WA with normally distributed inlet temperature variation (average of 5 runs of 100 servers each).

be stored. Even if all of the wax in VMT-WA is melted and the hot group extended, VMT-WA cannot absorb more energy than VMT-TA at the ideal GV setting. The ideal setting may vary as workload composition or daily load levels change.

This shows that the primary advantage of VMT-WA: it is robust. In a scenario where the operators can predict load accurately day to day, they can actually change the GV to the optimal value each day. However, with VMT-TA they must choose a conservative value because the risk of selecting a value too low is extreme. With VMT-WA, the risk is more balanced.

D. Impact of Inlet Temperature Variation

Real datacenters often have some variation in inlet temperature between servers due to airflow [47]. In this section, we consider the impact of server inlet temperature variation on VMT-TA and VMT-WA, and plot the average cooling load reduction from 5 runs with 100 servers each.

In Figure 19, we plot the peak cooling load reduction using VMT-TA for inlet temperature standard deviations of 0, 1, and 2 °C (95% within ±0, 2 and 4 °C of the mean) as the GV setting is swept from 16 to 28. We observe that at GV=22 (the peak without inlet temperature variation), no inlet temperature variation provides the best reduction. Below GV=21 or above GV=24 however, non-zero standard deviations offer slightly better load reduction than no variation due to the distribution, but still significantly less than near the optimal GV value.

In Figure 19, we plot the peak cooling load reduction using VMT-WA across the same range of temperature variations. We observe that like VMT-TA, outside of the best GV range a small deviation provides the same or slightly better reduction. At the peak we also observe a trend where a non-zero standard deviation increases the GV at which peak reduction is achieved. Even with STDEV=2 (95% within ±4 °C), the peak cooling reduction still reaches 10.9%.

We see then that VMT jobs placement is still effective at reducing total cooling load, even in a less uniform environment. The optimal choice of GV increases slightly in this case (because it is better to miss high than miss low); however, we also continue to see that VMT-WA is much more robust with respect to the choice of GV.

E. TCO Benefits of VMT

Lastly, we quantify the potential benefits that come from using VMT to reduce the peak cooling load using a methodology published in prior work [19]. The two primary benefits provided by a reduced peak cooling load are both derived from cooling oversubscription: either that datacenter can now achieve the same throughput with a smaller cooling system, or more servers can be added to increase throughput under the same cooling system. Both provide significant cost savings.

Both VMT-TA and VMT-WA achieve a peak cooling load reduction of 12.8% in a cluster of 1,000 servers, versus less than 0.2% with TTS alone. Considering the 25 MW datacenter from Section IV-A, a fully subscribed cooling system would need to remove 25 MW of thermal energy from the datacenter at peak load. (The following cost-savings include cost estimates to deploy wax into every server in the datacenter.)

Decreasing the peak cooling load 12.8% reduces the peak cooling load of the datacenter from 25 MW to 21.8 MW and enables a 12.8% smaller cooling system. This provides a cost savings of $2,690,000 over the lifetime of the datacenter based upon cooling system cost estimates [14].

(Note that deploying an n-paraffin wax with a melting temperature near 30 °C for TTS to achieve the same peak cooling load reduction would cost on the order of $10 million, four times more than the money with VMT including the cost of deploying commercial wax.)

For a more conservative approach, a datacenter using VMT-WA may choose undertake only a 6% reduction in the cooling system to account for load variation. A 6% reduction in the size of the cooling system still provides a cost savings of $1,260,000.

Alternatively, the reduced peak cooling load may be used to add more servers to the datacenter under the same cooling system size. Using VMT-TA or VMT-WA with the best peak cooling load reduction, VMT enables 14.6% more servers: 146 additional servers per cluster or 7,339 additional servers in a 25 MW datacenter. The conservative 6% percent application of VMT-WA also provides substantial benefit, enabling 6.4% more servers: 64 additional servers per cluster or 3,191 additional servers in the datacenter without increasing the cooling capital expenditure.

