Going beyond CPUs: The Potential of Temperature-Aware Solutions for the Data Center

Justin Moore, Ratnesh Sharma, Rocky Shih
Jeff Chase, Chandrakant Patel, Parthasarathy Ranganathan
Hewlett Packard Labs

Abstract

While there has been a lot of work on temperature-aware architectures at the processor level, their potential at higher levels of the system has largely been unaddressed. In this paper, we propose a temperature-aware *design for data center systems and solutions. This approach shows significant potential both in reducing the costs associated with cooling and in increasing system availability by avoiding failures due to thermal failover. We discuss the high-level architecture specifically focusing on modeling, measurement, metrology, and mechanism and policy design, and present preliminary results from a prototype implementation.*

1. Introduction

Several recent studies have looked at temperature-aware architectures at the processor level [Heo+2003][Srinivasan+2003][Skadron+2003][Brooks+2001][Huang+2000]. However, their potential has largely been unaddressed at higher levels of the system. In particular, at the level of the data center, a temperature-aware design of the systems and solutions architectures can be useful to address the problems of power and heat management in future data centers.

For example, the power consumption of a data center is increasingly becoming a large component of the operational costs of a data center. A significant fraction of this power is often for the infrastructure needed to cool the compute equipment. It is estimated that for every watt of power consumed by the compute infrastructure, another half to one watt is needed to operate the cooling infrastructure [Patel+2001, Uptime2000]. As data centers include more power-dense compute and networking equipment, this problem is likely to be further exacerbated. For a 30,000 square feet data center with 1000 standard computing racks each consuming 10KW, the cooling resources can consume anywhere from 5 to 10 MW of power. At $100/MWhr, this would be a cost of $4-$8 million a year just for cooling. Furthermore, given that the rate of increase in the power density of the data center is outpacing that of the cooling infrastructure improvements, these costs are only likely to increase in the future.

Similarly, from the point of heat management, current data center operators typically expend a large amount of effort to improve operational efficiency and avoid system downtime due to thermal failover. For example, a general rule of thumb is that a typical rack needs to have the air at its inlet temperature be in the range of 20-30 degrees Celsius to avoid thermal redlining. Similarly, every 10 degree increase over 21 degrees Celsius causes long-term electronics' reliability to go down by 50% [Uptime2000]. Additionally, with sensitive compute equipment, the cooling infrastructure also needs to ensure stable relative humidity; otherwise, humidity swings can damage the equipment. The problem is further exacerbated by the potential for increased failure rates of components in the cooling system by virtue of their dependence on moving parts.

Past work addressing these problems have focused on power-aware solutions that optimize for the power consumption of the compute equipment in the server clusters within a data center [Chase2001, Pinheiro2001, Elnozahy2002, Rajamani2003]. However, these optimizations do not focus on, and are typically unaware of, the temperature changes in the data center. Recently thermo-mechanical control to optimize and dynamically provision the cooling resources in a data center has been proposed. A "smart" data center built using flexible air conditioning building blocks and a distributed sensing network directs coolant based on inferred heat dissipation patterns [Patel+2003]. While cooling control can be effective in reducing energy consumption, a complementary approach that provides finer granularity of control with minimal data center physical infrastructure changes is needed.

