Cool Job Allocation: Measuring the Power Savings of Placing Jobs at Cooling-Efficient Locations in the Data Center

Cullen Bash and George Forman
Hewlett-Packard Labs, Palo Alto, CA 94304

Abstract

Data center costs for computer power and cooling are staggering. Because certain physical locations inside the data center are more efficient to cool than others, this suggests that allocating heavy computational workloads onto those servers that are in more efficient places might bring substantial savings. This simple idea raises two critical research questions that we address: (1) How should one measure and rank the cooling efficiency of different places in a data center? (2) How substantial is the savings? We performed a set of experiments in a thermally isolated portion of a real data center, and validated that the potential savings is substantial and therefore warrants further work in this area to exploit the savings opportunity.

1. Introduction

The total cost of ownership of a fully managed data center with a 1.3 megawatt cooling capacity is approximately $18 million per year (e.g. 100 fully loaded 13KW racks with 4000 1U servers) [7]. About 15% of the cost is for operation and maintenance of the environmental control system. This partly reflects that cooling resources are over-provisioned to cover worst-case situations. The temperature at the air inlet of all servers must be kept below a target threshold, $\leq 28^{o}$ C for example, even when all servers are 100% busy.

Local variations in airflow and server heat generation impact the efficiency of cooling different places within the data center. Air conditioning units on the periphery supply cool air into an under-floor plenum. The cool air is delivered to the room via ventilation tiles on the floor located in between the two rows of equipment racks. The equipment racks are oriented such that their air intakes are facing the "cold aisle" with the vent tiles. Hot spots in the top middle of the row result from recirculation of hot air exhausted on the opposite side of the server racks. The temperature of the exhaust air is primarily a function of equipment power consumption which is driven by server design and computational workload. Hot spots are a ubiquitous problem in air-cooled data centers, and drive the environmental control system to work much harder to ensure that no server is fed hot air (i.e. air at a temperature greater than the target threshold).

Related work [6,8,9,15] considered the placement of computational workload to alleviate these local hotspots and provide failure mitigation. Various algorithms have been developed to guide the placement of resources according to the external environment, but none yet have considered the varying ability of the air conditioning units to cool different places in the room. For example, a location might appear to be a good place because it is currently cold, but it may be difficult to cool, such as in a corner of the room in a location far removed from an air conditioning unit. Secondly, while simulations have been used to prove the concept of the various approaches, validation in a real data center under realistic workload conditions has not previously been attempted.

Section 2 describes a new practical metric to grade cooling-efficiency, which involves both the current air temperature and the historical ability of the computer room air conditioners (CRACs) to cool the location along with information about local airflow conditions. This metric can then be used to rank the different places in the data center, providing a preference for where to place heat-generating computational workload. The great complexity of managing a data center makes any additional considerations for cooling efficiency unwelcome. But the adaptive enterprise vision is that next-generation data centers will have management control software that will provide increased levels of automation and can more easily integrate cooling considerations into their policies. Of course, adding such software complexity to future data centers—as well as the research on how best to do it—is only warranted if the savings are sufficiently substantial. We address this strategic research question empirically:

We use our efficiency metric of Section 2 in a practical experiment described in Section 3 that measures the total power consumed by a thermally isolated portion of our data center under different control policies. The experiment assumes that computational workload, such as batch jobs, can be placed or moved within a data center based on cooling efficiency. Although this is not the practice today, it could be achieved easily enough by having job schedulers take a server preference list

Figure 1. Utilization of HP rendering service for 30 days (x-axis) over 345 servers (y-axis): white= busy.

into account when allocating large new jobs onto the servers, or else by future data centers that leverage Virtual Machine technology to dynamically migrate running jobs from one server to another in order to improve cooling efficiency.

The experiment results are described in Section 4. Briefly, we observed ~⅓ savings in the cooling power required, despite only having control of a fraction of the computers in the isolated data center. The ensuing discussion in Section 5 includes a translation of this savings into an estimate of the dollar savings for a modern, large scale data center. Depending on usage and other factors discussed in that section, it could easily exceed $1,000,000 savings per year.

