Distributed Battery Control for Peak Power Shaving in Datacenters

Baris Aksanli and Tajana Rosing
Computer Science and Engineering Department
UC San Diego
La Jolla, CA
{baksanli, tajana}@ucsd.edu

Eddie Pettis
Google Inc.
Mountain View, CA
epettis@google.com

***Abstract*— Datacenters are large cyber-physical systems with continuous performance and power measurements, and real-time control decisions related to workload placement, cooling and power subsystems etc. In our work we focus on the non-ideal UPS system used to shave peak power demands. Our novel distributed battery control design has no performance impact, reduces the peak power needs, and accurately estimates and maximizes the battery lifetime. We demonstrate that models which do not take into account physical characteristics of batteries overestimate their lifetime by 2.4x. In contrast, our design is within 3.3% of the centralized battery control in terms of battery lifetime with 10x reduction in the communication costs, while shaving 23MWhrs/week of energy in a 10MW datacenter, equivalent to adding 8760 more servers at no additional power cost.**

Keywords*— *datacenters, peak power shaving, batteries, distributed control

I. INTRODUCTION

Warehouse-scale datacenters can be viewed as large scale cyber-physical systems, consisting of computational components (e.g. servers), and support subsystem which ensures correct server operation (e.g. cooling subsystem, uninterruptible power supply - UPS). While quite a bit of work has been published on job scheduling and resource management among servers, and on cooling subsystem control, the topic of using batteries present as a part of a UPS system to reduce peak power is very new. Furthermore, the few papers that recently did address this topic [1], [2], [3], neglected to consider more realistic physical characteristics of the batteries, and, as a result, estimated benefits were too optimistic - by more than a factor of two. This illustrates the need to correctly model both the physical properties of the system (in this case batteries), along with the cyber components.

Datacenters often enter long-term power contracts with usage limits based on the expected peak to limit the cost of energy. The fundamental problem with power provisioning involves using as much power as possible without exceeding a fixed power budget. The overages charged at market prices may be five times more expensive than the contracted rates [4]. Although individual servers may reach peak power during normal operation, entire clusters of servers rarely operate at peak power simultaneously [5]. Several studies have proposed peak shaving (also called power capping) to increase power utilization [1], [2], [3]. This involves reducing the contracted power level and preventing utility-facing (or breaker-facing) power consumption from exceeding the contracted power with no cost to performance.

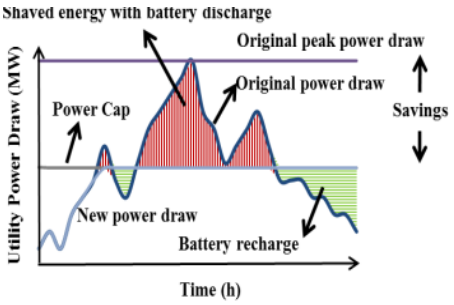


Figure 1. Peak shaving with batteries

Many mechanisms have been proposed to prevent servers from exceeding the provisioned power, including dynamic voltage and frequency scaling (DVFS) [5] [6], virtual machine power management [7], online job migration [8] [9], and batteries [1] [2] [3]. Batteries are particularly useful because they remove the performance overhead associated with meeting the power budget. Figure 1 illustrates peak shaving with batteries. The horizontal axis represents a 24-hour interval. The vertical axis is the aggregate power consumption. The upper horizontal line shows the original peak power demand and the lower one represents the power cap. If the demand is higher than the power cap, the batteries discharge to provide energy. They recharge during low power demand in preparation for the next peak. The extra energy required to recharge the batteries should be adjusted so that it does not create power cap violation. The difference between the original peak power draw and the power cap corresponds to energy savings. Alternatively, we can add more servers to the datacenter with the original power cap instead of saving energy [3]. In our work we also present the energy savings and number of additional servers we can put within the same power budget of a 10MW datacenter when using batteries for peak power shaving. Google's 10MW datacenter with 45 containers and 40000 servers is a good example of such a large scale deployment [10]. We show that our approach shaves

978-1-4799-0623-9/13/$31.00 ©2013 IEEE

23MWh/week of energy or enables us to add 8760 more servers within the same power budget.

