Oracle White Paper

"MegaWatt-Hours Avoided"

Kenny Gross & Kalyan Vaidyanathan
Oracle Quality Division
Oracle San Diego Physical Sciences Research Center
3/7/2015

This White Paper answers 2 questions for data center architects and for computing industry Sustainability Technologists:

(1) How are the most energy-efficient data centers, those with the very latest HVAC/CRAH systems and newest server designs, wasting MW-Hrs that are not helping with computing goals in any way (and wasting even more in older data centers)?

and

(2) **How does the new Energy-Aware Data Center (EADC) portfolio of Oracle innovations provide a new (but well proven) telemetry infrastructure that measures, monitors, and eliminates the MW-Hrs of (previously not visible) energy wastage?**

ABSTRACT:
Data center thermal/power control strategies to date have never had visibility into the thermal dynamics and energy dynamics inside the IT server assets. As such, data center ambient temperatures are presently controlled without insight into substantial energy wastage mechanisms that are present inside all enterprise server assets and storage assets that contain spinning hard disk drives (HDDs). These parasitic energy-wastage mechanisms include:

(1) nonlinear power draw from the server fan motors
-- for many enterprise servers these days, the fan motors use more power than the CPUs, and much more than the memory; moreover, the aggregate fan motor power goes up with the *cubic power* of the fan RPMs
(2) nonlinear "leakage power" inside the CPU chips
--leakage power is completely wasted (doesn't help with the computing), and goes up *exponentially* with chip temperature
(3) substantial energy wastage due to ambient vibrations in the servers (from internal fans and blowers) and the metal racks housing the servers (from nearby AC equipment, PDU internal fans, and flow-induced vibrations from fluid and fluid/air chillers).

The above mechanisms are present and waste substantial energy even in the newest and most energy-efficient data centers. (By "wasting" energy: none of the above parasitic energy-wasting mechanisms contributes to the customer's computing objectives in any way). Moreover, the heat generated by the foregoing nonlinear wastage mechanisms imposes a double penalty on the data center owner/operator: Paying once for the wasted power, and then paying again to remove that additional wasted heat from the assets. This white paper documents the sources of parasitic energy wastage in modern data center assets, and then shows how Oracle's proven Intelligent Power Monitoring (IPM) telemetry can be seamlessly integrated with the HVAC controls so that now the envelope of Intelligent Power Management will

encompass the chip level through the box level, rack level and through the HVAC systems. The resulting "CPU to CRAC" intelligent thermal-aware and energy-aware real-time power optimization minimizes large, nonlinear energy wastage mechanisms inside the data center assets, mechanisms to which conventional data center HVAC controls have been completely blind, thereby achieving global energy optimization across the data center. This well-proven IPM telemetry technology (already "baked in" for hundreds of thousands of Oracle servers, and addable via a software patch for heterogeneous legacy systems in data centers) enables "MegaWatts Avoided" in any modern (or legacy) data center.

Introduction:

Every data center that the US Govt operates to date, just like most data centers in the world, control the Computer Room Air Conditioner (CRAC) systems to deliver a desired target ambient temperature at the inlet grilles of the server and storage assets. This is the single most wasteful aspect of present data center cooling (and at the same time, the simplest to eliminate, via the IPM solution documented below that requires no hardware additions in the data center).

When the HVAC systems are operated to meet a target inlet temperature at the inlet grilles for the servers, that inlet temperature is established with worst-case conservatism assumptions by the server vendors: i.e. with the assumption that every CPU chip, every DIMM memory module, and every internal HDD, are all running maxed out (at maximum thermal dissipation rates). Present data centers implicitely use this worst-case conservatism because they have had no way of knowing what the real thermal flux or power flux are inside the IT servers.

Inside the servers, there is substantial conservatism in the "thermal head room", i.e. the difference between the allowable temperatures for internal components and the actual real-time temperatures of those components. The thermal head room margins (THMs) are very conservative to ensure there will be sufficient cooling under worst-case scenarios, i.e. assuming the server has a maximal internal configuration (CPUs, all memory slots full, maximum disk drives and IO cards), that the customer is running the maximum possible workloads for CPU, memory, and IO, and that the data center may be at a high altitude (where the air is thinner and has less cooling capability). Of course 99% of the server assets in the world don't meet those conservative assumptions. By "closing the loop" between the HVAC controllers and Oracle's comprehensive, accurate, internal system telemetry, we are able to safely collapse those very large and wasteful headroom margins for data centers that are not at high altitudes and are running whatever real workloads the customer is running (vs assuming maxed-out workloads 24x7). "Closing the loop" with the HVAC controls saves substantial energy and, unlike chip power management strategies presently in vogue with enterprise server vendors (Dynamic Voltage & Frequency Scaling, memory throttling) the Intelligent Power Monitoring (IPM) solution outlined in this paper has no offsetting performance penalties.

Avoiding Energy Wastage from MisGuided ASHRAE Guidelines:

Note that even without directly connecting real-time THM metric info for intelligent control of CRAC setpoints, the principles outlined in this white paper permit very simple "offline" computation of the near-optimal ambient temperature setpoints for data centers. This simple computation alone can minimize substantial wastage of energy in data centers that have followed the well-meaning but now outdated thermal guidelines of the American Society of Heating, Refrigerating & Air-Conditioning Engineers. ASHRAE has for many years recommended that data center owners save energy by reducing air conditioning and warming up the data center. For the first 25 years of the computer industry, there was no relationship between the operating temperature of enterprise computing servers and their energy efficiency. But for the most recent four or five years, this is no longer the case. Extensive Oracle research has demonstrated that with the latest generations of enterprise computing servers there are now

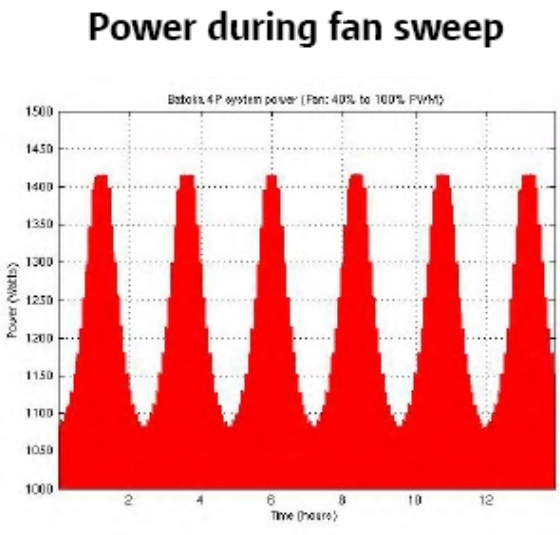
very temperature-sensitive "energy wastage" mechanisms in IT systems that not only waste significant energy in warm data centers, but also degrade compute performance. Oracle has developed and patented new temperature-aware algorithms that enable intelligent optimization of data center ambient temperatures to minimize or avoid these heretofore non-observable energy wastage mechanisms (discussed in detail below). Oracle's suite of "Energy Aware Data Center" (EADC) algorithms predict an optimal ambient temperature set point, decreasing energy wastage across the data center, significantly increasing overall compute performance, decreasing the carbon footprint for the data center owner/operator, while increasing return-on-assets for the IT server and storage systems throughout the data center.

