

THERMAL MANAGEMENT CONSIDERATIONS FOR GEOGRAPHICALLY DISTRIBUTED COMPUTING INFRASTRUCTURES

**Amip Shah, Cullen Bash, Martin Arlitt, Yuan Chen,
Daniel Gmach, Ratnesh Sharma, Chandrakant Patel**
Hewlett Packard Laboratories
Palo Alto, California, USA

ABSTRACT

This paper discusses an approach for optimizing the infrastructure thermal performance related to a geographically distributed computing service. Beginning by modeling the total energy costs associated with cooling a distributed environment, the cooling efficiency of a service is evaluated by superposing the piecewise IT workloads that may be delivered from various locations. We find that the total service-level thermal performance can be distinct from the facility- or infrastructure-level thermal performance, which requires a different global thermal management strategy relative to that of single-site environments. The approach is illustrated for a hypothetical example wherein a service is delivered from three different data centers in geographically diverse locations. Depending on the workload characteristics, the optimal distribution of resources across the data centers varies; but through dynamic resource allocation, it becomes possible to support the same service at increased energy efficiencies.

INTRODUCTION

Enterprise data centers, large computer warehouses that power much of the infrastructure underlying the Internet, can consume significant amounts of energy. For example, it has been estimated that in 2006, computer servers in data centers and the supporting infrastructure consumed approximately 1.5% of all electricity in the US [1]. Up to half of the total energy used in the data center can be attributed to the cooling infrastructure [1]. Therefore, much recent research has focused around improving the energy efficiency of the cooling infrastructure in data centers.

To reduce the above electricity costs, service providers are increasingly seeking diverse locations that offer lower electricity rates. Often, these locations are also chosen in geographically distributed time-zones, to take advantage of

lower off-peak electricity prices. Similarly, the growing trend of ‘cloud computing’ [2] involves federating multiple heterogeneous distributed data centers to offer a stack of IT resources upon which services can be hosted across multiple sites with variable (elastic) resources and capacity. Given such trends, the energy efficiency associated with the delivery of IT services from distributed sites (hereby referred to as ‘distributed computing’) will be a key factor in the next generation of computing applications. However, delivering distributed services may entail very different thermal management challenges than traditional single-site architectures. New ensemble-level thermal management strategies will be required to maximize the cooling efficiency of distributed computing infrastructures.

In this paper, we discuss an approach for optimizing the thermal performance related to a service delivered from a geographically distributed infrastructure. We begin by leveraging existing models for data center energy efficiency to estimate the global energy efficiency of a distributed service. This model takes into account various thermal management parameters, including the related component temperatures, the required inside air temperature, the outside air temperature, etc. We then approximate the thermal performance by superposing the piecewise IT workloads that may be delivered from various locations. In comparison to traditional approaches where the entire service is delivered from the same facility, we find that multi-site service-level cooling efficiency can be distinct from facility- or infrastructure-level cooling efficiency. Based on this finding, we hypothesize that – under appropriate conditions – the opportunity to achieve higher thermal efficiency in a distributed environment may be available. We illustrate the approach for such optimization through a hypothetical example wherein an IT service is delivered from three different data centers in geographically diverse locations. Each of these locations is assumed to house a different data center

architecture with different cooling efficiencies. We find the optimal distribution of resources across the data centers will vary depending on the workload characteristics and several parametric conditions related to site cooling efficiency. The paper concludes by reflecting upon additional considerations related to quality of service, time-of-use, etc. that may influence the proposed optimization scheme.

NOMENCLATURE

- A* Burdening coefficient related to cooling power in IT hardware per unit compressor power, Eq. (1a)
- B* Burdening coefficient related to power consumed in CRAC units per unit compressor power, Eq. (1b)
- C* Burdening coefficient related to power consumed in hydraulic pumps per unit compressor power, Eq. (1c)
- COP* Coefficient of Performance
- COP_G* Coefficient of Performance of the cooling ensemble
- CRAC* Computer Room Air Conditioning units
- d* Effective distance, relative to the degradation of quality of service when data must be transferred from one place to another
- D* Burdening coefficient related to power consumed in cooling tower per unit compressor power, Eq. (1d)
- i* Counter in summation
- n* Variable, nominally for number of modules
- PUE* Power Usage Effectiveness, Eq. (2)
- \dot{Q} Rate of heat transfer, Watts
- SHI* Supply Heat Index, Eq. (3a)
- TCO* Total Cost-of-Ownership
- \dot{W} Power consumption, Watts
- WPI* Workload Placement Index, Eq. (3)
- x* Dependent variable in functional expression, Eq. (7)
- y* Independent variable in functional expression, Eq. (7)

- α Arbitrary coefficient in function, Eq. (7)
- χ Workload placement function, Eq. (3a)
- ω Penalty function, due to loss in quality of service from migrating workload across locations

Subscripts

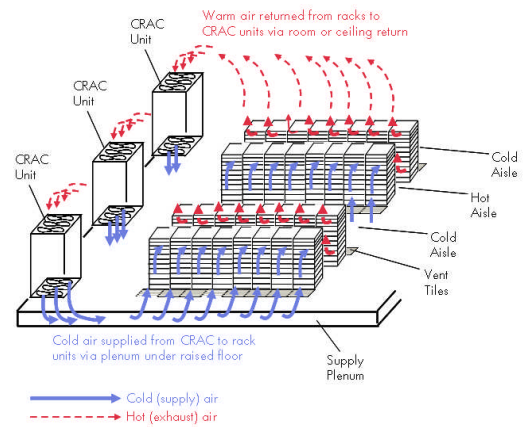
- avg average
- comp compressor
- dc data center
- CRAC related to Computer Room Air Conditioning units
- IT aggregate total for all IT equipment
- sys computer system or server