If the datacenter uses a centralized UPS for peak shaving, then all systems in the datacenter are switched to batteries until they exhaust their capacity or the peak subsides. This technique is useful primarily for peaks that are a few minutes long due to low battery capacity [4]. Recent trends in warehouse-scale datacenters focused on distributed UPS architectures, where individual servers [10] or collection of racks [11] have their own UPS. The distributed design shaves power more effectively due to its finer granularity but only works for datacenters willing to commit to a non-standard power architecture [3]. It also requires a control system to select discharging batteries carefully because management that does not take account of physical properties of batteries, may reduce battery lifetime and increase the overall cost. But, this coordination requires significant communication overhead that may increase reaction time during sudden spikes, causing expensive overages.

We revisit the analyses for existing peak shaving designs using more realistic battery models. Existing centralized approaches discharge batteries in a "boolean" fashion: the entire datacenter power domain is fully disconnected from the utility power and supplied from the UPS. This requires batteries to discharge at much higher currents than rated, lowering both their lifetime and the actual capacity they can deliver. Simplistic models overestimate battery lifetime by 2.4x and the actual capacity by 1.2x.

Distributed UPS design addresses boolean discharge problem at the whole datacenter level by providing the ability to discharge only a subset of batteries at a time, but a battery that is selected for discharge still operates in boolean fashion. The previous designs [2] [3] leveraged battery models that cannot capture the negative effects of this boolean mode and thus overestimated the peak shaving capabilities by up to 63%. In addition, that work does not address how to manage the coordination among the distributed batteries. The coordination is required to both reduce the communication overhead and to maximize the battery lifetime. Centralized controller for distributed batteries performs well in terms of both peak shaving and battery lifetime, but may take multiple seconds to respond to a power spike. To solve this problem, we present a distributed battery control mechanism that achieves the battery lifetime and power shaving performance within 3.3% and 6% of the best centralized solution with only 10% of its communication overhead. This power shaving enables 23MWh/week energy shaving or 8760 additional servers within the same power budget when scaled to a typical 10MW datacenter.

II. Related Work

Energy and power management is a major problem for datacenter operators because of high demand and job criticality. Approaches taken include applying power shaving mechanisms such as DVFS [5] [6] and virtual machine-based power management [7] to move the jobs where energy is less expensive [8] [9]. However, all of these solutions negatively impact performance, e.g. DVFS slows down the applications; consolidation and migration both incur network delays.

In contrast, batteries have been proposed to reduce the peak power of datacenters with no performance overhead. Govindan et al. [4] suggest using existing batteries within the centralized UPS. However, the UPS can shave only peaks of a few minutes long because it powers the entire datacenter and uses lead-acid batteries (LA). Wang et al. [12] investigate additional options, such as flywheels and ultra-capacitors. Palasamudram et al. [2] and Kontorinis et al. [3] use overprovisioned distributed batteries to sustain peaks of several hours. Even though there is finer grained control of the battery output, each battery powering a single server requires high discharge current, known to decrease both the effective battery capacity and the useful battery lifetime [13]. A key problem is that these publications do not model and manage physical capabilities of batteries well, as they do not capture the negative effects of high discharge currents and thus overestimate the battery lifetime. The distributed UPS implementations do not study the overhead of managing the distributed batteries at large scale. We show in this paper that this is necessary and significant.