Nonlinear Energy Wastage Mechanisms Inside Enterprise Servers: Fan Motor Power and CPU Leakage Power

To cope with Moore's Law and remove ever-increasing heat burdens from enterprise servers, fan motors inside servers have become increasingly powerful and increasingly numerous. Even the smallest servers in the data centers now incorporate multiple fans, and larger servers have very many fans. Server internal power supplies are now coming with internal fans, and each server contains multiple power supplies. Consequently it is not uncommon for larger servers to have 15-20 fans inside. For many enterprise servers today, the aggregate fan motor power is greater than the CPU power, and much greater than the memory power. The bigger problem is that for all mechanical fans inside enterprise servers, the fan motor power increases with the cubic power of the fan RPMs. (See example, Fig 1). If the fan RPMs double (which is well within their operational range), ***the fan motor power goes up by a factor of eight***. This means that as internal computing workloads go up, and/or as the temperature of ambient inlet air rises, the server fans consume very much additional power, and this is power that does not contribute to the customer's computing requirements. Moreover, the waste heat from the fan motors creates an additional burden that the CRAC units must remove from the data center.

The power dissipated by modern computer chips comprises two components: switching power and leakage power. Up until just 4-5 years ago, enterprise computing chips consumed power for only for "switching power" (the flipping of gates as computations are performed). Switching power is “well behaved” and is money well spent, insofar as the switching power is directly proportional to the compute work being done by the CPUs. Leakage power (which is considered wasted power because it does not support the computational workload on the CPU) has traditionally been negligible. In the most recent enterprise servers, because of relentless miniturization inside the computing packages, leakage power has become significant, using up to 40% of the power budget for recent servers and projected to continue growing for future generations of servers. This leakage power is not well behaved at all, and in fact it increases exponentially with the temperature of the CPU chips. (See example, Fig. 2). Consequently, when the ambient temperature is cool enough to keep the CPU chips lower than the threshold for leakage power, all power consumed by the CPUs directly supports the customer's computational demands. As soon as the ambient temperature rises to the point where the CPU chips enter the "leakage zone", there is a wasted energy component, called the leakage component, which grows exponentially with further increases in ambient temperature.

Fig 1. Fan power rises with cubic power of RPMs



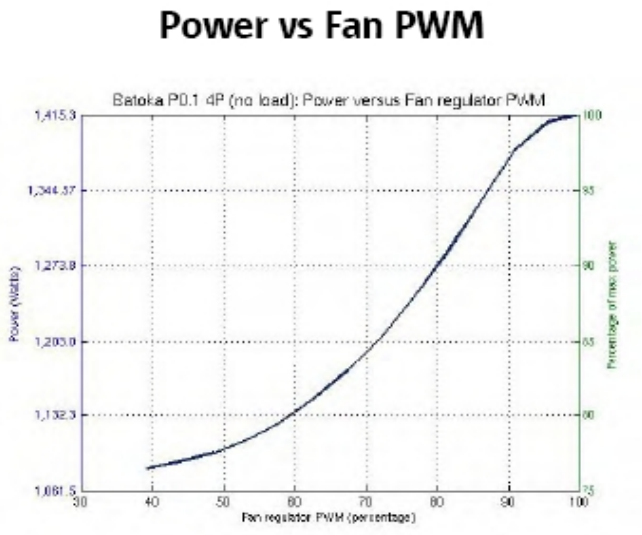
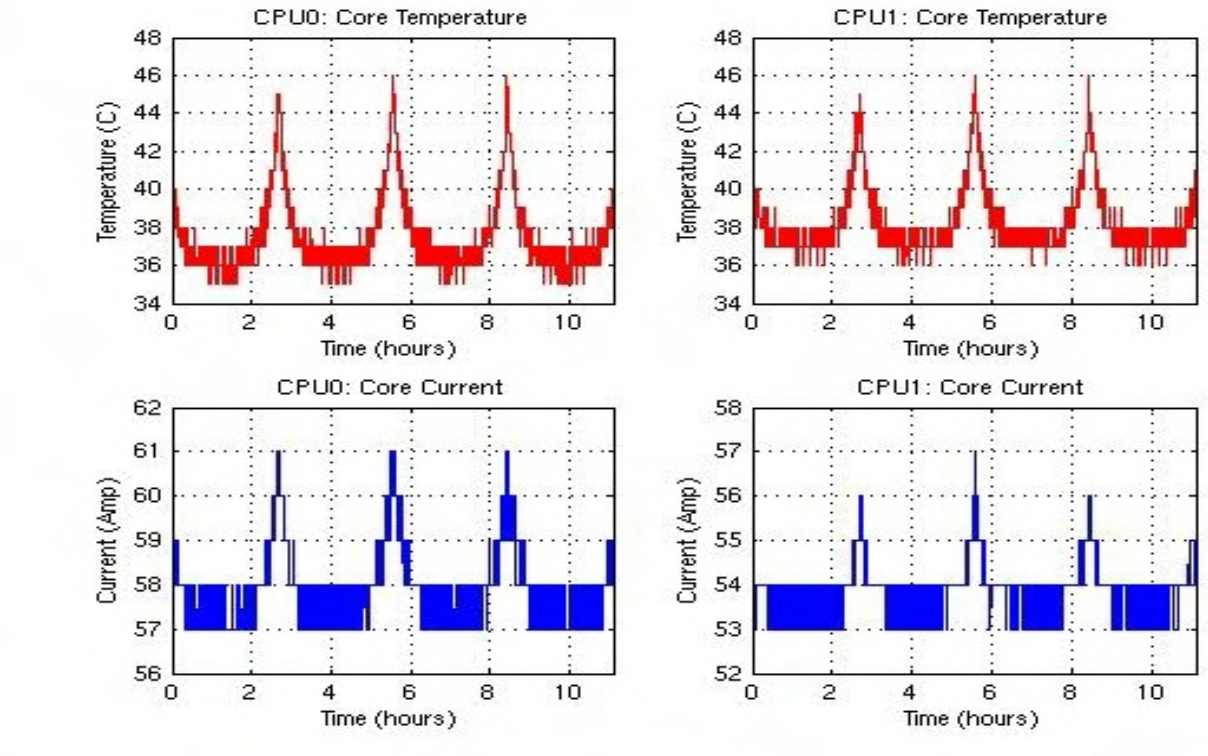