CONVENTIONAL DATA CENTER COOLING

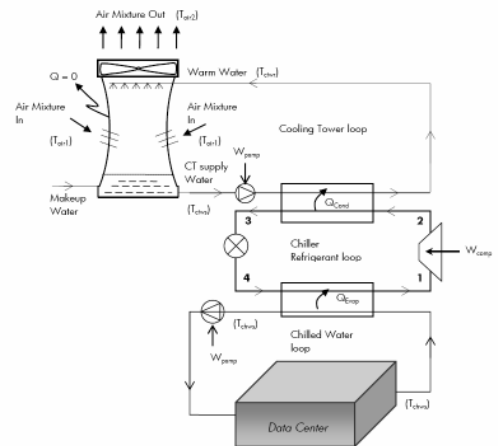
Figure 1(a) shows a typical raised-floor air-cooled data center architecture. Cold air is supplied from a Computer Room Air Conditioning (CRAC) unit, which is delivered through

perforated vent tiles to rows of racks containing numerous computer servers. The cold air is drawn in from a ‘cold aisle’ to the rack units, where it picks up the heat dissipated by the computer equipment. This warm air is then exhausted into a ‘hot aisle’, where it is returned to the CRAC units for refrigeration via room or ceiling return. Although other configurations exist [3, 4], the raised-floor architecture of Fig. 1(a) is most common and will therefore be the focus of the remainder of this paper.

For the typical data center described above, the cooling power consumption can be categorized into thermodynamic work (required to remove the heat dissipated by the IT equipment, returned via the hot aisle) or flow work (required to move the fluid within the data center and through the systems). This work is delivered through an infrastructure comprised of numerous sub-systems (e.g., from the chip to the system, from the system to the racks, from the racks to the CRAC, and then beyond the CRAC to the chiller and cooling tower).



(a)



(b)

Fig. 1. Typical cooling architecture within data centers. (a) Raised-floor cooling architecture to remove heat inside data center [5], and (b) Cooling infrastructure to remove heat outside data center [6].

Figure 1(b) illustrates one configuration of this burdened cooling infrastructure, wherein the hot air returned to the CRAC is cooled by a chilled water stream in a closed heat exchanger (an alternative configuration may involve the use of a compressor within the CRAC as part of a refrigeration cycle). As the heat is transferred from the air stream to the chilled water stream, the temperature of the water increases; therefore, to close the chilled water loop, heat is removed from the water via thermodynamic work in a chiller refrigeration cycle. The refrigerated water is returned to the CRAC unit, while the heat absorbed by the refrigerant is rejected to a secondary water stream. The secondary loop ultimately transfers the heat to the outside environment in a cooling tower.

Measuring Data Center Thermal Performance

Classical thermodynamics suggests evaluation of the cooling infrastructure in terms of the Coefficient of Performance (COP), which measures the amount of heat removed by the infrastructure per unit of power input to the cooling infrastructure. However, application of simple thermodynamic models to the complex data center cooling infrastructure can be challenging. Therefore, Patel *et al.* [7, 8] have suggested evaluating the data center cooling infrastructure in terms of a ‘grand’ COP, which considers the individual COP of the cooling solutions utilized at the chip, system, rack, data center and facility levels. This approach has also been utilized by subsequent researchers [9-11] for analytically evaluating data center thermal performance. Following the original work of Patel *et al.* [7, 8], the ‘grand’ COP for the infrastructure of Fig. 1 can be expressed as follows:

$$COP_G = \frac{\dot{Q}_{dc} / \dot{W}_{comp}}{1 + A + B + C + D} \quad (1)$$

where:

A is the ratio of the total power consumed by the cooling infrastructure across chips, systems, and racks to the power consumed in the chiller compressor, i.e.,

$$A = \frac{\sum(\dot{W}_{rack} + \sum(\dot{W}_{sys} + \sum(\dot{W}_{chip})))}{\dot{W}_{comp}} \quad (1a)$$

B is the ratio of power consumed by the blowers in the CRAC units to the chiller compressor power, i.e.,

$$B = \frac{\sum \dot{W}_{CRAC}}{\dot{W}_{comp}} \quad (1b)$$

C is the ratio of the power consumed by primary and secondary pumps to the chiller compressor power, i.e.,

$$C = \frac{\sum \dot{W}_{pumps}}{\dot{W}_{comp}} \quad (1c)$$

and D is the ratio of power consumed by blowers and pumps at the cooling tower to the chiller compressor power, i.e.,

$$D = \frac{\sum \dot{W}_{cooling\ tower}}{\dot{W}_{comp}} \quad (1d)$$

More recently, the IT industry has begun to consider the energy efficiency of data centers in terms of the Power Usage Effectiveness (PUE) [12, 13], which can be defined as follows:

$$PUE = \frac{\dot{W}_{dc}}{\dot{W}_{IT}} \quad (2)$$

where the total power consumption in the facility is essentially a combination of the power required by the IT equipment; power required by the cooling equipment; and losses related to the delivery of power. Typical PUE values for existing data centers are often in the range of 1.8-2.2 [14, 15]. Data centers with PUE as low as 1.2 have been suggested [16], although there are some concerns related to how PUE measurements may be taken in certain infrastructures [13].

Eq. (2) can be reduced to:

$$PUE = 1 + \frac{I}{COP_G} + \frac{\dot{W}_{power\ delivery}}{\dot{W}_{IT}} \quad (2a)$$

Thus, particularly for facilities where losses due to power delivery are approximately constant or negligible, maximizing the COP_G will lead to minimal PUE (i.e., highest energy efficiency). We assume this relationship between PUE and COP_G in the rest of this paper.

Global Data Center Workload Management

Of particular relevance to the present study is past work related to energy-efficient global workload management for data centers [17-26]. This prior work generally falls into two categories: first, work related to scheduling and management to ensure sufficient availability, performance and quality of service; and second, work related to optimization of resource use across a network of globally distributed systems. Of particular relevance Patel *et al.* [17], which considers allocation of resources across different data centers based on thermal efficiency; and work by Shah and Krishnan [18], which explores optimal distribution of resources across different data centers for minimal economic and environmental burden. Specifically, Patel *et al.* [17] suggest taking into account the thermal efficiency of a given data center in terms of the following Workload Placement Index (WPI):

$$WPI = \max(f(\omega_i \chi_i)) \quad (3)$$

where

$$\chi_i = \frac{COP_i}{SHI_i} \quad (3a)$$

and ω_i is a locality index which is dependent on the distance between data centers and related network, data transmission, and quality of service considerations. ω_i represents a counter weight for purely energy-driven placement decisions that involve long distances between the source of workload and destination data center, and can be based on several weighing schemes such as inverse-square law formulations or logarithmic decay. SHI (Supply Heat Index) is a metric that measures the amount of hot exhaust which is recirculated back to the inlet, and is thus a proxy for thermal efficiency within the data center airspace.

