TACOMA: Server and workload management in Internet data centers considering cooling-computing power trade-off and energy proportionality[1,2]

Zahra Abbasi, Arizona State University, Tempe, AZ
Georgios Varsamopoulos, Arizona State University, Tempe, AZ
Sandeep K. S. Gupta, Arizona State University, Tempe, AZ

A two-tier Internet data center management scheme, TACOMA, with *thermal-aware server provisioning* (TASP) in one tier, and *thermal-aware workload distribution* (TAWD) in the other is proposed. TASP and TAWD co-ordinate to maximize the energy savings by leveraging the workload dynamics, at coarse and fine time scale, respectively. TACOMA is aware of the QoS constraints, the energy proportionality of servers, and the potential trade-off between cooling and computing power. The obtained energy savings are a combination of suspending idle servers, using servers at their peak efficiency, and avoiding heat recirculation.

Categories and Subject Descriptors: D.4.7 [**Operating Systems**]: Organization and Design—*Distributed systems, Hierarchical design, Interactive systems*; D.4.8 [**Operating Systems**]: Performance; H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*Distributed systems, Performance evaluation*

General Terms: Design, Management, Performance

Additional Key Words and Phrases: Data center, energy saving, thermal aware server provisioning, thermal aware workload distribution, energy-proportional systems, and cooling computing power trade-off.

1. INTRODUCTION

Internet Data Centers (IDCs) are increasingly hosting interactive and delay-sensitive web based applications. The energy consumed by IDCs has been skyrocketing, mainly due to the increasing demand on Internet services. A recent study estimates that the energy cost will exceed the cost of the original capital investment of data centers by 2012 [US Department Of Energy 2010]. In response, the research community and industry propose hardware and software solutions to increase the energy efficiency of data centers: (i) computer vendors are developing *servers that are more energy-proportional* [Barroso and Holzle 2007; Varsamopoulos and Gupta 2010], and

[1]Extension of Conference Paper: this paper extends the conference version titled "Thermal aware server provisioning and workload distribution for Internet data centers", published in HPDC 2010 with: i) revised the entire work including TASP and TAWD modeling and solutions under modern non-energy proportional systems, iii) analytical and simulation results for cooling-computing power trade-off under server provisioning, iii) modeling and solutions specifically for homogeneous data centers, and iv) more realistic performance modeling.

[2]This work has been partly funded by NSF CRI grant #0855527, CNS grant #0834797 and Intel Corp.

Authors' address: School of Computing, Informatics, and Decision Systems Engineering, Ira A. Fulton Schools of Engineering, 699 South Mill Avenue Suite 517, Tempe AZ 85281, USA.
Email: {zahra.abbasi, georgios.varsamopoulos, sandeep.gupta}@asu.edu
Web: http://impact.asu.edu/.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies show this notice on the first page or initial screen of a display along with the full citation. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, to redistribute to lists, or to use any component of this work in other works requires prior specific permission and/or a fee. Permissions may be requested from Publications Dept., ACM, Inc., 2 Penn Plaza, Suite 701, New York, NY 10121-0701 USA, fax +1 (212) 869-0481, or permissions@acm.org.
© 0 ACM 1544-3566/0/-ART0 $10.00
DOI 10.1145/0000000.0000000 http://doi.acm.org/10.1145/0000000.0000000

(ii) there has been considerable research demonstrating the effectiveness of software-based management of workload and equipment in computing and cooling energy savings of data centers [Tolia et al. 2008; Chen et al. 2008; Chase et al. 2001; Kusic et al. 2009; Tang et al. 2008; Mukherjee et al. 2009].

The *power usage effectiveness* (*PUE*) of a data center, which is defined as the total power over the computing power [Rawson et al. 2008] is affected by the computing and the cooling power. A large PUE is a strong indication of large cooling power, since the cooling system is the biggest consumer of the non-computing power in a data center (followed by power conversion and other losses). According to US Department of Energy [2010], a modern data center's PUE is around 1.7, which means that $0.7/1.7 \simeq 41\%$ of the power is, in its most part, consumed in cooling the data center.

Server provisioning, i.e., dynamically adjusting the number of *active servers* and suspending unnecessary servers, has been the most prominent technique to reduce both computing and cooling power for IDCs [Kusic et al. 2009; Chen et al. 2008; Chase et al. 2001], since (i) the idle power is comparable to the maximum power for the current computing systems [Barroso and Holzle 2007; Varsamopoulos and Gupta 2010], and (ii) web servers have periods of low utilization due to the periodic nature of the workload [Barroso and Holzle 2007; Chase et al. 2001; Chen et al. 2008]. Analytical formalizations of sever provisioning focus mainly on the trade-off between the *quality of service (QoS)* and the *minimum number of active servers* to reduce the computing power [Chase et al. 2001; Chen et al. 2008; Kusic et al. 2009] and, expectedly, the associated cooling power. However, such an approach, namely computing power (only) aware server provisioning (CPSP), may not always be *effective* or *sufficient* with modern servers because of the ***cooling–computing power trade-off***: consolidating the workload on fewer servers tends to decrease the computing power (since modern servers are not energy-proportional) but also may create hot spots which typically demand greater cooling power [Faraz and Vijaykumar 2010]. Moreover, in some cases, this cooling power increase may outweigh the computing power decrease.

In contrast to CPSP, thermal aware server provisioning, which has been proposed recently, manages the set of active servers and can avoid the cooling-computing power trade-off [Faraz and Vijaykumar 2010; Abbasi et al. 2010; Varsamopoulos et al. 2010]. In general, the *active server set selection* affects the total power of the data center due to the non-uniform temperature distribution (because servers do not equally impact the temperature in the room, nor are the airflow patterns symmetric) and the servers' heterogeneity, in terms of their power and computing performance. Although recent studies show the effectiveness of incorporating thermal awareness into server provisioning for IDCs, formalization of the problem and its evaluation under data center configurations with different energy efficiency is still lacking. Also, the conditions that necessitate the use of thermal aware server provisioning instead of CPSP to ensure the energy efficiency of server provisioning schemes in data centers has not been studied yet; such a study is of vital importance, since under the cooling-computing power trade-off, a CPSP scheme not only may yield reduced energy savings, it may well increase the energy consumption instead.

In recent years, the energy efficiency, utilization over power consumption, of modern computing systems are becoming more diverse [Barroso and Holzle 2007; Varsamopoulos and Gupta 2010]: a non energy-proportional system does not have good energy efficiency at lower utilization levels (see Figs. 1(a), and 1(b)), whereas an ideal energy-proportional system has a constant energy efficiency (see Fig. 1(c)). The difference between the power behavior of "older" systems and "modern" systems (see Fig. 1(a) and Fig. 1(b)) has a big impact on the energy savings of server provisioning and workload distribution: a recent study shows that systems that are more energy-proportional reduce the energy-saving benefits of server provisioning [Varsamopoulos et al. 2010].

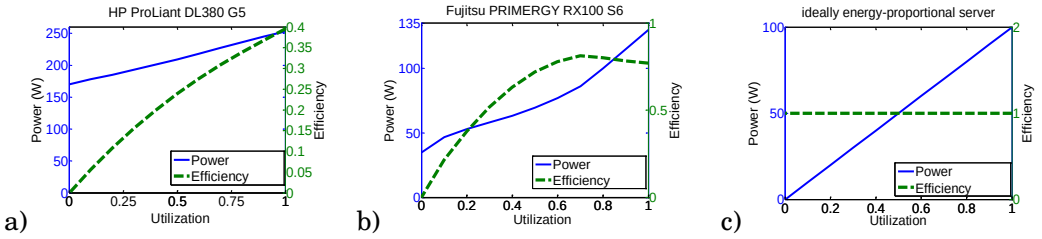

Fig. 1. Energy efficiency for various energy proportionality cases: (a) non proportional system with straight power curve; (b) near-zero idle with early sudden increase in power; (c) ideal energy-proportional system.

Also, as a side effect, the peak energy efficiency may not be at the maximum utilization (see Fig. 1(b)), and therefore an absolute minimum number of active servers may not be the most energy-efficient set. Note that, under server provisioning, servers may still be under-utilized due to the need to *over-provision* to respect the QoS requirements. For such a situation, a workload distribution scheme conscious of the energy proportionality behavior of servers can ensure that active servers operate close to their peak energy efficiency.

1.1. Challenges

The above discussion reflects the need for a holistic approach in server and workload management that should take into account the following factors: (i) the nature of the workload, (ii) the energy proportionality of systems, (iii) the heterogeneity of systems, (iv) the impact of server provisioning on the cooling power, and (v) the non-uniform temperature distribution within a data center. Developing such a holistic solution has to address the following challenges:

— **The trade-off between QoS and energy efficiency:** Intensity variation in the web traffic has been witnessed by several researches [Bohrer et al. 2002; Chen et al. 2005]. To ensure QoS, the active server set should be determined according to the peak traffic to be observed during the operation of that active set [Chen et al. 2008]. However, this results in under-utilization of the servers at all other times in that duration when the traffic is not at its peak. This under-utilization will most likely operate the active servers at a low energy efficiency point. In addition to this trade-off, there is the challenge of accurately predicting the peak traffic; usually the peak traffic is over-estimated to ensure the QoS.

— **The trade-off between cooling and computing power:** A recent study documented a trade-off between idle power and cooling power that can manifest when performing server provisioning to decrease total idle power [Faraz and Vijaykumar 2010]. This is due to the consolidation of the workload to one “side” of the data center, where the systems are utilized at or near their peak level; potentially resulting in hot spots due to generation of lot of localized high-temperature heat, and consequently increasing the cooling demand.

— **Ensuring computationally efficient thermal aware server provisioning:** The problem of server and workload management is exacerbated when considering the impact of active server selection on the cooling energy. Such consideration makes the server selection problem nonlinear because of the nonlinear relationship between computing and cooling power and the nonlinear nature of cooling system’s coefficient of performance (CoP), i.e., the ratio of the power extracted over the power required by the cooling unit [Moore et al. 2005]. These nonlinear effects increase the solution time for the online problem of server and workload management.

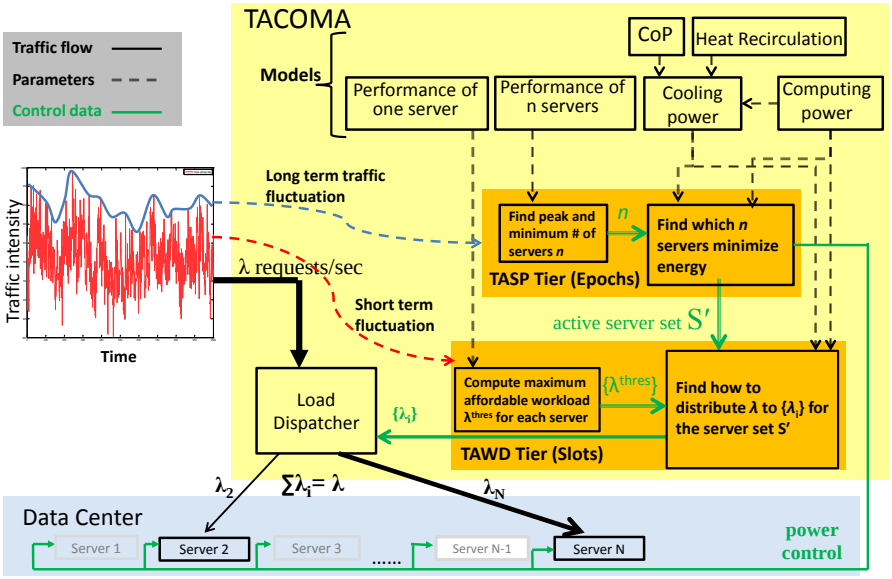

Fig. 2. TACOMA a two-tier architecture for thermal aware server and workload management of IDCs. The first tier, TASP, on each epoch determines the minimum number of active servers and chooses the active server set S' to minimize the total energy, and the second tier, TAWD, on each slot decides on the workload distribution, λ_is, of the active servers to minimize the energy consumption of active servers.

1.2. Overview of approach, contributions and results

Given the aforementioned requirements and challenges, the goal of this paper is *to study the impact of the aforementioned factors on data center energy consumption and develop workload and server management algorithms which take these factors into account to reduce the total energy consumption while maintaining the performance requirements*. To this end, we develop Two-tier Architecture for Cooling cOmputing energy Management Architecture (TACOMA) with Thermal Aware Server Provisioning (TASP) to run at "epochs" (∼1 hr) and Workload Distribution (TAWD) to run at "slots" (∼5 sec) as depicted in Fig. 2 that matches the long term and short term fluctuation of the web traffic. It runs online algorithms to manage servers (i.e. on/off control) and distribute workloads among active servers. The main contributions of the paper can be summarized as follows:

— ***Formalizing conditions for the existence of cooling-computing power trade-off***. The paper proves some lemmata to identify the parameters that affect the cooling–computing power trade-off, namely energy proportionality, energy efficiency of the data center (in terms of PUE), and the size of the active server set with respect to the available servers (§2.3). The lemmata provide an easy-to-solve analytical method to test the occurrence of cooling-computing power trade-off due to workload consolidation. Our study shows that low PUEs disfavor the occurrence of the trade-off, where as computing systems with good energy proportionality favor the occurrence of the trade-off (§2.3 and §7.2). Further, the occurrence of cooling-computing power trade-off depends on the active server set and can be avoided by thermal aware server provisioning.

— ***Formalizing TASP and TAWD*** for both homogeneous and heterogeneous data centers. In the first tier, TASP adjusts the number of active servers to the incoming workload and chooses thermal and power efficient servers as active servers to avoid

cooling-computing power trade-off and save energy. The server selection problem is formulated as a Mixed Integer Programing (MIP) for homogeneous data centers and as a non-linear min-max programming for heterogeneous data centers (§5). In today's IDCs, the performance requirements is expressed as percentile bound on the worst-case response time. The worst case response time, that is, the response time at peak traffic, is estimated using a GI/G/n queuing model [Allen 1990]. At the start of an epoch, its peak traffic is estimated in two steps: first, a Kalman filter estimates the average traffic (i.e., request arrival rate) and then a regression technique extracts the peak traffic with respect to the average traffic. In the second tier, TAWD, predicts the workload arrival rate at the start of each slot using a Kalman filter. It then determines the workload distribution of the active servers considering the maximum affordable workload arrival rate of each servers and their energy efficiency. The energy and thermal aware workload distribution is formulated as a linear program.

— ***Fast heuristic solutions to TASP and TAWD problems***. Due to the NP-hardness of TASP problems (mainly because of thermal awareness and computing power awareness, see Appendix A), heuristic solutions, namely TASP Least Recirculated Heat (TASP-LRH) and Computing Power LRH hybrid (TASP-CPLRH) are proposed (§6.1). TASP-LRH, an approximation solution, runs a rank-and-sort technique using the LRH metric and TASP-CPLRH, a heuristic solution, sorts the equipment according to their computing power efficiency and then applies LRH ranking in each group of servers with the same efficiency. TAWD also employs LRH and CPLRH metrics to quickly decide on the workload distribution of the active servers.

— ***Performance evaluation of TASP and TAWD***. TASP and TAWD are evaluated using realistic web traces, a thermal model of the ASU HPCI data center and various realistic power profiles of modern servers. The power savings of TASP approaches are compared against CPSP, and a recent thermal-aware server provisioning scheme, namely PowerTrade-d [Faraz and Vijaykumar 2010]. Results show that ***TASP with load balancing (i.e., no TAWD) yields increased energy savings of up to 20% with respect to CPSP and PowerTrade-d (§7.3)***. The higher energy saving of TASP compared to CPSP comes from considering the impact of server consolidation on the cooling energy. Also, compared to PowerTrade-d, TASP's energy savings comes from improving the algorithm that tries to minimize the sum of cooling and computing energy. Moreover, heuristic solutions to TASP yield energy savings that are similar to the optimal solutions (about 2-6% worse). Performing TAWD on "modern" servers that exhibit non-linear power curve can yield up to 12% energy savings compared to no server provisioning with performance-oriented load balancing. When ***TAWD is used in conjunction with TASP, the energy savings can reach 40% with respect to CPSP and PowerTrade-d (§7.4)***.