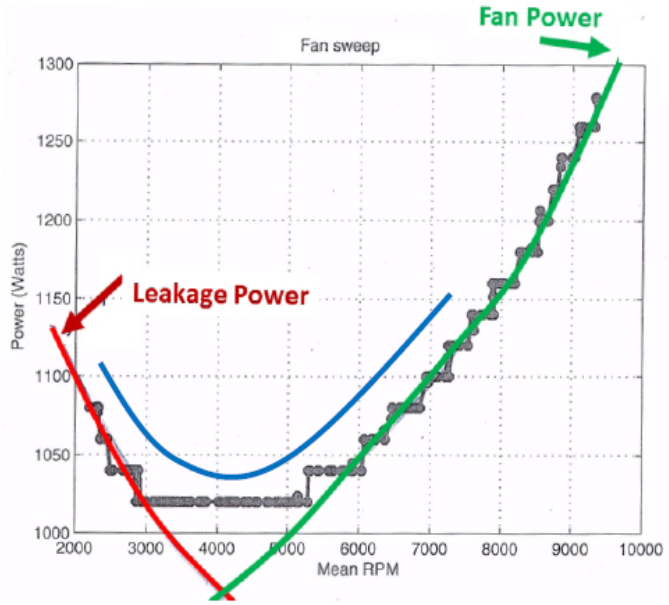
Fig. 2

CPU leakage current: Exponential with CPU Junction Temp (Data shown: System Idle, Temp variation from small fan speed variation)



Because of the foregoing nonlinear physics processes going on inside the enterprise servers, there is a three-way relationship between ambient air temperature outside the server, fan motor power, and CPU leakage power. By integrating Oracle's patented Intelligent Power Monitoring telemetry from inside the servers with the CRAC controls outside the servers we are now able to continuously "seek and settle" at the minimum of the "V-shaped" energy function: On one side of the "V" is the cubic relationship between fan speed and power. On the other side of the "V" is the exponential relationship between chip temperature and leakage power. The resulting "CPU to CRAC" intelligent thermal-aware and energy-aware real-time power optimization minimizes large, nonlinear energy wastage mechanisms inside the data center assets, mechanisms to which conventional data center HVAC controls have been completely blind, thereby achieving global energy optimization across the data center. See Fig. 3, below.

Figure 3: Optimal Server Fan Control



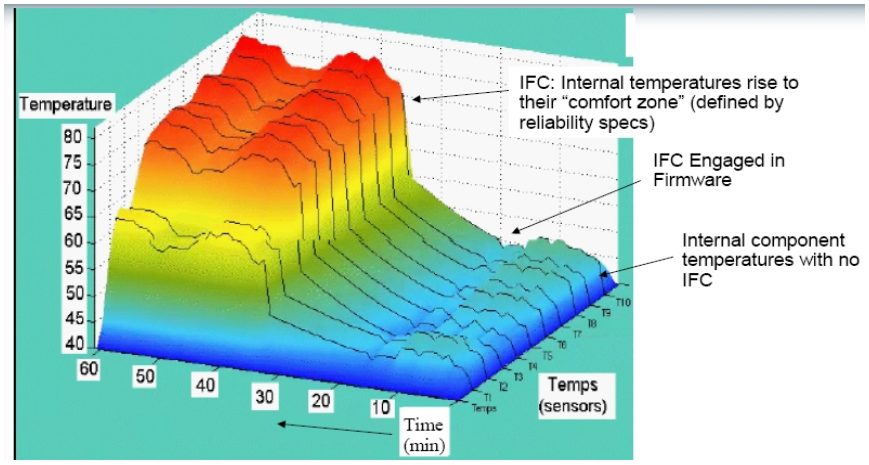
Continuously "seeks and settles" at optimal energy minimum, balancing cubic fan power vs exponential leakage power

Attains optimal server energy without penalizing server performance (as does CPU frequency scaling, clock gating, and DIMM memory throttling).

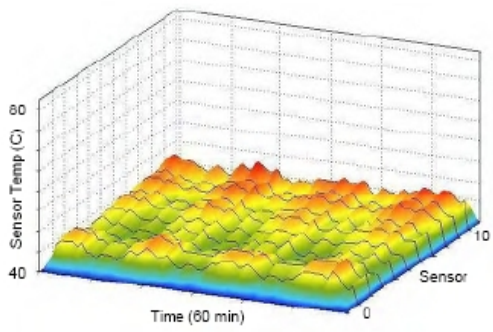
*Oracle SDCC Patent:"Optimal Fan Controller for Eco-Efficient Computer Servers," U.S. Patent 8,108,697 (Jan 31, 2012).

Continuous internal system telemetry enables Intelligent Fan Control to collapse excessive thermal headroom margins the industry has heretofore been blind to, mitigating wasted energy from fan motors, lowering vibrations (which grow linearly with fan RPMs) and acoustics (which grow with the 5th power of fan RPMs).

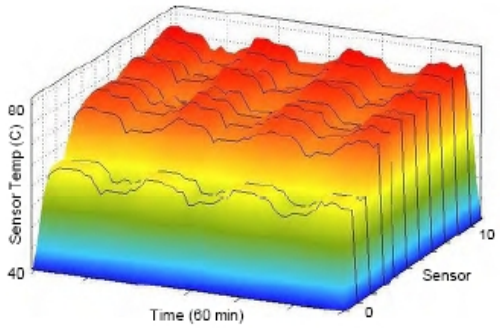
Fig. 4 IFC collapses the traditionally very conservative thermal headroom margins safely, keeping all components within their optimal reliability zone:



Intelligent Fan Control



- Without Intelligent Fan Control
  - Fans run at fixed nominal speed
  - System Power 1,300W



- With Intelligent Fan Control
  - Fans adjusted by IFC
  - System Power 1,000W

Growing Vibrational Challenges in the Data Center:
Vibrations Directly Impact and Degrade Server Energy Efficiency

For commodity hard disk drives (HDDs), areal densities have been growing exponentially at a rate faster than Moore's law. The write head on HDDs is now trying to hit a tiny track that is only 20 nanometers wide (1/1000th the width of a human hair) while floating just 7 nanometers above the surface of a platter is spinning at up to 15,000 RPMs.