Using the above scheme, Patel et al. [17] illustrate how decisions could be made about the most energy-efficient data center in an interconnected network (grid) to process a given workload. Shah and Krishnan [18] subsequently extended this WPI-based approach to include considerations of environmental burden and cost in choosing the optimal data center, but also considered the possibility of dynamically reallocating and distributing a given workload across multiple data centers. However, the literature is lacking in approach to assess the *thermal* efficiency of distributed workloads. The present work seeks to fill this gap in the state-of-the-art.

GEOGRAPHICALLY DISTRIBUTED COMPUTING

Three trends in computing today motivate the present work. First, the total cost-of-ownership (TCO) model for many IT services – particularly in a cloud computing environment – is typically differently structured than traditional enterprise computing. In traditional environments, personnel, depreciation and related operational costs are 50% or more of the TCO. By contrast, the cost of power and cooling is typically as high as 80% of the cloud TCO. Thus, there is greater pressure to design data centers with higher COP_G . For example, a common trend in data centers is to utilize aisle containment [27, 28] to eliminate recirculation, leading to reduced SHI and higher χ_i . Similarly, data centers are becoming extremely aggressive in the use of outside air for cooling, which allows for elimination of the chiller in the data center infrastructure ($C, D \rightarrow 0$ so that higher COP_G is possible). This also often requires operation at higher temperatures, which in turn implies improved redundancy in the software stack to compensate for potentially higher hardware failure rates.

Second, workloads are often being supported by a larger number of data centers. In traditional environments, a data center is provisioned for maximum capacity, even if that capacity is only required occasionally. For example, in the case of online retailers, computing capacity required on peak shopping days – such as the weeks before Christmas or Black Friday in the United States – could easily be orders-of-magnitude larger than the capacity required during the rest of the year. As these peak shopping periods are often the highest revenue generating periods for the business, the cost of running out of capacity may be too great to risk. So, online retailers commonly over-provisioned their data centers. As a result, their

data centers spent most of the year running at 10% to 20% of their maximum capacity. By contrast, in just the same manner that distributed power generating plants pool together resources to deliver power more efficiently to end users through an interconnected grid (rather than having a power plant dedicated to supporting each neighborhood), service providers are now pooling together distributed resources from different environments to support user needs on demand. Thus, an online retailer can now simply ‘rent’ additional capacity for peak shopping periods and build a much smaller data center for year-round use to save on depreciation costs (or even eliminate the need for their own data center entirely). However, due to this elasticity, the ability to simply bring additional computing capacity online or offline whenever demand is forecasted to spike or diminish can be quite beneficial. From a thermal management perspective, this implies a larger number of discrete systems in the control volume, which in turn suggests a higher number of degrees of freedom in the system (i.e., higher value of i but also greater dependence on ω_i).

Third, interactive and virtualized workloads are becoming increasingly common relative to physical (non-virtualized) batch processing jobs. Interactive and virtualized workloads often tend to be more volatile and dynamic in nature [29]. As a result, for a geographically distributed service, there may be more diversity of demand across geographic, diurnal, and seasonal considerations in a distributed environment.

Thermal Performance Model

We now consider the impact of the above trends on thermal performance of distributed computing. We begin by considering that a given service may be supported from a variety of data center ‘modules’. Each of these modules are likely discrete, and may be identical (such as sets of containers [30]), similar (such as different zones within a data center) or heterogeneous and diverse (such as different geographically distributed data centers). Then, the overall thermal performance of a distributed service can be given by:

$$COP_G = \frac{\sum_{i=1}^n \dot{Q}_i}{\sum_{i=1}^n \dot{W}_i} \quad (4)$$

where n is the number of modules supporting the service. By contrast, for the same number of data centers operating in an unsynchronized environment, we have:

$$COP_G = \max \left(\frac{\dot{Q}_i}{\dot{W}_i} \right) \quad \forall i = 1 : n \quad (5)$$

The main difference between Eq. (4) and Eq. (5) is that in the distributed services model, we assume a modular computing environment and then compute the thermal performance based

on the actual efficiency within each module. Ideally, the distribution across each module has been allocated to derive the maximum *global* efficiency. On the other hand, in the traditional model, we assume it is more efficient to consolidate workloads as much as possible into one facility (presumably the most efficient location) based on maximum *local* efficiency. While global workload allocation mechanisms may exist within traditional environments, we assume that the decision will be made based on local data center efficiency because traditional computing environments do not generally contain the necessary architectural elements required to support distributed workloads. Thus, at any given point of time, the COP_G expression of Eq. (5) will be accurate.

For the above formulations, an additional degree of freedom becomes available in the distributed services model: how resources are allocated across the different modules. Thus, the problem for optimizing thermal performance in distributed computing can be characterized as:

$$\max(COP_G \omega) \quad (6)$$

subject to:

$$\sum_{i=1}^n \dot{Q}_i = \dot{Q} \quad (6a)$$

$$\dot{W}_i = \frac{\dot{Q}_i}{COP_{G_i}} \quad (6b)$$

For simplicity, we define the penalty function ω in terms of the aggregate (total) linear distance d between all of n data center modules and the end user, weighted by the IT load in each module and relative to some average (expected) distance d_{avg} :

$$\omega = \frac{\sum_{i=1}^n d_i \dot{Q}_i}{d_{avg} \sum_{i=1}^n \dot{Q}_i} \quad (6c)$$

Thus, if more modules are utilized in delivering the distributed service and/or these modules are spaced farther apart, a higher penalty will be incurred because of added cost or loss in quality associated, which gets reflected in Eq. (6c). Alternatively, if most of the load is situated in a data center that is located far away, the performance will be penalized. The definition of Eq. (6c) may be overly simplistic for many instances; future work will be required to explore appropriate definitions of ω that consider characteristics such as the rate of decay, quality of service, etc. Nonetheless, we believe that Eq. (6c) provides a useful starting point upon which the dependence of IT workload on geographic distribution can be explored.