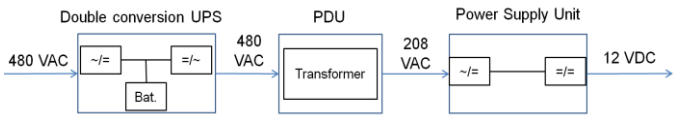
III. Distributed Battery Control

In this section, we first revisit the architectures of the existing designs. We show that their peak shaving capabilities are not accurately calculated without modeling the actual battery behavior observed in the peak shaving context. There are two battery placement architectures: centralized and distributed. The centralized design uses batteries within the datacenter-level UPS and does not require additional power equipment. A common power delivery hierarchy for this design is shown in Figure 2-a. When peak shaving occurs, UPS powers the entire datacenter, discharging the batteries at high rate. According to Peukert's Law, this drains battery capacity quickly [14]. Furthermore, the AC-DC-AC double conversion reduces UPS efficiency and decreases useful battery capacity by up to 20%.

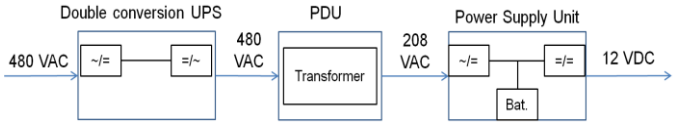
The distributed design co-locates the servers and batteries and eliminates the DC-AC battery power conversion [2] [3]. A sample design is shown in Figure 2-b. Each server may be switched to battery independently. This leads to finer grained control of the aggregate battery output because only a fraction of the servers are powered via battery at any given time. Together, conversion efficiency and fine-grained control permit hours of peak shaving compared to a few minutes of the traditional centralized designs.

A. Issues with the Previous Designs

Even though the distributed design achieves finer grained control, each battery still needs to power the entire server with high discharge currents. The existing distributed architectures do not account for the negative effects of high discharging rate. Figure 3 shows the peak shaving capability of the distributed design with and without a detailed battery model. We assume that each server is attached with a 20 Ah LA battery. A power cap is defined for each server at 255W. This



a) State-of the art power distribution with centralized battery



b) State-of-the-art power distribution with distributed batteries
PDU = Power Distribution Unit
UPS = Uninterruptible Power Supply

Figure 2. Centralized vs. distributed battery placements

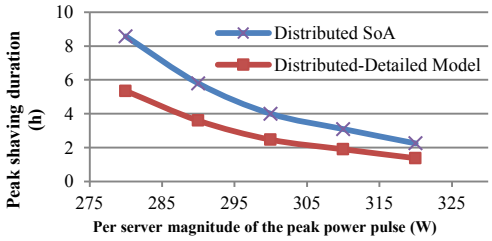


Figure 3. Peak shaving capabilities of the distributed design

value reflects the average maximum server power that the datacenter can allocate. For example, some servers can go higher than 255W that have more stringent needs, and some can have less than that. But the average power should not be more than 255W. This defines the datacenter power cap by restricting the total datacenter power consumption to 255W*#servers. The horizontal axis illustrates a range of peak server powers. The vertical axis represents the peak shaving duration. The upper curve estimates peak shaving duration with a simplistic battery model and the lower curve uses a more exact model, both outlined in section III.B. We see that the power shaving duration can be overestimated by up to 62% without a detailed battery model.

The ability to discharge batteries independently is crucial in the distributed design. However, since not all batteries are discharged at the same time, they may have very different discharge patterns, depending on server load. This variation results in capacity imbalances between them. Figure 4 represents this variation one and two years after the batteries are deployed when selecting batteries randomly each time battery power is needed. The outermost circle represents the nominal battery capacity. The innermost circle corresponds to the end of the battery's useful life. We consider a battery dead when it can use only 80% of its nominal capacity [15]. Each battery is denoted by a ray extending from the center. The length of the ray indicates the battery capacity. The line between the nominal and dead capacity indicates the ideal battery lifetime at each age. This graph illustrates that remaining battery capacities significantly deviate from the ideal. This deviation increases over time, resulting in early battery replacements, increasing the battery related costs. We

Figure 4. The maximum battery capacity with random battery selection

may reduce this variation by selecting batteries more effectively. This requires coordination between the batteries, which may have delays on the order of seconds depending on network congestion. Large delays can lead to miscalculating the total available battery capacity, reducing the peak shaving.