Low level vibrations inside storage servers/arrays can significantly degrade IO throughput. Examples in slides below.

The fans in each new generation of server systems are more powerful to cope with Moore's law. As servers are deployed in a metal rack with other servers that also have variable speed fans, ambient vibrations throughout the rack rise.

Customers are bolting supplementary Liebert and APC air conditioners right onto the metalracks. Additional vibration sources in Power Distribution Units (PDUs), and flow-induced-vibration (FIV) in fluid cooled cabinets and fluid/air heat exchangers.

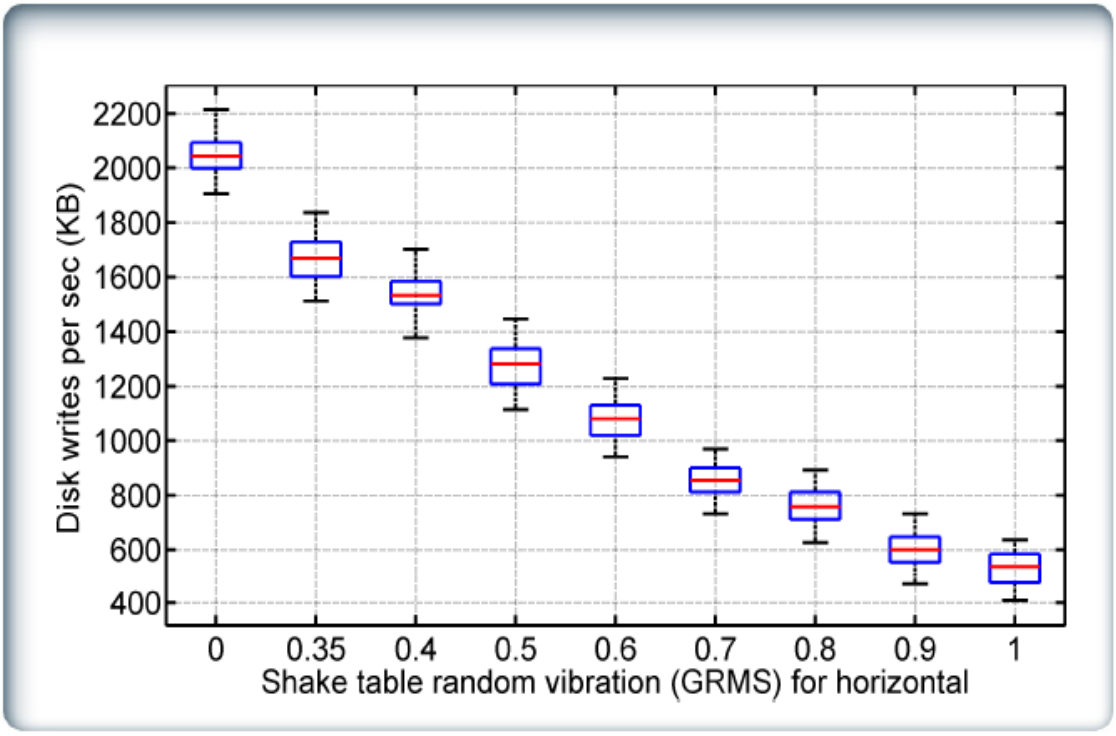
Ambient vibrations have a significant impact on performance when servers are running IO-intensive workloads, which include important enterprise applications such as Oracle OLTP, Data Warehousing, web serving, video streaming, CRM, and ERP processing.

Because of the extreme sensitivity of present generation HDDs to low level vibrations, the total integrated power (i.e. energy) needed to complete a fixed customer workload (e.g. updating a Multi-TB database) goes up as the ambient vibration levels go up.

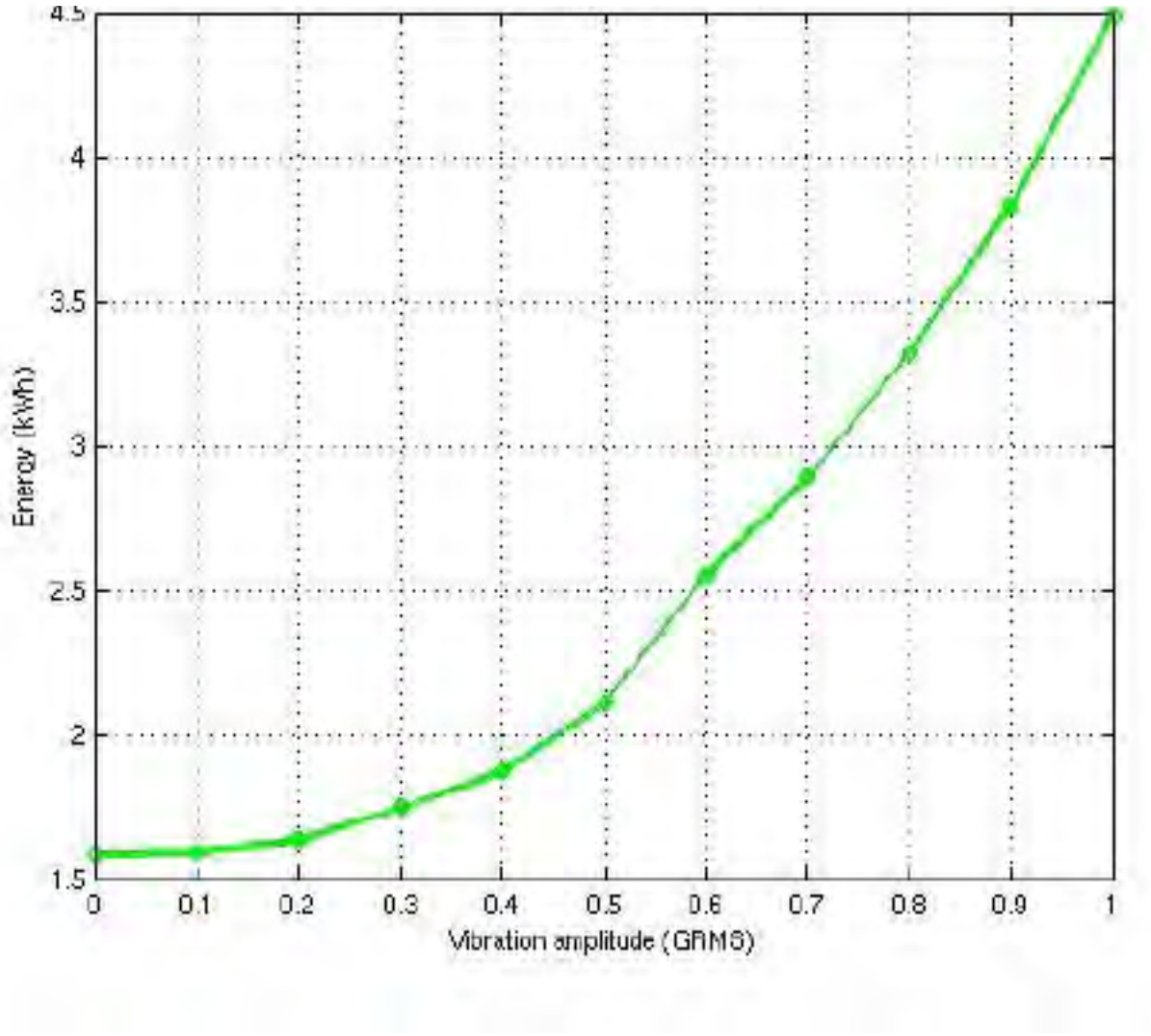
For example, if IO performance degrades by 20% due to elevated ambient vibrations for a server in a metal rack full of other servers, then it takes 25% longer for the customer workload to complete.