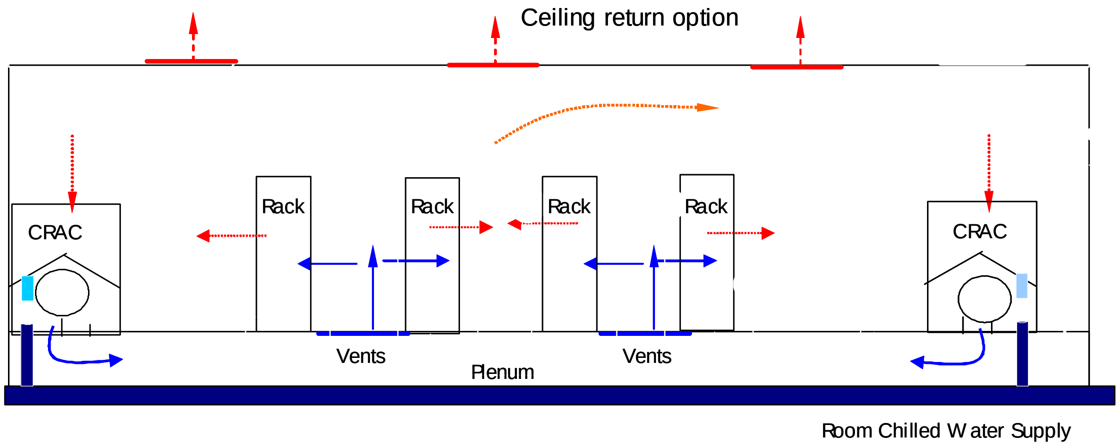

Figure 1: Organization of a typical data center

In contrast to these papers, our work proposes the notion of *temperature-aware design for data center systems and solutions*. Specifically, we propose the design of data center systems and solutions that monitor the temperature of the environment and adapt, at the resource control and workload migration levels, to lower the total costs of ownership (TCO) and improve operational efficiency.

The rest of the paper is organized as follows. Section 2 presents some background on state-of-the-art data center configurations. Section 3 discusses the potential for temperature-aware system design and management for energy reductions and increased uptime. Section 4 discusses the high-level architecture of our solution and addresses the challenges with modeling, monitoring, and metrology and the mechanisms and policies to facilitate temperature-aware data center architecture. Results for a prototype implementation are also presented. Section 5 concludes the paper and discusses ongoing and future work.

2. Background

Data center organization: Figure 1 shows how a typical data center is organized. The compute infrastructure is organized in several rows on a raised plenum, each row consisting of several racks, with each rack including several servers. The aisles between the racks are organized into *hot aisles* and *cool aisles* [Uptime2000] to minimize mixing and increase cooling efficiency. The servers in the individual racks are organized so that their inlet faces the cool aisles and their outlet faces the hot aisles (assuming sideways air flow). The cool aisles typically have vented tiles through which cool air is circulated.

The cool air is provided by CRAC (computing room air conditioning) units or air-handling units (AHU). These CRAC units take in the re-circulated exhaust hot air and cool the air over a refrigerated or chilled water cooling coil to approximately 10-17C. The air movers in the CRAC unit pump the cold air into the plenum through which it is sent to the cool aisles through the vented tiles. Some data centers may additionally include a ceiling return plenum to further avoid mixing of hot and cold air [Patel+2003]. In many data centers, the cooling units are connected to a separate chiller plant and cooling tower that provides the cooling through liquid-to-liquid heat exchange (or the condenser may be a roof-top heat exchanger).

Cooling provisioning/thermal failure avoidance: The power consumption associated with the cooling infrastructure typically constitutes one third to half the total power consumption of the data center. This includes both the “flow work” (associated with moving the air in the room) and the “thermodynamic work” (associated with change in phase and heat transfer) at various components of the system – at the chiller plant and at the CRAC units [Patel+2003].

The cooling is typically provisioned to provide adequate cooling for a factor over and above the worst-case scenario. The worst-case cooling needs are often computed based on the nameplate heat dissipation of the compute equipment with a suitable de-rating metric for safety. To avoid downtime due to thermal redlining, a data center technician or operator manually observes the temperature in various aisles