B. Detailed Estimates of Battery's Physical Condition

UPS-based peak shaving requires accurate estimates of battery's physical condition. This section provides estimates of battery's depth-of-discharge (DoD), available capacity after recharging and discharging, and a method for calculating useful capacity over time. The available battery capacity at a given time is defined as the state-of-charge (SoC) and reported as a percentage of the maximum capacity. State-of-health (SoH) quantifies the maximum capacity over time as a percentage of the initial capacity.

Battery lifetime modeling has been extensively studied before, especially in the context of mobile devices, e.g. [16] [17]. We combine a few models as a part of our work. Coulomb counting method presented in [18] describes the relation between DoD level and SoH using. The impact of using high discharge current rates on SoH is studied further in [13]. We also include Peukert's law which states that the effective capacity of a battery decays exponentially depending on discharging current [14]. The main benefit of the model we present is its simplicity and ability to easily leverage it in a large scale installation as it requires only voltage and current readings for all the calculations. We start describing our model by first calculating released capacity during a discharge event:

$$C_{released} = \Delta t * I_{discharge} \quad (1)$$

where Δt is the length of the time interval and $I_{discharge}$ is the discharge current. Then, we compute the DoD as the released capacity over the effective capacity:

$$C_{eff} = C_R * \left(\frac{C_R}{I_{discharge}*H}\right)^{k-1} * \frac{SoH}{100} \quad (2)$$

where C_{eff} is the effective capacity when using $I_{discharge}$ and C_R is the rated capacity. H is the rated discharge time in terms of hours (normal hours) and is obtained from the data sheets [14]. Peukert's exponent, k, reflects battery chemistry. The typical value is 1.15 for lead-acid (LA) and 1.05 for lithium iron phosphate (LFP) batteries [19]. C_{eff} is also scaled with an SoH value (defined in equation (3)) to reflect the capacity loss. The DoD is subtracted from the SoC at the end of each interval. When the discharge ends, we save the total DoD value, DoD_{final} as (100-SoC)%.

The effective capacity decreases with higher discharging currents. Figure 5 shows this effect on 20Ah LA and LFP

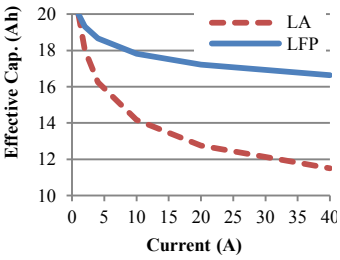


Figure 5. Effective capacity of 20Ah LA and LFP batteries in our model

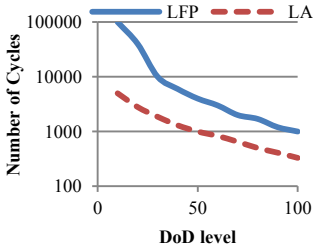


Figure 6. Cycle life of LA and LFP batteries rated at 20h [27], [28]

batteries. The horizontal and vertical axes show the effective battery capacity and discharging current respectively. The LA battery has significantly less capacity at high current because it has greater nonlinear behavior, represented by a larger Peukert exponent. At 40A, the LA battery loses 42% of its nominal capacity, but the LFP battery loses only 15%.

The battery SoH is updated after a complete recharge/discharge cycle [18]. This update depends on the battery chemistry, effective capacity and DoD_{final}. The number of available cycles decreases with larger DoD_{final}. We use a lookup table for each battery chemistry to define the effects of DoD_{final} shown in Figure 6. This data is available in battery datasheets. In Figure 6, the horizontal axis shows the DoD level for charge/discharge at 20h discharge rate, which is defined as the current that drains the battery in 20h. The vertical axis illustrates the number of cycles a battery can provide for a particular DoD level. As DoD increases, the cycle count decreases exponentially.