We round out this introduction with a final item of motivation. The proposed savings depends considerably on the utilization of the data center, e.g. when the data center servers are ~100% busy or ~100% idle there is no flexibility about where to place workload. Thus, the potential for savings depends on the data center being only partially utilized a substantial fraction of the time. Although one cannot argue that this is the case in most data centers, we find various evidence that this is the case in at least some data centers:

1. Reports from the field indicate that many customer data centers run at fairly low utilization most of the time. Indeed, this has recently led to research and services in server consolidation via virtualization technology [12].
2. Anecdotal evidence of several academic batch job servers and our experience with those within HP Labs suggest that, although there are periods when all servers are continually busy (e.g. conference submission season), many other times the offered workload is sporadic.
3. As a final anecdote, we examined the utilization of the HP Labs movie rendering service used by DreamWorks in the production of the movie Shrek II, and again found substantial periods of middling utilization. Refer to the visualization in Figure 1, where a black pixel indicates a server was idle for an entire 5 minute interval, and is white otherwise. Over this 30 day period corresponding to Nov. 2004, we see many times when only a portion of the servers were busy.

Further analysis of this data, as well as additional discussion and color photographs, is available in the longer technical report version of this paper at: www.hpl.hp.com/techreports/2007/HPL-2007-62.html

2. Measuring Cooling Efficiency

Initial work regarding measurement and optimization of data center cooling efficiency was centered around the modeling and characterization of airflow and heat transfer in the data center [3,10,13]. The work relied upon numerical simulations to improve the placement of cooling resources via the manipulation of vent tiles, CRAC unit placement and server placement. Additional work has focused on the optimization of the fundamental equations of fluid mechanics and thermodynamics within racks [11] and data centers [14] to minimize a given cost function and improve operational efficiency. Although much progress has been made in this area, the modeling techniques involved are time consuming and have to be re-run as data center operation changes with time, either due to changes in workload distribution or physical configuration.

More recent work has focused on real-time control systems that can directly manipulate the distribution of cooling resources throughout the data center according to the needs of the computer equipment. One such system, called Dynamic Smart Cooling, uses a network of temperature sensors at the air inlet and exhaust of equipment racks [2]. Data from the sensors is fed to a controller where it is evaluated. The controller can then independently manipulate the supply air temperature and airflow rate of each CRAC in the data center. In order to accomplish this efficiently, the impact of each CRAC in the data center must be evaluated with respect to each sensor. The result of such an evaluation will define the "regions of influence" of each CRAC unit. This information can then be used to determine which CRACs to manipulate when a given sensor location requires more or less cool air. Such a system has been shown to operate much more efficiently than traditional control systems that contain sparse temperature sensing, usually only at the inlet of each CRAC, and rudimentary operating algorithms that do not consider local environmental conditions [2].

The regions of influence are defined with respect to a metric called the Thermal Correlation Index (TCI) shown in Equation 1. It quantifies the response at the i^{th} rack inlet sensor to a step change in the supply temperature of the j^{th} CRAC. TCI is a static metric based on the physical configuration of the data center.

Since it does not contain dynamic information, it can be thought of as the steady-state thermal gain at the sensor to a step change in thermal input at the CRAC. The regions of influence that are defined by the TCI metric are stable with time, but are functions of data center geometry and infrastructure (e.g. vent tile arrangement) as well as CRAC flow rate uniformity.

$$TCI_{i,j} = \frac{\Delta T_i}{\Delta T_{crac,j}} \quad (1)$$

The process by which TCI is evaluated can be performed numerically or in-situ in the data center with the deployed sensor network. In-situ measurements are more accurate while numerical simulations can be done off-line and enable parametric analysis.

