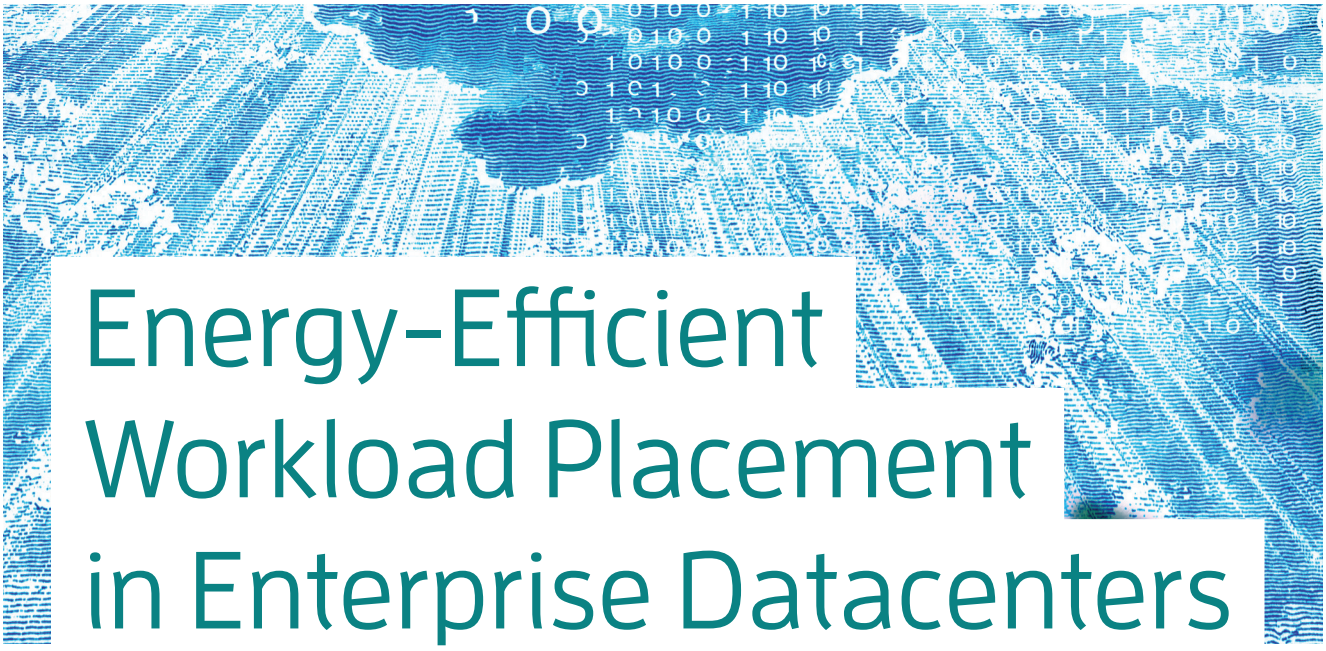

Energy-Efficient Workload Placement in Enterprise Datacenters

Quan Zhang and Weisong Shi, Wayne State University

Power loss from an uninterruptible power supply can account for 15 percent of a datacenter's energy. A rack-level power model that relates IT workload and its power dissipation allows optimized workload placement that can save a datacenter roughly $1.4 million in annual energy costs.

Rising electricity costs are making energy efficiency a critical concern for datacenters, which devour massive amounts of energy annually. According to the Natural Resources Defense Council, US datacenters expended 91 billion kWh of electricity in 2013, with a projected increase to 140 billion kWh annually by 2020 (www.nrdc.org/energy/data-center-efficiency-assessment.asp). At these consumption rates, energy costs will continue to be a major contributor to a datacenter's total cost of ownership (TCO).