We calculate the impact of each cycle on SoH by normalizing the effect of one cycle with DoD_{final} value over the lifetime. The lifetime is defined as the interval in which battery SoH is between 100% and the SoH value that determines when the battery is dead, SoH_{dead}. It is generally assumed to be 80% [15]. If the battery has $Cycles_{DoD\_final}$ cycles available with DoD_{final} value, the SoH of battery is updated as [13]:

$$SoH = SoH - (100 - SoH_{dead}) * \frac{1}{Cycles_{DoD\_final}} * \frac{C_R}{C_{eff}} \quad (3)$$

Battery management unit normally monitors and manages battery voltage and current, making it easy to implement our model which just requires these two measurements. In contrast, the simple battery model used by previous work [1] [2] does not calculate C_{eff}. Instead, it uses nominal battery capacity, C_R, to compute DoD, resulting in up to 42% overestimated discharge duration. It also does not account for the effects of decreasing SoH on the C_{eff}, which further increases errors.

TABLE I. BATTERY MODEL VALIDATION

Battery	Error
Li-Ion$_5$	4.3% ± 2%
Li-Ion$_6$	5.8% ± 3.6%
Li-Ion$_7$	3.8% ± 2.7%

Model verification: We use battery data available from the NASA Ames Prognostics Data Repository [20] to validate the accuracy of our model. The repository includes the results of the experiments that charge/discharge the 2Ah Li-ion batteries at different currents and temperatures. Each result set consists of the complete charge-discharge profiles of a single battery until it reaches end-of-life. We use the results of 3 batteries, tested at room temperature, to check our model and compare the SoH values at the end of each charge/discharge cycle. Table I shows that our model has a 4.6% average error compared to the battery measurements.

C. Distributed Control Mechanism

The distributed architecture permits finer grained control than centralized architectures because server batteries may be discharged independently. This process requires intelligent selection of batteries during each power peak. In section III.A, we demonstrate that simple battery selection algorithms may distribute power load unevenly and induce high variations in battery SoH. This variations lead to premature battery replacements because capacity is reduced sooner than expected. Therefore, the distributed design requires a mechanism that monitors battery health and selects batteries in a way that minimizes this variation.

The distributed controller first estimates the number of batteries to discharge during each peak power pulse as follows:

$$N_{batteries} = \left\lceil \frac{P_{demand} - P_{Threshold}}{V_{battery} * I_{discharge}} \right\rceil \quad (4)$$

where $\lceil . \rceil$ is the ceiling function, P_{demand} is the peak power demand at a given time, $P_{threshold}$ is the peak power threshold to be maintained, $V_{battery}$ is the single battery voltage and $I_{discharge}$ is the single battery discharging current. We use 12V batteries [10] and set $I_{discharge}$ to 23A. Since the servers use the battery power without AC-DC conversion, the battery incurs no conversion losses in the server. In our experiments, the measured server peak power is 350W and power supply unit (PSU) efficiency is 80%. Therefore, the server actually uses 280W, which corresponds to 23A discharging current.

An ideal controller for the distributed design should poll every server to gather data on server power demand, battery SoC and SoH. This process requires message exchanges through the datacenter network. However, the controller becomes subject to communication delays between the thousands of servers and large background traffic. Previous work shows that the switch delay can increase by over 100x with excessive queuing in the switches [21].

Our new method groups the batteries into multiple distributed controllers to address the communication complexity. Table II lists the possible group sizes and shows the corresponding level in the datacenter power hierarchy. The two extremes represent fully localized control, at each individual server, and the datacenter level, which is equivalent to fully centralized control. In between are rack level, PDU, which consists of approximately 10 racks, and cluster level, which is about the size of a typical datacenter container. We

TABLE II. GROUP SIZES

Hierarchy Level	Size of a group
Server	1
Rack	20-50 [4]
PDU	200 [3]
Cluster	1000 [3]
Data center	Multiple clusters

