Server Demand Response via Automated Hardware Management

Micah Sweeney and Mukesh Khattar, Electric Power Research Institute (EPRI)
Enrique Castro-Leon, Henry Wong and Derek Collier, Intel
Jay Madden and Paul Delaney, Southern California Edison
G.P. Li and Chris Battista, Calit2, University of California, Irvine, CA

ABSTRACT

As one of the most energy-intense categories of commercial buildings, data centers have long been considered a good candidate for energy efficiency improvements. Yet the demand response (DR) capabilities offered by data centers have not been fully realized. Prior efforts to demonstrate DR in data centers have identified several techniques to reduce demand, yet each requires manual control to achieve reductions.

This study seeks to demonstrate server DR by reducing power consumption of the IT equipment through built-in power management tools. In response to a utility DR signal, power management software can issue a power reduction policy to server hardware, which curtails its power use (for example by reducing voltage and/or frequency of the central processing unit [CPU]). By reducing the demand of servers directly, it is expected that the power of support systems—including cooling equipment power and electrical losses in power delivery—will reduce correspondingly for a magnified effect, albeit with a time delay.

In collaboration with an electric utility, an end user and an industry partner, the authors have deployed a proof-of-concept demonstration in both a laboratory setting and a small, production data center to evaluate this DR technique. The study seeks to understand the capabilities of this technique in terms of demand reduction, time to respond, impact to users, and demand rebound, as well as demonstrate its ability to respond to automated DR signals from a utility. Preliminary results indicate that servers may be useful for fast DR—including ancillary services to the grid such as frequency regulation—due to their immediate response. This paper summarizes the concept, design, implementation and analysis of server demand response, to provide the reader with an understanding of the possibilities and DR expectations with this approach.

Introduction

Data centers house the information and communications technology (ICT) equipment that supports the modern world. From small network closets to hyper-scale facilities that support the cloud, data centers are estimated to consume about 2% of the electricity currently used in the United States (NRDC 2014). As one of the most energy-intense categories of buildings, data centers have presented ample opportunity for energy efficiency (EE) improvements, yet the demand response (DR) capabilities of these facilities have not been sufficiently investigated.

Demand response refers to a reduction in the consumption of electricity during periods of peak demand or when the power supply is constrained and reliability of the grid is at risk. Customers are incentivized to reduce or curtail electricity demand during the DR event. Typically customers reduce their electricity demand by curtailing non-critical loads, adjusting air conditioning set-points, or scaling back an industrial process. Time-of-use electricity rates are an example of an incentive for customers to use less electricity during typical peak hours when rates

are higher. However, in many modern DR programs, utilities send a signal to the customers when a reduction in demand is needed and customers respond by reducing demand as per individual agreement with the utilities. In typical DR programs, the signal is sent one day in advance to give customers ample time to plan operations. In a newer version of fast DR, the signal can be sent as little as a few hours before the event, and yet in the market for frequency and voltage regulation, the event notification may be only minutes ahead of the event. Utilities enter into voluntary agreements with large customers to take advantage of end-use devices to balance supply and generation on the power system. Demand response has been successfully used to control lighting, water heaters, HVAC (heating, ventilation, and air conditioning) equipment, in addition to certain large, industrial facilities whose process is not time sensitive.

Despite their sizeable electrical load, data centers have not been considered a good candidate for DR due to the flat nature of their load profile and their reluctance to risk downtime. Although data centers exhibit a stable load to the grid, technologies over the past ten years have shown the ability to reduce this load and sustain operations over a specified period of time. Previous research efforts have identified several opportunities to reduce demand in data centers (Ghatikar et al. 2010, Ghatikar et al. 2012), yet data centers have not been utilized as a grid resource in this way. With large, stable, and opportunistically adjustable load, data centers offer a significant opportunity in balancing demand on the grid. In addition, most servers carry built-in sensors and controllers that enable power adjustments. An opportunity exists to leverage this telemetry, although it is little used in practice today.