To improve datacenters' energy efficiency, researchers and practitioners have focused on reducing IT equipment power consumption, which is typically 30 percent of a datacenter's energy cost. The most popular approaches apply dynamic voltage/frequency scaling (DVFS) to reduce power dissipation from the CPU and memory subsystem;[1,2] consolidate servers by assigning tasks to fewer servers and shutting down idle ones;[3,4] or evenly allocate workloads among servers through load balancing.[5,6] Other approaches are based on hardware-resource use, such as subsystem power models for specific computer components,[7–9] and system power models for nonvirtualized and virtualized environments.[10–12]

Reducing IT equipment power consumption certainly has merit, but these strategies ignore another large contributor to energy cost: power losses from an uninterruptible power supply (UPS), which account for an additional 15 percent of overall energy cost.[13] To address this area, we created a rack-level power model that maps workload directly to its power dissipation and formulated a mathematical solution that chooses an optimal workload allocation to minimize IT equipment power consumption and power loss from UPSs. Using a TCO model, we then analyzed potential electricity cost savings.

Our experimental results show that the rack-level power model precisely matches measured power, with an error rate of ±2.5 percent or less. For a datacenter that hosts 50 racks (1,000 servers) with 10 applications, our simulation showed a potential power savings of up to 5.2 percent relative to that with a uniform workload allocation. This percentage translates to $1.4 million in annual

 0018-9162/16/$33.00

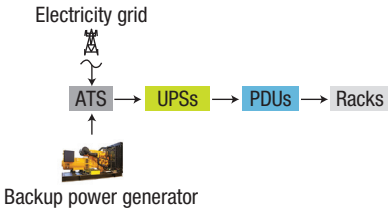

FIGURE 1. Simplified power flow in a typical datacenter. At the highest layer, the utility power and backup power, such as a diesel generator, pass through uninterruptible power supplies (UPSs) through an automatic transform switch (ATS) and go through power distribution units (PDUs) to different racks.

TABLE 1. Uninterruptible power supply (UPS) power loss with two workload distributions.

UPS configuration		Workload distribution 1			Workload distribution 2		
		Loaded capacity (%)	Power loss (W)	Total loss (W)	Loaded capacity (%)	Power loss (W)	Total loss (W)
N+1	Rack 1	87.50	1,094	2,047	74.30	1,026	2,079
	Rack 2	54.50	952		67.75	1,053	
2*N*	Rack 1	43.75	1,863	3,432	37.15	1,723	3,407
	Rack 2	27.25	1,569		33.88	1,684	

energy cost savings for a 76-MW datacenter with a power-usage effectiveness (PUE) of 1.7.

UPS AND ENERGY USE

Figure 1 shows a simplified power flow in a typical datacenter. From the racks, power is distributed through strips to individual servers, all of which have their own power supplies.

Because a UPS represents a single failure point, datacenters often use redundant UPSs in both centralized and distributed topologies to ensure that hosted services are always available. A redundant configuration can be a single *N* system, comprising one UPS module, or multiple *N* systems, comprising parallel modules whose capacities are matched to the critical-load projection. A centralized topology typically deploys UPSs at the facility level; a distributed topology deploys them at the rack or server level.[14] The choice of configuration depends on a datacenter's failure frequency.

Two popular redundant configurations are parallel, or *N*+1, and system-plus-system, or 2*N*. An *N*+1 redundant configuration consists of parallel, same-size UPS modules, and the spare power is at least equal to the critical-load capacity. A 2*N* redundant configuration—the most reliable and expensive design—can tolerate every conceivable single failure point.

Each configuration has a unique power-loss behavior. In an *N*+1 configuration, power loss decreases when the IT power load increases; in a 2*N* configuration, power loss increases as IT power load increases. Thus, for a rack-level UPS configuration, neither fewer servers running at full speed nor more servers running slower with uniform workload distribution will always save power because lowering UPS output load leads to lower conversion efficiency.

This observation about power-loss behavior was foundational to our work. Enterprise datacenters generally run fewer applications—sometimes only one across the entire datacenter. Google's datacenters, for example, run Web 2.0 and software as a service (SaaS). In these cases, a single datacenter has a large workload, no virtualization, and tens of thousands of physical servers. Thus, workload placement is critical in separating the often millions of user requests across racks.