Another attribute of TCI is that it describes the efficiency by which any given CRAC can provide cooling resources to any given server. We therefore use it in the development of a more general workload placement index that we term the "Local Workload Placement Index" described by Equation 2 as follows:

$$\text{LWPI}_i = \frac{(\text{Thermal Margin})_i + (\text{AC Margin})_i}{(\text{Hot Air Recirculation})_i} = \frac{(T_{set} - T_{in})_i + \sum_j \left[(T_{SAT} - T_{SAT,\min}) * TCI_i\right]_j}{(T_{in,i} - T'_{SAT,i})} \quad (2)$$

where the numerator quantifies the thermal management and air conditioning margin at sensor location i and the denominator quantifies the amount of hot air recirculation at the server (this is related to the Supply Heat Index described in [3]). Specifically, T_{set} is the desired computer equipment inlet temperature setpoint within Dynamic Smart Cooling, T_{in} is the current inlet temperature measured within the server or with an externally deployed sensor network, T_{SAT} and $T_{SAT,min}$ are the supply air temperature and minimum allowable supply air temperature of the air conditioning unit(s) respectively. Both are reported by the CRAC. The Thermal Correlation Index $TCI_{i,j}$ represents the degree to which CRAC j can provide cooling resources to the computer equipment at sensor i. Finally, $T'_{SAT,i}$ is the temperature of the air delivered through the vent tiles in close proximity to the i^{th} server and is a strong function of the supply air temperature (T_{SAT}) of the CRACs that serve the region in which the i^{th} sensor resides. As defined, the metric is a ratio of local (i.e. server level) thermal management and air conditioning margin to hot air recirculation and can therefore be used to gauge the efficiency of cooling resource placement and, by extension, workload placement.

Figure 2. Example server utilization, 8am burst jobs

3. Experiment Methodology

Fundamentally, any sort of workload might be placed so as to optimize cooling efficiency. We chose to focus on an opportunity that may be practical for widespread use in the near term: the placement of CPU-intense batch jobs. An obvious experiment scenario is to have jobs arrive occasionally, and to allocate each to the most cooling-efficient server available. It would remain only to choose job arrival rates and a distribution of job durations. One could then measure the power savings of cooling-efficient placement vs. today's cooling-oblivious placement. Though uncomplicated, this scenario is naïve. In batch processing systems it is common that a user enqueues a large number of jobs in a burst. For example, in the server utilization diagram in Figure 2, a large burst of jobs arrives at 8am, making all the servers go from idle (black) to busy (white). As each job completes, servers are kept busy by the supply of enqueued jobs. When the queue finally goes empty (~10am), each server runs its last allocated job to completion and then goes idle (the last being at 3pm). This type of pattern is evidenced repeatedly in the 30 day snapshot in Figure 1. Thus, in a practical deployment, the savings of cooling-aware placement will likely be realized only after the work queue is drained and servers begin to go idle.

The duration of this 'wind-down phase' can be substantial, especially if the variance in job lengths is large, as we often observe in practice. As a practical example, the job lengths in the NASA iPSC benchmark [5] have a coefficient of variation of ~350% (CV=std.dev/mean), and for a recent machine learning experiment by the second author, the CV was 130%. Thus, the placement of the last few long jobs determines which servers will remain busy long after the others have finished. It is here in the wind-down phase that we focus our experiments. We will compare today's cooling-oblivious, first-come first-served (FCFS) placement vs. a smart cooling-aware placement that puts the longer running jobs on the more efficient servers to cool, given the schedule shown in Figure 2.

In practical implementations, this could be achieved either by (1) having rough estimates of job lengths so that an efficient schedule can be devised, or (2) dynamically migrating long running jobs to the more efficient servers via virtualization technology, such as Xen. We initially attempted the latter, which is perhaps

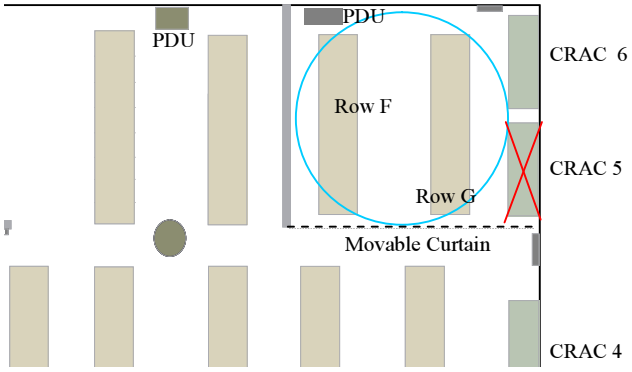

Figure 3. Experimental test bed.