This means that all the components inside the server are consuming power for 25% longer (fan motors, memory, HDDs, ASICs, IO cards, and PSUs).

Fig. 6: Typical Storage Array Disk IO Throughput vs Ambient Vibration



Energy Consumed Updating 10TB Database vs Server Environmental Vibration



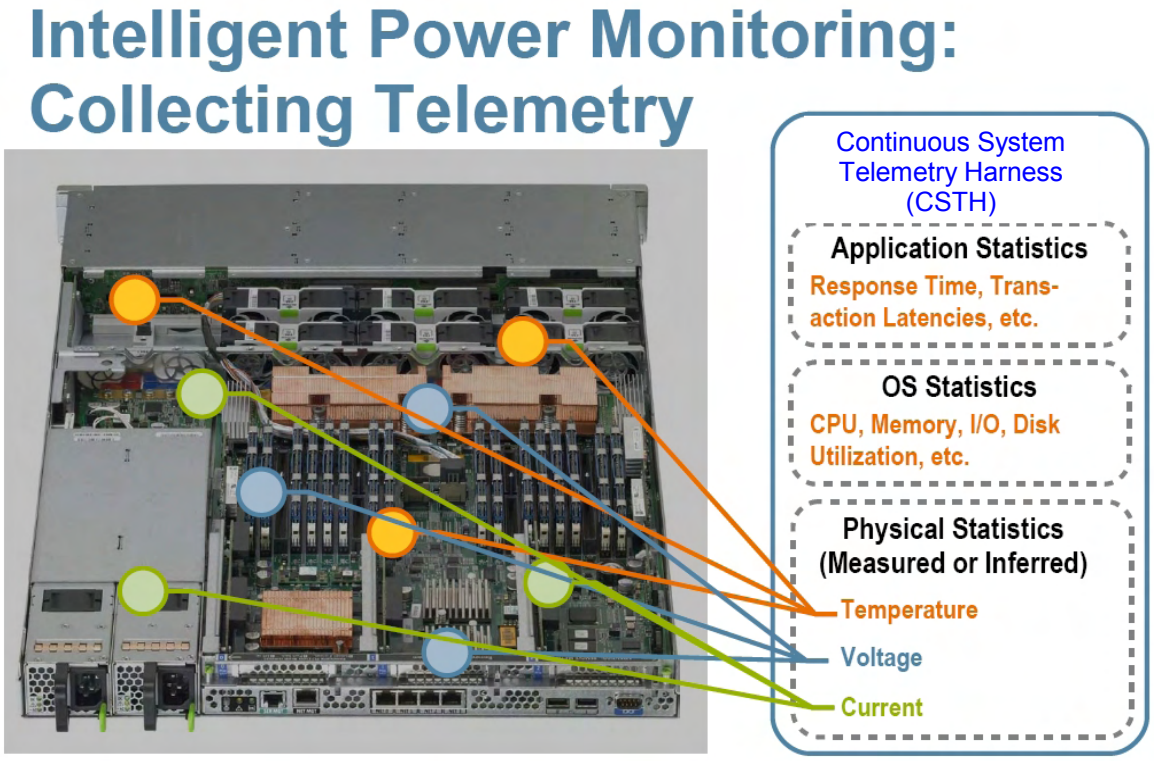
Further Reading:
Oracle Publication:
"Fan-Speed-Aware Scheduling of Data-Intensive Workloads," *Proc. Int'l. Symp. on Low Power Electronics and Design (ISLPED12),* Redondo Beach, CA (Jul 2012).

Oracle's Intelligent Power Monitoring (IPM) software

-- Monitors temperatures, voltages, currents, fan speeds, throughout interior of enterprise servers