TASP and TAWD are also validated in a small scale carton box environment that emulates heat recirculation within a data center room. Results show that even in this small test bed where the total power consumption of systems is about 500 W, TASP and TAWD decrease the temperature by almost two degrees with respect to thermally oblivious server provisioning and load balancing under the same workload. These results are available at `http://impact.asu.edu/publication/ExpReport-TASPTAWD.pdf`.

1.3. Organization of this paper

The rest of the paper is organized as follows: §2 gives an overview on the energy proportionality and the thermal aware scheduling and provides an analytical study on the occurrence of cooling-computing power trade-off due to server consolidation. §3 reviews the related work. §4 formally describes TACOMA and gives an overview of the performance and power consumption modeling of a data center. TASP and TAWD

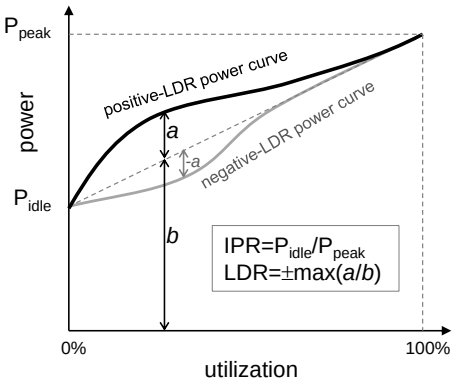

Fig. 3. Pictorial definition of Idle to Peak power Ratio (IPR) and of Linear Deviation Ratio (LDR).

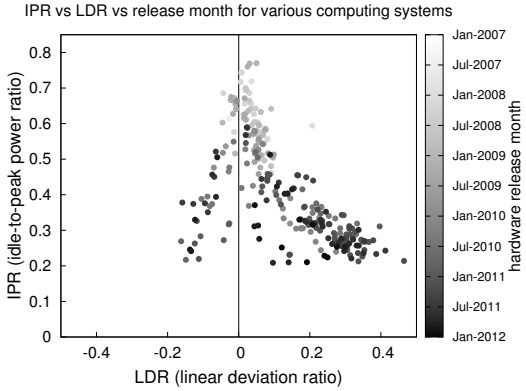

Fig. 4. Scatter-plot of IPR vs IPR of over 290 computing systems, color-coded by release date, based on SPECPower_ssj2008's public benchmark data (`http://www.spec.org/power_ssj2008/results/`).

problems and their analytical formulation are described in §5. Further, §6 presents the analysis of these problems and introduces the optimal and heuristic solutions for TASP and TAWD. Furthermore, §7 presents the simulation-based evaluation of TASP and TAWD under different energy proportionality levels of servers and different PUEs of data centers and discuses the results. Finally, we discuss the implementation challenges of TACOMA and future in §8 and in §9.

2. BACKGROUND ON ENERGY PROPORTIONALITY AND THERMAL AWARENESS

This section discusses the effect of energy proportionality and thermal impact of servers on energy savings of the workload and server management in data centers.

2.1. Energy proportionality of systems

One characteristic of computing systems, until recently, has been the high energy consumption when they are idle (see Fig. 1(b)). Variable performance technologies, such as dynamic duty cycling, considerably dampen the power consumption of recent computing systems at near-idle utilization levels [Barroso and Holzle 2007] (see Fig. 1(c)). However, the power-utilization curve is not as linear. Varsamopoulos and Gupta observe this behavior and introduce two metrics to measure how energy-proportional a system is (see Fig. 3): (i) Idle to Peak Ratio (IPR) measures how close to zero the idle power is, and (ii) LDR measures the linearity of the power curve (i.e., how close the power curve is to hypothetical linear curve connecting idle power to peak power) [Varsamopoulos and Gupta 2010]. *Ideal energy proportional* servers have zero IPR and LDR. Systems with power consumption of linearly increase over utilization have zero LDR and the power model of the form: $p = \alpha u + \omega$, where IPR=$\frac{\omega}{\alpha+\omega}$, and $0 \leqslant u \leqslant 1$ denotes the utilization. Recent trends show that existing computing systems cover a large area of the IPR-LDR spectrum [Varsamopoulos and Gupta 2010]. An updated scatter plot with SPECPower_ssj2008 published benchmark results up to December 2011, shows that (i) as systems lower their IPR, their LDR gets larger, and (ii) there are no systems, at least among the ones tested, with an IPR less than 0.2 (see Fig. 4). This paper uses these two metrics to study how the energy proportionality of systems can affect the performance of energy aware workload and server management.

2.2. Thermal aware scheduling

This section gives an overview on how the cooling energy can be affected by the workload and server management. In contemporary air-cooled data centers, the computing servers (blades) are organized in chassis which in turn are organized in racks, with the racks being organized in rows. The rows are arranged so that, in each aisle between two rows, either front panels or back panels are facing each other; this is called the hot aisle / cold aisle arrangement. In this paper, we use *chassis* and *computing node* interchangeably. Computing nodes consume power according to their hardware characteristics and computational workload, and heat the air in the room. Cooling is provided by the Computing Room Air Conditioning (CRAC), which pushes the cold air through the perforated tiles in the raised floor [Moore et al. 2005]. The temperature of the supplied cooled air, denoted as T_{sup}, should be low enough so that the inlet temperature of the computing nodes does not go beyond the red line temperature (T_{red}) which is specified by the manufactures. In the ideal case, all the hot air should directly go back to the CRAC; but, in practice, some of the hot air recirculates back to the computing nodes. The amount of heat recirculation depends on the physical layout and the airflow pattern in the data center room. Usually, the heat recirculation is not uniform, i.e., different parts of data center have different contribution in the heat recirculation. Tang et al. model the heat recirculation coefficients for all pairs of nodes in a data center, considering the data center layout and thermodynamic conditions: $\boldsymbol{D}$=$\{d_{ij}\}_{N\times N}$, where N is the total number of nodes, and each d_{ij} denotes the fraction of heat that flows from node j to node i [Tang et al. 2008]. Assume $\boldsymbol{p}^{comp}$ is a vector denoting the computing power consumption of servers, then T_{in}, the inlet temperature vector of servers can be written as:

$$T_{in} = T_{sup} + D\boldsymbol{p}^{comp}. \tag{1}$$

Eq. 1 shows that the inlet temperatures of servers depend on the power consumption of nodes, and consequently on the amount of workload they are assigned. Indeed, workload assignment is hidden in the computing power vector ($\boldsymbol{p}^{comp}$). On the other hand, cooling energy depends on T_{sup}. Cooling energy of the CRAC can be modeled by its *coefficient of performance* (CoP), which is the ratio of the heat removed (i.e., computing energy) over the work required to remove that heat (i.e., cooling energy). A higher CoP means more efficient cooling, and usually the higher the required supplied temperatures (T_{sup}), the better the CoP is. In other words, CoP is usually monotonically increasing function of the supplied temperature, e.g., for an HP data center CoP reported as $CoP(T_{sup}) = 0.0068T_{sup}^2 + 0.0008T_{sup} + 0.458$ [Moore et al. 2005]. However, according to Eq. 1, the highest CRAC output temperature is limited by the servers' *redline temperature*. Therefore, T_{sup} can be *at most* equal to:

$$T_{sup} = T^{red} - \max(D\boldsymbol{p}^{comp}), \tag{2}$$

where the function $\max$ ensures that the supplied temperature of CRAC does not exceed the redline temperature of the hottest equipment. Respectively, the cooling power, denoted by P^{AC}, can be written as a function of the CoP of the supplied temperature:

$$P^{AC} = \frac{P^{comp}}{CoP(T^{red} - \max(D\boldsymbol{p}^{comp}))}, \tag{3}$$

where P^{comp} denotes the total computing power. Eq. 3 suggests that for *a given load*, the cooling power can be potentially improved by efficient workload distribution. Intuitively, this is possible if the workload distribution is thermally balanced among the servers, meaning that a higher portion of the workload is assigned to the servers that have the least contribution on the heat recirculation. We refer to those servers as *ther-*

mally efficient servers. Since, the main contributors of the total power in a data center are the computing and the cooling power, a high PUE implies the high heat recirculation in a data center.

Cooling power is also affected by server consolidation. The computing power, p, of a server is a function of the idle power and the peak power, and the server's utilization level, such that by using a server consolidation scheme, e.g., CPSP, the total idle power decreases. However, due to workload consolidation effect of the scheme, the computing power of active servers also increases, which may increase the value of $\max(D\boldsymbol{p}^{comp})$ in Eq. 3, i.e., hot spots may be created which increase the required cooling power. In other words, under a server consolidation scheme such as CPSP, a trade-off between cooling and computing (idle) power can occur [Faraz and Vijaykumar 2010]. The following section provides analysis to illustrate the conditions for this trade-off to occur.

2.3. Cooling-computing power trade-off for server consolidation

In this section we present a formal definition of the cooling-computing power trade-off and prove some lemmata to: (i) clarify the conditions when the trade-off occurs, and (ii) provide an easy-to-solve analytical method to examine the occurrence of the trade-off. We also provide some discussion on the trade-off's implication on the energy saving of the workload and server consolidation schemes. The lemmata are given for the special case of server/workload consolidation, namely balanced workload consolidation as defined below[3].

Definition 2.1. In a homogeneous data center, any workload consolidation from N servers with an equal (balanced) utilization of u down to $n < N$ servers with a balanced utilization of $u' > u$ such that the total utilization remains unchanged, i.e., $Nu = nu'$, is called *Balanced Workload Consolidation, BWC*(N, u, n, u'). This definition also assumes that the $N - n$ servers that are being relieved of the workload are switched-off.

Definition 2.2. For a fixed workload if a consolidation in a data center decreases the computing power from P^{comp} to $P^{comp\prime}$, and the cooling power increases from P^{AC} to $P^{AC\prime}$, then a *cooling-computing power trade-off* is said to exist.

Definition 2.3. In a matrix $D = \{d_{ij}\}_{N\times N}$, a row i in which $\exists j, k :\ d_{ik} \neq d_{ij}$, is called a *non-uniform* row. A matrix with only non-uniform rows, is called a *strictly row-wise non-uniform* matrix.

Definition 2.4. A row-wise uniform matrix has no non-uniform rows.

LEMMA 2.5. *Consider a homogeneous data center $\mathcal{D}$ with cooling whose power adheres to Eq. 3, strictly row-wise non-uniform and positive heat recirculation matrix of $D = \{d_{ij}\}_{N\times N}$, and a CoP that is a strictly monotonically increasing function of the supplied temperature T_{sup}. There exist a BWC (N, u, n, u') and an associated $\varepsilon > 0$ for $\mathcal{D}$ such that BWC causes cooling-computing power trade-off for IPR $< \varepsilon$.*

PROOF. Let $P^{comp} = N(\alpha u + \omega)$, and $P^{comp\prime} = n(\alpha u' + \omega)$ be the computing power before and after the consolidation. Similarly, let P^{AC} and $P^{AC\prime}$ be the cooling power before and after the consolidation. According to Definition 2.1 $P^{comp} > P^{comp\prime}$. Hence, according to Definition 2.2 to show the existence of a cooling-computing power trade-off for a consolidation, it is enough to show $P^{AC} < P^{AC\prime}$. A constructive proof follows.

[3]In this section,"′" denotes parameter values after consolidation, "″" and "$*$" denote the parameter values of a particular consolidation scheme.

According to Eq. 2 the CRAC's supplied temperature before consolidation, T_{sup}, is as follows: $T_{sup} = T_{red} - d(\alpha u + \omega)$, where $d = \|D\|_\infty = \max\left\{\sum_{j=1}^{N} d_{1j}, \ldots, \sum_{j=1}^{N} d_{Nj}\right\}$[4]. After the consolidation BWC(N, u, n, u') chooses the active server set $S' \subseteq \{1, \ldots, N\}$, $|S'| = n$, leading to $T'_{sup} = T_{red} - d'(\alpha u' + \omega)$, where $d' = \|D\backslash S'\|_\infty$.

Choose a BWC(N, u, n, u'), say $\mathcal{C}$, with $d' > \frac{n}{N}d$ and let $d' = \frac{n}{N}d + \delta$, where $\delta > 0$. Note that such a BWC exists due to the row-wise non-uniformity of the heat recirculation matrix (Definition 2.3). Following shows that $\mathcal{C}$ satisfies the lemmas' conditions:

— **Hypothesis 1.** For the given $\mathcal{C}$, ($\exists \varepsilon_1 > 0$: IPR $< \varepsilon_1$) $\Leftrightarrow T_{sup} > T'_{sup}$. Note $T_{sup} > T'_{sup}$ is a necessary condition for the occurrence of the trade-off. We can prove Hypothesis 1 as follows:

$$\begin{aligned} T_{sup} > T'_{sup} &\Leftrightarrow d(\alpha u + \omega) < d'(\alpha u' + \omega) \qquad \text{[from Eq. 2]} \\ &\Leftrightarrow d(\alpha u' \frac{n}{N} + \omega) < (d\frac{n}{N} + \delta)(\alpha u' + \omega) \\ &\Leftrightarrow d\omega < \frac{n}{N} d\omega + \delta(\alpha u' + \omega) \\ &\Leftrightarrow \frac{\omega}{\alpha + \omega} d(1 - \frac{n}{N}) < \frac{\delta(\alpha u' + \omega)}{\alpha + \omega} \\ &\Leftrightarrow \text{IPR} < \frac{\delta(\alpha u' + \omega)}{d(1 - \frac{n}{N})(\alpha + \omega)}. \end{aligned} \tag{4}$$

Let $\varepsilon_1 \equiv \frac{\delta(\alpha u' + \omega)}{d(1 - \frac{n}{N})(\alpha + \omega)}$. Since $n > N$, it follows that $\varepsilon_1 > 0$, establishing Hypothesis 1.

— **Hypothesis 2.** For the given $\mathcal{C}$, ($\exists \varepsilon_2 > 0$: IPR $< \varepsilon_2$) $\Leftrightarrow P^{AC} < P^{AC'}$. Note $P^{AC} < P^{AC'}$ is a sufficient condition for the occurrence of the trade-off. Due to $Nu = nu'$, and the linearity of the power consumption model, it follows: $P^{comp} = n(\alpha u' + \omega) + (N - n)\omega$. We can prove Hypothesis 2 as follows:

$$\begin{aligned} P^{AC} < P^{AC'} &\Leftrightarrow \frac{n(\alpha u' + \omega) + (N - n)\omega}{CoP(T_{sup})} < \frac{n(\alpha u' + \omega)}{CoP(T'_{sup})} \\ &\Leftrightarrow \frac{n(\alpha u' + \omega)/(\alpha + \omega) + (N - n)\omega/(\alpha + \omega)}{CoP(T_{sup})} < \frac{n(\alpha u' + \omega)/(\alpha + \omega)}{CoP(T'_{sup})} \\ &\Leftrightarrow \frac{n(\alpha u' + \omega)}{\alpha + \omega} + (N - n)IPR < \frac{CoP(T_{sup}) n(\alpha u' + \omega)/(\alpha + \omega)}{CoP(T'_{sup})} \\ &\Leftrightarrow \text{IPR} < (\frac{CoP(T_{sup})}{CoP(T'_{sup})} - 1)\frac{n(\alpha u' + \omega)}{(N - n)(\alpha + \omega)}. \end{aligned} \tag{5}$$

Let $\varepsilon_2 \equiv (\frac{CoP(T_{sup})}{CoP(T'_{sup})} - 1)\frac{n(\alpha u' + \omega)}{(N - n)(\alpha + \omega)}$. According to Eq. 4 and monotonicity property of the CoP function, it follows that $\frac{CoP(T_{sup})}{CoP(T'_{sup})} > 1$. Also $N > n$, which establishes $\varepsilon_2 > 0$.