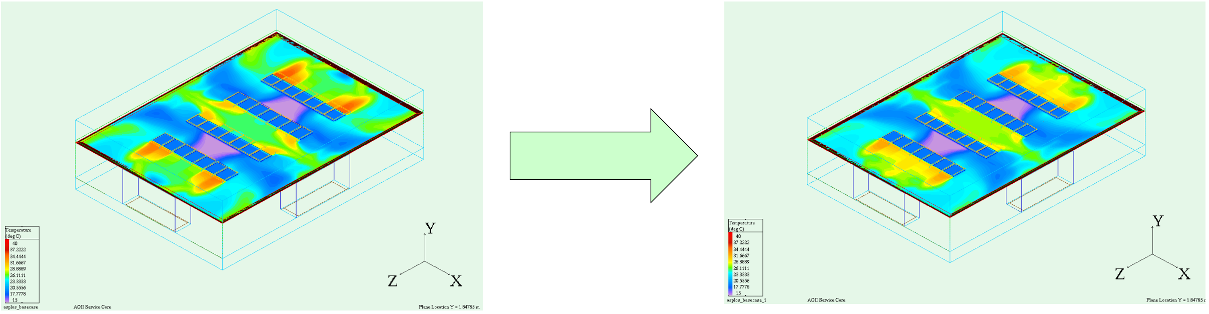

Figure 2: Temperature-aware data center optimizations for reduced energy costs

and flags potential local hotspots. In the event of such local hot spots in the environment, local parameters (e.g., tile vent positioning) are first modified and if the hot spot still persists, global parameters such as CRAC unit supply temperature and air flow rates are modified. The latter optimizations once again lead to gross overprovisioning of the cooling. CRAC units have their own control, but these are often based on a temperature sense point at the intake to the unit and so the responses are again biased by aberrant peaks in the system.

3. The Potential of Temperature-Aware Data Center Architectures

Our temperature-aware data center architecture monitors the temperature of the environment and adapts, at the resource control and workload migration levels, to lower energy costs and improve availability during failure scenarios.

3.1 Reducing cooling costs

A data center has, for any given layout and capacity of the cooling infrastructure, a thermal profile inherent to that organization. Essentially, at a spatial level, some locations in the data center can "tolerate" greater heat dissipation than others because of the nature of the air flow and cooling capacities; similarly, some locations produce greater heat than others because of the nature of the workload and resource properties. For a given set of workloads and compute resources, an intelligent resource provisioning and workload deployment algorithm that matches the heat production to this thermal profile has the potential to consume the lowest amount of energy in the cooling infrastructure.

Figure 2 shows the potential of temperature-aware data center optimizations to reduce the energy costs associated with cooling. The pictures represent the thermal maps based on modeling the computational fluid dynamics of a conventional data center, using a CFD tool [Flovent from Flomerics Corporation]. These results are extracted from experiments conducted by Sharma et al. [Sharma+2003] for energy aware workload distribution. The data center modeled is 11.7mx8.5mx3.1m and has a 0.6m deep raised-floor plenum. The compute racks are 40U high (1U=45mm) and contain 20 servers. There are seven racks per row and four rows in the data center for a total of 560 servers. The servers are modeled to consume about 75% of their 600W nameplate power rating, and provide a 15C increase in temperature with a volumetric flow rate of 34 liters/sec. Four CRAC units are located at each wall, and are modeled as heat extraction devices with characteristic outlet temperatures of 15C and limiting cooling capacities of 90kW each.

Figure 2a (on the left) shows the contour plot of temperature at a height of 1.85m above the floor. Aisles 1, 3, and 5 (numbering starting from the end of the room) are hot aisles and aisles 2 and 4 are cold aisles. As can be seen from the figure, though the workload distribution (and consequently, the power and heat dissipation) is assumed to be uniform, this does not lead to a corresponding uniform distribution of the temperature. There are several "hotspots" indicated by the regions in red, and several "cold spots" indicated by the regions in darker green and blue. Within each aisle, there is a significant variation in the temperature at the inlets and outlets. For example, at the two aisles in the ends of the rooms, there are severe hot spots at the two ends of the rows and relatively cold spots in the middle of the rows. The maximum air temperature in the data center was 36.5C.