This paper presents recent efforts that EPRI has undertaken with a major supplier in the ICT industry to examine the use of server power management as a form of DR. In managing load at the server level, it is expected that the demand of the entire data center will be impacted due to the dependence of overhead energy use on server load in terms of both cooling load and losses in the power delivery chain. The objective of this project is to demonstrate the technical feasibility of server power management as a demand response technique. This will be shown with a proof-of-concept experiment whereby server power is automatically limited by software in response to a DR signal from outside the data center. This paper describes the approach, experimental setup, and preliminary results from laboratory testing conducted as part of an evaluation sponsored by Southern California Edison, and supported by an industry liaison consisting of Intel, Schneider Electric, IPKeys, and Calit2 at the University of California in Irvine.

Background on Demand Response

In the simplest sense, demand response (DR) is defined as a reduction in end-use electricity consumption in order to balance supply and demand on the grid. DR has been used to relieve strained capacity of grid assets on high demand days, saving utilities (and by extension their customers, often called ratepayers) cost from operating expensive “peaker” plants. In some parts of the U.S. this can occur on the hottest afternoon of the summer when nearly all air conditioning units are running. Yet in other parts of the country this can be seen on cold winter mornings when electric heaters and water heaters turn on to provide comfort and hot showers for millions of customers. In either case, DR is a valuable mechanism to mitigate the cost of peak demand on the grid.

More than simply avoiding the fuel cost of operating inefficient generators, DR allows grid operators to reduce the amount of infrastructure required to deliver electricity on peak days, in terms of both generation and transmission capacity. On rare occasions, DR can be called upon

to curtail power usage when a generator trips offline, due to an unforeseen circumstances or other disruption of planned operation. Increasingly, DR is being seen as a tool to mitigate the inherent fluctuation of renewable energy sources such as wind and solar photovoltaics. The quick variation of these generation sources has spawned markets for several *ancillary services* to the grid, including frequency regulation, voltage control (provided by reactive power support), and reserve operating capacity (historically known as *spinning reserves*).

Historically, DR was accomplished by a utility making a phone call to one of its large customers with whom it has agreed a price, as well as and quantity and duration of power reduction (in kW or MW). Ideally, such a request would have minimal impact to productivity or comfort. Examples of traditional DR opportunities include industrial processes that are not time dependent, thermostat set-points for HVAC, and water heaters with storage capacity. In recent years, several technologies have been developed to provide automated DR as a result of signals from the grid. With increasing technological capabilities, DR is being called upon as a valuable resource in grid markets to provide many of the same capabilities as peaker plants and grid-scale energy storage.

To allow a greater number of distributed electrical loads to respond to DR signals en masse, an industry consortium led by the Demand Response Research Center (DRRC) at Lawrence Berkeley National Laboratory (LBNL) has developed the Open Automated Demand Response (OpenADR) communications standard. The standard defines a communications data model by which a utility or independent system operator (ISO) can send DR signals to electric customers. The intent of the standard is to enable electric loads, such as building and industrial control systems—to make pre-programmed actions in response to power system conditions without manual intervention.

Compared to traditional, manual DR methods, data center workloads bring the promise of automation in DR closer to reality. For instance, curtailment of server workloads is thought to be accomplished two or three orders of magnitude faster than mechanical process workloads—i.e. in seconds compared to hours. Such a quick response is expected of servers due to the superior computational power and complexity that they offer over application-specific load controllers (e.g. building energy management systems). In addition, server management functions continuously monitor the status and performance of hardware and software in real-time and have the ability to make complex decisions on the ability of the server to respond to DR signals in terms of available capacity and the relative criticality of workload at that time. With the potential for such quick response, automation of decision making at the load becomes a necessity. To that end, fast-acting data center workloads might be used in lieu of the most disruptive load shedding policies for contingency management.

Data Center Energy Use

Industry experts estimate that data centers accounted for 91 billion kWh (TWh) of electricity consumption in 2013—about 2% of all the electricity consumed in the U.S. annually and growing at more than 6% year over year (NRDC 2014). As the modern economy increasingly depends on the digital services that data centers provide, the amount of energy consumed by this industry is expected to increase. Due to the critical nature of these services, downtime can be very costly. A 2013 study of data center outages (commissioned by Emerson Network Power) found that unplanned downtime in large data centers costs $7,900 per minute on average (Ponemon 2013).