Let $\varepsilon = \min(\varepsilon_1, \varepsilon_2)$. Since $\varepsilon_1 > 0$ and $\varepsilon_2 > 0$, it follows that $\varepsilon > 0$. □

The following lemma shows that a necessary condition for the occurrence of cooling-computing power trade-off is non-uniform and non-zero heat recirculation.

[4] $\|D\|_\infty$ is the *infinity norm* of a matrix D. Conventionally, the definition sums the absolutes of the matrix elements, which does not affect the usage in this paper. Also, we expand the notation to $\|D\backslash S'\|_\infty$,i.e., "*infinity norm of D* ***under*** S'," which equals to the submatrix of D that has only the columns that correspond to the servers in active set S', i.e.,$\|D\backslash S'\|_\infty = \max\{\sum_{j \in S'} d_{1j}, \ldots, \sum_{j \in S'} d_{1j}\}$.

LEMMA 2.6. *In a data center with row-wise uniform heat recirculation $D = \{d_{ij} = d_i\}_{N\times N}$, homogeneous servers of LDR=0, cooling power of Eq. 3, and a CoP of strictly monotonically increasing function of supplied temperature T_{sup}, a BWC(N, u, n, u') never causes cooling-computing power trade-off.*

PROOF. According to Definition 2.1, the cooling-computing power trade-off occurs if $P^{comp} > P^{comp\prime}$, and:

$$P^{AC} < P^{AC\prime} \Leftrightarrow \frac{P^{comp}}{CoP(T_{sup})} < \frac{P^{comp\prime}}{CoP(T'_{sup})}.$$

Since, $P^{comp} > P^{comp\prime}$, the necessary condition for the occurrence of the trade-off is $CoP(T_{sup}) > CoP(T'_{sup})$, which means that there must exist $h > 0$ such that $T_{sup} = T'_{sup} + h$ (due to strictly monotonically increasing of CoP as a function of supplied temperature, $h > 0$). We prove the lemma through contradiction as follows.

According to Eq. 2 the CRAC's supplied temperature before consolidation, T_{sup}, is as follows: $T_{sup} = T_{red} - d(\alpha u + \omega)$, where $d = \|D\|_\infty$. We denote as $d_* \equiv d/N$ the maximum element in the matrix. After the consolidation BWC(N, u, n, u'), the supplied temperature becomes $T'_{sup} = T_{red} - d'(\alpha u' + \omega)$. Since, $Nu = nu'$ we have the following:

$$\begin{aligned} &T_{sup} = T'_{sup} + h \Leftrightarrow T_{red} - Nd_*(\alpha u + \omega) = T_{red} - nd_*(\alpha u' + \omega) + h \\ &\Leftrightarrow d_*(\alpha n u' + N\omega) = d_* n(\alpha u' + \omega) - h \\ &\Leftrightarrow d_* N\omega = d_* n\omega - h \Leftrightarrow h = d_* n - d_* N. \end{aligned} \tag{6}$$

Since $n < N$, $h < 0$, which contradicts the fact that $h > 0$. □

THEOREM 2.7. *Consider a homogeneous data center with cooling whose power adheres to Eq. 3, a CoP that is a strictly monotonically increasing function of the supplied temperature T_{sup}, a strictly positive heat recirculation matrix D, there exist a BWC(N, u, n, u') and an associated $\varepsilon > 0$ to cause cooling-computing power trade-off when IPR $< \varepsilon$ if and only if there are n elements in one non-uniform row whose sum is greater than n times the maximum element in the uniform rows.*

PROOF. Let $d = \|D\|_\infty$ before consolidation, and $d' = \|D\backslash S'\|_\infty$ after some arbitrary BWC(N, u, n, u') chooses the active server set S'. We prove the two directions of "if and only if" in separate parts:

"if". Given that there are n elements in a row whose sum is greater than n times the maximum element in the uniform rows, it is possible to construct the BWC(N, u, n, u') such that it yields $d' > \frac{n}{N}d$, by setting the active server set S' to contain the n servers that correspond to those n elements in that non-uniform row. By use of the proof steps in Lemma 2.5, we can easily show that there exists an $\varepsilon > 0$ such that if IPR $< \varepsilon$, the constructed BWC causes cooling-computing power trade-off.

"only if". Given that there are no n elements in any non-uniform row whose sum is greater than n times the maximum element in the uniform rows, d' must be one of the uniform row sums. Therefore $d' = \frac{n}{N}d$. By use of the proof steps in Lemma 2.6, we can easily show that the arbitrary BWC(N, u, n, u') causes no cooling-computing trade-off under any IPR.

□

The above theorem provides a link between the IPR of servers and the structure of the heat recirculation matrix to the *existence* of cooling-computing power trade-off. Specifically, the inequality in Eq. 5 provides a sufficient condition for the existence

of the cooling-computing power trade-off. The lemma below provides a way to find a balanced workload consolidation that is close (by a bound) to the one that minimizes the supplied temperature. In general, the lower the T'_{sup} of a consolidation is at a certain IPR, because of the trade-off, the more likely that consolidation is to still cause the trade-off if the IPR is increased.

LEMMA 2.8. *Consider a homogeneous data center with zero-LDR servers* $S = \{1, 2, \ldots, N\}$ *and positive heat recirculation coefficient matrix* $D = \{d_{ij}\}_{N\times N}$*. The lowest supplied temperature,* T'_{sup*}*, by any BWC*(N, u, n, u')*, is bounded by the inequality:*

$$T_{red} - Nd''(\alpha u' + \omega) \leqslant T'_{sup*} \leqslant T_{red} - d''(\alpha u' + \omega),$$

where $d'' = \|D\backslash S''\|_\infty$*,* $S'' \subseteq S$*,* $|S''| = n$*, and* $\forall j \in S''$ *and* $\forall k \notin S''$*,* $\sum_{i=1}^{N} d_{ij} \geqslant \sum_{i=1}^{N} d_{ik}$*.*

PROOF. The CRAC's supplied temperature when using workload consolidation, denoted by T'_{sup}, depends on the active server set. Let S'_k, for $1 \leqslant k \leqslant \binom{N}{n}$ denote the k^{th} active server set of size n that can be chosen from N servers. According to Eq. 2 there is a d'_k associated with every S'_k such that $T'_{sup,k} = T_{red} - d'_k(\alpha u + \omega)$, where $d'_k = \|D\backslash S_k\|_\infty$. Note that S'' is one of the $\{S'_k\}$ active server sets, where $d'_k = d''$. Consider a balanced workload consolidation that chooses the active server set $S_*{}'$ which causes the lowest supplied temperature T'_{sup*}, its associated d'_* satisfies the following:

$$d'_* = \max\left\{d'_1, \ldots, d'_k, \ldots, d'_{\binom{N}{n}}\right\} \geqslant d''. \tag{7}$$

According to definition of S'', it follows that $\sum_{i=1}^{N}\sum_{j=1}^{N} d_{ij} < \frac{N}{n}\sum_{i=1}^{N}\sum_{j\in S''} d_{ij}$. Also according to well-known *norm inequalities* we have that:

$$d'_* \leqslant \sum_{i=1}^{N}\sum_{j=1}^{N} d_{ij} \leqslant \frac{N}{n}\sum_{i=1}^{N}\sum_{j\in S''} d_{ij} \leqslant \frac{N}{n}\left(n\,\|D\backslash S''\|_\infty\right) = Nd''. \tag{8}$$

Eqs. 7 and 8 establish the lemma. □

In the above lemma S'' can be found through calculating the summation of columns over matrix D, sorting them in ascending order and picking n columns whose sum is maximum. This scheme is used for implementing CPSP in the evaluation section.

Based on the above we conclude the following:

(1) *The lower the yielded* T'_{sup} *of a consolidation at a certain IPR, because of the cooling-computing power trade-off, the higher the IPR has to increase before that consolidation ceases to cause the trade-off.*
(2) *High IPR and low PUE disfavor the occurrence of cooling-computing power trade-off, whereas low IPR and high PUE favor the occurrence of the trade-off.* Assume[5] PUE=$1 + 1/CoP(T_{sup})$, then high PUE is indication of high heat recirculation. High heat recirculation increases the value for $\frac{CoP(T_{sup})}{CoP(T'_{sup})}$ where T_{sup} and T'_{sup} denote the supplied temperature of without and with workload consolidation. According to Eq. 5 higher ratios of $\frac{CoP(T_{sup})}{CoP(T'_{sup})}$ expand the range of IPR values (i.e., we can find higher a ε) where the trade-off may occur.

[5]Note that PUE characterizes the overall power efficiency of a data center with respect to its computing power whereas CoP characterizes the efficiency of CRAC unit. Specifically, PUE=$\frac{\text{Total power}}{\text{computing power}}$. However, for simplification, we ignore non-computing power other than that of the cooling system.

(3) *Performing a balanced workload consolidation may be detrimental in improving the energy efficiency of a data center.* According to Lemma 2.5, there can be cases where a workload consolidation causes cooling-computing power trade-off, then it follows that if the magnitude of cooling power increase outweigh computing power decrease, workload consolidation is detrimental in improving the energy efficiency of a data center compared to no workload consolidation. This may happen due to the high heat recirculation in data centers and low IPR of servers.
(4) *A balanced workload consolidation scheme that chooses the active server set among thermal efficient servers, will mitigate the cooling-computing power trade-off.* Such a scheme will choose servers to minimize heat recirculated to the servers. Hence, $\frac{CoP(T_{sup})}{CoP(T'_{sup})}$ decreases, disfavoring the occurrence of cooling-computing power trade-off.

3. RELATED WORK AND MOTIVATION

This section gives an overview of previous work and the way that we benefit from or improve them.

3.1. Dynamic server provisioning

In dynamic server provisioning, the idea is to adjust the number of active servers in a server farm to the offered workload and suspend the rest. The effectiveness of this method is based on a) the traffic intensity difference between periods of workload *lows* and *peaks*, the latter being about two to three times as intense as the former [Chen et al. 2008; Barroso and Holzle 2007; Ranganathan et al. 2006], and b) current computing systems (and system components) consume significant amount of power at idle compared to being turned off or throttled down [Chen et al. 2008]. Related works show that dynamic server provisioning saves significant energy under real web workload. However, the existing analytical formulations [Chase et al. 2001; Chen et al. 2008] mainly account for the trade-off between providing quality of service and energy saving, while overlooking (i) the heterogeneity of systems, (ii) the power proportionality of systems, (iii) the thermal efficiency of servers, and, for some of them, (iv) the challenges associated with the long term workload prediction. *In this paper we enhance dynamic server provisioning to take into account the overlooked parameters in the previous works in order to improve energy savings*.

Dynamic server provisioning poses some challenges. First, it takes energy to switch servers on and off. Second, while servers are being switched on, they are unavailable and any requests sent to them will be lost; while servers are being switched off, any requests running on them may be lost before they finish. Some solutions have already been proposed to address these problems [Kusic et al. 2009; Chen et al. 2008]. Our approach avoids the above repercussions by performing infrequent and proactive switching. Dynamic server provisioning requires to selectively suspend individual servers; modern data centers are increasingly likely to support this functionality.

3.2. Thermal aware server provisioning and workload scheduling

Thermal aware scheduling at both the chip-level and the data center-level has been proposed in some works [Moore et al. 2005; Tang et al. 2008; Mukherjee et al. 2009; Zhou et al. 2010]. The idea is to decrease the heat generated by microchips and servers through workload scheduling. We focus on the thermal aware scheduling at the data center level which can be seen as complementary to the thermal aware workload scheduling in chip-level. Moore et al., and Bash and Forman show that thermal aware workload placement can save energy in data centers [Moore et al. 2005; Moore et al. 2006; Bash and Forman 2007]. Tang et al. and Mukherjee et al. model the heat

that, inefficiently, is recirculated among the servers; using this model, they propose spatio-temporal thermal-aware job scheduling algorithms for high performance computing data centers [Tang et al. 2008; Mukherjee et al. 2009]. Thermal aware scheduling for IDCs is also investigated in some papers. Sharma et al. introduce thermal load balancing and show that dynamic thermal management based upon asymmetric workload placement can promote uniform temperature distribution and reduce cooling energy [Sharma et al. 2005]. We also develop TAWD which is not only aware of servers' impact on the temperature distribution, but also is aware of the power behavior of servers. Parolini et al. provide analytical formulation to manage the workload distribution among servers which relies on the expected value of the traffic over time [Parolini et al. 2010]. In contrast, our TASP and TAWD are designed to incorporate the fluctuating nature of web workloads for server provisioning and workload distribution. Thermal aware server provisioning for IDCs is also studied by Faraz and Vijaykumar. Authors propose *PowerTrade-d*, a dynamic thermal aware server provisioning which trades the idle power and the cooling power for each other [Faraz and Vijaykumar 2010]. They argue, reducing the active server set size may not always reduce the total power, as it may increase the cooling power. PowerTrade-d manages the trade-off through a dynamic refinement process such that whenever a change in the size of active server set is required, extra servers are activated or deactivated one by one to ensure the desired balance between cooling power and idle power. *Similar to PowerTrade-d, our TASP and TAWD are aware of both the cooling and the computing power and are targeted to minimize the total power, however the refinement process to minimize the total power is analytically formulated eliminating the need for dynamic refinement*. We implement PowerTrade-d and use it as a reference algorithm to evaluate our approaches.

3.3. Related work on energy proportionality based workload management

The performance of computing power (e.g., the behavior of power versus utilization) of systems and its use in the power-aware workload scheduling has not been sufficiently addressed in related works. Chun et al. and Tsirogiannis et al. through some experiments observe that different platforms have different energy proportionality characteristics and exhibit different power performance with respect to the type of workload (e.g., I/O intensive, CPU intensive and transactions), and that the CPU power is not linear with utilization [Chun et al. 2010; Tsirogiannis et al. 2010]. Krioukov et al. design a power-proportional cluster by considering the power proportionality of systems consisting of a power-aware cluster and a set of heterogeneous machines, namely NapSAC [Krioukov et al. 2010]. The results in this work confirm that considering the computing power behavior of systems helps to save energy, however, their design account for a simple configuration of data center where systems are statically sorted according to their average power efficiency. No theoretical formulation of the problem is provided, nor is thermal efficiency of the servers considered. *We design a workload management that is aware of the energy proportionality behavior of modern systems which have significant variation of energy efficiency over utilization*.

4. SYSTEM MODEL

We assume a virtualized heterogeneous data center, where all servers are capable of running any web application albeit at a different speed and power consumption. We further assume a workload that consists of short requests (or transactions). These transactions usually take up to few hundred milliseconds to complete (e.g., HTTP requests). Moreover, we assume that the arrival rate exhibit both short term and long term temporal fluctuations similar to what have been observed for IDC arrival rate [Bohrer et al. 2002; Chen et al. 2008].

This section models the operational architecture, the energy consumption as well as the delay performance.