ANALYZING UPS POWER LOSS

In a double-conversion UPS, power loss occurs when power transforms from AC to DC for battery storage and again from DC to AC for delivery to racks and servers. Power loss is also tied to UPS topology.

Our focus is on power loss in a rack-level distributed UPS topology, where power loss depends on UPS efficiency and redundant configuration choice, and loaded capacity (real-time power load) depends on IT workload. For an *N*+1 configuration, loaded capacity varies from 0 to 100 percent; for a 2*N* configuration, maximum loaded capacity is only 50 percent, as the total power load is evenly allocated to two UPSs. UPS efficiency depends on the technology used.

To gather evidence that optimal workload distributions for various UPS configurations differ, we looked at data from two racks in the Wayne State University datacenter; one rack had 20 fully loaded servers and the other had 20 idle servers. We then collected data from two workload distributions:

- distribution 1 had 20 fully loaded servers on the same rack and all idle servers on the other rack; and
- distribution 2 had 12 fully loaded servers and 8 idle servers on one rack and the remaining mix of 8 loaded and 12 idle servers on the other rack.

Table 1 shows the UPS power losses of different UPS configurations and workload distributions. For the *N*+1 configuration, distribution type 1 has lower power losses; for the 2*N* configuration, distribution 2 has lower power losses.

Figure 2 shows a UPS efficiency curve based on data we collected from our two workload distributions. Typically, lower UPS efficiency leads to higher UPS power losses, but in datacenters, IT equipment dictates loaded capacity and thus UPS power loss.

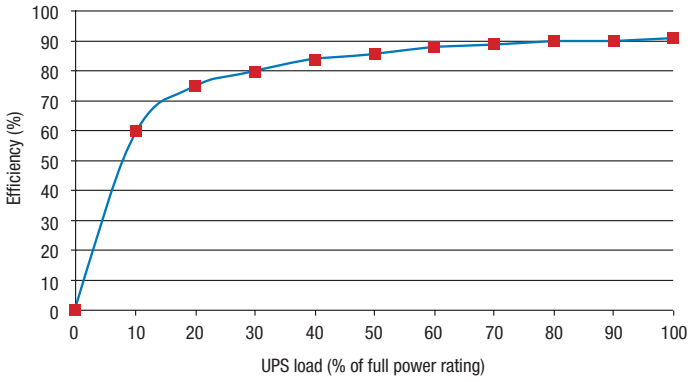

FIGURE 2. UPS efficiency curve based on data collected from Wayne State University's datacenter. To find the relationship between IT equipment power and UPS power loss, we used a UPS with a power rating of 8 kW for a rack with 20 servers. All servers had the same measured peak power of 350 W. The idle power of 20 servers was 4,366 W.

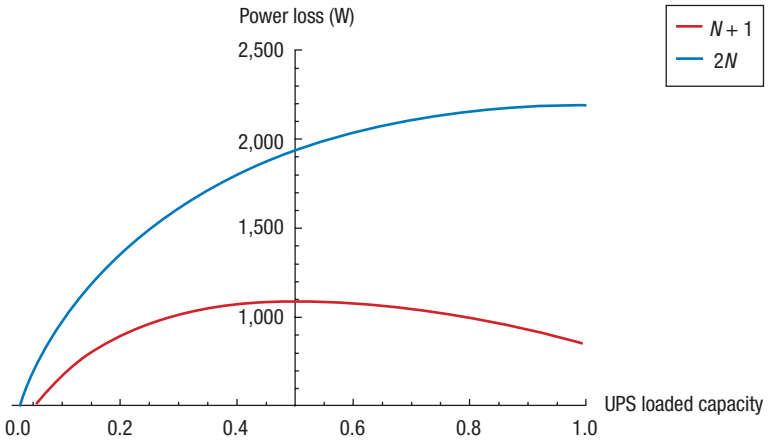

FIGURE 3. UPS power loss of a single rack with *N*+1 and 2*N* configurations. For the *N*+1 configuration, power loss increases when loaded capacity is less than 50 percent and decreases when it is higher than 50 percent. For the 2*N* configuration, power loss continuously increases with loaded capacity.