To avoid costly downtime, data centers typically rely on several power protection devices to maintain uninterrupted services. For extended interruptions in utility power, many data centers utilize an emergency generator typically fueled by diesel. To prevent servers from shutting down when utility power dips (known as *voltage sags*), data centers use uninterruptible power supplies (UPS). In addition, energy storage (batteries, ultracapacitors, or flywheels, often integrated into the UPS) can be used to provide "bridge power" in the time required for the emergency generator to come online. Each of these devices incurs electrical losses, which contribute to the thermal load when located in air-conditioned space.

Within data centers, energy is consumed by three primary subsystems: the ICT equipment, cooling systems, and power delivery equipment (losses from uninterruptible power supplies [UPS], power distribution, etc.). The primary metric used by industry to measure data center efficiency is PUE (power usage effectiveness), which has been championed by The Green Grid. PUE is defined as the ratio of total facility energy use divided by the energy use of the ICT equipment. This metric gives an indication of the relative amount of energy used to support ICT equipment, with ideal PUE of 1.0 when all energy goes to drive the ICT equipment and none for supporting equipment such as cooling of power distribution losses. Several industry studies over the past years have shown that average PUE is somewhere between 1.8 and 2.0, indicating that for every watt of ICT equipment power there is 0.8 to 1.0 watt of infrastructure equipment to support it.

Some data centers have been found to exhibit a load profile that is quite "flat", with very little change in power use over 24 hours. For continuous availability of services, ICT equipment is usually powered continuously, presenting a continuous electrical load to the grid. Thus the power demand of the ICT equipment and the power delivery equipment is nearly constant. In addition, the cooling load is relatively constant—unlike typical commercial buildings—since the ICT equipment is always energized and drives the cooling requirement. Yet cooling equipment exhibits somewhat of a change in load profile due to the dependence of cooling system efficiency on outdoor conditions. Specifically, data center cooling equipment will draw more power when outdoor temperature is high, showing both diurnal and seasonal variation. This variation is even more dramatic in data centers that utilize some form of economization, since these systems reduce mechanical cooling when outdoor conditions allow.

Approach

Previous studies have explored the potential of DR in data centers, and identified several opportunities (Ghatikar et al., 2012; Liu, 2013; Irwin, 2011). Most practical implementations of data center DR have focused on adjusting cooling system settings during peak periods. ASHRAE (the American Society of Heating, Refrigerating, and Air-Conditioning Engineers) has provided guidelines for the temperature and humidity of air used to condition data centers, and recommends air temperatures as high as 27°C (81°F) for continuous cooling of ICT equipment. In addition, ASHRAE specifies several ranges of "allowable" conditions, which follow manufacturers' recommendations for conditions at which ICT equipment can operate for short periods of time. Yet increasing cooling system set-points is perceived by some data center operators as a threat to equipment life, and is limited in usefulness for DR by the increase in ICT equipment power at elevated temperatures. Moreover, the ASHRAE guidelines recommend a

maximum of 5°C (9°F) rise in data center temperature per hour for data centers with tape drives, which are commonly used for archiving data (a 20°C [36°F] per hour rise is allowed in data centers with disk drives).

To avoid the challenges presented by increasing temperature set-points, this study seeks to manage the demand of server equipment itself using a power capping feature provided by a major chip manufacturer. In addition, since data center infrastructure has continued to increase in efficiency, the ability to manage server power allows a much greater range of control. Moreover, it is expected that reductions in server power translate to reductions in infrastructure power, e.g. by reducing cooling load and losses in power distribution equipment. In this way, managing server power has the potential to impact the demand of the entire facility.