more elegant because it can be difficult to obtain job length estimates. Unfortunately, due to ownership constraints, we could not get Xen installed on enough servers to make any significant impact on the room temperature, considering the many other computers present. Forced to resort to method (1), we developed a simple FCFS scheduler that placed the longest schedule on the most efficient server and the progressively shorter schedules on the less efficient servers in sequence. We used the pre-determined job lengths of 126 jobs from a previous experiment having CV 130%. This created the schedule shown in Figure 2.

But assuming one takes power savings seriously, there is another factor to consider: putting servers that are not being used into a low power state, e.g. shutting them off. This is simple enough to implement, and with quick hibernation available in future servers, it will become quite easy to effect. As we shall show, this complements efficient placement nicely, and used together, a great deal of power savings can be had.

To conclude, the experiment design follows a lesion study model, determining the power savings of cooling-efficient placement together with server shutdowns, as well as the marginal benefit of each technique by itself. The detailed protocol will be discussed after we introduce the test bed.

3.1 Test Bed

We are fortunate to have available to us a thermally isolated portion of an active data center at HP Labs, Palo Alto, depicted in Figure 3. We isolated the research area (upper right quadrant) via a heavy plastic curtain and closeable air baffles beneath the floor plenum. This area is cooled by two redundant Computer Room Air Conditioning units (CRACs), but for these experiments we turned off CRAC 5. Both CRACs 5 and 6 have two operational modes. One mode utilizes the facility's chilled water system to remove heat from the air via an internal heat exchanger while the other uses a vapor compression refrigeration system internal to the unit. When operating in the latter mode, power consumption of the unit can be directly measured and was therefore used throughout the experimental phase of this work. This CRAC was controlled via Dynamic Smart Cooling by using temperature sensors in the test bed. More modern Proliant-class HP servers have inlet temperature sensors built into each server.

We obtained control of 54 of the 76 NetServer LP2000r servers in the 6 racks in Row F (marked in Figure 3). Although other users had control of the remaining NetServers and many servers in the other row, we monitored their power consumption and discarded measurements affected by any substantial change.

3.2 Experiment Protocol

0. Rank the servers by their $LWPI_i$ value, computed from the temperature sensors and TCI efficiencies.
1. Determine the FCFS job schedules from the batch job predicted run-times, and place the longest running schedules on the most efficient servers.
2. At 8am, all servers go busy for two hours, giving the data center ample time to come to a thermally steady state, and reflects the situation after an arbitrary number of hours of being fully busy.
3. As each server completes between 10am and 3pm, it shuts down (simulating a low power mode).
4. As servers shut down and contiguous regions of servers around a sensor are all off, the acceptable temperature limit for that sensor is increased by about +5°C—enough to essentially remove it from control while still providing minimal cooling.
5. Measure the server and CRAC power consumption during the wind-down phase: 10am to 3pm (efficient placement has no effect when all servers are busy or all are idle).
6. On separate days, repeat the above experiment without server shutdowns, without efficient placement, and without either—for the baseline mimicking current behavior of batch services.

4. Experiment Results

Table 1 shows the experimental results in terms of average power consumed over the duration of the experiment and the savings with respect to the baseline setting. Power consumption of the servers in row F of Figure 3 is reduced by 30% when the test machines are shut down after their jobs have each completed. Naturally, the *server* power consumption is unaffected

by load placement. By contrast, power consumption of the air conditioning equipment is reduced by 8% via cooling-aware placement alone, 15% via shutdown alone, and 33% when both cooling-aware placement and shutdown are employed. Overall, the total power savings is reduced 32% when both techniques are used.

Table 1. Kilowatts consumed by each setting.