Given the UPS efficiency curve in Figure 2, we used a natural logarithmic function to fit the curve and Mathematica to calculate the UPS power loss. Figure 3 shows the power loss of UPSs with *N*+1 and 2*N* UPS configurations.

MODELING ENERGY-EFFICIENT PLACEMENT

On the basis of the data in Table 1, we formulated an optimization problem to minimize the total power of IT equipment and UPS power loss through the use of a rack-level power model that directly maps the rack's workload to its power dissipation. We used the model along with our workload-placement calculations to solve the optimization problem.

Rack-level power modeling

Our rack-level power model uses workload information, such as throughput and instructions per second (IPC), as direct inputs. The model's target application is an enterprise datacenter with nonvirtualized servers, each of which hosts only one application. We assume that the CPU is running at a fixed speed without dynamic tuning.

We express the rack-level power model as

$$P_i(w) = P_i^{\text{IDLE}} + \sum^{j} \alpha_i^j \times w_i^j, \quad (1)$$

where P_i^{IDLE} is the rack's idle power and the summation of

$$\sum^{j} \alpha_i^j \times w_i^j$$

is the total power introduced by all workloads on this rack. α_i^j is the coefficient that represents watts per performance of workload *j* on the *i*th rack. α_i^j has different units for different applications and hardware.

For CPU-intensive applications, α_i^j could be watts per instruction; for memory-intensive applications, it could be watts per byte; and for Web services, watts per request might be a solid indicator of system efficiency. Workload profiling provides historic knowledge that can be used to choose the appropriate α_i^j metric.

In our experiments, we profiled an application in four steps. We first measured the idle power of the *i*th rack as P_i^{IDLE}. We then fully loaded the rack to get a performance upper bound for this application. As a third step, we gradually increased the workload, making the rack run at different power levels, and recorded the rack power. Finally, we calculated the average value (performance per watts) of all sample points, which we used as α_i^j. We repeated this process for different types of applications to get the corresponding α_i^j value for the *i*th rack.

Optimization problem

Our optimization problem was for a datacenter that hosts multiple applications

simultaneously, with each server hosting only one application at a time, and a workload that can be dynamically assigned to a different server subset. In addition, we assumed that one UPS is connected to only one rack deployed in either an *N*+1 or a 2*N* redundant configuration. Because UPS power loss varies significantly with IT power load, clearly any workload change or revised distribution will affect IT equipment power loss.

Our goal was to minimize both the total IT equipment power and wasted rack-level UPS power. We chose the optimal workload allocation given the equality and inequality constraints of

- *performance*, which means the summation of all racks' workload should be equal to the total workload from all users;
- *capacity*, which means the hardware resource requirement should be less than each rack's maximum hardware capacity; and
- *power*, which means the total rack power satisfies the specified power-capping requirement (by operator or hardware).

Given these constraints, the mathematical formulation of the optimization problem is

$$\text{Minimize} \quad \sum_i \frac{P_i}{\eta\left(P_i\right)} \tag{2}$$

as long as

$$\sum_i w_i^j = w^j, \tag{3}$$

$$\sum^j \frac{w_i^j}{C_i^j} \le C_i, \text{and} \tag{4}$$

$$P_i^{\text{IDLE}} + \sum^j \alpha_i^j \times w_i^j \le P_i^{\text{CAP}}, \tag{5}$$

where P_i is the power of the *i*th rack and $\eta(P_i)$ is the conversion efficiency when the UPS has the IT power load of $\eta(P_i)$. Equation 3 ensures that the performance requirement is satisfied for workload *j*. In Equation 4, C_i^j is the capacity limitation for workload *j* on rack *i*, and C_i is the hardware limitation for rack *i*. In Equation 5, P_i^{CAP} is the capping power for the *i*th rack.