Equation (6) provides the desired optimization approach for maximizing the thermal performance in delivering a distributed service. The algorithm relies on finding the optimal distribution of workload across a given set of modules ($\dot{Q}_1 \dots \dot{Q}_n$), relative to an IT-driven performance penalty for moving workloads around (ω), and the cooling power required to manage the workload in each location ($\dot{W}_1 \dots \dot{W}_n$). The cooling power can be determined from experimentation or through modeling [7-11].

Equation (5) and Eq. (6) also provide a set of boundary conditions where thermal efficiency of service delivery from a traditional single-site model will be identical:

- If all the modules in the distributed environment are homogeneous in terms of cooling efficiency and can be treated identically (i.e., $COP_{G_1} = COP_{G_2} = \dots = COP_{G_n}$);
- if the load is uniformly distributed across all the modules (i.e., $\dot{Q}_1 = \dot{Q}_2 = \dots = \dot{Q}_n$);
- if the distributed modules are spaced at the same distance from the user (i.e., $d_1 = d_2 = \dots = d_n$) and the single-site data center is approximately at the same distance (i.e., $\omega_i = 1$);
- if changes to the IT load are synchronous across modules; then the thermal efficiency of the distributed service infrastructure will be equal to the thermal efficiency of a single site. It should be noted that the above conditions are sufficient but not necessary; the efficiencies for both systems *may* be comparable even if one of the above conditions is violated, but in most cases, violation of the above conditions will lead to differing efficiencies.

The next section illustrates application of the model for a given set of workloads, and compares the efficiency in delivering a given service over distributed sites relative to delivering the same service from a traditional single-site environment.

EXAMPLE

We consider a case study with five different IT infrastructure configurations. The first (A) is a traditional ‘mass market’ enterprise data center in Houston, Texas with an annually averaged PUE of 1.9 and a maximum compute capacity of 3-MW. The second (B) is a ‘best-in-class’ large-scale enterprise data center located in the UK with an annually-averaged PUE of 1.2 and a maximum compute capacity of 3-MW. The third (C) is a distributed data center environment, where three smaller 1-MW data centers with PUE ranging between 1.2 and 1.7 are brought online from Houston, Bangalore (India) and the UK. The fourth (D) is a set of thirty 100-kW modular containerized data centers with a PUE of 1.2 running industry-standard hardware, all located inside a warehouse in Houston; while the fifth (E) is the same set of 100-kW containerized data centers but evenly distributed across the three locations discussed earlier and placed outdoors.

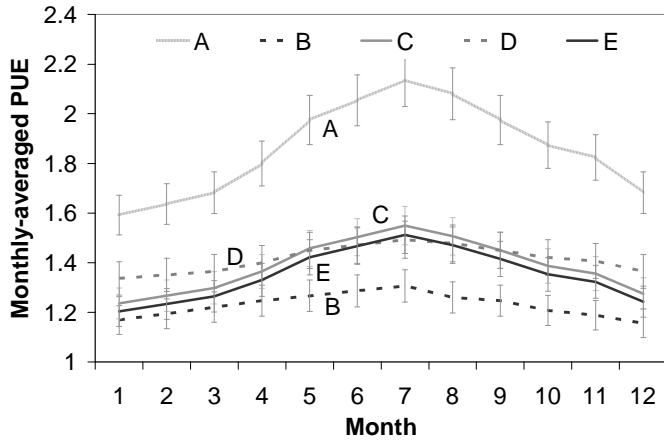


Fig. 2. Average PUE values for the different IT infrastructures considered. Lower PUE corresponds to a more energy-efficient infrastructure.

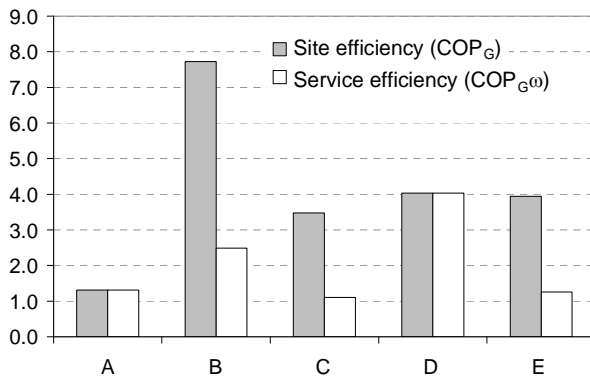


Fig. 3. Site versus Service Efficiency for the different IT infrastructures considered. Higher service efficiency is better.

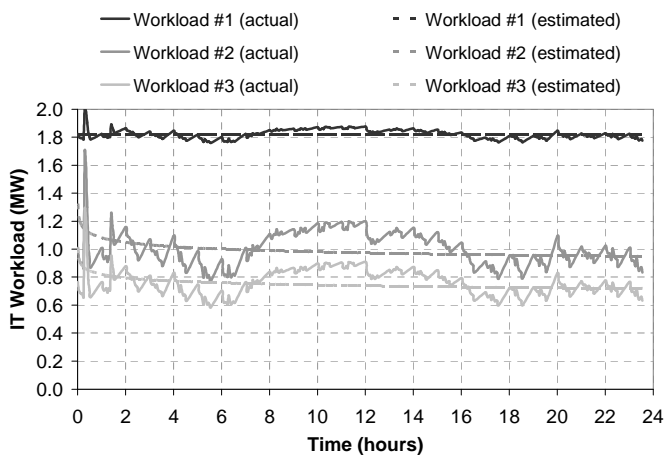


Fig. 4. Workload traces considered.