4.1. TACOMA architecture

We develop TACOMA, a two-tier, global-view, centralized control software architecture as shown in Fig. 2. Tier 1 (T1), the server provisioning tier, iteratively decides the active server set in coarse-time intervals called *epochs*. Let $S=\{s_i\}$, $i=1 \dots N$ be the server set. The T1 controller at the beginning of each epoch estimates $n \leqslant N$ minimum number of required active machines and chooses the active server set S', where $S' \subseteq S$, $S' = \{s_i, 1 \leqslant i \leqslant N\}$, and $n \leqslant |S'| \leqslant N$. Due to the overhead of removing the servers from the active set, e.g., power control and releasing reserved computing resources, an epoch is assumed to be around half hour. Fig. 2 shows an example that for a given workload in an epoch, T1 controller determines that two servers are required to be active. Then T1 based on power and thermal efficiency of servers specifies active servers for which the least energy consumption is incurred and minimum server requirement is met (in this example servers S_2 and S_N are selected and servers $S_1, S_3, S_4, \dots S_{N-1}$ are suspended). The controller of Tier 2 (T2), the workload distribution tier, operates at fine-time intervals called *slots* (around 1-10 seconds) and decides on the distribution (i.e., partitioning) of the workload among the active servers. Let the average request arrival rate at the m^{th} slot be λ_m; T2 controller determines λ_{im}, for all $s_i \in S'$ where $\sum_{s_i \in S'} \lambda_{im} = \lambda_m$ such that the performance requirement is met and the energy consumption is minimized. In other words, number of active servers determined by T1 is adjusted by the peak traffic during an epoch. However, in most slots during an epoch, traffic is much lower than the peak traffic. In these cases, T2 shifts traffic toward energy efficient servers and tries to utilize them at near their maximum energy efficiency level. For example consider Fig. 2 where T2 is going to distribute workload of a given slot to S_2 and S_N where the intensity of traffic is lower than the total capacity of servers S_2 and S_N. In this case, T2 gives more traffic to the energy efficient server (S_N) and utilizes it near its maximum energy efficiency level in such a way that it is not overloaded.

4.2. Data center energy consumption model

Energy consumption of a data center is the sum of its cooling and computing energy:

$$E^{total} = E^{comp} + E^{AC}. \tag{9}$$

Typically, there are other sources of energy consumption in a typical data center such as lighting that are: (i) irrelevant to computing energy and (ii) are insignificant. Hence, the paper disregards them. To calculate computing energy, we model each servers' computing power as follows:

$$p_i = \omega_i + f(u_i),$$

where ω_i denotes power consumption of server i at idle state, and f is a function that maps the utilization to the power consumption. We use two linear and nonlinear model for the function f. The linear model is the basis of our analytical study and is chosen for the sake of simplicity and driving optimal solutions. In this model the total power consumption of active servers ($\forall s_i \in S'$), at a slot m, having utilization of u_{im} can be written as:

$$P_m^{comp} = \sum\nolimits_{s_i \in S'} (\omega_i + \alpha_i u_{im}), \tag{10}$$

where α_i represents extra power consumption at full utilization for each server i. Applying Eqs. 3 and 10 to Eq. 9, the total energy consumption at m^{th} slot of the length t

becomes:

$$E_m^{total} = \left(1+\frac{1}{\text{CoP}\left(T^{\text{red}}-\max\left(\boldsymbol{D}(\boldsymbol{w}+\boldsymbol{a}\odot\boldsymbol{u}_m)\right)\right)}\right)\sum_{i=1}^{N}(\omega_i+\alpha_i u_{im})t, \tag{11}$$

where, $\boldsymbol{w}$ and $\boldsymbol{a}$ denote vector form of the scalar computing power parameters and $\boldsymbol{u_m}$ represents vector form of utilization of servers. Also the operator $\odot$ is defined such that $\boldsymbol{a}\odot\boldsymbol{u_m}$ is vector $\langle\alpha_i u_{im}\rangle$.

We also use a nonlinear model of the function f in our simulation study (§7.4), which is based on the approximation of the power-utilization curve of recent systems.

4.3. Performance modeling

We use $GI/G/n$ queuing model to determine how many active servers to provision for each epoch as well as how much workload to assign for each active server during slots to satisfy the performance requirement of requests (i.e., response time). The model is a n-server queuing system serving requests with generalized arrival and service time distribution and has been previously used in the literature for QoS-cognizant server provisioning [Faraz and Vijaykumar 2010; Urgaonkar et al. 2008; Kusic et al. 2009]. The behavior of a $GI/G/n$ system can be captured using the following approximation model [Allen 1990]:

$$\bar{t}_{res} = \frac{1}{\mu} + \frac{u'_n}{\mu(1-u)}\frac{C_A^2+C_B^2}{2n}, \tag{12}$$

where $\bar{t}_{res}$ represents the average response time, μ denotes the average service rate, C_A and C_B denote the coefficient of variation (i.e., standard deviation divided by the mean) of request inter arrival time and service time respectively, and u' is as follows:

$$u' = u^{\frac{n+1}{2}} \text{ for } u \leqslant 0.7 \text{ and } u'_n = \frac{u^n+u}{2} \text{otherwise.}$$

Let λ be the average arrival rate, the parameter u denotes the average utilization of a server and can be written as $\frac{\lambda}{n\mu}$. Given the average arrival and service rate, the above model can be used to calculate the minimum number of required active servers to achieve the desired response time.

Consider an epoch e where the average arrival rate of requests in its L slots is as follows: $\{\lambda_1, \lambda_2, \ldots, \lambda_L\}$. To ensure that the response time is met in all the slots of e[6], the number of active servers for e should be determined according to $\lambda^{peak} = \max\{\lambda_1, \lambda_2, \ldots, \lambda_L\}$. Using the existing prediction mechanisms, the average arrival rate λ of an epoch over all the slots can be predicted. It represents the smooth variation of workload over time. We use Kalman filtering to predict λ. However, the arrival rate during peak time (i.e., λ^{peak}) is not easy to predict.

Analyzing some Internet traces, we observe that there is a strong correlation between the average arrival rate λ and the peak arrival rate λ^{peak} in an epoch (§7.1.1). For this analysis, we use *Minitab* and perform a regression analysis [Montgomery et al. 2001] on two weeks of the web trace of 1998 FIFA World Cup [Arlitt and Jin 2000] and two days of a web trace of Microsoft Hotmail [Thereska et al. 2009] and build a regression linear model of the form:$\lambda^{peak} = \beta_1\lambda + \beta_0$. We observe that R-square predicted (i.e., R-sq(pred)) for both the workloads are high, (77.45%, and 93.83% for Hotmail and FIFA traces, respectively), and the corresponding *P-value* for both of the *t-test* and the

[6]Today's data centers performance criteria focus on the worst case or 95th-percentile response time. For that, the active server set is provisioned according to the response time at peak traffic which is the worst case during an epoch.

test for significance of regression (i.e., F0) are almost zero (it was smaller than 0.0001). These results indicate that λ^{peak} can be linearly predicted through λ.

Service rate depends on servers' physical characteristics and the size of requests. To consider the heterogeneity of the servers and model the service rate of servers, we define a *reference computing capacity* and a *reference request size*. The *reference computing capacity* represents the computing capacity of the server with the minimum capacity which is normalized to one. The computing capacity c_i of server i is expressed as an integer factor of the reference computing capacity. This assumption is reasonable, since the computing capacity of modern multi-core servers are mainly affected by their core counts. The *reference request size* represents a block of data where the reference server can handle requests of that size with the rate of μ^{ref}. Therefore, the service rate for each server can be modeled as a function of c_i, μ^{ref} and the requests size. For server i, c_i, and μ^{ref} are constant, and the statistics of requests size is updated over time, such that at the beginning of each epoch the most recent statistics are used.

To calculate the maximum affordable workload by each server, a server is modeled as $GI/G/1$ whose average response time $\bar{t}_{res}$ is:

$$\bar{t}_{res} = \frac{1}{\mu_i} \frac{u_i}{1-u_i} \frac{C_{Ai}^2 + C_{Bi}^2}{2}, \tag{13}$$

where the index i ties the queuing system variable to server i. Using this model, the maximum affordable workload arrival rate of serve i at m^{th} slot, λ_{im}^{thres}, to achieve the desired response time can be calculated by computing the u_{im}^{thres} corresponding to the desired response time and using $\lambda_{im}^{thres} = \mu_i u_{im}^{thres}$.

5. PROBLEM FORMULATION OF TASP AND TAWD

The formulation of TASP and TAWD as optimization problems is based on combining the aforementioned models to express the energy consumption as a function of the active server set and the workload distribution. For both TASP and TAWD there are separate formulations for homogeneous and heterogeneous data centers. Although the homogeneous case is a sub-case of the heterogeneous one, its formulation is presented for the following reasons: (i) it can be solved by a heuristic with a known approximation bound, and (ii) that heuristic is used to build a heuristic for the heterogeneous case.

5.1. T1: Thermal aware server provisioning (TASP) problem

We first formalize the general (heterogeneous data center) case and then present the formalization for the homogeneous data center case.

Given a data center with the server set S of N servers, for an epoch of length Lt, where L is the number of t-length slots in an epoch, the average arrival rate λ, the peak arrival rate λ^{peak}, and the minimum number n of required servers, choose the active set $S' \subseteq S$, where $n \leqslant |S'| \leqslant N$, that minimizes the total energy E^{total}.

Let the binary vector $\boldsymbol{x}$ denote the choice of the servers in the active set. TASP can be represented in the following optimization problem of *finding the vector $\boldsymbol{x}$*:

$$\text{minimize } (1+\frac{1}{\text{CoP}(T^{red}-\max(\boldsymbol{D}(\boldsymbol{x}\odot(\boldsymbol{w}+\boldsymbol{a}\odot\boldsymbol{u}))}) \sum_{i=1}^{N} x_i(\omega_i+\alpha_i u)Lt,$$

subject to:

$$\sum_{i=1}^{N} c_i x_i \geqslant n \qquad \text{[performance constraint]} \tag{14}$$

$$u = \frac{\lambda^{peak}}{\mu \sum_{i=1}^{N} c_i x_i} \qquad \text{[load balancing constraint]}$$

$$x_i \in \{0,1\}, \ \ \forall i = 1 \ldots N.$$

The above problem is nonlinear and has $N+1$ variables. The vector $\boldsymbol{x}$ of size N is binary and the variable u denoting the average utilization of servers is a continuous variable. Note that the load balancing constraint implies that the formulation assumes that the workload will be equally balanced among active servers. Also, note that the problem is on minimizing the total energy and not on minimizing the active server set size. It may happen that an optimal solution for one is not optimal for the other.

5.1.1. TASP problem for homogeneous data centers. In a homogeneous data center, where all nodes have the same computing efficiency (i.e., $c_i = c$, $\mu_i = \mu \ \forall i$), and the same computing power efficiency (i.e., $\alpha_i = \alpha$, and $\omega_i = \omega \ \forall i$), the TASP problem can be simplified as follows. Assume that $|S'|$ is fixed, e.g., $|S'| = n$, which means that we know that we need $|S'|$ servers but the question is which servers we need. In this case, the summation part in the objective function in Eq. 14, representing the computing energy, becomes independent of the selection of servers and therefore a constant. The only part that depends on the server selection is $\max D(\boldsymbol{x}\odot(\boldsymbol{w}+\boldsymbol{a}\odot\boldsymbol{u}))$[7]. Hence, the objective function in Eq. 14 can be simplified as follows:

$$\text{minimize } \ \max D(\boldsymbol{x}\odot(\boldsymbol{w}+\boldsymbol{a}\odot\boldsymbol{u})).$$

Let matrix $A = \{a_{i,j}\}_{N\times N}$ be such that $a_{i,j} = d_{i,j}(\omega_j + \alpha_j u)$, $\forall i,j = 1 \ldots N$, where $u = \frac{\lambda^{peak}}{|S'|\mu}$. By applying the same constraints as for the heterogeneous case, TASP can be written as the following optimization problem:

$$\begin{array}{ll} \text{minimize max} & A\boldsymbol{x} \\ \text{subject to} & \sum_{s_i \in S'} x_i = |S'| \\ & x_i \in \{0,1\}. \end{array} \tag{15}$$

Note that each element $a_{i,j}$ of A represents the temperature rise of server i from server j. Also the sum of row i is the total temperature rise of server i. Although the form of the problem is simplified, it is not simplified computationally. This formulation is based on the assumption that $|S'|$ is fixed, however this is not true. §6 explains how to use this problem iteratively to solve TASP for the homogeneous data center case. Appendix A provides a proof on the NP-hardness of this problem.

5.2. T2: Thermal aware workload distribution (TAWD) problem

Given an active server set S' for an epoch with length Lt, and the average request arrival rate of λ_m at the m^{th} slot, determine λ_{im}, $\forall s_i \in S'$, $\forall m = 1 \ldots L$,

[7]Note that if we allow $|S'|$ to vary, then the problem can not be simplified, since utilization varies (u) computing energy is no longer a constant value and both the computing and cooling energy should be involved in the minimization problem.

Table I. Symbols and definitions

Symbol	Definition
N	number of total computing nodes
n	minimum number of required active nodes
t	length of slots (in second)
L	number of slots in a given epoch
i	index of nodes
m	index of slots
f	function to map utilization to power
ω_i	idle power of node i
α_i	peak power of node i minus ω_i
d_{ij}	heat dissipated from node j to node i
p_{im}^{comp}	power consumption of node i at m^{th} slot
c_i	computing capacity of node i w.r.t. reference
C^A	coefficient of variation of workload arrival rate
C^B	coefficient of variation of request size
$\bar{t}_{res}$	average response time
λ_{im}	avg. request arrival rate of node i
$\lambda_{im}^{\text{thres}}$	maximum affordable workload of node i
λ	avg. request arrival rate at an epoch
λ^{peak}	peak request arrival rate at an epoch

Symbol	Definition
μ^{ref}	reference service rate
μ	average service rate
S	set of all servers
S'	active server set
$\boldsymbol{w}$	vector $\{\omega_i\}_N$
$\boldsymbol{a}$	vector $\alpha_{i\,N}$
$\boldsymbol{u}_m$	utilization vector $u_{im\,\{N\}}$
$\boldsymbol{p}_m$	computing power vector $p_{im\,\{N\}}$
$\boldsymbol{r}$	servers' ranking vector
$\boldsymbol{D}$	heat recirculation matrix $\{d_{ji}\}_{(N\times N)}$
E^{total}	total energy consumption of data center
E^{AC}	total cooling energy
TASP	Thermal Aware Server Provisioning
TAWD	Thermal Aware Workload Distribution
CPSP	Computing Power aware Server Provisioning
NoSP	No Server Provisioning
LRH	Least Recirculated Heat
CPLRH	Computing Power LRH hybrid

where $\sum_{s_i\in S'}\lambda_{im}=\lambda_m$, such that the desired response time is achieved and the energy consumption for that slot is minimized.

Let $\lambda_{im}^{\text{thres}}$ be the maximum workload arrival rate that server i at slot m can afford such that its average response time does not exceed the reference response time (see Eq. 13). TAWD can be represented as the following optimization problem on *finding the utilization vector of $\boldsymbol{u}$*:

minimize E_m^{total} on S' in Eq. 11,
subject to:

$$0 \leqslant u_i \leqslant \frac{\lambda_{im}^{thres}}{\mu_i}, \forall i | s_i \in S' \text{ [performance constraint]} \tag{16}$$
$$\sum_{s_i\in S'} u_i\mu_i = \lambda_m \text{ [capacity constraint].}$$

TAWD for the homogeneous case can be simplified as a min-max combinatorial optimization problem, similar to §5.1.1, and due to continuous variables (i.e., u_i) it can be formulated as a linear program.

All problems in this section are NP-hard (except for TAWD for the homogeneous case). For this reason, we describe some heuristics in the next section.

6. TASP AND TAWD SOLUTIONS

6.1. TASP Solutions

To solve the nonlinear optimization problem of TASP, we propose the following approaches.

6.1.1. TASP solution for homogeneous data center. Two methods, namely TASP-MIP and TASP-LRH, are presented to solve TASP in the homogeneous data center case, as follows:

— TASP-MIP: The optimal active server set can be calculated by solving the mixed integer program of TASP for the homogeneous case (see Eq. 15) in an iterative fashion. As mentioned in §5.1.1, MIP modeling of the problem is only true when $|S'|$ is fixed (because the variable u is then fixed, the computing energy is constant, and hence the

latter can be removed from the optimization problem). Therefore, to cover all cases ($|S'| = n \dots N$), we run MIP in a loop starting from $|S'| = n$, i.e., the minimum number of required servers, toward $|S'| = N$, and stopping when the total energy no longer reduces (this means that increasing number of active servers no longer decrease the total energy).