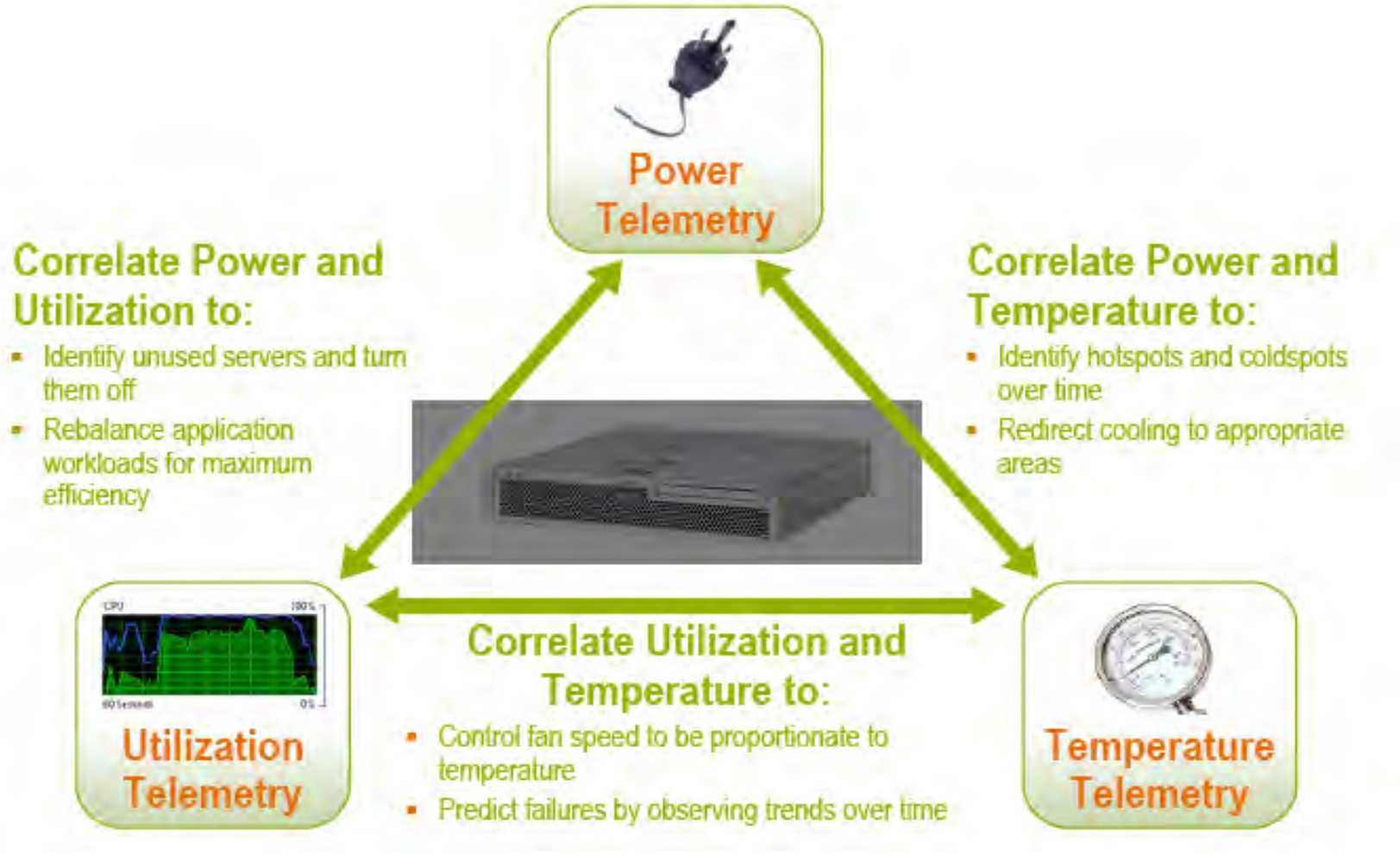
-- Provides continuous, accurate, 3D thermal contours in all dynamically executing servers and storage arrays

-- "Close The Loop" between internal system component temperatures and external HVAC controls



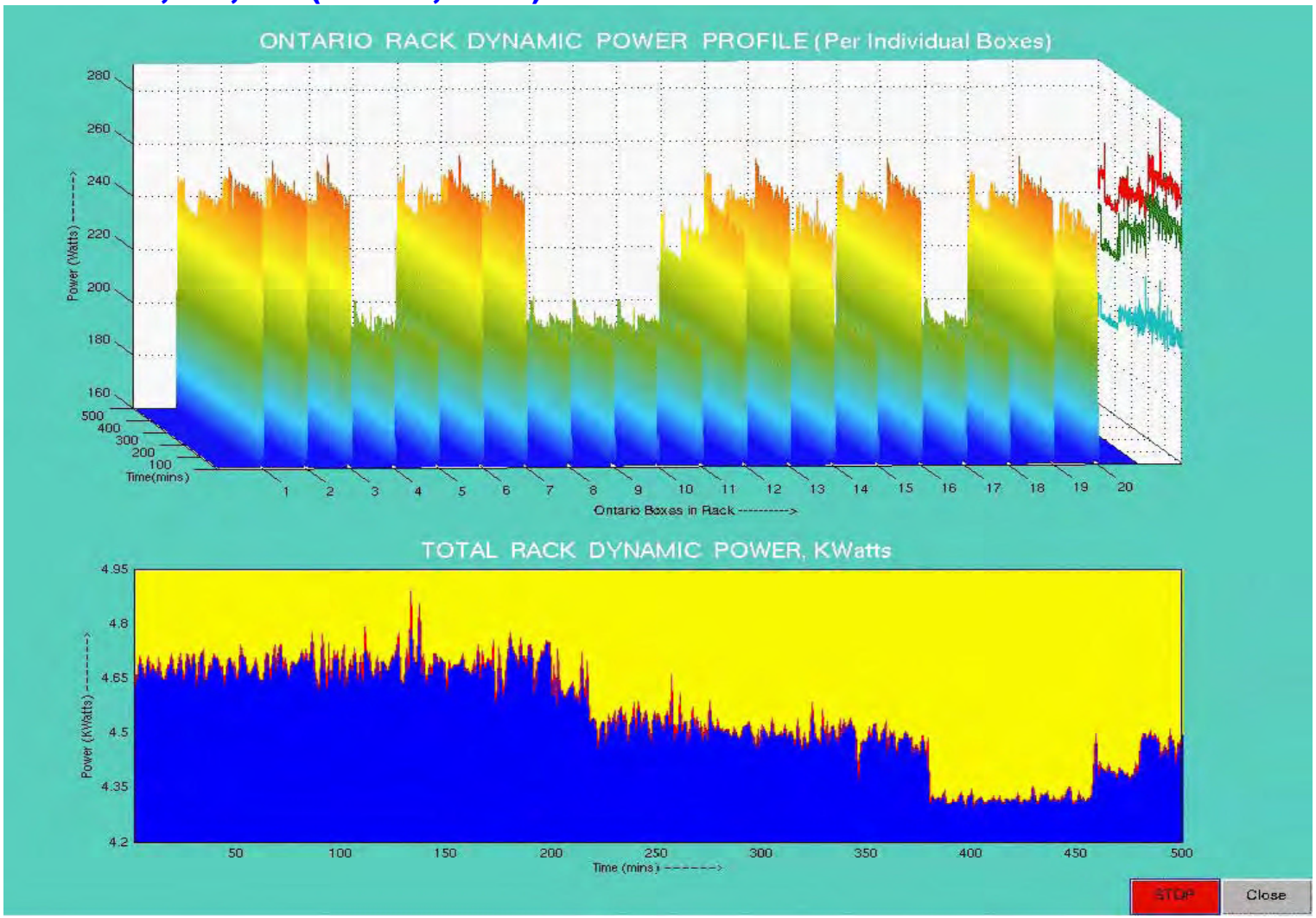
Actionable Power Savings from IPM Service