Figure 2 shows the simulated fluctuation in monthly PUE averages for the different cases considered (lower PUE is better). The base case is a 3-MW uniform IT load that remains constant over time and is being used for a site of users in San Francisco, with an average expected distance corresponding to a site in Houston. The variation in PUE stems predominantly from changes in the outside air temperature, which influences the data center in primarily two manners. First, for conventional infrastructures, the effectiveness of heat transfer through the chiller depends partly on the temperature at which heat is rejected to the environment. So, as discussed by Breen et al. [10], any fluctuations in the external environment will influence the thermal performance through the data center infrastructure. Second, for infrastructures depending on outside air, higher outside temperatures will generally require an increase in mass flow rate to maintain the same system operating temperature. Thus, the increased airflow speeds needed will require more power and affect the PUE.

As might be expected, scenario B – where the data center has the lowest annually-averaged PUE year-round owing to a favorable climate location – is most efficient in terms of site PUE, while scenario A (which has the most out-of-date and inefficient infrastructure of the cases considered) is the least efficient. The remaining three scenarios – with distributed computing – are quite similar in terms of site thermal efficiency. Thus, based purely on site thermal efficiency, one might consider scenario B (a large-scale enterprise data center in a favorable climate) to be the best choice; while scenario A (large-scale enterprise data center in typical climate) to be the worst choice.

However, from a *service* efficiency standpoint, as shown in Fig. 3, the results can be quite different. (Per Eq. 6, higher service efficiency is better.) Even though site B has the highest infrastructure thermal performance due to local climatic benefits, when the distance of the site from the users (ω) is taken into consideration, the overall efficiency goes down for the site in the UK – by approximately a factor of 3 for case B – so that this is no longer the ideal IT configuration. Instead, case D – where low-PUE containers are deployed in a location closest to the user base – is found to be the ideal configuration.

Effect of Varying Workload

The above examples consider a fixed and uniform IT workload. In reality, the IT workload at each site may vary over time, particularly for distributed computing configurations. Figure 4 shows a set of actual workload traces from different computing environments, and functional forms that are used to approximate these workloads. The functional fits are of the form:

$$y = \sum_{i=1}^n \alpha_i x_i \quad (7)$$

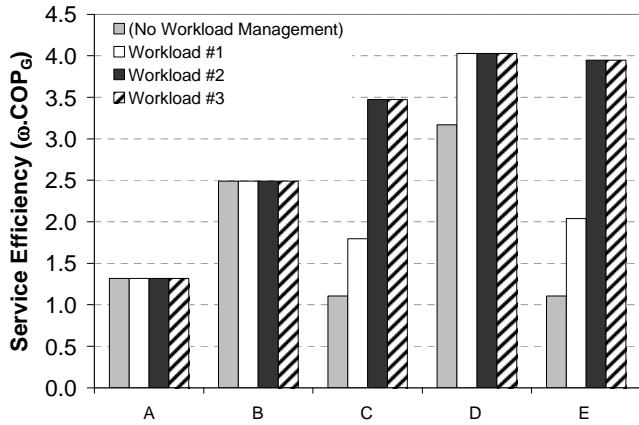


Fig. 5. Impact of workload migration on service efficiency for different IT workloads.

Although simplistic, the functional forms are only used to obtain an analytical form for representing the IT workload. For workload #1 (enterprise setting), the maximum absolute error over a 24-hour trace by using the functional form is about 9.5% with an average error of 1.5%. For workload #2 (virtualized enterprise), which is far more ‘bursty’ in nature (with large and frequent peak-to-trough valleys), the maximum absolute error over a 24-hour trace by using the functional form is about 33% with an average error of 10%. For workload #3 (virtualized enterprise with active power management, as may be typical in a distributed setting), the maximum absolute error over a 24-hour trace by using the functional form is about 32% with an average error of 10%. In both of the virtualized settings, the maximum error only comes in while evaluating the period of high utilization; outside of these periods of higher utilization, the functional form does a reasonable job of predicting the *average* utilization, which is of most interest in the present work. Thus, even though the functional form of Eq. (7) is not entirely accurate, it is deemed sufficient for the present work (the reason for choosing the specific functional form in Eq. (7) will be discussed shortly).

The workloads of Fig. 4 present a methodology to consider the effect of varying demand on thermal efficiency in distributed settings. Particularly, two major parameters are varied in the above: (i) the utilization within the data center; and (ii) the burstiness of the workload. First, the peak utilization across all of the above workloads is around 2-MW, which is about 33% lower than the maximum capacity of the data center. Secondly, the average burstiness of the workload is somewhat captured in the functional form of Eq. (7). We find that there is an optimal value of n , beyond which increasing the number of independent variables in the functional form does not yield improved accuracy along both the maximum and average errors (arbitrarily, by more than 1%). For example, for workload #1, we find that $n=1$ is optimal; for workload #2, $n=3$ is optimal; for workload #3, $n=2$ is optimal. While this is true for the workloads considered, the generality of this claim needs to be

investigated further. Additionally, increasing values of n may still improve the accuracy of the curve-fit (in fact, the curves should become exact as $n \rightarrow \infty$). To bound the solution space, we propose that a threshold value of n exists beyond which further improvements to the accuracy of the curve-fit will be incremental, in this case (for the given workloads) less than 1% for each interval of the size $n=1$. Additional work is required to examine this in more detail.

With the above limitations, we conjecture that the threshold value of n discussed above is representative of the dimensionality which provides the maximum scope for optimality in thermal management. Thus, for $n=1$, having more than one data center is unlikely to yield any improvements in energy efficiency of the thermal management system; for $n=2$, two data centers is the ideal number of nodes delivering the service; etc. This hypothesis is the reason for choosing a functional fit in the form of Eq. (7): by capturing the piecewise nature of the workload, it becomes possible to identify appropriate schemes for workload allocation across each of n data centers. For the workload #2 and #3, as example, we find that a module provisioned roughly at 750-kW capacity (with the capability to scale by about 100-kW) is sufficient to capture the demand from $t=16$ to $t=24$; adding another data center with up to 500-kW capacity meets demand from $t=2$ to $t=16$; and a third data center with about 200-kW of additional capacity satisfies the demand from $t=0$ to $t=2$ (this last data center is not required for workload #3, which is why we obtain $n=2$).