— TASP-LRH: This solution is similar to TASP-MIP, except that instead of solving MIP iteratively, a well-known *N-approximation* algorithm [Aissi et al. 2006] is used (to solve Eq. 15), where N is the total number of servers (the approximation bound N can be found similar to the proof in Lemma 2.8). This algorithm works as follows: (i) calculate the sum of each column in matrix A in Eq. 15, (ii) sort the sums in descending order, and (iii) pick $|S'|$ servers (as per the iteration's number) whose corresponding sum is the lowest, and set their corresponding x_i to 1. The sum of each column is the contribution of the corresponding server to the heat recirculation. We refer to this algorithm as Least Recirculated Heat (LRH) in the rest of the paper. According to LRH, we can define the thermal efficiency rank of servers (i.e., the contribution of servers to the heat recirculation), denoted by $\mathbf{r^{LRH}}$, as the summation of columns of matrix A which can be written as follows:

$$r_i^{LRH} = \sum_{j=1}^{n} d_{ji}(\alpha_i u + \omega_i), \tag{17}$$

where u is calculated according to the load balancing constraint of Eq. 14. Using this metric to select the active servers is equivalent to the aforementioned method. LRH does not find the optimal solution because it minimizes the total heat recirculated instead of minimizing the maximum per-server heat recirculated.

Example 1: To understand how MIP and LRH work in a particular iteration, consider a simple example. Assume there are three servers and we need two active servers. Assume the matrix A in Eq. 15 is as follows:

$$A = \begin{bmatrix} 0.3 & 0 & 0 \\ 0.2 & 0.1 & 0.1 \\ 0 & 0.2 & 0.3 \end{bmatrix}.$$

The LRH metrics (i.e., summation of columns) of these three servers are: $[0.5, 0.3, 0.4]$. Therefore, according to TASP-LRH, servers two and three are chosen as the active servers. Consequently the temperature rise vector is $[0, 0.2, 0.5]$ which is calculated from summing the rows of A after removing Column 1. However, the optimal active server set chosen by TASP-MIP consist of servers one and three where the temperature rise vector is $[0.3, 0.3, 0.3]$. It can be seen in this example that the maximum temperature rise for TASP-LRH is higher than that for TASP-MIP.

According to Eq. 17, the computation of LRH for one server is $O(N)$. Since there are N servers and the upper bound of iterations is $O(N)$, the overall complexity of TASP-LRH is $O(N^3)$.

Note both TASP-LRH and TASP-MIP are balanced workload consolidation (Definition 2.1) when they employ load balancing for workload distribution scheme.

6.1.2. TASP solutions for heterogeneous data center. For the heterogeneous case, we use TASP-MiniMax, a numerical solution, and TASP-CPLRH, a heuristic, as follows:

— TASP-MiniMax: The optimization problem of Eq. 14 is a nonlinear min-max optimization problem. Min-max problems can be solved numerically in MATLAB using sequential quadratic programming (SQP). SQP is polynomial but it can only find a

local optimum. Since the TASP problem is a discrete min-max[8] problem (Eq. 14), TASP-MiniMax computes a solution in the continuous domain, and then discretizes the vector to the closest discrete feasible solution:

> **Algorithm TASP-MiniMax: 1.** In every epoch, solve the problem in Eq. 14 using a min-max solver such as SQP on the continuous domain; obtain vector $\boldsymbol{u}$. **2.** Sort $\boldsymbol{u}$ in descending order, then chose enough of the corresponding highest-value servers (each element in $\boldsymbol{u}$ corresponds to a server) as the active server set to satisfy the capacity constraint of Eq. 14.

The high complexity of TASP-MiniMax limits its use in the online selection of active servers—the time complexity of QP alone is $O(m^3M)$, where m is the number of variables and M is the size of input. However, due to providing a good approximation, it is used for comparison in the evaluation section.

— TASP-CPLRH: This solution is similar to TASP-LRH, except that at each iteration, instead of LRH, Computing Power LRH hybrid (CPLRH) is used, described as follows. Servers are ranked and grouped according to their computing efficiency (p_i^{comp}/c_i), where each group contains servers of the same computing efficiency. Within each group servers are ranked and sorted according to LRH (i.e., Eq. 17).

Example 2: To understand how TASP-CPLRH works, consider again matrix A in Example 1. Further, assume that server one is the most computing efficient server and servers two and three have the same computing efficiency. Hence, if the active server set of size two is required servers one and two are chosen.

TASP-LRH and TASP-CPLRH are of low complexity and they are recommended for actual use in data centers. Algorithm 1 gives the pseudo-code to be run in such cases. It illustrates the actions that TACOMA would perform, i.e., workload prediction at each epoch (*OnEpochTimeOut*), determining the active server set (*TASP*) and deciding on the workload distribution (*TAWD*). Specifically, procedure *TASP* is coded to be compatible with TASP-LRH and TASP-CPLRH, and contains a loop to determine the active server set. TAWD will be discussed in detail in §6.2.

6.1.3. CPSP: the baseline algorithm. Most of the previous research only takes into account the computing efficiency in the active server selection. Therefore, we use the power-aware yet thermally oblivious version of CPSP as the baseline algorithm to evaluate the efficiency of TASP, wherein servers are ranked solely based on their computing power efficiency, $r_i^{CPSP} = p_i^{comp}/c_i$ and n lowest-ranking servers are chosen as S'. Choosing among multiple, equally ranked servers is implementation dependent. Note that CPSP for homogeneous data centers satisfies Definition 2.1, i.e., it performs a balanced workload consolidation.

6.1.4. PT-d+. We also compare our solutions to PowerTrade-d [Faraz and Vijaykumar 2010]. PowerTrade-d avoids total energy increase due to the cooling-computing power trade-off by dynamically and iteratively resizing the active server set by one server at a time and checking that the overall energy is not increased. The size of active servers is determined by 'SurgeGuard' which uses a $GI/G/n$ model to estimate the average size and then augments it by *server-reserve*. *Server-reserve* is the number of servers that are reserved and added to the active server set prediction. The original PowerTrade-d uses the "inverse-temperature" scheme, that tries to balance the servers' inlet temperatures. We, however, implement "PT-d+" using MinHR [Moore et al. 2005] instead of the inverse-temperature scheme, as its performance is better [Tang et al. 2008]. The

[8]We use the term *min-max* to refer to the class of problems, and *MiniMax* as the name of the specific algorithm defined in this paper.

Algorithm 1 Thermal aware server provisioning and workload distribution

procedure INITIALIZATION( )
  Set up epoch controller timer:*OnEpochTimeOut* (i.e., Lt)
  Set up slot controller timer:*OnSlotTimeOut* (i.e., t)
end procedure

procedure ONEPOCHTIMEOUT( )
  Estimate avg. arrival rate (λ) and calculate λ^{peak}
  Calculate the minimum number of required servers n(Eq. 12)
  S'= call TASP(n,λ^{peak})
  Set the supplied temperature T_{sup} (Eq. 2)
  Switch on and off servers as required
end procedure

procedure TASP(n,λ^{peak})
  E^{tot} and $E^{tot}_{previous} \leftarrow$ A very large number
  S' and $S'_{previous} \leftarrow$ Current active server set.
  while $n \leqslant N$ and $E^{tot}_{previous} \leqslant E^{tot}$ **do**
    $S'_{previous} \leftarrow S'$
    $S' \leftarrow n$ lowest ranking server (using LRH or CPLRH)
    $E^{tot}_{previous} \leftarrow E^{tot}$
    $E^{tot} \leftarrow$ total energy of S', for $u = \frac{\lambda^{peak}}{n\mu}$ using Eq. 11
    $n \leftarrow n + 1$
  end while
  return S'
end procedure

procedure ONSLOTTIMEOUT( )
  Estimate upcoming workload arrival rate (λ_m)
  Calculate $\boldsymbol{\lambda_m^{thres}}$ using Eq. 13
  $\boldsymbol{u}$=Call *TAWD*(λ_m,$\boldsymbol{\lambda_m^{thres}}$)
  Rank servers according to CPLRH or LRH
end procedure

procedure TAWD(λ_m,$\boldsymbol{\lambda_m^{thres}}$)
  $\{\beta_1, \beta_2, \ldots, \beta_n\} \leftarrow \{0, 0, \ldots, 0\}$
  Index servers in S' by ascending r^{LRH} order using j as index
  for $j = 1 \ldots n$ **do**
    $\beta_j = \min(\frac{\boldsymbol{\lambda_m^{thres}}[j]}{\lambda_m}, 1 - \sum_{k=1}^{n} \beta_k)$
  end for
  $\lambda_{jm} = \beta_j \lambda_m \; \forall \boldsymbol{s}_j \in S'$
end procedure

procedure UPONJOBARRIVAL
  if estimated_$\lambda_m \leqslant$ observed_λ_m **then**
    Dispatch job according to TAWD policy
  else
    Dispatch jobs according to Load Balancing
  end if
end procedure

MinHR scheme is similar to LRH with the difference being that MinHR uses experimental measurements instead of a model to rank the servers.

6.2. TAWD solutions

Energy-efficient workload distribution, as described in §5.2, has to compensate for the energy waste resulting from using peak-optimal active servers at non peak times during each epoch. Since TAWD algorithm should run in every slot (a few seconds), it is not practical to use high-complexity solutions. Therefore, LRH and CPLRH ranking are used to shift workload to the most thermally efficient servers in the homogeneous and heterogeneous cases, respectively, and with the following goals: (i) reducing the cooling energy by utilizing thermally efficient active servers, and (ii) reducing the computing energy by operating the active server at high energy efficiency utilization level.

In *load balancing*, the workload is distributed among servers such that their utilization levels are *equalized* (i.e., *balanced*). In contrast, in TAWD servers with low LRH are more likely to be utilized close to their maximum affordable workload (while satisfying the performance constraint, Eq. 16) and servers with higher LRH are likely to be utilized to a lesser degree. Consider an example of 50 homogeneous servers with load balanced 30% average utilization. Also assume that there is no performance-oriented threshold on the utilization level of servers, i.e., they can reach 100% utilization. Then, applying TAWD, $\frac{50 \times 30\%}{100\%} = 15$ servers with the lowest LRH get 100% utilization each and the rest remain idle. Note that servers are usually more energy efficient at high utilization levels (see Fig. 1(c)). Hence, TAWD achieves both the aforementioned goals whereas MinHR and inverse-temperature achieve only the first goal.

Function TAWD in Alg. 1 illustrates how λ_{jm} for each server is calculated. At the beginning of each slot, the average arrival rate, λ_m, is predicted using some method (in the simulation section we use Kalman Filtering), then the maximum affordable workload for each server, λ_{jm}^{thres}, is calculated using Eq. 13, and finally the workload arrival rate, λ_{jm}, is calculated using LRH.

Table II. Server power profiles.

Type	Server model	ω	α	IPR	LDR	Type	Server model	ω	α	IPR	LDR
1	Ideal	0	330	0	0	4	PowerEdge1855	182	50	0.66	0
2	IBM 350M2	100	200	0.33	0	5	IBM dx360M3	100	230	0.3	0.3
3	PowerEdge1955	242	90	0.73	0						

Compared to load balancing, TAWD exhibits a significantly increased likelihood of performance violation because it utilizes many servers close to their maximum affordable workload, λ_{jm}^{thres}, which can be breached due to prediction errors and the stochastic nature of workload arrival and its dispatching. Therefore, the dispatching algorithm UPONJOBARRIVAL in Alg. 1 is configured so that it dispatches workload according to TAWD as long as the observed arrival rate is less than the predicted arrival rate (meaning that TAWD does not overload thermally efficient servers), and it dispatches jobs according to load balancing otherwise.

7. EVALUATION

We evaluate our approaches through simulations to illustrate the cooling-computing power trade-off and to estimate the energy saving benefits of TASP and TAWD in a typical data center. The evaluation section is organized in four parts. First, the conditions where a workload consolidation causes cooling-computing power trade-off is illustrated and its implication on the performance of server provisioning and workload distribution approaches (TAWD, TASP-MIP, PT-d+, and CPSP) are evaluated under constant workload rates, variable PUE and energy proportionality values (§7.2). §7.3 and §7.4 evaluate the energy efficiency of the schemes under realistic dynamic workloads using linear power model and nonlinear power model, respectively. Finally, the possible performance violation of the schemes is evaluated in §7.5.

7.1. Data center Profiles

The data center profile consists of the thermal profile, i.e., the heat recirculation matrix, and the servers' power and performance models.

The heat recirculation matrix used in this simulation is derived from a CFD model of the ASU HPCI data center with physical dimensions $9.6\,\mathrm{m}\times 8.4\,\mathrm{m}\times 3.6\,\mathrm{m}$, and having two rows of industry standard 42U racks arranged in a typical cold aisle and hot aisle layout [Tang et al. 2008; Mukherjee et al. 2009]. The cold air is supplied by one computer room air conditioner, with a flow rate of $8\,\mathrm{m}^3/\mathrm{s}$. The cold air rises from the raised floor plenum through vent tiles, and the exhausted hot air returns to the air conditioner through ceiling vent tiles. There are ten racks and each rack is equipped with five 7U (12.25 inch) 10-server chassis. We use the following CoP function $CoP(T_{sup}) = 0.0068T_{sup}^2 + 0.0008T_{sup} + 0.458$ [Moore et al. 2005]. Further, we use different power profiles of servers as listed in Table II. For all data center configurations, the same heat recirculation and number of servers is used. Note that, the heat recirculation model is irrelevant to the energy proportionality of servers. Power profile Type 1 in Table II is an example of an ideal energy-proportional server. The power profile characteristics of Types 2 and 5 are derived from `www.spec.org/power_ssj2008/results/`. Finally, power profiles of Types 3 and 4 are derived from the measurements for some I/O intensive jobs [Mukherjee et al. 2009]. These profiles are appropriate for HTTP-like Internet requests.

To model the utilization of servers, we assume a reference service rate μ^{ref} of around 4000 requests per second (for a server with two cores) with a *reference request size* on the data of 512 bytes. The reference response time is set to 6 ms [Chen et al. 2005]. Our simulation environment is developed using MATLAB 2009. We use GNU Linear

Programming Kit (*GLPK*) to solve MIP in Eq. 15 and use *fminimax* solver provided by MATLAB for solving TASP-MiniMax. To do a fair comparison for power saving, PT-d+ is configured to use the same methodology to determine the minimum active server set size as in TASP (i.e., n). To evaluate the performance violation effect of TASP, we compare it with SurgeGuard [Faraz and Vijaykumar 2010] in §7.5. Algorithms CPSP, PT-d+ and No Server Provisioning, namely NoSP, are used as baseline algorithms to evaluate TASP and TAWD schemes. Specifically for CPSP, when it has to choose among servers of equal power efficiency, the ones with largest LRH are chosen (i.e., using Lemma 2.8). This configuration choice is made to show how badly CPSP can perform compared to thermal aware server provisioning (note that a thermal aware CPSP would be equivalent to TASP-LRH and TASP-CPLRH). We use various combinations of server provisioning and workload distribution schemes, and whenever TAWD is used, it is denoted as a subscript, e.g., $\text{NoSP}_{\text{TAWD}}$ and $\text{TASP-LRH}_{\text{TAWD}}$, when load balancing is used the subscript is empty, e.g., NoSP.