Insight from Correlated Data



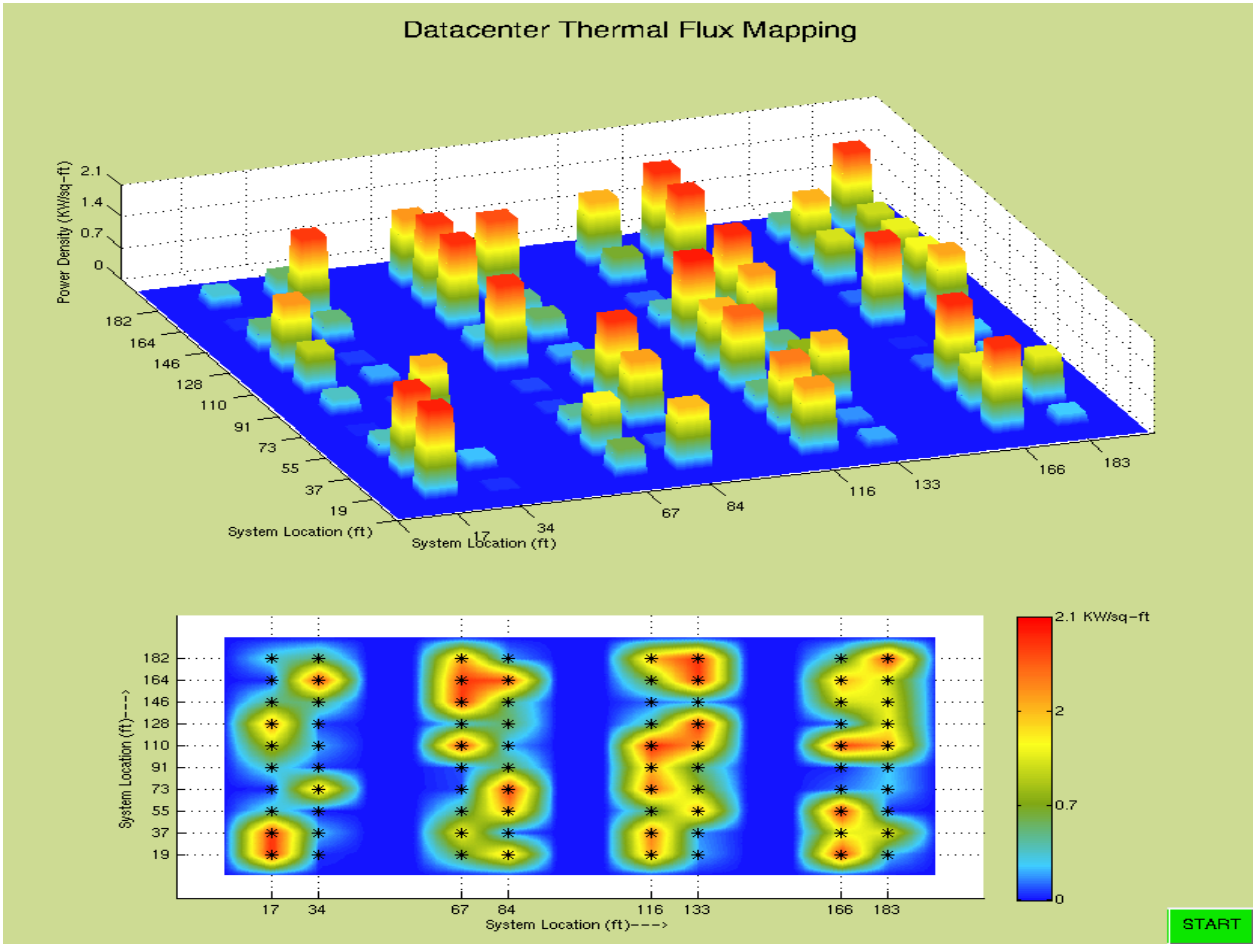
Real Time Thermal Profiling: Rack of Servers
No Hardware Power Meters / PDUs Required

Oracle Patent: "Real Time Power Harness: Power Monitoring for Computers via Telemetry," U.S. Patent # 7,191,411 (Mar 27, 2007).



Data Center Thermal Flux Mapping

Continuous real time dynamic thermal flux and power flux inside the server assets.

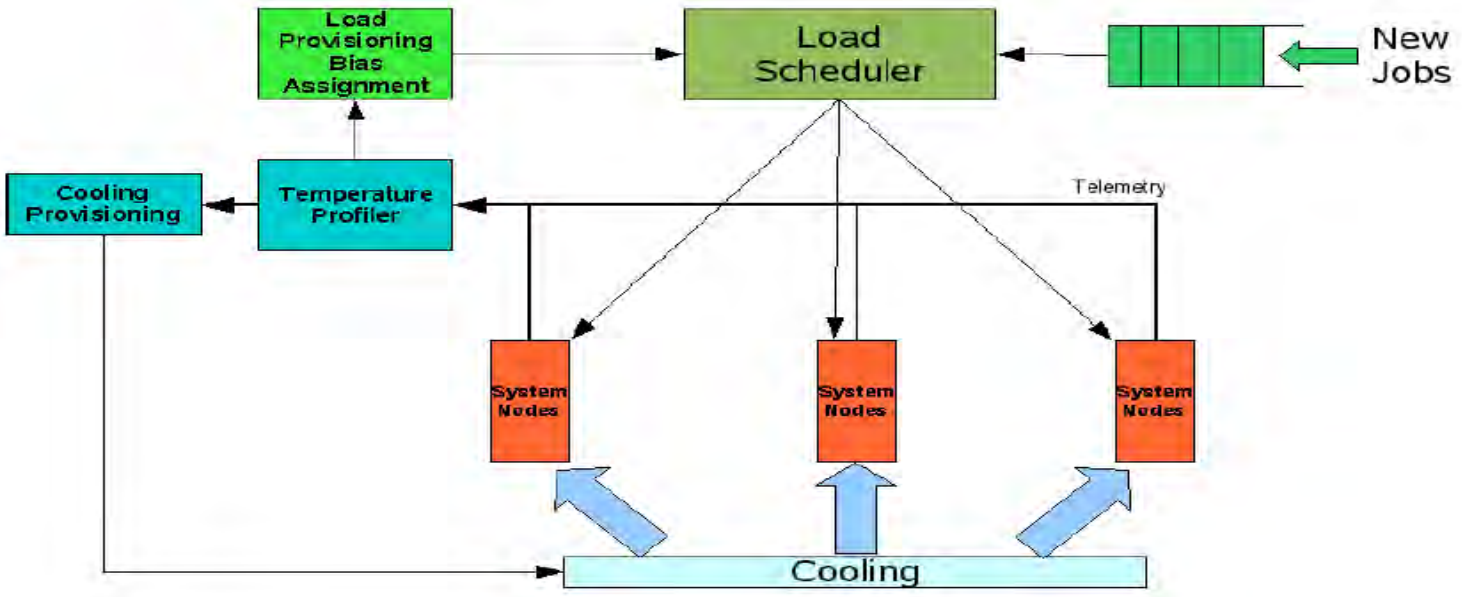


*Oracle Patent: "Datacenter Spatial Thermal Flux Mapping Via Telemetry," U.S. Patent # 7,549,070 (June 16, 2009).