The authors chose to utilize Intel® Node Manager technology to manage server power from the hardware level. Available on Intel's latest generation of E3, E5, and E7 processors, this technology reports server power and temperature over industry standard communications protocols. This information can be monitored with data center infrastructure management (DCIM) tools available from a variety of vendors to manage assets and operations across the data center. On top of monitoring server power and temperature, this technology features the ability to manage server power by setting individual power limits on each server. This is accomplished by regulating server performance (P) and throttle (T) states, adjusting the voltage and frequency of the CPU (P states) and/or introducing delays between processor execution cycles (T states).

AMD offers the same feature called TDP Power Cap on its latest CPUs (TDP refers to *thermal design power*, which represents the maximum amount of heat that must be removed from the CPU under real-world workloads). The authors chose to evaluate Intel chips in this study due to their large market segment share in the server market.

Power management is accomplished using the baseboard management controller (BMC), a microcontroller built into the server motherboard to manage the interface between management software and the server hardware. The BMC communicates with other devices via intelligent platform management interface (IPMI), a standard communications interface between management software and computer systems. Each server manufacturer has branded its own BMC functionality: Dell iDRAC (integrated Dell remote access controller), HP iLO (integrated lights out), and others from Cisco, IBM, etc.

To investigate the feasibility of this technology for DR, this study seeks to demonstrate three separate functions. First, Node Manager should be shown to have the ability to effectively limit server power. Second, it should be shown that this technology can reduce the power draw of the server under typical workload over a useful period of time, from a few minutes to several hours. Note that conventional DR requires that average power of a number of loads be curtailed over a peak demand period that lasts several hours. Finally, the technology should be able to initiate a server power cap in response to an external signal from a utility or other power system operator (for example an ISO or independent system operator).

To examine the capabilities of this technology to accomplish the functions outlined above, the authors implemented nearly identical test conditions in a laboratory environment and an operational data center, using five Dell servers equipped with Intel processors compatible with Node Manager technology. Power capping is managed by Intel Data Center Manager, a DCIM tool to monitor and managing power and thermal data within the data center. Finally, a custom version of EISSClient DR communications software was delivered by its supplier (IP Keys) to provide end-node response to OpenADR signals from a utility or power system operator.

Laboratory Testing

To evaluate server power capping in a controlled manner, EPRI acquired five Dell servers with nearly identical specifications to those installed at a field site. The models were selected to supply the needs of the field site in terms of application, operating system (OS), memory, and storage. EPRI purchased nearly identical hardware for testing in its IDEA (Innovations in Datacenter Efficiency Advances) Laboratory in Knoxville, TN, matching the number and types of components with those in the field. (Although the number, type, and speed of disk drives were matched, the storage capacity was not duplicated due to its limited impact on drive energy use.) Each of these servers is equipped with a single CPU (central processor unit) from the Intel E5-24xx series.

To enable power capping on Dell servers, each was equipped with a power supply unit (PSU) that supports the PMBus (power management bus) communications protocol. The BMC was configured for IPMI over LAN (local area network) to enable remote management. The remote power management feature on Dell servers required an upgraded iDRAC license (Enterprise) that added an additional license fee per server. In this test, Intel Data Center Manager (DCM) was installed on a separate server on the network for monitoring and management of the servers under test (SUT).

The SUT were installed in EPRI's IDEA Laboratory for testing in a controlled environment (shown in Figure 1). Each was installed with Windows Server 2012 and configured for remote monitoring and management from a local server hosting Intel DCM software, which recorded server power and temperature reported from each machine. One of the servers was selected for preliminary benchmarking and power cap testing. The specifications of this server are listed below in Table 1.

Figure 1. Servers under test in EPRI's IDEA Lab

Table 1. Specifications of server used for benchmarking and power cap testing

Model	**Dell PowerEdge R520**
Processor (single)	Intel Xeon E5-2430v2 (2.5 GHz, 6 cores, 80 W)
Memory	Two (2) 8GB (1600 MT/s, low-voltage)
Storage	Three (3) 500GB 7.2k SATA HDD Two (2) 300GB 10k SAS HDD

To verify the basic functionality of power capping, LINPACK benchmarks were used to load the SUT. This benchmark is used to report the peak performance of a computer in terms of floating point operations per second (FLOPS), and a version of LINPACK is used to rank supercomputers on the TOP500 list. This testing utilized a version of the LINPACK benchmark that was optimized and compiled by Intel for use with its processors.