	Baseline	Placement	Shutdown	Both
Servers	16.2	16.2	11.4	11.4
% savings		0%	30%	**30%**
CRACs	25.2	23.2	21.4	16.9
% savings		8%	15%	**33%**
Total	41.4	39.4	32.8	28.3
% savings		5%	21%	**32%**

The savings afforded by cooling-aware placement of workloads without shutting servers down is due to the change in the distribution of heat that results in the reduction of recirculation of hot air into the inlet of the racked equipment. Recall that recirculation is a component of LWPI. This recirculation increases air conditioning costs, thus placement alone provides savings primarily in the cost to provision air conditioning resources. Shutting down machines, however, provides both savings at the power delivery level (i.e. power delivered to the computers) and the air conditioning level. The latter is due to the fact that the air conditioning system need not expend the energy to remove the heat formerly dissipated by the inactive computer equipment. When both placement and shutdown are used, added benefit is derived from the fact that clusters of machines in close proximity to each other are shut down as load is compacted to the most efficient places in the data center. These inactive clusters result in zones that can tolerate warmer air than active clusters and the cooling distribution can be adjusted accordingly (e.g. via Dynamic Smart Cooling) resulting in an additional 18% savings in the air conditioning costs from baseline over that achieved by shutting down machines without regard for placement (the shutdown scenario). Indeed, the air conditioning savings of including cooling-aware placement more than doubles the savings of shutdown alone.

5. Discussion

To help convey the practical impact of these results, we work through a simple computation to translate this savings into dollars, and then we discuss issues one may face in practical deployment. Finally, we give a remark on how difficult it is to perform this sort of research on a real, physical data center.

The results indicate that the application of job allocation based on environmental factors can significantly reduce the overall power consumption of the data center. As an example, if we consider a typical large-scale data center with a power consumption of 2.5 MW by the computational equipment (~190 13 kW racks) and a cooling load factor of 2.2 (defined as the ratio of the amount of heat being removed by the amount of power consumed by the air conditioning system to remove the heat), the total power consumption of the data center is 3.6 MW. (Note the load factor of 2.2 matches our experimental conditions and is a conservative assumption given that many data centers operate with load factors much lower than this—i.e. worse.) If we further assume that the data center is partially active 70% of the time per an analysis of data from Figure 1, and that the savings we observed in our experiment (32%) can be extended to the rest of the data center, at an energy cost of $0.15/kW·hr the energy savings will result in an operational savings of more than $1,000,000 per year. Naturally, a rough computation such as this is only an illustration.

Our experiments avoided several complications that may need to be surmounted for practical deployment, of course. One issue is that after servers have been shut down, they must be booted up again when new jobs arrive. It takes only a moment for Dynamic Smart Cooling to provide cooler air to such servers, but the delay of the reboot process is comparatively lengthy and undesirable. Future servers will have fast methods for low power or hibernation modes. Until available, one could trade off some cooling efficiency in order to avoid some of the boot-up delays by (a) never shutting off some of the most cooling-efficient servers, and/or (b) imposing a minimum idleness delay before shutting down any server.

We proposed to place workload either by requiring estimates of job length in advance (which can be difficult to obtain in most general settings), or else by migrating long-running jobs during the wind-down phase via virtualization technology. We believe this migration would be quite practical for most types of CPU-intense batch jobs, with only a sub-second suspension in computation. Note, however, that the job's memory needs to be migrated across the high-speed data center network. If the jobs are very memory intense, or if many migrations are requested in a short time window, the volume of network traffic may begin to pose a substantial cost and delay. Thus, practical controllers may occasionally need to temper their

eagerness for cooling efficiency in order to avoid network overload.

In our experiments, we only considered homogeneous servers, i.e. no matter which server is selected to run a job, the same amount of heat is generated. But over time, real data centers may accumulate a mixture of servers of different generations. Thus, optimal placement decisions may also need to take into account the differing amount of heat generated by different servers. And with widely different CPU speeds, the placement decisions will also affect how long the jobs take to complete. This leads to a complex area of optimization that mixes cooling efficiency considerations with traditional scheduling. Furthermore, economics may play a role: although it may be most efficient to run a user's jobs on a small set of old, slow servers that produce little heat, the user may be willing to pay more for a higher class of service that returns their results sooner at additional expense. These issues are beginning to be explored [4].