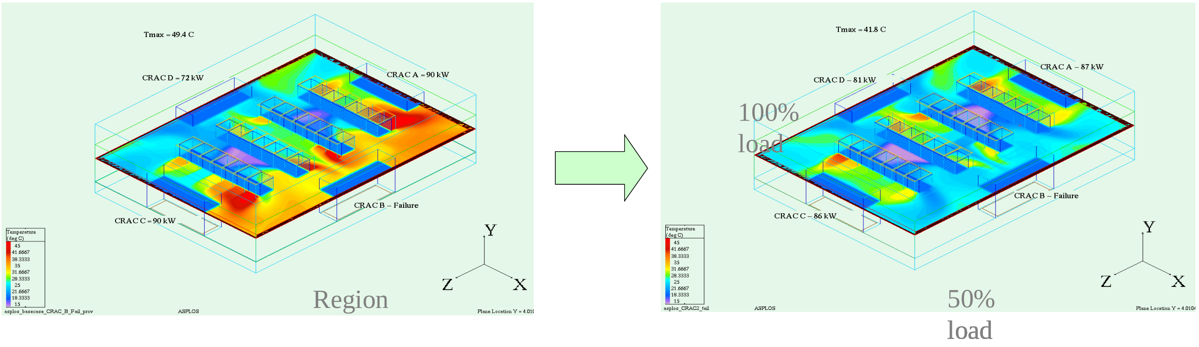

Figure 3: Temperature-aware data center optimizations for increased availability

Figure 2b (on the right) shows the contour plot for the same data center after applying temperature-aware optimizations to reduce cooling energy costs. In this case, the power consumption of each rack was modified to better match the thermal profile of the room. Workload placement was carried out based on thermo-fluids policy derived from CFD-based thermo-fluids model of the data center. Higher power consumption loads were moved to the cooler portions of the data center (middle of the racks) and lower power consumption loads were moved to the hotter portions of the data center (end of the racks). The specific redistribution was based on a simple heuristic that considered the inlet temperature at each rack in a row. The power dissipated in the i^{th} rack of the row (denoted by P_i) is assumed to be inversely proportional to the excess temperature rise associated with the thermal profile of that rack.

As can be seen from Figure 2b, the temperature-aware power redistribution can significantly change the thermal contour profile of the data center. The temperature distributions are significantly more uniform compared to those in Figure 2a. In particular, compared to the more than 10C variation in the hot aisle temperature in the earlier case, the variation in the hot aisle is now close to 2C. The absolute temperature values are also much lower with the maximum temperature of the entire data center reducing to 32.4C compared to the earlier 36.6C. The lower maximum temperature means that the CRAC units can be driven at a much higher air discharge temperature (in this case 18C instead of 15C) to obtain the same performance with respect to system thermal redlining avoidance. Increase in the CRAC unit air discharge temperature can lead to an increase in the co-efficient of performance of the refrigeration system. All these combine to achieve close to a 25% reduction in total energy costs associated with the cooling infrastructure.

3.2 Improving availability and system uptime

A data center is often vulnerable to overloading of the cooling infrastructure due to mechanical cooling unit failures or from supply/demand spikes due to higher workload usage or hot weather.

Figure 3 shows the thermal profile of our example data center during one such instance. In this case, we assume the one of the CRAC unit fails. As seen from Figure 3a, the temperature around the failed unit increases rapidly (often within 60-90 seconds), leading to rack inlet temperatures as high as 49.4C. This leads to thermal redlining and consequent failover of the first three racks of all the rows. The loss of compute power for the workloads running on these 240 servers can be quite disruptive.

Figure 3b shows how the system would respond in the context of a temperature-aware data center design. The results show the scenario when, on detection of the CRAC unit failure, part of the workload is moved from the first three racks closest to the failed unit to the racks farthest away. As can be seen from the Figure, this leads to a comparatively much better thermal distribution, and barring a few minor hot spots, all servers continue to operate in non-redlined situations. The maximum temperature in this case is almost 8C lower than the maximum in the previous case.

Temperature-aware data center designs that allow the movement of power dissipation (through either workload or resource control) enable much faster responses to thermal catastrophes than with corresponding approaches that use changes to mechanical parts such as changing vent tile flows, etc.

4. Architecting a Temperature-Aware Data Center Solution

We can divide the key challenges with architecting a temperature-aware data center solution into the following categories – (1) measurement and monitoring, (2) metrology to determine metrics that capture thermal capacity, (3) mechanisms and policies for resource control and workload migration that enable dynamic temperature-based optimizations.