In essence, the LINPACK benchmark measures the performance of a computer when solving an n by n system of linear equations, requiring a vast number of floating point addition and multiplication calculations be performed. With such a computationally intense workload, the LINPACK benchmark represents an extreme, worst-case processing requirement and offers the most challenging workload a server would ever experience. As such, testing with this workload is expected to illustrate the level of curtailment attainable at the high end of virtualized cloud workloads (above 80% CPU utilization). Traditional enterprise workloads load the CPU to less than 20% utilization, and in many cases less than 10%. It is expected that the amount of power curtailment available with LINPACK is greater than available from more conventional workloads.

This server was used to test the basic functionality and impact of power capping and its impact with the LINPACK benchmark. Two benchmark parameters were chosen to represent a light workload (1,000 by 1,000 system of equations) and a heavy workload (30,000). Each benchmark was run without power restrictions, followed by a test under the most restrictive power cap. Under the latter scenario the CPU power is held to its lowest operating level.

Table 2 shows the results of these tests, indicating the time required to complete each benchmark, measured GFLOPS (gigaFLOPS) score, and average power of the server over the duration of the benchmark. These results show that power capping successfully limits the power use of the server to 72 W, yet impacts LINPACK performance metrics by an order of magnitude for both workload levels. For reference, this server exhibits idle power about 60 to 62 W.

Table 2. LINPACK benchmark results under maximum power cap

		Baseline	**Power Cap**	**Change**
Light Load	Time (s)	0.016	0.341	22x
	GFLOPS	42.8	2.0	
	Power (W)	**110**	**72**	35%
Heavy Load	Time (s)	157	4151	26x
	GFLOPS	114.5	4.3	
	Power (W)	**135**	**72**	47%

To determine the impact of varying power cap levels on performance, these benchmarks were repeated under several additional settings. Preliminary testing of this server revealed peak power of 152 W under heavy workload, which was limited to 72 W under the most aggressive power cap. With these characteristics in mind, the two benchmarks were repeated with increasingly stringent power limits, beginning with a limit of 152 W and increasing the cap (decreasing server power) by 10 W for each successive test. Figure 2 shows the results of this test, with benchmark score plotted as a function of average server power during the benchmark. Performance is seen to exhibit a non-linear dependence on power, falling off dramatically at more aggressive power limits. Interestingly, benchmark performance was seen to exceed

baseline performance for less stringent power limits. In fact, the performance of the heavy workload was found to increase under power limits of 132 W, 142 W, and 152 W limits, exceeding performance without a power cap. Moreover, the 152 W limit was shown to increase power consumption over the test period above the baseline case.

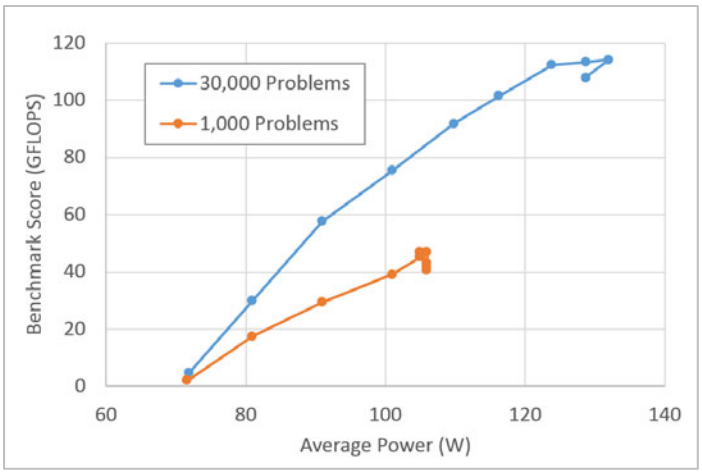

Figure 2. Performance score (GFLOPS) of two LINPACK benchmarks under increasing power cap levels

To understand the dependence of performance on power caps of various size, a series of six benchmarks was run under five power cap levels and compared with unconstrained performance. The five power cap levels chosen were 72 W (minimum power), 92 W, 112 W, and 132 W. Under each of these power limits, the following benchmark sizes were run: 1000, 5000, 10000, 20000, 30000, and 35000. Figure 3 shows benchmark performance against workload, demonstrating similar performance curves under each power cap level, with the most aggressive limits having the greatest impact (negative) on performance.