With the above considerations, an additional degree of freedom is availed in the thermal management: how much workload should be allocated to each data center at a given time. The impact of this additional degree of freedom is shown in Fig. 5.

For cases A and B, there is no difference on the service efficiency since there is only one data center available. For workload #1 (traditional enterprise workload), the service efficiency is improved by between 62%, 27% and 84% respectively over the baseline (no migration) for cases C, D and E. With the capability to migrate workload, both cases C and E achieve a higher service efficiency than case A. The increase in service efficiency is driven primarily by reducing ω through improved distribution of workload; and in case D, the ability to reduce the total number of unused modules. For example, in case C, the data center in Houston is generally maintained at maximum capacity; the data center in UK is maintained at partial capacity; and the data center in Bangalore is not used at all. During periods of inefficiency – such as the hot summer months – the workload is correspondingly shifted to the UK data center. The resulting distribution is such that on average, the data center in Houston supports 66% of the workload; the data center in the UK supports about 33% of the workload; and the data center in Bangalore supports about 1% of the workload. Thus, even though the Bangalore data center is more efficient on average than the Houston data center, the benefit of a data center in close proximity to the user outweighs the benefits of improved energy efficiency within the chosen

models. Cases D and E follow a similar distribution pattern for workload #1.

For workload #2, the bursty nature of the workload provides the opportunity for even further optimization related to time-of-use. In addition, the lower magnitude of workload provides the opportunity for further consolidation into more efficient locations. Combined, these two considerations enable elimination of nearly all inefficient cooling locations within the given ecosystem. Specifically, during periods of low utilization, the most efficient service delivery configuration is selected in the same manner as earlier. As the absolute magnitude of workload is lower, the utilization of a second-choice location is no longer necessary at all times. This leads to a further improvement of 94% for cases C and E relative to the improvement already seen in workload #1. For case D, no further reduction in the number of sites was possible; therefore, no further gains in efficiency are seen for workload #2. Across all of the cases, the differences between workload #2 and workload #3 were sufficiently small that no additional gains in efficiency were observed. For all of these bursty workloads, each of the distributed or modular computing scenarios (C,D,E) were more efficient than any of the single-location infrastructures.

The above parametric study suggests some key observations in terms of thermal management efficiency for distributed computing scenarios. First, higher levels of virtualization in distributed environments may provide the opportunity for improved energy efficiency. However, the degree to which workload migration will provide benefits depends on a number of factors:

- the workload characteristics,
- the number of sites (modules) available for migration,
- the environmental and design characteristics of each module (e.g., outside air temperature; PUE; etc.),
- the compute capacity of each size of each module relative to the maximum and average workload, and
- distance and frequency of workload migration.

That is, while virtualization technology makes migration easier, the delay tolerance of a workload is what offers the highest potential for exploitation. For example, a batch job requiring 24 hours of CPU time running inside a virtual machine could be moved over a 24-hour period across multiple sites depending on time zones, external ambient temperature, etc. to improve thermal efficiency. (Additional considerations, such as completion time or quality of service, may prohibit such movement of workload; these considerations are discussed shortly.) The ideal scenario for distributed computing purely in terms of energy efficiency would be to have a large number of distributed modules located close to each other and fairly bursty workloads. Generally, energy efficiency gains stemming from increasing the number of sites available for migration will only be realized up to some threshold point beyond which further degrees of freedom will not provide additional savings. (The optimal number of sites will depend on the workload

characteristics.) In addition, such efficiency considerations need to be balanced against quality of service considerations. If, as a first-order evaluation, quality of service is assumed to degrade approximately linearly with distance, we find that net savings will only be achieved for highly virtualized interactive-type workloads corresponding to relatively high average utilization; for relatively uniform workloads at low average utilization, it may be more advantageous to consolidate workload into the most efficient sites since the efficiency gains obtained by migrating workload are relatively small compared the disadvantage of moving those workloads over long distances.

A key assumption in the above analysis is that only a single customer site (in San Francisco) is being supported by the distributed services offered. In practice, for many applications, the user base may be globally distributed and therefore access will be sporadic as well as spread out over the course of the day (due to time differences). The optimal thermal management strategy in this case will be slightly different than the above scenario.

Effect of Distributed User Base

Instead of being concentrated at a single location, we now consider a user base that is normally distributed in terms of distance from each of the chosen sites. Such a distribution is arbitrarily chosen for illustrative purposes; future work will evaluate the appropriateness of such an assumption. For such a scenario, for any given site, the probability that a user will be closer to a given site increases as the number of sites increases. That is, as the number of sites approaches infinity in the limit, $\omega \rightarrow 1$. Thus, we assume that quality of service in such an environment will *improve* as the number of sites increases. However, the extent of improvement will also depend on the distance between sites, as workload may need to be migrated from site-to-site to reach the closest user at any given point. So, for distributed services with distributed user base, we suggest the following evaluation of ω :

$$\omega = \frac{\sum_{i=1}^n d_i}{n^2} \quad (8)$$

Equation (8) considers the average distance between all the modules ($\sum d_i / n$) and then divides that distance by the number of nodes. To verify that ω indeed reaches 1 in the limit, consider the special case of n data centers distributed at unit distance from each other. For each such data center, we assume that the workload must be migrated twice: once from the initial data center to a module nearest to the user, and then back to the initial data center so that another user can expeditiously access the workload. Then, for large values of n with regards to a user located at the first data center sequentially accessing data from each module:

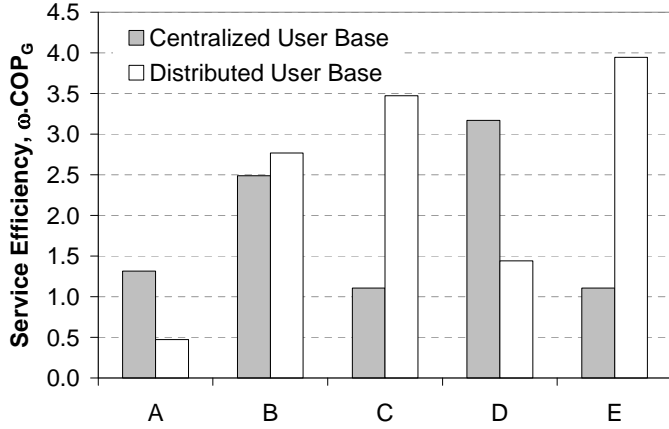


Fig. 6. Service efficiency for a distributed user base.

$$\omega \sim \frac{2(1+2+\dots+n)}{n^2} = \frac{n(n+1)}{n^2} = 1 + \frac{1}{n} \quad (8a)$$

so that clearly $\omega \rightarrow 1$ as $n \rightarrow \infty$. The formula in Eq. (8a) can be shown to be true similarly for greater values of d as well.

On the other extreme, if $n = 1$ (i.e., single data center), then Eq. (8) is reduced to:

$$\omega = d_1 \quad \text{for } n = 1 \quad (8b)$$

That is, users who are situated far away from the site (large d_i) will pay a higher penalty and see lower service efficiency. For a random sample of users, the penalty will be normally distributed; so that for an infinitely large sample, ω will have an expected value that is proportional to the mean (expected) distance.

Figure 6 shows the service efficiency for the different IT infrastructure configurations considered in the case study, assuming a distributed user base with an expected distance at the centroid of the three data center sites considered. For infrastructures with a single site in Houston (A, D), the efficiency becomes slightly worse than the base case (customer in San Francisco) because the average distance to a user is larger. On the other hand, for the single site in the UK (B), the average user is relatively closer than someone in San Francisco, so the service efficiency actually improves slightly. For the cases with geographically distributed data centers (C and E), the service efficiency improves quite significantly as there is now a location in closer proximity to each user, leading to significantly improved ω - yielding an overall service efficiency that is almost identical to the average site efficiency (i.e., the penalty of a distributed computing environment is essentially eliminated). Thus, for distributed services expecting highly distributed user bases, it may be advantageous to deploy a large number of distributed modules where each is close to a user base. Clearly, there are some trade-offs associated with deploying too many modules, as then the average distance between the modules (per Eq. 8) may unduly increase,

potentially off-setting any gains associated with increasing the size of the distribution network.

CONCLUSION

In this paper, we considered some unique characteristics associated with thermal management for geographically distributed computing infrastructures. By contrasting trends in distributed architectures with traditional single-site architectures, key parameters that are expected to become more significant in distributed services are identified. These included the type of workload being supported, the distribution and size of the sites supporting the workload, as well as the distribution of users accessing the workloads. We find that there is no single architecture that definitively provides the best thermal management efficiency in the distributed environment. For example, in interactive applications with significant variation in usage patterns but high resulting utilization rates, a distributed network of energy-efficient modular data centers provides optimal efficiency; but for applications with constant usage and low utilization rates, a single site in an environmentally friendly location near the largest user base may be more efficient. Furthermore, our case study revealed that the penalty for choosing an inefficient architecture could be as much as a factor of 2X. Thus, there is a great incentive for end-to-end customized design that considers the integrated hardware, software, and applications stack within distributed environments.

We note, however, that all results in the present study are based on a simplistic formulation of quality of service as a function of distance between data centers. More sophisticated modeling of this penalty function is required. In particular, a better understanding of the relationship between service efficiency and workload migration will be helpful in obtaining improved estimates. Nonetheless, by identifying key parameters of relevance in distributed computing environments, we believe this work provides a useful starting point for more detailed future investigations.

REFERENCES

- [1] Koomey, J. G., 2007, "Estimating Total Power Consumption by Servers in the U.S. and the World," available <http://enterprise.amd.com/Downloads/svrpwrusecompletefinal.pdf>
- [2] Armbrust, M., Fox, A., Griffith, R., Joseph, A.D., Katz, R.H., Konwinski, A., Lee, G., Patterson, D.A., Rabkin, A., Stoica, I., Zaharia, M., 2009, "Above the Clouds: A Berkeley View of Cloud Computing," Technical Report No. UCB/EECS-2009-28, University of California, Berkeley, CA.
- [3] Shrivastava, S., Sammakia, B., Schmidt, R.R., Iyengar, M., 2005, "Comparative Analysis of Different Data Center Airflow Management Configurations," Proc. ASME/Pacific Rim Technical Conference and Exhibition on Packaging and Integration of Electronic and Photonic Systems, MEMS and NEMS (InterPACK), San Francisco, CA.
- [4] Iyengar, M., Schmidt, R., Sharma, A., McVicker, G., Shrivastava, S., Sri-Jayantha, S., Amemiya, Y., Dang, H.,