Measurement and metrology: An important challenge in enabling temperature-aware optimizations at the data center level is the ability to measure environmental conditions and correlate them with the relevant computational parameters. Past work on coordinated monitoring and control of large scale computing infrastructures have traditionally focused on IT-level metrics such as CPU utilization, etc. Such tools need to be extended to include environmental sensors, physical location and spatial and topological relationships with respect to support systems such as cooling and power distributions. Additionally, given that the number of sensors is often an order of magnitude higher than the number of computing nodes, the infrastructure has to scale to much higher levels than previously required.

Metrology: The second aspect of the design has to with the metrology involved in defining relevant metrics that will allow us to quantify the effectiveness of our management policies. These metrics should allow us to create a thermal profile of the data center and guide us on where to place load – and the corresponding heat – within the data center. These metrics need to adequately capture the thermodynamics in a typical data center including effects such as mixing and short-circuiting as well as consider the interactions between compute resource utilization, power consumption, and the need for heat extraction. Additionally such metrics need to be easily computable from measurable real-world parameters.

Mechanisms and policies: The final aspect of the design is the mechanisms and policies for resource control and workload migration to enable heat migration. This can leverage the huge body of work on mechanisms and policies for distributed resource management.

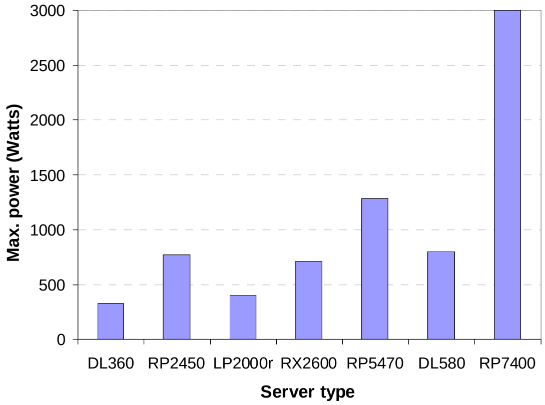

Figure 4: Heterogeneity in prototype data center

For example, at the mechanism level, this can include flexibility in the individual system elements. For example, processor power states (on/off, voltage scaling, etc.) can provide a simple mechanism to match the power consumption required at a certain location. Other such optimizations include memory bank control, disk spin down for storage, and being able to use system power states as defined by standards like ACPI.

For legacy data centers that typically include a heterogeneous collection of servers, the diversity in the power-performance space offered by the heterogeneity can offer a mechanism for power control as well. For example, Figure 4 shows the variation in nameplate power for the different classes of servers in the prototype data center. The first four classes of servers (DL360, RP2450, LP2000r, and RX2600) represent dual-processor servers based on IA32, PA-RISC, and IA-64 processors. The RP5470 and DL580 are four processor machines based on PA-RISC and IA32 processors and the RP7400 is a 8-processor SMP based on PA-RISC processors.

At the software level, techniques such as virtual machines (e.g., vmware, Xen), process migration (e.g., Zap), service migration, and request redirection (e.g., TCP handoff, Linux virtual server) can be used to direct workloads to individual systems to better match to better optimize the power distributions to match the thermal profiles. At the policy level, the system can implement a control algorithm based on a range of options from simple scheduling heuristics to more complex control theory-based or market-based