The gains from reduce cooling capacity or greater overprovisioning come from using VMT to reduce the peak (annual) cooling load as evaluated in this paper. There may be additional benefits offered by the ability to control the melting temperature

day-to-day, such as leveraging less expensive off-peak power or green power when cooling energy can be temporally shifted as well.

VI. Related Work

A number of works have proposed to leverage the thermal energy storage capacity of PCMs in the computing domain. Computational sprinting [48], [49], [50], [51], [52] proposes to place a small amount of PCM in contact with the CPU to enable brief "sprints" of fast operation that exceed the safe sustained power levels, but is less useful for datacenters where increased activity lasts for multiple hours at a time. TTS proposes to use wax [19], [24] to passively reshape the thermal profile, but cannot be widely deployed or adapted for many workload mixtures. Other work related to TTS has proposed to use a PCM for emergency overprovisioning [53], and to use an adversarial approach to mitigate conflict for shared resources in datacenters with limited power and cooling utilities [54].

VMT uses a similar approach to TTS, but propose to accomplish the thermal reshaping using both latent energy storage in wax as well as thermal aware job placement to maximize stored energy. In contrast to TTS, which places wax in servers and passively waits for conditions to be amenable to melt wax, TTS actively places jobs to maximize thermal storage and peak cooling load reduction.

Prior work on load balancing [55], [56], [57], [58], [59], [60] used workload placement to improve performance, energy consumption, and/or cooling efficiency. VMT implements workload placement to unbalance power consumption and thus temperatures at the cluster level, however for many workloads these load balancing techniques may still be useful to coordinate jobs within the hot and cold groups and to distribute hot and cold servers spatially to balance load across multiple cooling systems.

Similarly, job consolidation has been considered in prior work to reduce power consumption [61], [62], [63], [64], [65], however this approach requires extra server capacity that may not be available during the peak hours. Job consolidation can be used alongside VMT during the off hours, as long as jobs are not consolidated to a level where they melt wax before the peak hours.

Prior work in thermal aware job placement leverages spacial aware of hot and cold spots in the datacenter to increase efficiency [66]. Tang et al. manage the inlet temperature distribution and place jobs accordingly for maximum power efficiency [67], and Xu et al. propose to relocate jobs between geographically dispersed datacenters to maximize cooling efficiency [68], [69], [70], [71]. These are parallel or compatible work with potential benefits when used alongside VMT.

Power over subscription is another area where prior work proposed to use batteries to manage peak hours and/or power emergencies [14], [72], [73], [74], [75]. Most of these leverage uninterruptible power supply (UPS) batteries already present in datacenters, and their techniques complement VMT well as hot jobs both draw the most power and release the most heat.

Prior work for thermal overprovisioning proposed to use a variety of sensible heat storage mechanisms. Several works propose to use water tanks for thermal energy storage as the thermal density of water is much greater than air, and the water may be chilled during off hours to prepare for peak hour load [76], [77], [78], [79]. VMT is not strictly applicable to techniques that rely on sensible energy storage, rather than latent energy storage, but these techniques are compatible with VMT.

VII. Conclusion

In this work we introduced Virtual Melting Temperature (VMT), a technique to control the thermal load of a datacenter using workload placement in conjunction with Phase Change Material (PCM)-enabled Thermal Time Shifting (TTS). We presented two algorithms, Thermal Aware VMT (VMT-TA) and Wax Aware VMT (VMT-WA), that manage workload placement in order to maximize melting wax, and thus maximizing energy storage with Thermal Time Shifting (TTS). Both policies group hot jobs together to create warm spots in a subset of servers (which may be distributed throughout the datacenter to maintain balanced power distribution), melting more wax in this subset than if job temperatures were evenly distributed. VMT-WA goes a step further by relocating jobs as wax in the hot group becomes fully melted.