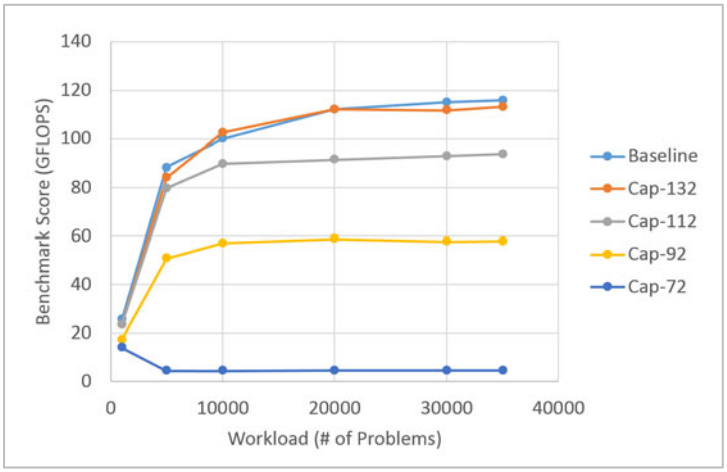

Figure 3. Performance score (GFLOPS) of various LINPACK benchmarks under select power cap levels

Figure 4 shows the power level of each benchmark over the five power cap conditions, showing that each cap restrained server power below the specified limit. These results demonstrate that power capping is able to limit the average demand over the period required to complete each benchmark. Compared to the baseline power of 127 W, the strictest power cap (72-W limit) was shown to reduce server demand by as much as 55 W, signifying a 43%

reduction. Yet given the significant impact to server performance, such a limit may only be feasible in extreme situations (e.g. when grid stability is threatened). Less stringent limits yielded lower savings: 35 W (28%) for 92-W cap, 17 W (13%) for 112-W cap, and virtually no savings for the 132-W cap. However, due to the high level of load that the LINPACK benchmark places on the CPU, it is expected that these results would only be observed for real-world applications with very high CPU utilization.

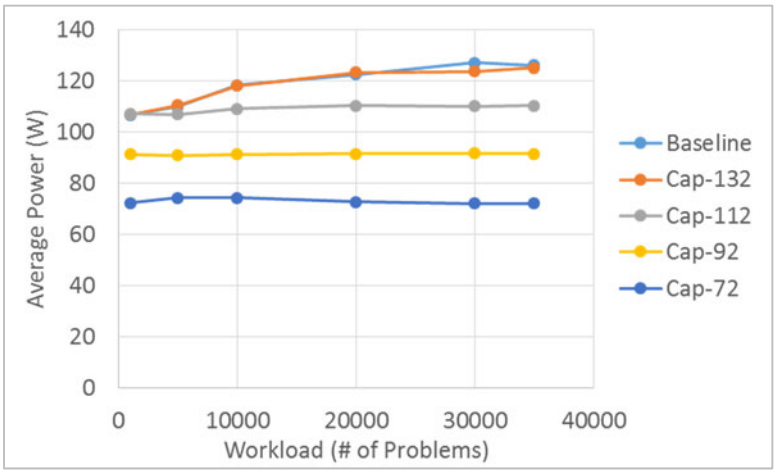

Figure 4. Server power (W) for LINPACK benchmarks under several power cap levels

Although the power cap was shown to maintain server power below specified limits, the impact on performance was found to dramatically increase the time required to complete each benchmark. Figure 5 shows the amount of time to complete each workload under the five power limits, with a logarithmic scale marking the duration of benchmark tests. These results show that the maximum power limit increased time to complete each benchmark by an order of magnitude over the baseline and 132-W cap benchmarks.