TABLE III. POLICIES TO CONTROL BATTERY GROUPS

Policy	Communication
Random	Local
LRU(Iterative) [3]	Local
Max-SoH-local	Local
Max-SoH-global	Global
Max-SoH-limited-comm.	3 groups
Max-SoH-more-limited-comm.	2 groups

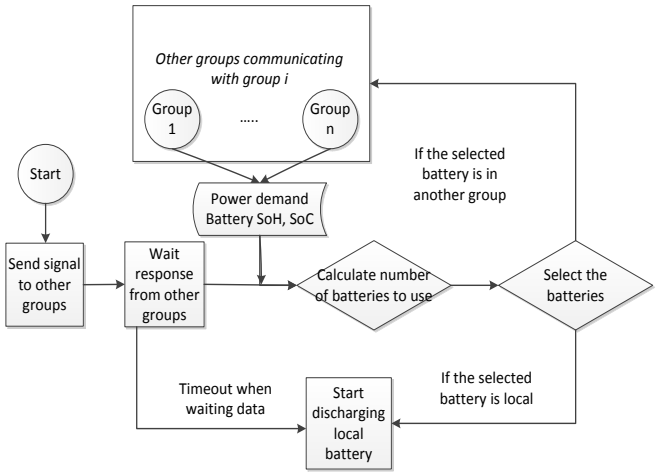


Figure 7. Battery selection with communication based policies

chose these hierarchy layers as they correspond to the typical organization found in the datacenter's power hierarchy.

Each level of the controller implements one of the policies shown in Table III to select a battery. Random, Least-Recently-Used (LRU) and Max-SoH-local policies make a local decision regarding which battery to use for peak power shaving from their immediate group. Random policy selects a random battery from available ones. LRU, also used in [3], always selects the next available battery from its local list. Max-SoH-local chooses the available battery with the greatest SoH value. We assume that the controllers do not know or predict the length of the upcoming peak power pulse. Hence, selecting the battery with the greatest SoH value is the best a controller can do because it minimizes the probability that the selected battery empties during the peak power pulse. These policies result in lower latency with smaller groups, but their knowledge about total power demand and battery status is limited.

We implement three other Max-SoH policies to address this problem. They are similar to Max-SoH-local, but controllers can communicate with other ones during a decision process. The Max-SoH-global policy represents a centralized controller and uses all data available in the system. Although this controller can make the best decision, it leads to large communication delays and becomes a single point of failure. Max-SoH-limited and –more-limited communication policies are limited to two and one other groups. Each group's partners are assigned statically based on power and network infrastructure. We compare these policies with the local ones to demonstrate the trade-off between the communication overhead and power shaving, and battery lifetime performance.

Figure 7 shows the peak shaving and the battery selection process of a single group when communicating with others. The number of sharing groups depends on the policy. The controller first awaits power consumption and battery data from its sharing groups. It next computes the peak power that can be shaved by finding the number of batteries required and selects the batteries to use. Local batteries discharge immediately. Remote batteries require explicit signals to their controller. We use a timeout when waiting for the data from other groups to avoid problems, including miscalculating the total available battery capacity. The timeout may decrease the quality of selection since less data will be present.

IV. METHODOLOGY

We present our experimental setup and describe how we use it to evaluate different battery designs and control implementations. We start with power measurements and the description of the workloads we use. Since our measurement infrastructure is not of sufficient size to compare to effects observed in large scale datacenters, we also design and describe our simulation platform. Lastly, we present a number of test cases that illustrate the improvements with a detailed battery model and the benefits of using distributed control to manage the batteries. Our results show that both power shaving capabilities and battery lifetime is overestimated with a simple battery model. Distributed and hierarchical control decreases the communication overhead of the centralized solution by 10x while staying within 6% and 3.3% of the centralized control performance in terms of power shaving and battery lifetime, respectively.