We evaluated these algorithms with a scale out study using a simulated cluster of 1,000 servers enabled with paraffin wax over a two day trace covering a mixture of 5 datacenter workloads with different thermal profiles. We found that both VMT-TA and VMT-WA job placement algorithms provide significant benefits VMT-WA, while slightly more complex to implement than VMT-TA, also incorporates workload movement to create a built-in safety factor against temperature and workload variation. Overall, VMT enables up to a 12.8% reduction in the peak cooling load that corresponds that to over $2.6 million in savings over the life of a datacenter, or adding up to 7,339 additional servers running under the same fixed cooling budget.

Acknowledgements

The authors would like to thank the anonymous reviewers for their suggestions and feedback. This work was sponsored by the National Science Foundation (NSF) under grants IISVEC1539011, CAREER SHF-1553485, and CNS-1652925. AMD, the AMD Arrow logo, and combinations thereof are trademarks of Advanced Micro Devices, Inc. Other product names used in this publication are for identification purposes only and may be trademarks of their respective companies.

References

[1] L. A. Barroso and U. Hölzle, "The datacenter as a computer: An introduction to the design of warehouse-scale machines," *Synthesis lectures on computer architecture*, vol. 4, no. 1, pp. 1–108, 2009.

[2] Switch, "SUPERNAP Las Vegas digital exchange campus." https://goo.gl/mmi1km, 2016. Online; accessed 16-Nov-2016.

[3] A. Ghiasi, R. Baca, G. Quantum, and L. Commscope, "Overview of largest data centers," in *Proc. 802.3 bs Task Force Interim meeting*, 2014.

[4] R. Miller, "The billion dollar data centers." https://goo.gl/bhfteC, Apr 2009. Online; accessed 11-Nov-2016.

[5] Y. Sverdlik, "Here's how much energy all us data centers consume." https://goo.gl/UA8u97, Jun 2016. Online; accessed 11-Nov-2016.

[6] J. Maida, "Increase in data center construction will significantly augment the global data center precision air conditioning market until 2020, says technavio." https://goo.gl/9uX1Gj, Jun 2016. Online; accessed 11-Nov-2016.

[7] L. A. Barroso and U. Hölzle, "The case for energy-proportional computing," 2007.

[8] D. Abts, M. R. Marty, P. M. Wells, P. Klausler, and H. Liu, "Energy proportional datacenter networks," in *ACM SIGARCH Computer Architecture News*, vol. 38, pp. 338–347, ACM, 2010.

[9] K. T. Malladi, B. C. Lee, F. A. Nothaft, C. Kozyrakis, K. Periyathambi, and M. Horowitz, "Towards energy-proportional datacenter memory with mobile dram," in *ACM SIGARCH Computer Architecture News*, vol. 40, pp. 37–48, IEEE Computer Society, 2012.

[10] J. Leverich, M. Monchiero, V. Talwar, P. Ranganathan, and C. Kozyrakis, "Power management of datacenter workloads using per-core power gating," *IEEE Computer Architecture Letters*, vol. 8, no. 2, pp. 48–51, 2009.

[11] D. Wong and M. Annavaram, "Knightshift: Scaling the energy proportionality wall through server-level heterogeneity," in *2012 45th Annual IEEE/ACM International Symposium on Microarchitecture*, pp. 119–130, IEEE, 2012.

[12] V. Jimenez, F. Cazorla, R. Gioiosa, E. Kursun, C. Isci, A. Buyuktosunoglu, P. Bose, and M. Valero, "Energy-aware accounting and billing in large-scale computing facilities," *IEEE Micro*, vol. 31, no. 3, pp. 60–71, 2011.

[13] D. Wong and M. Annavaram, "Implications of high energy proportional servers on cluster-wide energy proportionality," in *2014 IEEE 20th International Symposium on High Performance Computer Architecture (HPCA)*, pp. 142–153, IEEE, 2014.

[14] V. Kontorinis, L. E. Zhang, B. Aksanli, J. Sampson, H. Homayoun, E. Pettis, D. M. Tullsen, and T. Simunic Rosing, "Managing distributed ups energy for effective power capping in data centers," in *Computer Architecture (ISCA), 2012 39th Annual International Symposium on*, pp. 488–499, IEEE, 2012.