References

1. Andrzejak, A., Arlitt, M., and Rolia, J. "Bounding the Resource Savings of Utility Computing Models." HPL Technical Report, HPL-2002-339.
2. Bash, C.E., Patel, C.D., Sharma, R.K., "Dynamic Thermal Management of Air Cooled Data Centers", Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems, San Diego, CA, 2006
3. Bash, C. E., Patel, C., Sharma, R. K., 2003, "Efficient Thermal Management of Data Centers – Immediate and Long-Term Research Needs," International Journal of HVAC & R Research, Vol. 9, No 2.
4. Burge, J., Ranganathan, P., and Wiener, J. "Cost-aware Scheduling for Heterogeneous Enterprise Machines (CASH'EM)." HPL Technical Report HPL-2007-63.
5. Feitelson, D. G., and Nitzberg, B. "Job characteristics of a production parallel scientific workload on the NASA Ames iPSC/860." In Job Scheduling Strategies for Parallel Processing, D. G. Feitelson and L. Rudolph (Eds.), Springer-Verlag, 1995, Lect. Notes Comput. Sci. vol. 949, pp. 337-360
6. Moore, J., Chase, J., Ranganathan, P., Sharma, R. "Making Scheduling 'Cool': Temperature-Aware Workload Placement in Data Centers", USENIX 2004
7. Patel, C.D., Shah, A., "Cost Model for Planning, Development and Operation of a Data Center", HPL Technical Report, HPL-2005-107(R.1)
8. Patel, C.D., Sharma, R.K., Bash, C.E., Graupner, S., "Energy Aware Grid: Global Workload Placement Based on Energy Efficiency", Proc. of the ASME International Mechanical Engineering Congress and R&D Expo Nov 15-20, 2003; Washington D.C
9. Patel, C.D. "Smart Chip, System, Data Center – Dynamic Provisioning of Power and Cooling from Chip Core to the Cooling Tower", Temperature Aware Computing Workshop (TACS), Int'l Symposium on Comp. Architecture ISCA-2005, Madison, WI, 2005.
10. Patel, C. D., Bash, C. E., Belady, C., Stahl, L., Sullivan, D., "Computational Fluid Dynamics Modeling of High Compute Density Data Centers to Assure System Inlet Air Specifications," Proc. of the Pacific Rim/ASME Int'l Electronic Packaging Tech. Conf. and Exhibition (InterPACK), Kauai, HI, 2001.
11. Rolander, N., Rambo, J., Joshi, Y., Mistree, Y., 2005, "Robust Design of Air Cooled Server Cabinets for Thermal Efficiency," Paper IPACK2005-73171, Proc. of the ASME Int'l Electronic Packaging Tech. Conf. & Exhibition (InterPACK), San Francisco, CA, 2005.
12. Rolia, J., Cherkasova, L., Arlitt, M., and Andrzejak, A. 2005. A capacity management service for resource pools. In Proc. of the 5th Int'l Workshop on Software and Performance (Palma, Illes Balears, Spain, July 12 - 14, 2005). WOSP '05. ACM Press, NY, 229-237.
13. Schmidt, R., "Effect of Data Center Characteristics on Data Processing Equipment Inlet Temperatures," Paper IPACK2001-15870, Proc. of the Pacific Rim/ASME Int'l Electronic Packaging Technical Conf. and Exhibition (InterPACK), Kauai, HI, 2001.
14. Shah, A., Carey, V., Bash, C., and Patel, C. "Exergy-Based Optimization Strategies for Multi-Component Data Center Thermal Management," parts I and II, Papers IPACK2005-73137–8, Proc. of the ASME Int'l Electronic Packaging Technical Conference and Exhibition (InterPACK), San Francisco, CA, 2005.
15. Sharma, R.K., Bash, C. E., Patel, C. D., Friedrich, R.J., Chase, J.S. "Balance of Power: Dynamic Thermal Management of Internet Data Centers", IEEE Internet Computing, January 2005