7.1.1. Workload model. We use two web traces: (i) a synthetic trace based on e-commerce web benchmark suite of SPECweb2009 (`www.spec.org`) and 1998 FIFA World Cup web trace [Arlitt and Jin 2000], and (ii) Microsoft Hotmail web trace [Thereska et al. 2009] of one week. To get traces from SPECweb2009, we set up an Apache web server on a Dual-core Intel Xeon LV system and an additional system as the SPECweb traffic-generating client and run SPECweb for 24 hours. SPECweb is apt at exhibiting short term traffic fluctuation. To make it exhibit long term traffic fluctuation, the parameter *"SIMULTANEOUS_SESSION"* needs to be dynamically adjusted. For that purpose, in the 24-hour SPECweb experiment, the long term traffic fluctuation is taken from 24 hours of FIFA's web log and *SIMULTANEOUS_SESSION* is adjusted accordingly in each 30-minute epoch. We also use one week of Microsoft HotMail web trace [Thereska et al. 2009]. In both traces, the traffic intensity (i.e., λ) is scaled up to match the capacity of the simulated data center. The coefficient of variations of the workload traces for SPECweb and HotMail is 1.5, and 0.95, respectively. Hence, they are representative of high and low fluctuating workloads.

Two Kalman filters, the first for TAWD and the second for TASP, are trained for five slots and epochs, respectively. Subsequently, they respectively predict the average rates at each slots (λ_m) and at each epoch (λ). To estimate the peak arrival rate over an epoch (i.e., λ^{peak}), we used $\lambda^{peak} = \beta_1\lambda + \beta_0$ to estimate the peak arrival rate using the average predicted rate. We estimate β_0 and β_1 using regression on the last four days of the first week for the HotMail trace and a portion of the World Cup trace. The regressed values are: for SPECweb $\lambda^{peak} = 2.97\lambda + 54973$, and for HotMail $\lambda^{peak} = 1.522\lambda + 1000000$.

7.2. Cooling-computing power trade-off and schemes' performance under various energy efficiency of data centers

This section investigates under what conditions of the data center utilization, the PUE and the servers' energy proportionality, cooling-computing power trade-off occurs and what implication the trade-off has on the energy savings of the schemes. To this end, we simulate nine cases of homogeneous data centers, organized into three groups. The three groups use server Types 1, 2, and 3 respectively (Table II). In each group, we use a different PUE (we scale the heat recirculation matrix by half, one, and 1.5 to get different PUEs). PUE values shown in the figures reflect no server provisioning and when all the servers are utilized at 50%. In our simulation, the PUE varies from 1.2 to 2.25, a range that covers both modern and old data centers. We run simulations under a constant workload rate such that the utilization of the entire data center varies from 10% to 100% in 10% increment. Results are shown in Figs. 6-9.

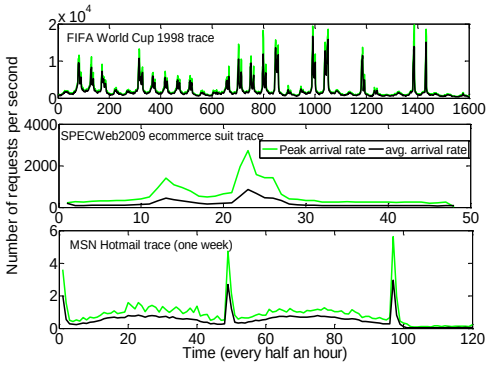

Fig. 5. Average and peak arrival rate in every half an hour over some realistic web traces.

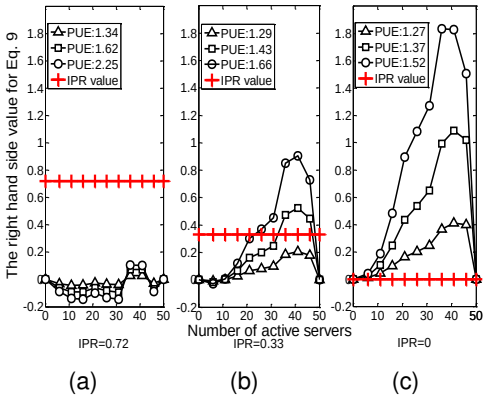

Fig. 6. The right hand side value of Eq. 5 over number of required active servers (i.e., n) for various PUE and IPRs, and $N = 50$ (according to Lemma 2.5, the cooling-computing power trade-off exists when y-axis value is greater than IPR value).

7.2.1. Cooling-computing power trade-off. Fig. 6 illustrates the conditions of Lemma 2.5 under which there exists a balanced workload consolidation (e.g., CPSP) that causes cooling-computing power trade-off. The right hand side value of Eq. 5 in Lemma 2.5 (i.e., ε_2) is plotted where the active server set is calculated using Lemma. 2.8 . According to Lemma 2.5 the trade-off exists if IPR $< \varepsilon_2$. The figure shows that the condition for the occurrence of the trade-off does not hold for high IPR servers, i.e., server Type 3 (see Fig. 6(a)), holds when PUE is high and servers have low IPR, i.e., server Type 2 (see Fig. 6(b)), and always holds for ideal energy-proportional servers (see Fig. 6(c)). The results are compatible with arguments in §2.3 which says low IPR and high PUE favors the occurrence of cooling-computing power trade-off. The next section shows the implication of the trade-off occurrence, on the energy saving of various server provisioning schemes. Figs. 7, 8, and 10 show the energy consumption of the schemes under the data center energy efficiency configuration corresponding to Fig. 6(a), Fig. 6(b), and Fig. 6(c), respectively.

7.2.2. Energy efficiency of the schemes. **Energy efficiency of the schemes for non energy-proportional servers (server Type 3):** Fig. 7, demonstrating the energy consumption of schemes under the data center energy efficiency configurations corresponding to Fig. 6(a), shows two trends: (i) server provisioning schemes yield tremendous power savings compared to NoSP, including the thermally oblivious CPSP, due to removal of idle power (compare NoSP and $\text{NoSP}_{\text{TAWD}}$ to CPSP, PT-d+ and TASP-MIP); (ii) higher PUE values magnify the savings of thermal awareness (compare NoSP to $\text{NoSP}_{\text{TAWD}}$, and CPSP to PT-d+ and TASP-MIP). Note that absence of the cooling-computing power trade-off (as illustrated by Fig. 6(a)) is sufficient condition for CPSP to always outperform NoSP. On the other hand, TASP-MIP yields more power savings than PT-d+ due to its optimality. Further, under constant workload, the performance of PT-d+ and TASP-LRH are almost the same, because both of them minimize the total heat recirculation. For plot legibility, we omit the plot for TASP-LRH. Also observe that power savings of TASP-MIP and PT-d+ compared to CPSP are maximized within the range of 40-70% utilization; since in this range of utilization, there are more

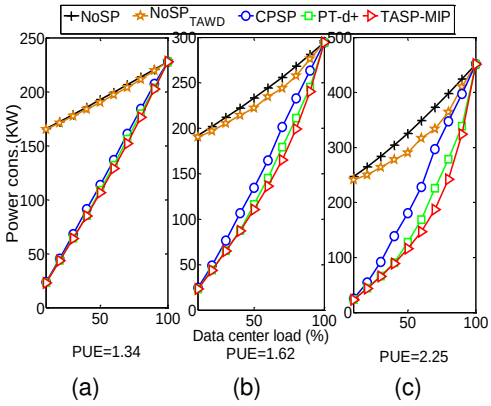

Fig. 7. Power consumption of schemes for different PUEs and under a homogeneous data center with server Type 3.

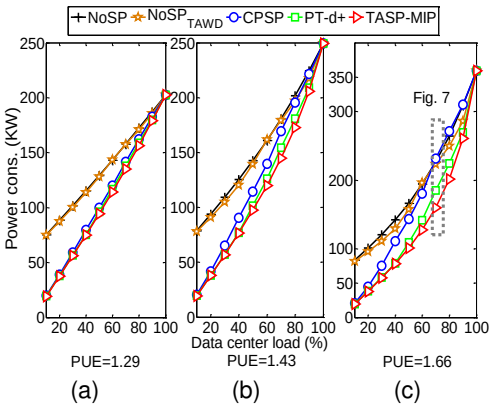

Fig. 8. Power consumption of schemes for different PUEs and under a homogeneous data center with server Type 2 (the box on the plot (c) shows the data center load condition for Fig. 9).

combination choices of servers, and the difference between a thermally efficient and a thermally inefficient active server set is maximized.

Energy efficiency of the schemes for modern low IPR servers (server Type 2): Fig. 8, demonstrating the energy consumption of schemes under the data center energy efficiency configurations corresponding to Fig. 6(b), shows that the power savings in this case have the same two trends of the previous case (compare Fig. 7 to 8) albeit at lower magnitude, except that for high PUE (PUE=1.66) and high utilization load (beyond 60%) CPSP consumes marginally more power than NoSP (see Fig. 8(c)). Under this scenario, cooling-computing power trade-off exists as illustrated by Fig. 6(b). The cooling-computing power trade-off increases the total power consumption of CPSP compared to NoSP which is consistent with arguments in §2.3. This behavior is shown in detail in Fig. 9 at 70% data center utilization, specifically, CPSP reduces idle power and saves 11% (20 W) compared to NoSP, however it creates hot spots which cause T_{sup} to decrease from 9.6 °C for NoSP to 7.2 °C; this T_{sup} decrease causes the cooling power to increase by 17% (25 W) which translates to a net increase of total power by 2% (5 W) compared to NoSP. Under the same condition of data center load and PUE, TASP-MIP and PT-d+ save power around 30% and 18%, respectively, compared to NoSP.

Energy efficiency of the schemes for ideal energy-proportional servers (server Type 1): Fig. 10 shows that when servers are ideally energy proportional, server provisioning (suspending servers) yields no power savings (compare $\text{NoSP}_{\text{TAWD}}$ and PT-d+), in contrast to the trends in the previous two paragraphs. Of particular interest is CPSP, which yields higher total power than NoSP for all three PUE cases. Actually, for any PUE$>$1, CPSP would yield higher total power than NoSP, because it does not decrease the computing power yet creates hot spots due to workload consolidation (recall that CPSP is configured to select thermally inefficient servers). In other words, according to Lemma 2.5, CPSP in this case causes cooling-computing power trade-off for all data center load and PUE conditions as illustrated by Fig. 6(c). In contrast to CPSP, TASP-MIP saves power for all three PUE cases, and its saving increases up to 13% with increasing PUE (Fig. 10(c)). The savings of TASP-MIP come from consolidating workload to the thermally efficient servers and not from suspending servers. Further, PT-d+ has marginal power savings due to its non-optimality.

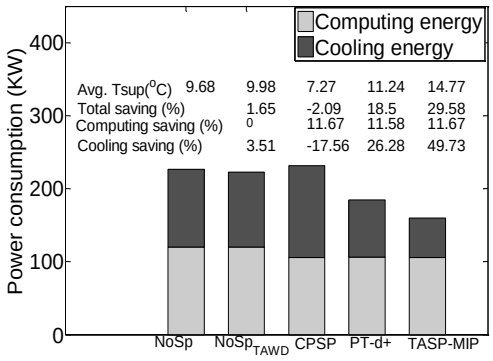

Fig. 9. Power saving of schemes under a homogeneous data center with server Type 2, load of 70%, and PUE=1.66. (see box in Fig. 8)

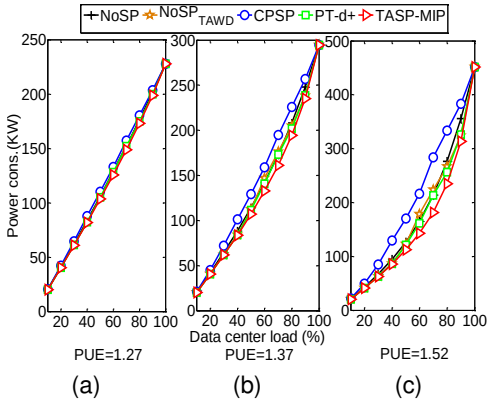

Fig. 10. Power consumption of schemes for different PUEs and under a homogeneous data center with server Type 1.

Table III. Power saving of TASP and CPSP with respect to NoSP under different data centers' energy efficiency.

		Server's energy proportionality case								
		3 (non energy proportional)			2 (modern servers)			1 (ideal energy proportional)		
PUE	load*	trade-off**	TASP	CPSP	trade-off	TASP	CPSP	trade-off	TASP	CPSP
high	high	false	20.1	6.1	true	11.9	-0.9	true	7.7	-7.8
	mid	false	60.1	37.5	true	36.0	8.7	true	13.1	-33.2
	low	false	83.3	78.4	false	62.3	53.7	true	5.5	-19.5
mid	high	false	13.0	6.3	true	7.1	1.4	true	3.6	-36
	mid	false	47.8	36.5	false	28.1	15.7	true	6.3	-13.0
	low	false	78.3	75.3	false	59.4	54.9	true	3.3	-9.0
low	high	false	8.4	6.4	false	4.3	2.6	true	1.1	-0.9
	mid	false	39.8	35.9	false	23.2	19.1	true	2.3	-3.8
	low	false	74.8	73.5	false	57.3	55.4	true	1.5	-3.3

* high load: $70\% <$load $\leqslant 100\%$, mid load: $30\% <$load $\leqslant 70\%$, and low load: load $\leqslant 30\%$.
** The true value for "trade-off" means that CPSP increases cooling power compared to NoSP.
Color coding: actual waste | $0\% \leqslant$ savings $< 5\%$ | $5\% \leqslant$ savings $< 20\%$ | $20\% \leqslant$ savings

7.2.3. Discussion on results. The simulation above supports the analytical arguments in §2.3 that non thermal aware server provisioning (i.e., CPSP) is *insufficient* and even *inefficient* (under some energy proportionality and PUE of data centers) in increasing the energy efficiency of a data center. The necessary conditions for inefficiency of CPSP is explained by Lemma 2.5 which are easy to test. The summary of simulation results shown in Table III indicates that CPSP can increase the total power consumption compared to NoSP when servers are moderately or ideally energy-proportional (i.e., for the case: high load and high PUE accentuate this tendency). This happens due to the cooling-computing power trade-off under ideal and modern energy-proportional servers as explained by Lemma 2.5 and illustrated by Fig. 6. Under modern servers, CPSP outperforms NoSP most of the time, however its savings is much less than that of TASP (see Table III). This means that although data centers improve their PUE and the energy proportionality of their servers, making the occurrence of the cooling-computing power trade-off unlikely, considerable energy savings will not be achieved unless thermal aware server provisioning schemes are used. The results show that in

the same conditions where CPSP underperforms with respect to NoSP, TASP yields considerable savings.

To summarize, servers are increasingly exhibiting ideal energy proportionality (IPR and LDR $\rightarrow 0$) and data centers' PUE are improving (PUE $\rightarrow 1$). Consequently, our results indicate CPSP will no longer be efficient (see case: ideal energy-proportional and low and mid PUE in Table III) or sufficient (see case: modern servers and low and mid PUE in Table III). On the other hand, TASP can address this inefficiency.

7.3. Performance of TASP and TAWD under realistic workload

To evaluate TASP and TAWD under realistic workloads, we setup one homogeneous and one heterogeneous data center. The servers for heterogeneous data center are configured similar to ASU HPCI data center (20 chassis of server Type 4 and 30 chassis of server Type 3), and the servers for the homogeneous case are configured according to the modern server Type 2. Distinct simulations are performed for 24 hours of both SPECweb and HotMail traces. The results presented in this subsection reflect the summed energy consumptions from the separate experiments on the two workload model.

Energy efficiency of TASP and TAWD for *homogeneous* data center case: Fig. 11, showing the percentile energy saving of various server provisioning scenarios under homogeneous data center, indicates that TASP schemes can significantly save energy compared to no server provisioning. It can be surmised from the figure that CPSP creates higher temperature hot spots compared to NoSP (since it demands cooler supplied temperatures), yet it decreases both the computing and the cooling energy compared to NoSP. Hence, it decreases the total heat produced (i.e., the total computing power). Also, in contrast to fixed workload scenarios where PT-d+ always performs better than CPSP, the total energy saving of PT-d+ is lower than CPSP. However, PT-d+'s cooling energy saving is more than CPSP. Since, under a fluctuating workload (as opposed to fixed workload), the trade-off may not be adequately computed by PT-d+ dynamically. Because, the variation of cooling energy can also be attributed to variation in workload and not just to adjusting of the active server set. Accordingly, our results show that under the same active server set size policy, and the same workload (HotMail) the average active server set size of Pt-d+ is 37, whereas it is 31 for CPSP and other TASP algorithms.