[15] "Data center cooling market size by application," May 2016.

[16] I. Chowdhury, R. Prasher, K. Lofgreen, G. Chrysler, S. Narasimhan, R. Mahajan, D. Koester, R. Alley, and R. Venkatasubramanian, "On-chip cooling by superlattice-based thin-film thermoelectrics," *Nature Nanotechnology*, vol. 4, no. 4, pp. 235–238, 2009.

[17] M. Iyengar, M. David, P. Parida, V. Kamath, B. Kochuparambil, D. Graybill, M. Schultz, M. Gaynes, R. Simons, R. Schmidt, *et al.*, "Server liquid cooling with chiller-less data center design to enable significant energy savings," in *Semiconductor Thermal Measurement and Management Symposium (SEMI-THERM), 2012 28th Annual IEEE*, pp. 212–223, IEEE, 2012.

[18] P. R. Parida, M. David, M. Iyengar, M. Schultz, M. Gaynes, V. Kamath, B. Kochuparambil, and T. Chainer, "Experimental investigation of water cooled server microprocessors and memory devices in an energy efficient chiller-less data center," in *Semiconductor Thermal Measurement and Management Symposium (SEMI-THERM), 2012 28th Annual IEEE*, pp. 224–231, IEEE, 2012.

[19] M. Skach, M. Arora, C.-H. Hsu, Q. Li, D. Tullsen, L. Tang, and J. Mars, "Thermal time shifting: Leveraging phase change materials to reduce cooling costs in warehouse-scale computers," in *Proceedings of the 42nd Annual International Symposium on Computer Architecture*, pp. 439–449, ACM, 2015.

[20] Y. Zhang, G. Prekas, G. M. Fumarola, M. Fontoura, I. Goiri, and R. Bianchini, "History-based harvesting of spare cycles and storage in large-scale datacenters," in *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation*, OSDI-12, (Berkeley, CA, USA), USENIX, 2016.

[21] K. Pielichowska and K. Pielichowska, "Phase change materials for thermal energy storage," in *Progress in Materials Science*, vol. 65, pp. 67–123, 2014.

[22] A. Sharma, V. Tyagi, C. Chen, and D. Buddhi, "Review on thermal energy storage with phase change materials and applications," vol. 13, pp. 318–345, Elsevier, 2009.

[23] D. Hale, M. Hoover, and M. ONeill, "Phase-change materials handbook," 1972.

[24] M. Skach, M. Arora, C.-H. Hsu, Q. Li, D. Tullsen, L. Tang, and J. Mars, "Thermal time shifting: Decreasing data center cooling costs with phase-change materials," *IEEE Internet Computing*, vol. 21, no. 4, pp. 34–43, 2017.

[25] J. Mars, L. Tang, R. Hundt, K. Skadron, and M. L. Soffa, "Bubble-up: Increasing utilization in modern warehouse scale computers via sensible co-locations," in *Proceedings of the 44th annual IEEE/ACM International Symposium on Microarchitecture*, pp. 248–259, ACM, 2011.

[26] M. A. Laurenzano, Y. Zhang, L. Tang, and J. Mars, "Protean code: Achieving near-free online code transformations for warehouse scale computers," in *Microarchitecture (MICRO), 2014 47th Annual IEEE/ACM International Symposium on*, pp. 558–570, IEEE, 2014.

[27] H. Yang, A. Breslow, J. Mars, and L. Tang, "Bubble-flux: Precise online qos management for increased utilization in warehouse scale computers," in *ACM SIGARCH Computer Architecture News*, vol. 41, pp. 607–618, ACM, 2013.

[28] "Paraffin wax listings on alibaba." http://www.alibaba.com/. Online; accessed 07-Nov-2016.