> **UPS POWER LOSS VARIES SIGNIFICANTLY WITH IT POWER LOAD, SO ANY WORKLOAD CHANGE OR REDISTRIBUTION WILL AFFECT IT EQUIPMENT POWER LOSS.**

The function f[$\eta(P_i)$] denotes the relationship between the UPS output power load and its corresponding conversion efficiency. $\eta(P_i)$ can be expressed as

$$\eta\left(P_i\right) = \alpha \times \ln \frac{P_i}{P_{\text{UPS}}} + b\,, \tag{6}$$

where P_{UPS} is the UPS input power, and α and *b* are fixed to match the conversion efficiency curve for different UPSs. In our evaluation, we chose a value of 0.1279 for α and 0.9343 for *b*.

EVALUATION RESULTS

To verify the power model, we conducted an experiment with 10 servers and two applications. The servers were eight Intel CPU servers and two AMD CPU servers; the applications were Y-Cruncher (www.numberworld.org/y-cruncher), a CPU-intensive application, and Yahoo Cloud Serving Benchmark (YCSB; http://labs.yahoo.com/news/yahoo-cloud-serving-benchmark), which simulates Web service requests to read and write to a database.

We ran the two applications separately to get α_i^j as described in Equation 1 and then estimated real-time power by running the two applications on all 10 machines simultaneously while randomly changing each application's workload during the test. The sample frequency is 1 Hz, which is sufficient for tasks that must execute over many hours or even days. We conducted the test on Intel and AMD machines separately and used linear regression to fit the datapoints. The P^{IDLE} of 10 machines was 2,183 W.

The workload of Y-Cruncher and YCSB are represented in digits per second and operations per second. The coefficients α of Y-Cruncher on Intel and AMD servers were 3×10^{-5} W/digits/s and 4×10^{-5} W/digits/s. The coefficients α of YCSB on Intel and AMD servers were 0.0024 W/operations/s and 0.0039 W/operations/s.

Figure 4 shows the measured power by a power meter and the estimated power using our rack-level power model. Error rates were within ±2.5 percent—a corresponding power-estimation error of less than 83 W. Moreover, our rack-level power model overestimated power consumption 82 percent of the time (246 out of 300 sample points). For underestimated cases,