Except for PT-d+, all other thermal aware server provisioning schemes (i.e., TASP-LRH, TASP-MIP and TASP-LRH$_{\text{TAWD}}$) decrease the cooling demand. Interestingly, the T_{sup} of both TASPs as well as Pt-d+ are higher than T_{sup} of even NoSP. This means that thermal aware server consolidation not only can prevent creation of hot spots, but also can alleviate the existing hot spots through assigning workload to thermally efficient servers.

The optimal algorithm, i.e., TASP-MIP saves around 20% energy with respect to CPSP, and TASP-LRH saves 14% energy compared to CPSP. This result indicates that the performance of LRH is higher than what can be proved theoretically. As mentioned in §6.1.1, LRH is a N-approximation algorithm (see Eq. 15), where N is the number of servers. However, according to our numerical analysis, the approximation ratio of LRH is 1.18 under the aforementioned simulation setup. Notice, all extra energy savings of TASP variants compared to CPSP comes from thermal aware sever provisioning and decreasing the data centers' cooling demand. It can be seen in the figure that the supplied temperature of TASP approaches increases upto 4 °C compared to CPSP, whereas their computing energy savings is same as CPSP's. Further, TAWD under no server provisioning yields marginal energy savings compared to NoSP. Furthermore, incorporating TAWD to the TASP-LRH marginally improves its energy saving benefit.

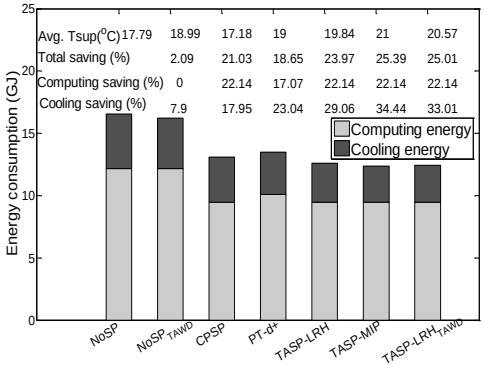

Fig. 11. Energy consumption of schemes under a homogeneous data center with server Type 2. $|S'| = 34$ for TASP.

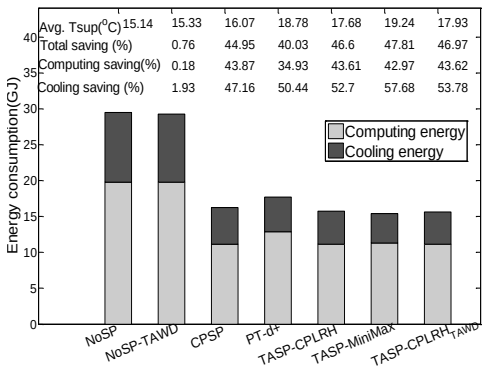

Fig. 12. Energy consumption of schemes under a heterogeneous data center with server Type 3 and 4. $|S'| = 25$ for TASP.

Energy efficiency of TASP and TAWD for *heterogeneous* data center case: similar to the homogeneous data center case, CPSP saves significant energy compared to NoSP in the heterogeneous case (see Fig. 12). In fact, its saving is much higher than the homogeneous case, since it chooses power efficient servers. It also demands higher T_{sup} than NoSP. The reason is in our data center configuration, power efficient servers accidentally were more thermal efficient. TASP schemes save upto 22% cooling energy compared to CPSP, which translates into 7% energy savings total. The results also indicates the higher performance of TASP schemes and even CPSP compared to PT-d+. However, PT-d+ causes a cooler supplied temperature than TASP-LRH and even less than TASP-LRH$_{\text{TAWD}}$. This is because PT-d+ uses thermal oriented load balancing using MinHR [Moore et al. 2005], which is more thermal efficient than TAWD. In the next section we see how performance of TAWD increases under nonlinear power curves of systems. Another important result in the figure is that heuristic TASP (i.e., TASP-CPLRH) competently saves energy compared to TASP-MiniMax solution. However, the results show that the performance of TASP schemes decrease compared to the homogeneous data center case (see Fig. 12). This refers to the size of the active server set. As shown in §7.2, the benefit of TASP schemes maximized when data center is half utilized (see Fig. 8). Since the average size of the active server set in the heterogeneous case is around 25 (out of 30 of high power efficient servers), TASP's energy savings is lower than the homogeneous data center case where its average active server set size is 34 (out of 50).

7.3.1. Discussion on results. Results indicate that heuristic TASP schemes can competently save energy with respect to optimal solutions. The results also show that the dynamic management of the cooling-computing power trade-off (i.e., the PT-d+ solution) which can be successfully manged under constant workloads, can not be adequately performed under realistic workloads. The reason is, under realistic workloads, the variation of cooling power can be because of variation in workload rather than adjusting the active server set size.

7.4. Performance of TASP and TAWD under nonlinear power curves

To study the effect of the nonlinearity of the power curve on the power consumption, we use a state of the art server, i.e., server Type 5, whose power curve exhibit nonlinearity. We fit the power curve to the polynomial function $f(u) = -4301u^6 + 15419u^5 -$

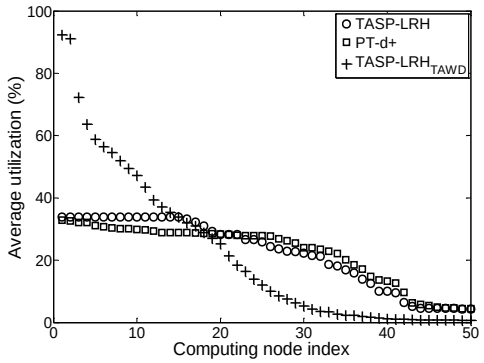

Fig. 13. Average utilization for computing nodes indexed by LRH.

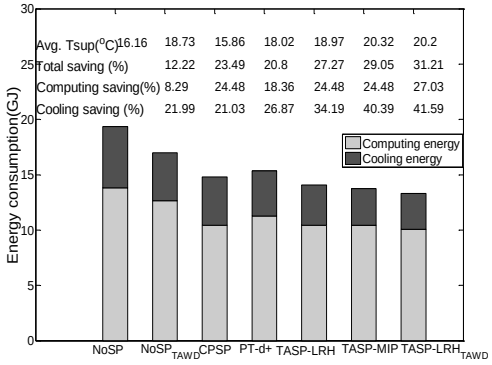

Fig. 14. Energy consumption of schemes under a homogeneous data center with server Type 5.

$21798u^4 + 15326u^3 - 5490u^2 + 1092u + 92.8$. We setup a homogeneous data center case of fifty chassis of this server and calculate the energy consumption of TASP and TAWD schemes under SPECweb and HotMail workload traces. The results shown in Fig. 14 interestingly show that under no server provisioning TAWD decreases the total energy by 12% which is much higher than the case where systems are assumed to have linear power curve (see Fig. 11). This saving is the outcome of consolidating the workload on thermal efficient servers and allowing them to operate at higher utilization where their computing energy efficiency is higher. Similarly, the energy saving benefit of TASP-LRH$_{\text{TAWD}}$ increases. The reason is, TAWD portion of TASP-LRH$_{\text{TAWD}}$ shifts the workload toward the more energy efficient servers and less to the other active servers compared load balancing scheme used in TASP-LRH (e.g., see Fig. 13 where active servers' average utilization is 35% by using TASP-LRH whereas average utilization for energy efficient servers is 80% by using TASP-LRH$_{\text{TAWD}}$). Since PT-d+ uses thermally oriented load balancing (i.e., MinHR based workload distribution), its energy savings is much lower than TASP-LRH$_{\text{TAWD}}$ which make thermally efficient servers to operate at their high energy efficiency level. Note that consolidation workload on the thermally efficient servers mitigates the hot spots. The average utilization of thermal efficient computing nodes in PT-d+ as shown in Fig. 13 are a little lower than LRH because the PT-d+' active server set size is larger than TASP-LRH. If active server set size of both PT-d+ and TASP-LRH would be the same, the average utilization of the most thermal efficient servers under PT-d+ would be higher than their average utilization under TASP-LRH.

7.4.1. Discussion on results. A significant observation is that non linear energy proportionality (observed in power-utilization curve of modern servers) can help in the savings of TASP and TAWD. Provisioning can consolidate the workload to fewer servers, thus increasing the per-server utilization with minimal increase in the energy consumption. Note that since consolidation of workload is applied on thermal efficient servers, hot spots are unlikely to occur and the total energy is always minimized.

7.5. Performance violations of TASP

This section evaluates TASP for predicting the minimum active server set size and its impact on performance violations, compared to SurgeGuard [Faraz and Vijaykumar 2010] as described in §6.1.4. Since, the average workload prediction scheme is not

Table IV. Performance Violations of TASP and SurgeGuard

Trace	Method	server-reserve*	$\|S'\|$	$\bar{t}_{res}$(ms)	Perf. viol.(%)	violated $\bar{t}_{res}$(ms)	saving**
HotMail	TASP	0	31	5.5	6.18	6.7	24.96
	SurgeGuard	5	25	6.1	40	6.7	22.2
	SurgeGuard	10	30	5.7	10.23	7.4	12.57
SPECweb	TASP	0	15	4.8	2.6	10.1	60
	SurgeGuard	5	12	5.1	8.02	9.5	59.8
	SurgeGuard	10	17	4.8	3.3	10.9	47.6

* *server-reserve* is expressed as the number of servers in reserve.
** energy saving percentage of TASP-LRH under the corresponding "method" compared to NoSP.

available for SurgeGuard we use Kalman filtering to predict the workload arrival rate. The performance violation is measured as the percentage of slots where the average response time of servers goes above the reference response time. Results for TASP-LRH with HotMail and SPECweb workloads shown in Table IV indicate that for almost the same active server set size, TASP saves more energy and yields much lower performance violations than SurgeGuard. The reason is TASP leverages the correlation between the peak and the average arrival rate during an epoch, whereas SurgeGuard uses the same static value of over provisioning (i.e., *server-reserve* for all epochs). Therefore, sometimes it underprovisions and sometimes it overprovisions. Even under TASP performance violation still happen due to unpredictable spikes in the workload. We conjunct that These violations can be predominantly eliminated by using a hybrid approach of proactive and reactive server provisioning (e.g., quickly activating servers whenever an unpredicted spike is observed). We leave the investigations of such a hybrid scheme for our future work.

8. DISCUSSION

In this paper, we presented TACOMA, an integrated management architecture suited for Internet Data Centers (IDCs). TACOMA distinguishes two functionalities: (i) deciding on which servers are needed during a small period to serve the predicted workload, and (ii) how to dynamically distribute the workload among the selected resources as it arrives. The first functionality is known as *server provisioning*, which is largely equivalent to *dynamic consolidation plus power management*. The second functionality is *workload distribution* (or *dispatching*). With these functionalities, TACOMA aims to infuse an elastic behavior to IDCs, providing a way to adjust the energy consumption to the given workload at the time. The achievable energy savings of TACOMA come from the synergy between predictive modeling, proactivity and the ability to individually shut down unused resources and utilize the needed ones near their peak efficiency. This includes both computing (i.e., servers) and non computing resources (i.e., CRACs).

TACOMA represents the proposition that holistic, integrated management of workload, computing and non computing resources can achieve several synergies, can address many energy-saving issues, and can perform optimized decisions. There are researchers that consider a data center to be a computer [Patterson ; Hoelzle and 2009]. As there is a strong correspondence of a data center to a desktop computer, there is an associated correspondence of TACOMA (and its likes) to a modern operating system (OS) (see Fig. 15): in collaboration with the firmware, the OS takes the newly created processes and assigns them to CPUs and to CPU slots, and decides on the fan speed and on the power state of the components. An OS-equivalent management software of a data center, much like TACOMA, takes the submitted workload (or jobs) and decides how to distribute them among the servers and perform power management of unused resources. For that reason, an OS-equivalent software would require the proper APIs and access hooks to manage the components of the data center, both in cyber (i.e., work-

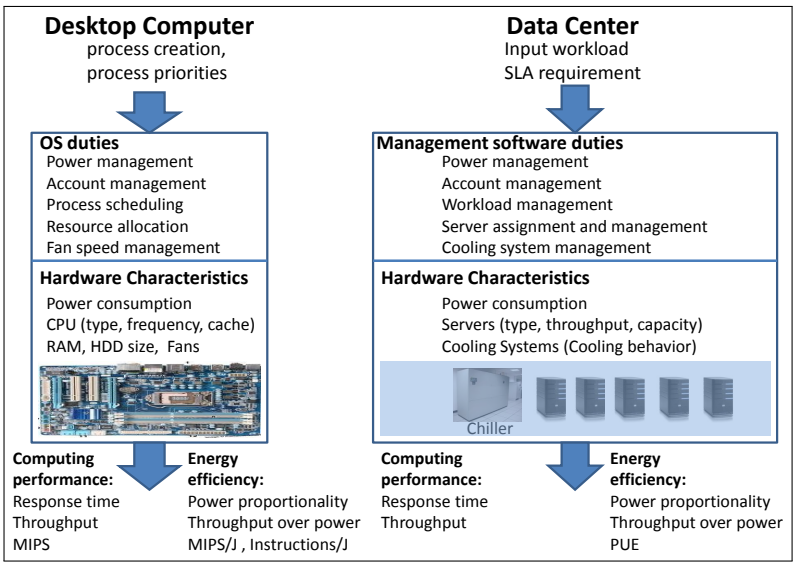

Fig. 15. Desktop computer system versus data center.

load and processing) and physical (i.e., power and cooling) aspects—a cyber-physical system needs a cyber-physical OS. The rest of this section discusses some some implementation issues and the usefulness of TACOMA.

8.1. Challenging issues

TACOMA is sensitive to the *predictive accuracy* of the models, on the *distributed* nature of data centers, and on the *ability* to sense and implement its functionality.

(1) **Predictive accuracy of models**. TACOMA heavily relies on several models and methods, including traffic prediction, performance estimation, heat recirculation modeling and power consumption modeling. As TACOMA needs to make regular and frequent decisions, the models and prediction methods have to be very lightweight. The models presented in this paper (e.g. Kalman filtering for traffic average, linear scaling of average to estimate mean, use of heat recirculation matrix to compute the temperature distribution at the servers' air inlet) have been carefully selected to meet the accuracy and speed objectives. Nevertheless, if TACOMA is viewed as an abstract architecture, it is easy to see that several models can be flexibly replaced with others that meet stricter criteria (on either time or accuracy).
(2) **Distributed nature of IDCs**: As presented, TACOMA considers the individual effect of each server to the power consumption of the data center. It is expected that as the server count increases, the computation of the solutions will take longer. For that reason, the heuristics have low polynomial complexity and can still compute efficiently even with large number of servers (see the next subsection). Moreover, data centers may have multiple CRAC units. This study accounts for a data center cluster which is cooled by one CRAC. We believe that the results are also applicable to data centers with multiple CRACs as long as the interference effect between two CRACs is negligible. A comprehensive study of a cloud consisting of multiple data centers [Abbasi et al. 2011], where each data center is allowed to have multiple CRACs, is left as future work.