Double Dynamic Provisioning (DDP)

At the rack/datacenter level, DDP provisions load preferentially to the cool spots via (thermal-aware virtualization workload mobility) and cooling preferentially to the hot spots.

Optimal energy efficiency while minimizing spatial and temporal thermal gradients (resulting in improved long-term reliability of data center assets).



*Oracle SDCC Patents: "Optimized Workload Scheduling For Improved Energy Utilization and Reliability for MultiCore Chip Servers," U.S. Patent 7,716,006 (May 11, 2010).
"Double Dynamic Provisioning Method and Apparatus for Optimal Datacenter RAS and Energy Utilization," Case ID SUN061298 (Oracle patent pending).
Telemetry-Enabled Energy-Aware and Temperature-Aware Scheduling for Optimal Server Energy Utilization and Reliability," Case No. SUN070443 (Oracle patent pending).

Eco Irony:

Eco efficient data center design calls for getting the air conditioning systems as close as possible to the sources of heat. Consequently AC's are being bolted onto the tops and sides of metal racks. IO performance decreases with vibration levels, taking longer to complete workload, consuming more energy and generating more heat for the AC's to remove.

Increasing Vibrational Chaos:

Lieberts with powerful 8-inch fans bolted to tops of metal racks.

In same datacenter, vertical APC AC's are bolted to sides of metal racks.

CRAC units are being located in the middles of DC rows/aisles.

Conclusion:

ASHRAE thermal guidlines advocating warmer data centers to save energy are well meaning, but now-days are very mis-guided and counter productive (except for data centers in year-round cold climates where it becomes possible to leverage free air cooling but without warming up the data center). With newer (i.e. introduced after 2011) enterprise server and storage technology, warming up the data center significantly impacts performance and wastes energy. For any Oracle customers contemplating warming their data center to 90 degF (or higher...new ASHRAE guidelines are proposing 113 degF), system performance and energy efficiency will be severely degraded. Use of Oracle's patented Intelligent Power Monitoring (IPM) innovations to optimize the setpoint for ambient inlet air temperature minimizes heretofore invisible parasitic energy wastage mechanisms (i.e. invisible in the sense that if one hooks up a power meter to the IT server, there is no insight into these mechanisms). ASHRAE guidelines call for warming up data centers to the highest ambient temperature levels allowable by hardware reliability guidelines. Doing so will result in very low PUE ratios for the data centers, but the fallacy in associating PUE ratios with energy efficiency lies in the fact that for all recent server and spinning-storage technology, warming up the data center will severely degrade both CPU and IO performance, while degrading IT energy efficiency by 50%-80% in terms of "work done per energy consumed). By cooling the data center to a value that minimizes the insidious penalizing energy-wastage mechanisms that are present in all air-cooled enterprise server and spinning-storage systems today, substantial energy is saved, significantly increasing overall compute performance, decreasing the carbon footprint for the data center owner/operator, while increasing return-on-assets for the IT assets throughout the data center.

External References

Recent refereed papers demonstrating the effects of the three "invisible" energy wastage mechanisms as data centers are warmed up: (1) CPU Leakage, (2) Aggregate Fan Motor Power, and (3) System Ambient Vibrations

"Simware: A Holistic Warehouse-Scale Computer Simulator," Georgia Tech Journal Paper, IEEE Compute Journal (0018-9162/12), 2012. *[Journal paper shows that as data center temperature rises, fan motor power in the IT assets rises significantly, wasting energy and distorting the facility PUE.]*

"Effects of Data Center Vibration on Compute System Performance," Julius Turner, ***Proceedings of USNIX Workshop on Sustainable Information Technology***, San Josa, CA (2010).

"Fan-Speed-Aware Scheduling of Data-Intensive Workloads For Optimal Performance and Energy Efficiency of Data Center Assets," C. S. Chan, Y. Jin, Y. K. Wu, K. C. Gross, K. Vaidyanathan, and T. S. Rosing, ***Proc. Int'l. Symp. on Low Power Electronics and Design (ISLPED12)***, Redondo Beach, CA (Jul 2012).

"Correcting Vibration-induced Performance Degradation in Enterprise Servers ," C. S. Chan, B. Pan, K. C. Gross, K. Vaidyanathan, and T. S. Rosing, Proc. ***ACM 2013 Intn'l GreenMetrics Conference***, Pittsburg, PA (Jun 2013).

"Data Center Vibration and Commodity Server Performance Analysis," B. Factor, A. Youssefi, C. Bahar, and G. Malek-Madani, Technology White Paper, Green Platform Corp (2012). www.greenplatformcorp.com/literature.html

"Impact of DVFS on n-Tier Application Performance," GA Tech and Fujitsu Laboratories, ***Proc. of ACM Conf. On Timely Results in Operating Systems (TRIOS'13)***, Nov 2013.

"Data Center Cold Aisle Set Point Optimization," Rubenstein and Faist, Microsoft Corp, ***Proc. IEEE ITHERM Conf***. (June 2014).

Demonstration of extreme sensitivity of modern spinning-storage systems to tiny vibration levels: 2-min video demonstration: www.youtube.com/watch?v=tDacjrSCeq4

"Some Like it Hot – Some Like it Cold," ***Industry Technology Public Debate***, Silicon Valley Leadership Group Data Center Efficiency Summit, San Jose, CA (Nov 2014) http://svlg.org/policy-areas/energy/data-center-efficiency-summit/2014-data-center-efficiency-summit/agenda Oracle Slide Deck by K. Gross and K. Vaidyanathan: http://svlg.org/wp-content/uploads/2014/11/WarmDataCentersWasteEnergy.pdf