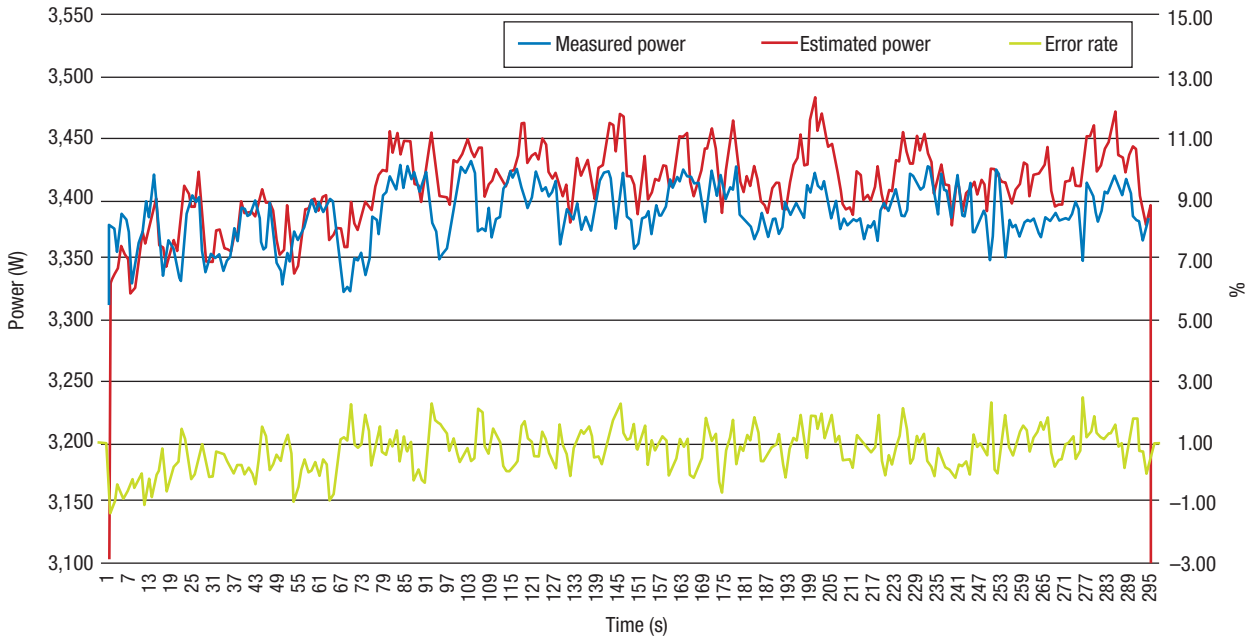

FIGURE 4. Real-time power estimation and error rate. Error rate is represented as (estimated power − measured power)/measured power. In some cases, the gap between measured power and estimated power is less than 47 W with an error rate of −1.4 percent.

the gap between measured power and estimated power was less than 47 W with an error rate of −1.4 percent.

These results are significant in light of the optimization problem's power constraint. High underestimation probability and error rate can lead to a violation of the rack-level power-capping requirement.

Simulation results

We compared the total IT equipment power and wasted power from UPSs in both our optimal workload allocation and a baseline case that evenly allocates workload among racks. We assumed that each rack hosts 20 servers, and that each rack's power is supplied by one or two UPSs with an *N*+1 or a 2*N* UPS configuration. Our simulation was for 50 racks running 10 applications simultaneously. We used Mathematica to perform the simulation, which ended when either the iterations exceeded a predefined threshold or the results converged to the requested precision.

Power-reduction comparison. Figure 5 shows the power reduction with our optimal workload allocation relative to the baseline allocation (evenly distributed workloads). For both the *N*+1 and 2*N* UPS configurations, the optimal workload allocation reduces power consumption by 1.23 percent to 5.20 percent. The *N*+1 configuration has a slightly higher power-reduction rate, but overall the rate gap between configurations is small for all loads.

Optimal workload allocation achieves the highest power reduction at the datacenter utilization of 50 percent, which is also the average level for most datacenters.[15] The degree of power reduction depends on the UPS efficiency curve and the alpha coefficient value (α_i^j) in Equation 1. As an extreme example, if UPS efficiency is constant, power reduction will be zero for all datacenter-utilization levels. That is, regardless of workload distribution, UPS efficiency (total UPS output power) is constant for a particular workload size. Because UPS efficiency is constant, UPS input power and power loss are also constant.

Because the alpha coefficient decides the power-increase rate for a specific application, it affects the datacenter's power consumption (UPS output power), which is why a different workload allocation might have a different power consumption for the same workload size.

Workload-type effects. To better understand how workload type affects power reduction, we changed the application mix while keeping total datacenter utilization at 50 percent. In this simulation, we divided the 10 applications into two categories—CPU-intensive and Web service—and then mixed the types with different proportions.

As Figure 6 shows, a greater proportion of CPU-intensive applications translates to higher power reduction, with the maximum reduction at 85 percent CPU-intensive applications and 15 percent Web service applications. These results are predictable: the more CPU-intensive applications there are, the more power goes to the rack. The higher rack power increases the UPS's loaded capacity, which could result in lower power loss.

Translation to energy-cost reduction

Datacenter power capacity varies considerably. As of 2011, it was estimated to be anywhere from 2 to 100 MW (www.greenpeace.org/international/Global

/international/publications/climate/2011/Cool%20IT/dirty-data-facilities-table-greenpeace.pdf), with half the datacenters surveyed falling between 20 and 76 MW. According to a 2014 Uptime Institute report (https://journal.uptimeinstitute.com/2014-data-center-industry-survey), the average datacenter PUE is 1.7. For a datacenter with a 76-MW capacity, IT equipment power consumption would be 44.7 MW. With the 5.2 maximum power-consumption reduction demonstrated in our simulation, the datacenter could reduce its annual energy cost by 44.7 MW per day, which is roughly $1.4 million (44.7 × 365 days × 24 h × 0.07 $/kWh × 0.052).