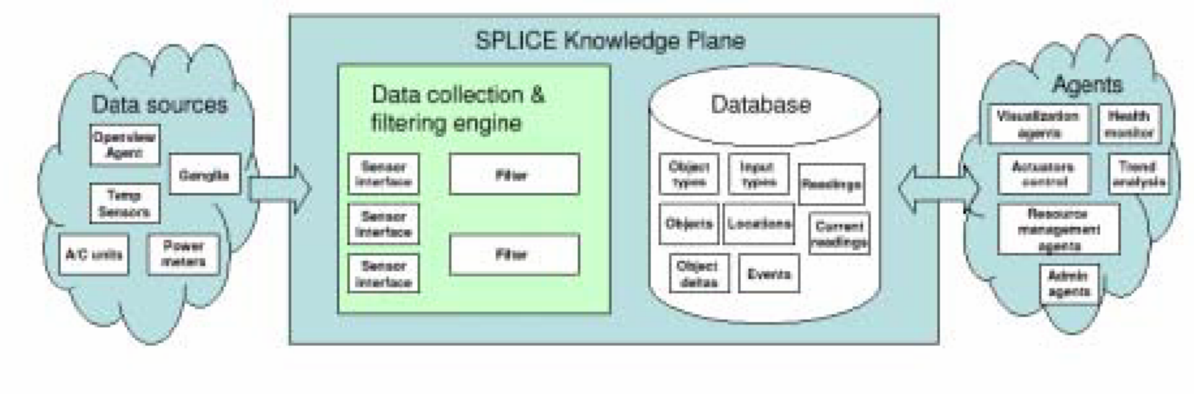

Figure 5: Splice measurement and monitoring infrastructure

implementations.

4.1 Prototype Details and Results

Measurement: In our prototype data center, we collect data from a rich monitoring infrastructure that includes a variety of sensors [Patel+2003]. We place sensors on racks to measure the servers' supply and exhaust air temperatures, in the aisles to observe the room-wide three-dimensional temperature distribution, at the CRAC units to measure temperatures at the air return and supply vents, and in the air distribution plenum to measure pressure. Power meters at each rack measure the power consumption. In addition to the power and temperature sensors, we log performance data for traditional metrics such as CPU performance, memory consumption, disk utilization, network bandwidth, etc, using HP OpenView.

A key element of our work is a database engine that filters sensor data and stores it in a relational database that supports a standard SQL query interface. This component – called Splice [Moore+2003] – combines all our readings and normalizes them to a common spatial and temporal frame of reference. Figure 5 presents an overview of the Splice architecture. Splice interfaces with a variety of heterogeneous data sources and implements a filtering engine to provide contextual compression. The database schema is developed to be scalable to a large number of sensors and also provides higher-level object views that include location and other topological relationships. The rest of our control system interfaces to the Splice infrastructure interfaces as an agent through the database.

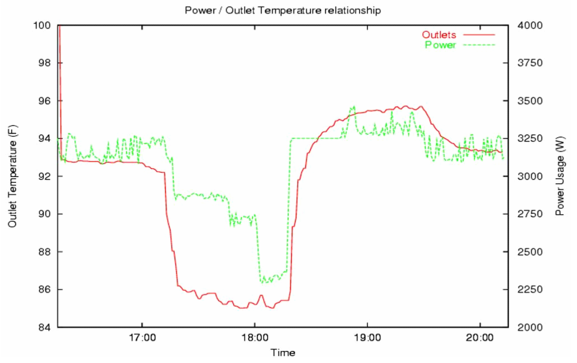

Figure 6: Correlated power and temperature

Figure 6 illustrates how Splice enables automatic correlation of resource usage with power consumption, heat dissipation and temperature changes. For example, an increase in the power consumption leads to a corresponding increase in the outlet temperature.

Metrology: For our prototype implementations, we define a dimensionless metric called the *Heat Density Factor* (HDF). This parameter is based on principle of the thermal multiplier θ_i formulated by Sharma et al.[Sharma+2003]. This metric builds on the results from numerous computational fluid dynamic analyses [Patel +2001] and thermo-fluids experiments conducted on data centers with diverse thermo-mechanical architectures and compute load distribution [Sharma+2001]. The energy efficiency of data center is a function of the quantity and location of the heat generated by our compute infrastructure and the complex air flow patterns. The complex airflow patterns that cause thermal imbalances are a function of thermo-mechanical architecture and are difficult to modify on a case-to-case basis. If the heat density of a location exceeds the cooling infrastructure's ability to remove that heat, the warm air will diffuse and mix with the cold air streams headed for the servers' air intake fans.