A. Power Measurements and Workloads Run

We use measurements from our datacenter container on campus to estimate the overall power cost for a larger scale datacenter. Our container has 200 servers consisting of Nehalem, Xeon and Sun Fire servers running Xen VM. We run a mix of commonly used benchmarks to measure power and performance of service and batch jobs on our servers. We use RUBiS [22] to model service-sensitive eBay-like workload with 90th percentile of response times at 150ms, and Olio [23] to model social networking workloads with response times ranging from 100ms up to multiple seconds, depending on the type of request (e.g. text post vs. video upload). Multiple Hadoop [24] instances are run as batch jobs. We measure performance at 10ms sampling rate and obtain power at 60Hz.

The measurements are used to create an event-based simulator that embeds the power information and the workload characteristics to simulate a larger datacenter environment. We model each 8-core server with an M/M/8 queuing model, and a linear CPU utilization based power estimate commonly used by others [5] [6]. Table IV shows that the average simulation error is well below 10% for all quantities of interest, with 3% average error for power estimates, while performance for

services has only 6% and MapReduce completion times are within 8% of measured values.

TABLE IV: VERIFICATION OF POWER AND PERFORMANCE MODELS

Parameter	Ave. Error
Avg. Power Consumption	3%
Services QoS	6%
Avg. MapReduce Comp. Time	8%

To understand the benefits of peak power shaving, we model the typical user request load for a full datacenter. We use a year of publicly available traffic data of two Google products, Orkut and Search, as reported in Google Transparency Report [25]. A week's worth of workload combinations based on the waveform shown in Figure 3 of [26] where Social Networking and Search workloads represent service jobs, and MapReduce is for batch jobs. Table V shows the workload parameters, while Figure 8 compares each job's contribution to the total datacenter load. The maximum load ratio is around 80% with average of 45%.

TABLE V. WORKLOAD PARAMETERS

Workload	Average Time	
	Service	**Interarrival**
Search [6]	50ms	42ms
Social Networking [23]	1sec	445ms
MapReduce [26]	2 min	3.3 min

B. Datacenter and Battery Simulation

Fine event granularity in simulation is computationally expensive, so we limit our datacenter simulation period to a week. We extract the datacenter power consumption along with the each battery's charge/discharge profile. These values are scaled to longer time intervals in order to analyze the required battery DoD, the discharge current profile and to get an estimate of the lifetime.

Figure 9 shows the DoD level variation with different level controllers over a week when DoD_{goal} is set to 60%. Higher level distributed controllers are more consistent. They use all the available battery capacity, because the battery power output can be distributed evenly across them. In contrast, the DoD value is uniformly distributed between 20% and 60% with a server level controller because individual server power profiles vary and there is limited coordination between the servers.

After analyzing short-term battery usage profiles, we use the battery model described previously and simulate only charge/discharge cycles to estimate the battery lifetime. We simulate several years of simulation time and consider a battery dead when its SoH goes below 80% [15]. We include both LFP [3] & LA [1], [2], [4] batteries in our study. The battery capacity is sized to the maximum volume that fits per server (40Ah and 20Ah respectively) with 12V nominal voltage [3].