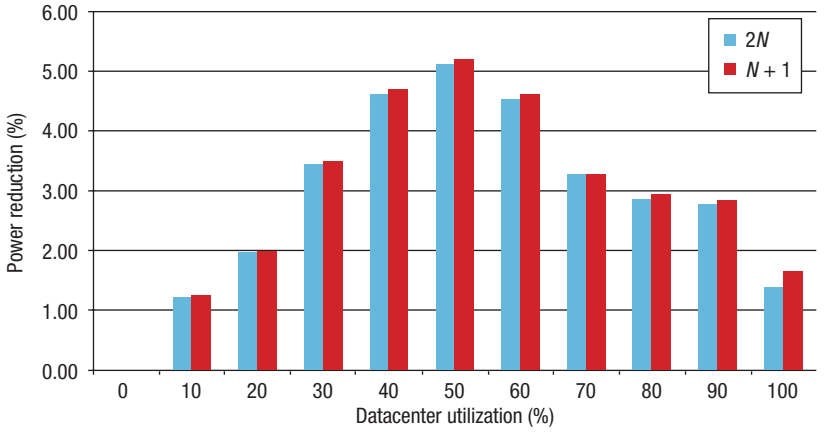

FIGURE 5. Power reduction with our optimal workload allocation (optimal power) relative to an evenly allocated workload (baseline) for an $N+1$ and a $2N$ UPS configuration. Power reduction is represented as (baseline power − optimal power)/baseline power. Power reduction is highest for both configurations when datacenter utilization is 50 percent.

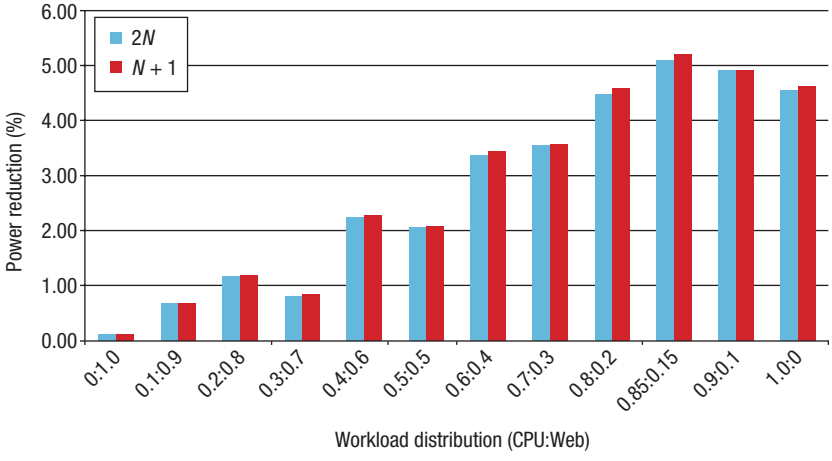

FIGURE 6. Power reduction with different proportions of CPU-intensive (CPU) and Web service (Web) applications when datacenter utilization is 50 percent. Power reduction is (baseline power − optimal power)/baseline power.

Our experiments with the rack-level power model show that UPS configurations significantly affect a datacenter's energy efficiency and that UPS power loss is different when IT workload changes. Along with our workload-placement calculations, the model minimizes IT equipment power consumption and UPS power loss with up to a 5.2 percent power reduction relative to an even workload-allocation strategy. In future work, we will focus on how DVFS and switching servers on and off can enhance UPS efficiency.

ACKNOWLEDGMENTS

This work is supported in part by National Science Foundation (NSF) grant CNS-1205338. We thank Wayne State University's Computing and Information Technology Department for its significant assistance and NextEnergy for collecting data and running experiments. The research reported in this article is based on work done by Weisong Shi while he was at NSF.