HDF for object *I* is defined as:

$$HDF_I = \frac{T_{avg} - T_{ref}}{T_I - T_{ref}}$$

where T_{avg}, T_I and T_{ref} are the average outlet (exhaust), object *I* outlet and CRAC unit supply temperatures, respectively. The average exhaust temperature can be for a single rack, a row of servers, or the whole data center. HDF can be calculated for individual servers, racks or data centers and is inherently scalable. HDF will indicate what the power consumption should be at object *I* in relation to the average power consumption for the reference object(s).

To calculate the HDF for a data center, we run a balanced (homogeneous) load on each server and allow the system to attain equilibrium. We then choose a "reference point"; our experience indicates that the average exhaust temperature and average power consumption over all the servers in the data center provides the best reference point. We then calculate the HDF for each server and populate a three-dimensional model of the data center with these values. This enables our control algorithm to define areas of arbitrary size – such as rows, racks, or three-dimensional volumes ("pods") – and calculate the optimal power allocation for that area. Our experiments show that HDF values obtained from different baseline utilization levels – i.e., all servers idle, all servers using one processor, and all servers using two processors – are identical, provided that the overall utilization does not exceed the cooling capacity of the data center.

Resource control: We studied a media rendering workload and implemented a batch scheduler that allocates workloads in a temperature-aware manner to reduce cooling costs. Our scheduler implements policies based on the HDF values determined above and factors in the discrete system states of the servers at idle and full CPU utilization.

Scheduler heuristic	**Savings over "worst" case**
Thermal-ceiling-discretize	0%
FIFO-row-based	2.3%
Thermal-row-level-bias	17.9%
Thermal-even-discretize	19.1%
Thermal-poaching	*25.0%*
Thermal-analog-best	**29.5%**

Figure 7: Prototype scheduler results

Figure 7 summarizes our results. The column on the left indicates the heuristic used and the column on the right represents the cooling savings possible from a worst-case schedule. For our experiments, the worst-case scheduling happened with a heuristic that used the HDF values to determine a ceiling-based schedule of the load between idle and full-power utilizations. A simple row-based scheduler that schedules all the load on the first set of rows presents the same kind of thermal imbalances as the worst-case schedule (within 3%). The three schedules based on intelligent use of the thermal policies do better, achieving 18% to 25% savings in the cooling costs. The thermal-row-level-bias heuristic tries to schedule the power to the middle rows while the thermal-even-discretize heuristic tries to evenly discretize the power consumption based on the analog HDF distribution. The best algorithm is the thermal-poaching algorithm. This heuristic tries to first approximate HDF distributions at a higher region level and then uses a systematic process of underweighting and overweighting servers around recent heat allocations. This achieves within 86% of the cooling savings attainable from a "best-case" analog HDF redistribution of the load.

5. Conclusions

As power and heat continue to be increasingly important challenges in the design of future systems, it is important to extend the notion of temperature-aware architectures beyond processors to higher levels of the system. In this paper, we explored the benefits from temperature-aware data center solution designs.

Based on CFD modeling of thermo-fluids in a representative data center, we showed that such temperature-aware designs can actually be quite beneficial in reducing the cooling costs incurred in a data center. For example, a simple heuristic to map the power consumption in a manner that was inversely proportional to the temperature distribution profile of the data center obtained a 25% reduction in cooling energy costs. Additionally, temperature-aware designs also enabled rapid responses to thermal failover situations and provide graceful degradation in such cases.

We also discussed our ongoing implementation of a temperature-aware design. Our experience indicates that the design needs to address the challenges of providing (1) a rich monitoring, measurement and data aggregation infrastructure, (2) a formal metrology-based approach to design and evaluate metrics that best capture the interactions between the system parameters, their power consumptions, and their impact on the environment, and (3) a rich feature set of mechanisms that provide resource control and workload migration coupled with a powerful control-based policy engine to implement the temperature-aware adaptivity.