(3) **Implications on implementation**. These implications are mainly on the impact of frequent switching servers on and off, and on whether there are multiple CRACs in a data center, or if there are multiple data centers to be managed together.
   — *Feasibility of sensing and actuating*: TACOMA requires the ability to sense the various cyber parameters (e.g., workload intensity and utilization levels) and physical parameters (e.g., temperatures and power consumption), and also to actuate the various decisions, e.g., to power down certain servers. Modern computing servers are equiped with built-in temperature and power sensors, and provide control interfaces, e.g., ACPI (`http://www.acpi.info/` and IPMI (`http://www.intel.com/design/servers/ipmi/index.htm`); modern CRACs have optional modules for remove monitoring and control over the network (network thermostats); modern power distribution units (PDUs) have similar networked capabilities for individual monitoring and control of their power plugs.
   — *Cost of switching*: Generally, suspending the servers can cause overhead in two ways: (i) affect the response time of workload due to resume time, and (ii) incur extra power consumption due to frequent switching. TACOMA is a proactive scheme that activates the required servers before they are needed, thusly eliminating the performance concerns. Switching a server is an action that happens once in an epoch ($\sim$1hr); modern systems are manufactured to support "deep sleep" modes that can go back to an active state rather quickly without adversely affecting the extected remaining lifetime of the equipment.
   — *Compatibility with performance metrics*: The performance (QoS) of IDCs is typically defined as some metric on delay, e.g. 95^{th}-percentile delay. We believe that the delay metric used in this paper, i.e., the expected delay at peak traffic, is compatible with the ones widely used. This is because traffic peaks account for less than 5% of the traffic, thus the presented optimization satisfies the 95th-percentile. Moreover, it is possible to introduce performance margins to the active server set calculation in order to satisfy arbitrary performance constraints.

8.2. TACOMA's value

TACOMA's value is governed by three criteria: *effectiveness*, *usability* and *validity*.

(1) **Usability**. Much of the usability has to do with the *implementation limitations*, as discussed above, and with TACOMA's *low run time*. The simulations were performed on MATLAB 2009b running on a 2.8GHZ Pentium D, 2MB cache and 2GB RAM with MS Windows 7. For a simulated data center of 50 servers, all algorithms took less than a second each to run and decide on either the active server set, for TASP variants, or the workload distribution, for TAWD variants. In preliminary experiments for a simulated data center of 500 servers, minimax took half an hour, where as all the other heuristics still took less than a second each to run.
(2) **Validity**. To show that TACOMA is a valid approach, we used one real web trace (hotmail) and a trace from a renown web trace generator (SPECWeb2009). The paper presents results from simulation-based experiments, where TACOMA manages to yield considerable energy savings. For further validation using not only real traces but also actual hardware and software, we plan to test TACOMA on BlueCenter, a small data center which is part of the BlueTool[9] [Gupta et al. 2012], a project with the purpose to provide an infrastructure for research on data centers.
(3) **Effectiveness**. TACOMA is effective if it is *accurate* for realistic workloads and if it *yields savings* for realistic hardware parameters.

[9]http://impact.asu.edu/BlueTool/

—*Accuracy under realistic workloads*: For the given realistic workloads, the prediction errors of averages and peaks are very low (see §4.3). Traffic prediction errors may cause overloading of servers (i.e., more requests arrive than what was expected). In case of overloading, TAWD switches to load balancing for that overloaded slot. Our simulations show that TAWD reverts to load balancing for only 0.5% of the times for the SPECweb2009 trace (that is 44 slots out of 8640), and 0.04% of the times for the Hotmail trace (that is 4 slots out of 8640).
—*Savings under realistic hardware parameters*: As we saw in §2.3, the effectiveness of TACOMA with respect to conventional methods depends on the thermal efficiency of the data center and on the energy proportionality of its servers. In Fig. 3, we observe that servers have low IPR, but no less than 0.2, and they have a varying LDR. Low IPRs favor the occurrence of cooling-computing power trade-off, whereas a varying LDR pushes toward heterogeneity of system's peak efficiency; both of these factors necessitate the use of thermal-aware management, even at low data center PUEs. For example, in Fig. 14, the data center has an initial PUE of about 1.4; this is yielded from dividing the leftmost column's height ($\sim$19 GJ) over the computing-only height ($\sim$14 GJ). For that kind of data center energy efficiency, TACOMA manages to yield over 31% energy savings, compared to 23% savings of CPSP.

9. CONCLUSIONS

This paper proposes *thermal aware server provisioning* (TASP) and *workload distribution* (TAWD) schemes integrated into a two-tier Thermal-Aware Computing and cOoling Management Architecture (TACOMA). The paper examines the performance of TACOMA over modern servers which tend to have *low idle power* and a *nonlinear power-utilization curve*. Our analytical study and simulation results show that non thermal-aware server provisioning schemes are *insufficient* and even *inefficient* in improving the energy efficiency of data centers with modern servers, as their impact on cooling power increases (due to the creation of hot spots by highly utilizing the active servers) may outweigh their impact on computing power decrease. This phenomenon is due to the cooling-computing power trade-off, which can manifest in air-cooled data centers. The paper provides theoretical work on the conditions that cause the trade-off to occur. TASP addresses inefficiencies such as this trade-off through formulating the server provisioning problem as a mixed integer program and a minimax optimization for homogeneous and heterogeneous data center cases, respectively. In both cases, the objective is to minimize the *sum of cooling and computing power*. Our results show that TASP solutions save upto 20% of the energy compared to previous non thermal aware and a newly published thermal aware server provisioning schemes. Further, the nonlinearity of the power-utilization curve of modern servers can be leveraged by the proposed TAWD to shift the workload dispatching toward the thermally and power efficient servers, and utilize them at their high energy efficiency level without creating hot spots. TASP and TAWD respect the performance constraints by considering the maximum affordable load of each server (according to a queuing model used) into the determination of the minimum number of required active servers servers, for TASP, and the determination of the workload distribution, for TAWD.

Acknowledgments

We thank the associate editor, Frederic Chong, and the anonymous reviewers for their insightful suggestions.

REFERENCES

ABBASI, Z., MUKHERJEE, T., VARSAMOPOULOS, G., AND GUPTA, S. K. S. 2011. DAHM: A green and dynamic web application hosting manager across geographically distributed data centers. *ACM Journal on emerging technology (JETC)*. Accepted for publication.

ABBASI, Z., VARSAMOPOULOS, G., AND GUPTA, S. K. S. 2010. Thermal aware server provisioning and workload distribution for internet data centers. In *ACM International Symposium on High Performance Distributed Computing (HPDC10)*. Chicago, IL, 130–141.

AISSI, H., BAZGAN, C., AND VANDERPOOTEN, D. 2006. Approximating min-max (regret) versions of some polynomial problems. In *Proceedings of the 12th International Computing and Combinatorics Conference COCOON*. 428–438.

ALLEN, A. O. 1990. *Probability, statistics and queuing theory with computer science applications.* Academic Press Inc.

ARLITT, M. AND JIN, T. 2000. A workload characterization study of the 1998 world cup web site. *Network, IEEE 14,* 3, 30 –37.

BARROSO, L. A. AND HOLZLE, U. 2007. The case for energy-proportional computing. *Computer 40*, 33–37.

BASH, C. AND FORMAN, G. 2007. Cool job allocation: Measuring the power savings of placing jobs at cooling-efficient locations in the data center. Tech. Rep. HPL-2007-62, HP Laboratories Palo Alto. August.

BOHRER, P., ELNOZAHY, E. N., KELLER, T., KISTLER, M., LEFURGY, C., MCDOWELL, C., AND RAJAMONY, R. 2002. The case for power management in web servers. 261–289.

CHASE, J., ANDERSON, D., THAKAR, P., VAHDAT, A., AND DOYLE, R. 2001. Managing energy and server resources in hosting centers. In *SOSP '01: Proceedings of the eighteenth ACM symposium on Operating systems principles*. ACM, New York, NY, USA, 103–116.

CHEN, G., HE, W., LIU, J., NATH, S., RIGAS, L., XIAO, L., AND ZHAO, F. 2008. Energy-aware server provisioning and load dispatching for connection-intensive internet services. In *NSDI'08: Proceedings of the 5th USENIX Symposium on Networked Systems Design and Implementation*. USENIX Association, Berkeley, CA, USA, 337–350.

CHEN, Y., DAS, A., QIN, W., SIVASUBRAMANIAM, A., WANG, Q., , AND GAUTAM, N. 2005. Managing server energy and operational costs in hosting centers. *SIGMETRICS Performance Evaluation Review 33,* 1, 303–314.

CHUN, B.-G., IANNACCONE, G., IANNACCONE, G., KATZ, R., LEE, G., AND NICCOLINI, L. 2010. An energy case for hybrid datacenters. *ACM SIGOPS Operating Systems Review 44,* 1, 76–80.

FARAZ, A. AND VIJAYKUMAR, T. 2010. Joint optimization of idle and cooling power in data centers while maintaining response time. *ACM SIGARCH Computer Architecture News 38,* 1, 243–256.

GUPTA, S. K. S., VARSAMOPOULOS, G., HAYWOOD, A., PHELAN, P., AND MUKHERJEE, T. 2012. *Handbook of Energy-Aware and Green Computing*. Number 45. Chapman and Hall/CRC, Chapter BlueTool: Using a computing systems research infrastructure tool to design and test green and sustainable data centers.

HOELZLE, U. AND , L. B. 2009. *The Datacenter as a Computer: An Introduction to the Design of Warehouse-Scale Machines* 1st Ed. Morgan and Claypool Publishers.

KRIOUKOV, A., MOHAN, P., ALSPAUGH, S., KEYS, L., CULLER, D., AND KATZ, R. 2010. NapSAC: design and implementation of a power-proportional web cluster. In *Proceedings of the first ACM SIGCOMM workshop on Green networking*. ACM, 15–22.

KUSIC, D., KEPHART, J. O., HANSON, J. E., KANDASAMY, N., AND JIANG, G. 2009. Power and performance management of virtualized computing environments via lookahead control. *Cluster Computing 12*, 1–15.

MONTGOMERY, D. C., PECK, E. A., AND VINING, G. G. 2001. *Introduction to Linear Regression Analysis*. John Wiley and Sons Inc.

MOORE, J., CHASE, J., AND RANGANATHAN, P. 2006. Weatherman: Automated, online, and predictive thermal mapping and management for data centers. In *IEEE International Conference on Autonomic Computing (ICAC)*. 155–164.

MOORE, J., CHASE, J., RANGANATHAN, P., AND SHARMA, R. 2005. Making scheduling "cool": temperature-aware workload placement in data centers. In *ATEC '05: Proceedings of the annual conference on USENIX Annual Technical Conference*. USENIX Association, Berkeley, CA, USA, 5–5.

MUKHERJEE, T., BANERJEE, A., VARSAMOPOULOS, G., GUPTA, S. K. S., AND RUNGTA, S. 2009. Spatio-temporal thermal-aware job scheduling to minimize energy consumption in virtualized heterogeneous data centers. *Computer Networks 53,* 17, 2888–2904.

PAROLINI, L., TOLIAZ, N., SINOPOLIY, B., AND KROGHY, B. H. 2010. A cyber-physical systems approach to energy management in data centers. In *ACM ICCPS '10*. Stockholm, Sweden, 168–177.

PATTERSON, D. A. The data center is the computer. *Communications of the ACM 51,* 1, 105.

RANGANATHAN, P., LEECH, P., IRWIN, D., AND CHASE, J. 2006. Ensemble-level power management for dense blade servers. In *Computer Architecture, ISCA '06. 33rd International Symposium on*. 66–77.

RAWSON, A., PFLEUGER, J., AND CADER, T. 2008. Green grid data center power efficiency metrics: PUE and DCIE.

SHARMA, R. K., BASH, C. E., PATEL, C. D., FRIEDRICH, R. J., AND CHASE, J. S. 2005. Balance of power: Dynamic thermal management for internet data centers. *IEEE Internet Computing*, 42–49.

TANG, Q., GUPTA, S. K. S., AND VARSAMOPOULOS, G. 2008. Energy-efficient thermal-aware task scheduling for homogeneous high-performance computing data centers: A cyber-physical approach. *IEEE Trans. Parallel Distrib. Syst. 19,* 11, 1458–1472.

THERESKA, E., DONNELLY, A., AND NARAYANAN, D. 2009. Sierra: a power-proportional, distributed storage system. Tech. rep., Microsoft Research.

TOLIA, N., WANG, Z., MARWAH, M., BASH, C., RANGANATHAN, P., AND ZHU, X. 2008. Delivering energy proportionality with non energy-proportional systems: optimizing the ensemble. In *Proceedings of the 2008 conference on Power aware computing and systems*. USENIX Association, 2–2.

TSIROGIANNIS, D., HARIZOPOULOS, S., AND SHAH, M. A. 2010. Analyzing the energy efficiency of a database server. In *Proceedings of the 2010 international conference on Management of data SIGMOD'10*. ACM, 231–242.

URGAONKAR, B., SHENOY, P., CHANDRAY, A., AND GOYAL, P. 2008. Agile dynamic provisioning of multi-tier internet applications. *ACM Transactions on Autonomous and Adaptive Systems (TAAS) 3,* 1, 1–39.

US DEPARTMENT OF ENERGY. 2010. Quick start guide to increase data center energy efficiency. Tech. rep., U.S. Department of Energy. September.

VARSAMOPOULOS, G., ABBASI, Z., AND GUPTA, S. K. S. 2010. Trends and effects of energy proportionality on server provisioning in data centers. In *International Conference on High performance Computing (HiPC2010)*. Goa, India, 1–11.

VARSAMOPOULOS, G. AND GUPTA, S. K. S. 2010. Energy proportionality and the future: Metrics and directions. In *Parallel Processing Workshops (ICPPW), 39th International Conference on*. IEEE, 461–467.

YU, G. AND KOUVELS, P. 1998. On min-max optimization of a collection of classical discrete optimization problems. *Optimization Theory and Applications 98(1)*, 221–242.

ZHOU, X., YANG, J., CHROBAK, M., AND ZHANG, Y. 2010. Performance-aware thermal management via task scheduling. *ACM Transactions on Architecture and Code Optimization (TACO) 7,* 1, 1–31.

A. NP-HARDNESS PROOF OF TASP PROBLEM FOR THE HOMOGENEOUS DATA CENTER

LEMMA A.1. *The TASP problem, i.e., Eq. 15, is NP-hard even if only two nodes are affected by heat recirculation of the data center's node : (i) all elements of matrix A are positive, (ii) A is an* $\{2 \times n\}$ *matrix, and finally (iii)* $\boldsymbol{x}$ *is a binary vector.*

PROOF. We reduce the 2-partition problem to TASP (similar to the NP-hardness proof of the min-max resource allocation problem [Yu and Kouvels 1998]). In the 2-partition problem, a set I containing positive integers ($y_i \in Z_+$) is given and the question is: is there a subset $I' \subseteq I$ such that $\sum_{i \in I'} y_i = \sum_{i \in II'} y_i$? 2-partition problem is known to be NP-hard even if the set I contains even numbers and partitioning is subject to equal cardinalities (i.e $|I'| = |I|/2$). Given a 2-partition problem, a TASP problem with A as $\{2 \times n\}$ matrix can be constructed as follows. let elements of A be:

$$a_{1j} = y_j$$
$$a_{2j} = \frac{2}{n} \sum_{i=1}^{n} y_i - y_j$$

Assume two partitioned subsets are restricted to have equal cardinalities. Thus exactly $n/2$ elements will be selected. If an optimal solution to the TASP problem, x^*, has:

$$x_i^* = 1 \;\; i \in I',$$
$$x_i^* = 0 \quad \text{Otherwise,}$$

and $|I'| = n/2$, then the total heat recirculated to node 1 is: $z_1 = \sum_{i \in I'} y_i$, and the total heat recirculated to node 2 is: $z_2 = n/2(2/n \sum_{i \in I} y_i) - \sum_{i \in I'} y_i = \sum_{i \in I/I'} y_i$. By definition $z_{TASP} = max\{z_1, z_2\}$. We conclude that there exist a 2-partition with $|I'| = |I|/2$ if and only if TASP has an objective value $z_{TASP} = \sum_{i \in I'} y_i = \sum_{i \in I} y_i/2$. □