REFERENCES

1. A. Gandhi et al., "Optimal Power Allocation in Server Farms," *Proc. ACM Int'l Conf. Measurement and Modeling of Computer Systems* (SIGMETRICS 09), 2009, pp. 157–168.
2. Q. Deng et al., "Coscale: Coordinating CPU and Memory System DVFS in Server Systems," *Proc. 45th IEEE/ACM Int'l Symp. Microarchitecture* (MICRO 12), 2012, pp. 143–154.
3. J.S. Chase et al., "Managing Energy and Server Resources in Hosting Centers," *ACM SIGOPS Operating Systems Rev.*, vol. 35, no. 5, 2001, pp. 103–116.
4. R. Nathuji and K. Schwan, "Virtual Power: Coordinated Power Management in Virtualized Enterprise Systems," *ACM SIGOPS Operating Systems Rev.*, vol. 41, no. 6, 2007, pp. 265–278.
5. Q. Tang et al., "Energy-Efficient Thermal-Aware Task Scheduling

ABOUT THE AUTHORS

QUAN ZHANG is a doctoral researcher in computer science at Wayne State University. His research interests include distributed systems, cloud computing, and energy-efficient computing. Zhang received a BS in computer science from Tongji University. He is a student member of IEEE. Contact him at quan.zhang@wayne.edu.

WEISONG SHI is a professor of computer science at Wayne State University. His research interests include energy-efficient computer systems and software, Internet computing, and mobile health. Shi received a PhD in computer engineering from the Chinese Academy of Sciences. He is an IEEE Fellow and a Senior Member of ACM. Contact him at weisong@wayne.edu.

for Homogeneous High-Performance Computing Data Centers: A Cyber-Physical Approach," *IEEE Trans. Parallel and Distributed Systems*, vol. 19, no. 11, 2008, pp. 1458–1472.
6. A. Verma, P. Ahuja, and A. Neogi, "Pmapper: Power and Migration Cost Aware Application Placement in Virtualized Systems," *Proc. 9th ACM /IFIP/USENIX Int'l Conf. Middleware* (Middleware 08), 2008, pp. 243–264.
7. R. Joseph and M. Martonosi, "Runtime Power Estimation in High-Performance Microprocessors," *Proc. 6th ACM/IEEE Int'l Symp. Low Power Electronics and Design* (ISLPED 01), 2001, pp. 135–140.
8. H. David et al., "Rapl: Memory Power Estimation and Capping," *Proc. 15th ACM/IEEE Int'l Symp. Low-Power Electronics and Design* (ISLPED 10), 2010, pp. 189–194.
9. J. Zedlewski et al., "Modeling Hard-Disk Power Consumption," *Proc. 2nd USENIX Conf. File and Storage Technologies* (FAST 03), 2003, pp. 217–230.
10. C. Lefurgy, X. Wang, and M. Ware, "Server-Level Power Control," *Proc. 4th IEEE Int'l Conf. Autonomic Computing* (ICAC 07), 2007, pp. 4–14.
11. D. Meisner, B.T. Gold, and T.F. Wenisch, "The Powernap Server Architecture," *ACM Trans. Computer Systems*, vol. 29, no. 1, 2011, pp. 3.1–3.24.
12. A. Kansal et al., "Virtual Machine Power Metering and Provisioning," *Proc. 1st ACM Symp. Cloud Computing* (SOCC 10), 2010, pp. 39–50.
13. E. Pakbaznia and M. Pedram, "Minimizing Data Center Cooling and Server Power Costs," *Proc. 14th ACM /IEEE Int'l Symp. Low-Power Electronics and Design* (ISLPED 01), 2009, pp. 145–150.
14. V. Kontorinis et al., "Managing Distributed UPS Energy for Effective Power Capping in Datacenters," *Proc. 39th Ann. IEEE/ACM Int'l Symp. Computer Architecture* (ISCA '12), 2012, pp. 488–499.
15. L.A. Barroso and U. Hölzle, *The Data Center as a Computer: An Introduction to the Design of Warehouse-Scale Machines*, Morgan and Claypool, 2009.

Selected CS articles and columns are also available for free at **http://ComputingNow.computer.org**.