Using these components, our design is able to match workloads to resources to best match the ensuing power consumption to the cooling capacities of the system. We are currently in the process of deploying our design on a prototype experimental test bed and our preliminary results with real-life workload traces are promising. As part of future work, we also plan to study the interactions of such temperature-aware optimizations with conventional power-aware optimizations and the interactions of both of these in the context of broader resource management existing in current data centers.

7. References

[Heo+2003] S. Heo, K. Barr, and K. Asanovic, Reducing Power Density through Activity Migration, ISPLED'03, 2003

[Srinivasan+2003] J. Srinivasan and S.V. Adve, Predictive Dynamic Thermal Management for Multimedia Applications, ICS'03, 2003

[Skadron+2002] K. Skadron, T. Abdelzaher, and M. R. Stan, Control-Theoretic Techniques and Thermal RC Modeling for Accurate and Localized Dynamic Thermal Management, HPCA 2002

[Huang+2000] A Framework for dynamic energy efficiency and temperature management, M. Huang and J. Renau and S.-M. Yoo and J. Torrellas, Proceedings of the 33rd Annual International Symposium on Microarchitecture, 2000

[Brooks2001] Dynamic thermal management for high-performance microprocessors, D. Brooks and M. Martonosi, Proceedings of the 7th International Conference on High Performance Computer Architecture (HPCA), January 2001

[Patel+2001] Patel C.D., Bash C.E, Belady C, Stalhl L, Sullivan D, Computational fluid dynamics modeling of high computer density data centers to assure system inlet air specifications." Proceedings of IPACK'01 (The Pacific Rim/ASME International Electronics Packaging Technical Conference and Exhibition), 2001

[Uptime2000] Alternating cold and hot aisles provides more reliable cooling for server farms, R. F. Sullivan, Uptime Institute, 2000

[Pinheiro+2001] "Load Balancing and Unbalancing for Power and Performance in Cluster-Based Systems", Eduardo Pinheiro, Ricardo Bianchini, Enrique Carrera, Taliver Heath, Proceedings of the Workshop on Compilers and Operating Systems for Low Power, September 2001.

[Chase+2001] Managing Energy and Server Resources in Hosting Centers by Jeff Chase, Darrell Anderson, Prachi Thakar, Amin Vahdat, and Ron Doyle. In 18th Symposium on Operating Systems Principles (SOSP), October 2001.

[Elnozahy+2002] Energy-Efficient Server Clusters, Mootaz Elnozahy and Mike Kistler, and Ram Rajamony, in Proceedings of the Second Workshop on Power Aware Computing Systems, Feb 2, 2002

[Rajamani+2003] K. Rajamani and C. Lefurgy, "On Evaluating Request Distribution Schemes for Saving Energy in Server Clsuters," Proc IEEE International Symposium on Performance Analysis of Systems and Software, 2003

[Patel+2003] Patel C.D, Bash C.E, Sharma R.K, Beitelmal M, Friedrich R, "Smart Cooling of Data Centers," Proceedings of IPACK '03 (International Electronic Packaging Technical Conference and Exhibition), 2003

[Flovent] Flovent version 2.1, 1999, Flometrics Ltd, 81 Bridge Road, Hampton Court, Surrey, KT8 9HH England

[Sharma+2003] Balance of Power: Dynamic Thermal Management for Internet Data Centers, Sharma, Ratnesh K.; Bash, Cullen E.; Patel, Chandrakant D.; Friedrich, Richard J.; Chase, Jeffrey S., Hewlett Packard Technical Report HPL-2003-5, 2003

[Moore+2004] A Sense of Place: Towards a Location-aware Information Plane for Data Centers, Justin Moore, Jeff Chase, Keith Farkas and Parthasarathy Ranganathan, Hewlett Packard Technical Report TR2004-27.

[Sharma+2002] Sharma R.K., Bash C.E., Patel C.D., "Dimensionless parameters for evaluation of thermal design and performance of large-scale data centers," Proceedings of the 2002 AIAA (The Eighth ASMEE/AIAA Joint Thermophysics and Heat Transfer Conference), 2002