[29] M. Ferdman, A. Adileh, O. Kocberber, S. Volos, M. Alisafaee, D. Jevdjic, C. Kaynak, A. D. Popescu, A. Ailamaki, and B. Falsafi, "Clearing the clouds: a study of emerging scale-out workloads on modern hardware," in *ACM SIGPLAN Notices*, vol. 47, pp. 37–48, ACM, 2012.

[30] D. Wang, C. Ren, and A. Sivasubramaniam, "Virtualizing power distribution in datacenters," in *ACM SIGARCH Computer Architecture News*, vol. 41, pp. 595–606, ACM, 2013.

[31] J. L. Henning, "Spec cpu2006 benchmark descriptions," *ACM SIGARCH Computer Architecture News*, vol. 34, no. 4, pp. 1–17, 2006.

[32] M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon, "I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system," in *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, pp. 1–14, ACM, 2007.

[33] "Upload files to Google Drive." https://goo.gl/lz3OKG, 2016. Online; accessed 15-Nov-2016.

[34] R. Chitta, *Kernel-based clustering of big data*. PhD thesis, Michigan State University, 2015.

[35] M. Malik and H. Homayoun, "Big data on low power cores: Are low power embedded processors a good fit for the big data workloads?," in *Computer Design (ICCD), 2015 33rd IEEE International Conference on*, pp. 379–382, IEEE, 2015.

[36] L. A. Barroso, J. Dean, and U. Holzle, "Web search for a planet: The google cluster architecture," *IEEE micro*, vol. 23, no. 2, pp. 22–28, 2003.

[37] E. Pinheiro, W.-D. Weber, and L. A. Barroso, "Failure trends in a large disk drive population.," in *FAST*, vol. 7, pp. 17–23, 2007.

[38] B. Schroeder, E. Pinheiro, and W.-D. Weber, "Dram errors in the wild: a large-scale field study," in *ACM SIGMETRICS Performance Evaluation Review*, vol. 37, pp. 193–204, ACM, 2009.

[39] N. El-Sayed, I. A. Stefanovici, G. Amvrosiadis, A. A. Hwang, and B. Schroeder, "Temperature management in data centers: why some (might) like it hot," *ACM SIGMETRICS Performance Evaluation Review*, vol. 40, no. 1, pp. 163–174, 2012.

[40] D. G. Andersen and S. Swanson, "Rethinking flash in the data center," *IEEE micro*, vol. 30, no. 4, pp. 52–54, 2010.

[41] T. J. Stachecki and K. Ghose, "Short-term load prediction and energy-aware load balancing for data centers serving online requests,"

[42] E. Gal and S. Toledo, "Algorithms and data structures for flash memories," *ACM Computing Surveys (CSUR)*, vol. 37, no. 2, pp. 138–163, 2005.

[43] A. M. Caulfield, L. M. Grupp, and S. Swanson, "Gordon: using flash memory to build fast, power-efficient clusters for data-intensive applications," *ACM Sigplan Notices*, vol. 44, no. 3, pp. 217–228, 2009.

[44] Intel, "Intel server board S3420GP server system SR1630GP server board SR1630HGP server chassis SC5650UP calculated MTBF estimates." Intel White Paper, 2009. Online; accessed 01-Oct-2017.

[45] M. K. Patterson, "The effect of data center temperature on energy efficiency," in *Thermal and Thermomechanical Phenomena in Electronic Systems, 2008. ITHERM 2008. 11th Intersociety Conference on*, pp. 1167–1174, IEEE, 2008.

[46] "Google Transparency Report." https://goo.gl/lz3OKG, 2011. Online; accessed 2011.

[47] J. Moore, J. S. Chase, and P. Ranganathan, "Weatherman: Automated, online and predictive thermal mapping and management for data centers," in *Autonomic Computing, 2006. ICAC'06. IEEE International Conference on*, pp. 155–164, IEEE, 2006.

[48] A. Raghavan, Y. Luo, A. Chandawalla, M. Papaefthymiou, K. P. Pipe, T. F. Wenisch, and M. M. Martin, "Computational sprinting," in *High Performance Computer Architecture (HPCA), 2012 IEEE 18th International Symposium on*, pp. 1–12, IEEE, 2012.