Figure 5 shows that the effects of capping on performance are nonlinear: a small impact on performance is observed until 92 watts. One explanation is that Node Manager takes out cycles off the top. A CPU running at less than 100 percent utilization has headroom left to soak up variations in demand. Below 92 watts there is no more headroom left, and we start seeing resource contention, especially with the OS process scheduler.

LINPACK presents a constant workload to the CPU. This underestimates a side effect of capping that does not show up as reduced performance: as CPU cycles are removed off the top, the CPU's ability to respond to upticks in workload demand is impaired. As an example, assume a virtualized workload is at 80% utilization. If the server is capped so it is not allowed to go beyond the current power level, a 10% uptick in workload will put the system into resource contention with significant performance degradation. Without the cap, utilization might go from 80% to 90% with a very little degradation in performance.

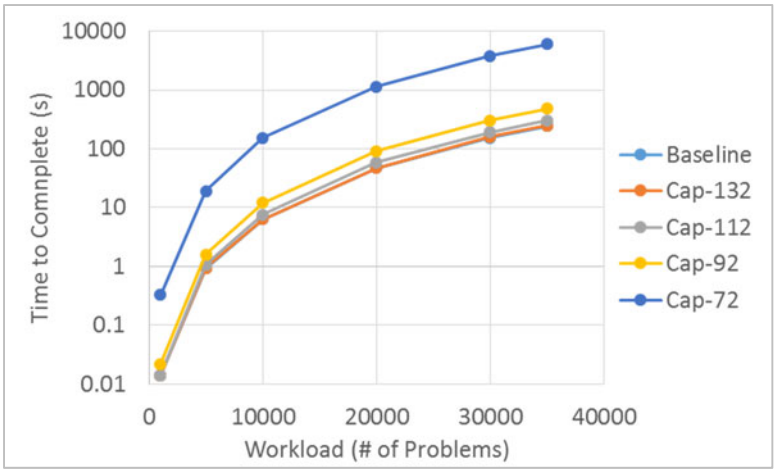

Figure 5. Time to complete LINPACK benchmarks under several power cap levels

Figure 5 may indicate that throughput-related workloads (e.g. transactions) have increased opportunity of power reduction while minimizing impact to the transaction throughput. The case would be especially applicable for constant workloads. Also, although the baseline performance might not change, capping will affect the system capability to respond to a demand uptick.

In terms of DR, perhaps the most useful metric to study is the total energy required to perform each benchmark routine. Table 3 shows the energy used to complete each benchmark, (J) calculated as the average power over the benchmark (W) times the average period of completion (s). These results echo earlier findings that more aggressive limits have greater impact on overall energy use, with the strictest power limit (72-W cap) increasing execution energy by an order of magnitude in all cases. It should be noted that the 132-W limit had minimal impact on energy use.

Table 3. Energy (J) used to complete LINPACK benchmark under various power cap levels

Workload	**1000**	**5000**	**10,000**	**20,000**	**30,000**	**35,000**
Baseline	1.5	103.0	751.9	5,720	19,672	30,725
Cap-132	1.5	104.1	750.7	5,795	19,710	31,341
Cap-112	1.5	113.5	836.9	6,476	20,924	33,728
Cap-92	2.0	144.3	1,097	8,410	28,028	44,292
Cap-72	23.6	1435	11,363	83,415	276,103	435,718

Conclusions

This paper describes the first phase of a project that EPRI has undertaken to demonstrate the feasibility of server power capping for electricity demand response. Preliminary testing using LINPACK benchmarks demonstrates that power capping can successfully limit the instantaneous power of a server.