- Chainer, T., Sammakia, B., 2005, "Thermal Characterization of Non-Raised Floor Air Cooled Data Centers Using Numerical Modeling," Proc. ASME/Pacific Rim Technical Conference and Exhibition on Packaging and Integration of Electronic and Photonic Systems, MEMS and NEMS, San Francisco, CA.
- [5] Shah, A., Patel, C., Bash, C., Sharma, R., Shih, R., 2008, "Impact of Rack-Level Compaction on the Data Center Cooling Ensemble," Proc. 11th Intersociety Conf on Thermal and Thermomech Phenomena in Electronic Systems, Orlando, FL.
- [6] Beitelmal, M.H., Patel, C.D., 2006, "Model-Based Approach for Optimizing a Data Center Centralized Cooling System," Technical Report No. HPL-2006-67, Hewlett Packard Laboratories, Palo Alto, CA.
- [7] Patel, C., Sharma, R.K., Bash, C.E., Beitelmal, M., 2006, "Energy Flow in the Information Technology Stack: Coefficient of Performance of the Ensemble and its Impact on the Total Cost of Ownership," Technical Report No. HPL-2006-55, Hewlett Packard Laboratories, Palo Alto, CA.
- [8] Patel, C., Sharma, R.K., Bash, C.E., Beitelmal, M., 2006, "Energy Flow in the Information Technology Stack: Introducing the Coefficient of Performance of the Ensemble," Proc. Intl. Mechanical Engineering Congress and Exhibition, Chicago, IL.
- [9] Iyengar, M., Schmidt, R., 2007, "Analytical Modeling of Energy Consumption and Thermal Performance of Data Center Cooling Systems – From the Chip to the Environment," Proc. ASME/Pacific Rim Technical Conference and Exhibition on Packaging and Integration of Electronic and Photonic Systems, MEMS and NEMS (InterPACK), Vancouver, BC.
- [10] Breen, T.J., Walsh, E.J., Punch, J., Bash, C.E., Shah, A.J., 2010, "From Chip to Cooling Tower Data Center Modeling: Part I, Influence of Server Inlet Temperature and Temperature Rise Across Cabinet," Proc. 12th Conf on Thermal & Thermomech Phenomena in Electronic Systems (ITHERM), Las Vegas, NV.
- [11] Walsh, E.J., Breen, T.J., Punch, J., Bash, C.E., Shah, A.J., 2010, "From Chip to Cooling Tower Data Center Modeling: Part II, Influence of Chip Temperature Control Philosophy," Proc 12th Intersociety Conf on Thermal and Thermomech Phenomena in Electronic Systems (ITHERM), Las Vegas, NV.
- [12] The Green Grid, 2007, "Green Grid Metrics: Describing Datacenter Power Efficiency," Technical Committee White Paper, available <http://www.thegreengrid.org>.
- [13] Hamilton, J., 2009, "PUE and Total Power Usage Effectiveness (tPUE)," <http://perspectives.mvdirona.com/2009/06/15/PUEAndTotalPowerUsageEfficiencyTPUE.aspx>
- [14] EPA, 2007, "Report to Congress on Server and Data Center Energy Efficiency Public Law 109-431," U.S. Environmental Protection Agency, Washington DC.
- [15] Greenberg, S., Mills, E., Tschudi, B., 2006, "Best Practices for Data Centers: Lessons Learned from Benchmarking 22 Data Centers," Proc. ACEEE Summer Study on Energy Efficiency in Buildings, Pacific Grove, CA.
- [16] "Microsoft's New Ireland Data Center Using Outside Air to Cool the Facility and Drive Energy Efficiency," available <http://blogs.msdn.com/see/archive/2009/09/24/microsoft-s-new-ireland-data-center-using-outside-air-to-cool-the-facility-and-drive-energy-efficiency.aspx>
- [17] Patel, C., Sharma, R., Bash, C., Graupner, S., 2003, "Energy Aware Grid: Global Workload Placement based on Energy Efficiency," Proc. ASME International Mechanical Engineering Congress and Exposition, Washington, DC.
- [18] Shah, A.J., Krishnan, N., 2008, "Optimization of Global Data Center Thermal Management Workload for Minimal Environmental and Economic Burden," IEEE Trans. Comp. and Packaging Technologies, Vol. 31, No. 1, pp. 39-45.
- [19] Foster, I., Kesselman, C., Nick, J.M., Tuecke, S., 2002, "Grid Services for Distributed System Integration," IEEE Computer, Vol. 35, No. 6, pp. 37-46.
- [20] Chandra, A., Gong, W., Shenoy, P., 2003, "Dynamic Resource Allocation for Shared Data Centers Using Online Measurements," Proc. International Workshop on Quality of Service (IWQoS), Vol. 27, pp. 381-398.
- [21] Buyya, R., Abramson, D., Giddy, J., 2000, "Nimrod/G: An Architecture of a Resource Management and Scheduling System in a Global Computational Grid," Proc. Fourth International Conference and Exhibition on High Performance Computing in the Asia-Pacific Region, Beijing, China.
- [22] Ranjan, S., Rolia, J., Fu, H., Knightly, E., 2002, "QoS-Driven Server Migration for Internet Data Centers," Proc. 10th IEEE Intl Workshop on Quality of Service, Miami Beach, FL.
- [23] Shan, H., Olikar, L., Smith, W., Biswas, R., 2004, "Scheduling in Heterogeneous Grid Environments: The Effects of Data Migration," 12th Intl Conf on Advances in Computing and Communications (ADCOM), Ahmedabad, India.
- [24] Ramakrishnan, K.K., Shenoy, P., Van der Merwe, J., 2007, "Live Data Center Migration across WANs: A Robust Cooperative Context Aware Approach," Proc. SIGCOMM Workshop on Internet Network Management, Kyoto, Japan.
- [25] Hermenier, F., Lorca, X., Menaud, J.-M., Muller, G., Lawall, J., 2009, "Entropy: A Consolidation Manager for Clusters," Proc. ACM/Usenix Int. Conference on Virtual Execution Environments (VEE), Washington, DC, pp. 41-50.
- [26] de Assunção, M. D., Buyya, R., 2009, "Performance Analysis of Allocation Policies for InterGrid Resource Provisioning," Information and Systems Technology, Vol. 51, No. 1, pp. 42-55.
- [27] Martin, M., Khattar, M., Germagian, M., 2007, "High Density Heat Containment," ASHRAE Journal, Vol. 49, No. 12, pp. 38-43.
- [28] Samadiani, E., Joshi, Y., Mistree, F., 2008, "The Thermal Design of a Next Generation Data Center: A Conceptual Exposition," Vol. 130, No. 4, Article No. 041104.
- [29] Gmach, D., Rolia, J., Cherkasova, L., Kemper, A., 2007, "Workload Analysis and Demand Prediction of Enterprise Data Center Applications," Proc. IEEE International Symposium on Workload Characterization (IISWC), Boston, MA.
- [30] HP Performance Optimized Datacenter (HP POD), <http://h20338.www2.hp.com/enterprise/cache/595887-0-0-0-121.html>