[49] A. Raghavan, L. Emurian, L. Shao, M. Papaefthymiou, K. P. Pipe, T. F. Wenisch, and M. M. Martin, "Computational sprinting on a hardwaresoftware testbed," vol. 48, pp. 155–166, ACM, 2013.

[50] A. Raghavan, L. Emurian, L. Shao, M. Papaefthymiou, K. Pipe, T. Wenisch, and M. Martin, "Utilizing dark silicon to save energy with computational sprinting," 2013.

[51] A. Raghavan, Y. Luo, A. Chandawalla, M. Papaefthymiou, K. P. Pipe, T. F. Wenisch, and M. M. Martin, "Designing for responsiveness with computational sprinting," *Micro, IEEE*, vol. 33, no. 3, pp. 8–15, 2013.

[52] L. Shao, A. Raghavan, L. Emurian, M. C. Papaefthymiou, T. F. Wenisch, M. M. Martin, and K. P. Pipe, "On-chip phase change heat sinks designed for computational sprinting," in *Semiconductor Thermal Measurement and Management Symposium (SEMI-THERM), 2014 30th Annual*, pp. 29–34, IEEE, 2014.

[53] M. A. Islam, X. Ren, S. Ren, A. Wierman, and X. Wang, "A market approach for handling power emergencies in multi-tenant data center," in *2016 IEEE International Symposium on High Performance Computer Architecture (HPCA)*, pp. 432–443, IEEE, 2016.

[54] S. Fan, S. M. Zahedi, and B. C. Lee, "The computational sprinting game," in *Proceedings of the Twenty-First International Conference on Architectural Support for Programming Languages and Operating Systems*, pp. 561–575, ACM, 2016.

[55] D. B. LD and P. V. Krishna, "Honey bee behavior inspired load balancing of tasks in cloud computing environments," *Applied Soft Computing*, vol. 13, no. 5, pp. 2292–2303, 2013.

[56] M. Alizadeh, T. Edsall, S. Dharmapurikar, R. Vaidyanathan, K. Chu, A. Fingerhut, F. Matus, R. Pan, N. Yadav, G. Varghese, *et al.*, "Conga: Distributed congestion-aware load balancing for datacenters," in *ACM SIGCOMM Computer Communication Review*, vol. 44, pp. 503–514, ACM, 2014.

[57] K. Li, G. Xu, G. Zhao, Y. Dong, and D. Wang, "Cloud task scheduling based on load balancing ant colony optimization," in *Chinagrid Conference (ChinaGrid), 2011 Sixth Annual*, pp. 3–9, IEEE, 2011.

[58] E. Pinheiro, R. Bianchini, E. V. Carrera, and T. Heath, "Load balancing and unbalancing for power and performance in cluster-based systems," in *Workshop on compilers and operating systems for low power*, vol. 180, pp. 182–195, Barcelona, Spain, 2001.

[59] D. M. Dias, W. Kish, R. Mukherjee, and R. Tewari, "A scalable and highly available web server," in *Compcon'96.'Technologies for the Information superhighway'Digest of Papers*, pp. 85–92, IEEE, 1996.

[60] V. Cardellini, M. Colajanni, and P. S. Yu, "Dynamic load balancing on web-server systems," *IEEE Internet computing*, vol. 3, no. 3, pp. 28–39, 1999.

[61] A. Verma, G. Dasgupta, T. K. Nayak, P. De, and R. Kothari, "Server workload analysis for power minimization using consolidation," in *Proceedings of the 2009 conference on USENIX Annual technical conference*, pp. 28–28, USENIX Association, 2009.

[62] M. Uddin and A. A. Rahman, "Server consolidation: An approach to make data centers energy efficient and green," *arXiv preprint arXiv:1010.5037*, 2010.