Despite the reduction in instantaneous server power that results from power capping, this technique does not increase practical efficiency—i.e. the amount of energy required to complete the LINPACK benchmarks. Such a result can be understood as the power limit preventing the server from reaching its most efficient state: peak utilization. Thus, for workloads that are computationally intense (i.e. place significant load on the CPU, like LINPACK), power capping is expected to increase the server's energy use due to the additional time required to execute the workload. This suggests that power capping may not provide reduction in average server power to complete a fixed workload over a longer period. Yet for computationally light (transaction-based) workloads, there is indication of opportunity for power savings with minimal impact to throughput.

Yet power capping was demonstrated to successfully limit the instantaneous power of multiple servers at a moment's notice. This capability may provide value to the grid, for example to respond to grid emergencies or to provide ancillary services—such as frequency or voltage regulation—with the very quick response required to compensate for variable generation sources on the grid (i.e. wind and solar). When fully implemented, this technology could provide a control mechanism to allow for broad and coordinated control between the utility and end user.

To determine the capability of power capping on limiting power use for real-world applications, it is recommended that additional testing be done to evaluate power capping with dynamically varying workload. It is recommended that several types of workload be tested, such as web hosting, email server, database, etc., so that the impact to workloads with different needs be evaluated (comparing processor-intensive workloads to memory or data-limited applications). Such a study should be performed in both laboratory and field conditions so that basic functionality, response to stochastic load, and user impacts can be quantified. EPRI intends to pursue these efforts in the future, in partnership with the industry collaborators that supported this work.

Finally, it is commonly seen in hierarchical controls systems that a control policy exposed at one level can become a "control knob" (i.e. a controllable parameter) for the level above. Currently, there are several gaps in the state of the art for managing server power for DR. First, a power system operator cannot be expected to decide how to limit each of its customers' servers. What's more, this cannot be expected using the tools that are currently available for local server management. The technology evaluated in this study elevates these tools to commands that can be called by software, but the operational knowledge to set power limits is still absent. Such information needs to be supplied by the data center operator, including the amount of power curtailment possible and performance indications such as response time, duration, and pricing signals. Certain curtailment requests may be more costly than others, which needs to be reflected through pricing signals. Given the lack of performance signals in current DR communications protocols, it is recommended that future efforts pursue the development of more robust pathways that allow an intelligent load to communicate its current state and make an informed decision about how to respond.

Acknowledgements

This work was made possible by sponsorship from Southern California Edison. EPRI would like to acknowledge the assistance from industry partners Intel, Schneider Electric, and Calit2 (California Institute for Telecommunications and Information Technology) at the University of California, Irvine, as well as contributions made by IPKeys in the development of its OpenADR appliance for use in this project.

References

2013 Census Report: Global Data Center Power 2013. DCD Intelligence, London, England: 2013.

2013 Cost of Data Center Outages. Ponemon Institute, Traverse City, MI: 2013.

Castro-Leon, Enrique. "IT-Driven Power Grid Demand Response for Datacenters." *IT Professional* vol. 18, no. 01 (2016), pp 42-49.

Data Center Efficiency Assessment: Scaling Up Energy Efficiency Across the Data Center Industry: Evaluating Key Drivers and Barriers. National Resources Defense Council (NRDC), New York, NY: 2014.

Ghatikar, Girish, M.A. Piette, S. Fujita, A. T. McKane, J.Q Han, A. Radspieler, K.C. Mares, D. Shroyer. *Demand Response and Open Automated Demand Response Opportunities for Data Centers*. California Energy Commission, PIER Program and Pacific Gas and Electric Company (PG&E), 2010.

Ghatikar, Girish, V. Ganti, N. Matson, M.A. Piette. *Demand Response Opportunities and Enabling Technologies for Data Centers: Findings from Field Studies*. Lawrence Berkeley National Laboratory (LBNL), 2012.

Irwin, David, N. Sharma, P. Shenoy. "Towards Continuous Policy-driven Demand Response in Data Centers." Published in *GreenNet '11*. Toronto, Canada. August 2011.

Liu, Zhenhua, A. Wierman, Y. Chen, B. Raon, N. Chen. "Data Center Demand Response: Avoiding the Coincident Peak via Workload Shifting and Local Generation." Elsevier, 2013.