[63] D. Meisner, B. T. Gold, and T. F. Wenisch, "Powernap: eliminating server idle power," in *ACM Sigplan Notices*, vol. 44, pp. 205–216, ACM, 2009.

[64] P. Padala, X. Zhu, Z. Wang, S. Singhal, K. G. Shin, *et al.*, "Performance evaluation of virtualization technologies for server consolidation," *HP Labs Tec. Report*, 2007.

[65] P. Apparao, R. Iyer, X. Zhang, D. Newell, and T. Adelmeyer, "Characterization & analysis of a server consolidation benchmark," in *Proceedings of the fourth ACM SIGPLAN/SIGOPS international conference on Virtual execution environments*, pp. 21–30, ACM, 2008.

[66] J. D. Moore, J. S. Chase, P. Ranganathan, and R. K. Sharma, "Making scheduling" cool": Temperature-aware workload placement in data centers.," in *USENIX annual technical conference, General Track*, pp. 61–75, 2005.

[67] 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 Transactions on Parallel and Distributed Systems*, vol. 19, no. 11, pp. 1458–1472, 2008.

[68] H. Xu, C. Feng, and B. Li, "Temperature aware workload management in geo-distributed datacenters," in *Proceedings of the 10th International Conference on Autonomic Computing (ICAC 13)*, pp. 303–314, 2013.

[69] H. Xu, C. Feng, and B. Li, "Temperature aware workload management in geo-distributed data centers," *IEEE Transactions on Parallel and Distributed Systems*, vol. 26, no. 6, pp. 1743–1753, 2015.

[70] Í. Goiri, T. D. Nguyen, and R. Bianchini, "Coolair: Temperature- and variation-aware management for free-cooled datacenters," in *ACM SIGPLAN Notices*, vol. 50, pp. 253–265, ACM, 2015.

[71] Z. Liu, M. Lin, A. Wierman, S. H. Low, and L. L. Andrew, "Geographical load balancing with renewables," *ACM SIGMETRICS Performance Evaluation Review*, vol. 39, no. 3, pp. 62–66, 2011.

[72] S. Govindan, A. Sivasubramaniam, and B. Urgaonkar, "Benefits and limitations of tapping into stored energy for datacenters," in *Computer Architecture (ISCA), 2011 38th Annual International Symposium on*, pp. 341–351, IEEE, 2011.

[73] S. Govindan, D. Wang, A. Sivasubramaniam, and B. Urgaonkar, "Leveraging stored energy for handling power emergencies in aggressively provisioned datacenters," in *ACM SIGPLAN Notices*, vol. 47, pp. 75–86, ACM, 2012.

[74] R. Urgaonkar, B. Urgaonkar, M. J. Neely, and A. Sivasubramaniam, "Optimal power cost management using stored energy in data centers," in *Proceedings of the ACM SIGMETRICS joint international conference on Measurement and modeling of computer systems*, pp. 221–232, ACM, 2011.

[75] S. Govindan, D. Wang, A. Sivasubramaniam, and B. Urgaonkar, "Aggressive datacenter power provisioning with batteries," *ACM Transactions on Computer Systems (TOCS)*, vol. 31, no. 1, p. 2, 2013.

[76] W. Zheng, K. Ma, and X. Wang, "Exploiting thermal energy storage to reduce data center capital and operating expenses," in *Proceedings of the IEEE 19th International Symposium on High-Performance Computer Architecture*, pp. 132–141, IEEE, 2014.

[77] Y. Zhang, Y. Wang, and X. Wang, "Testore: exploiting thermal and energy storage to cut the electricity bill for datacenter cooling," in *Proceedings of the 8th International Conference on Network and Service Management*, pp. 19–27, International Federation for Information Processing, 2012.

[78] K. Roth, R. Zogg, and J. Brodrick, "Cool thermal energy storage," *ASHRAE journal*, vol. 48, no. 9, pp. 94–96, 2006.

[79] D. Garday and J. Housley, "Thermal storage system provides emergency data center cooling," *White Paper Intel Information Technology, Intel Corporation*, 2007.