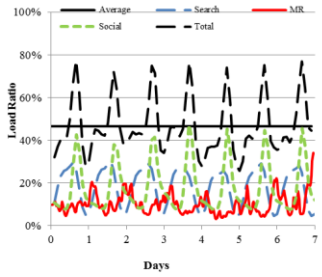


Figure 8. Datacenter workload mix

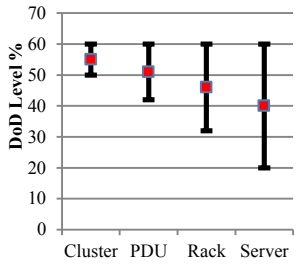


Figure 9. DoD level variation

V. RESULTS

A. Accuracy of the Battery Model

We start our evaluation by comparing the power capping capabilities of the state-of-the-art (SoA) battery placement designs with both LA and LFP batteries. The SoA centralized design adjusts the battery capacity to handle only emergency cases, which last only a few minutes. We assume that this design has a 3200 Ah LA battery as proposed in [10], [3] to support a single datacenter container. In distributed case, each server has a dedicated 20Ah LA or 40Ah LFP battery, the maximum possible given their volume, same as in [3]. These battery capacities are adjusted to match previous work. In Table VI, we compute how long the batteries can shave a fixed average peak power pulse per server with specified magnitude where the datacenter power cap is defined at 255W/server. We first apply the simplistic battery model used by recent SoA publications. This model accounts only for the total battery capacity and ignores the effects of high discharge currents and nonlinear behavior of different battery types [2] [4]. Table VI shows that the centralized design can shave a peak for only 7 minutes whereas the distributed design can successfully shave peaks of over 3 and 6 hours with LA and LFP batteries, respectively.

Next, we use the detailed battery model presented in section III.B to account for the battery type and the negative effects of high discharging currents. Surprisingly, the peak power shaving amount can be overestimated by 133% in the centralized design. The discharging current in the distributed design is still high, but the rate of the discharging current is lower relative to total battery capacity. This results in error of 64% for LA batteries, and 14% for LFP. LFP's error rate is up to 4.5x lower than the LA's because of its more linear discharge behavior. However, the error, a result of an inaccurate model and interaction with physical devices – the

TABLE VI. PEAK SHAVING CAPABILITIES OF THE SQUARE POWER PEAK WITH CENTRALIZED AND DISTRIBUTED DESIGNS USING LA & LFP BATTERIES.

Peak power/ server (W) – shaving %	Centralized – 3200Ah total capacity LA			Distributed – 20Ah/server LA – 40Ah/server LFP					
	Power capping duration (min)		Error (%)	Power capping duration (min)				Error (%)	
	Model			simple model		detailed model		LA	LFP
	Simple	Detailed		LA	LFP	LA	LFP		
300 – 15%	8	4	100%	240	480	148	423	62%	13%
310 – 17%	7	3	133%	186	372	114	327	63%	14%
320 – 20%	7	3	133%	135	270	82	237	64%	14%

batteries, is still significant to affect peak power shaving decisions, such as determining battery design or the total needed capacity.

TABLE VII. BATTERY LIFETIME ESTIMATION COMPARISON

	LA	LFP
Low current rated estimations	3 years	10 years
Our estimations	1.4 years	4.1 years

We use our battery model with our long term battery simulation to estimate the average lifetime of an LA and LFP battery when shaving peak power. Table VII compares our long-term battery lifetime estimates with previous work [3], [2]. Neglecting the effects of high current results in high error: as much as 210% and 240% longer battery lifetime estimates leading to severely underestimated battery costs and overstated cost savings due to peak shaving.

B. Performance of the Distributed Control

We next evaluate the performance of our communication based distributed controllers, which increase the overall battery lifetime by balancing the power demand across the batteries. We use 1000 40Ah LFP batteries [3] with configurations shown in Table II, with policies described in Table III. Tables VIII and IX summarize the comparison between different policies and group sizes in terms of peak shaving and average battery lifetime. To calculate the best peak power shaving for each configuration we first use the workload distribution shown in Figure 8 to create the power profile of the datacenter over a week. We initially set a power cap, e.g. 280W/server, and reduce it in each simulation experiment until we cannot guarantee that cap. We then compute the power shaving percentage with the amount of power shaved over the peak.

TABLE VIII. AMOUNT OF ENERGY SHAVED FOR A 10MW DATACENTER PER WEEK IN MWHRS & (% OF POWER SHAVED COMPARED TO THE PEAK)

Policies	**Datacenter partitioning**				
	1 cont.	**5 PDUs**	**10 PDUs**	**50 Racks**	**1000 Servers**
Local	30 (19%)	14.3 (16%)	11.2 (15%)	4.8 (12%)	2.5 (10%)
Max-SoH – glob.	30 (19%)	30 (19%)	30 (19%)	30 (19%)	30 (19%)
Max-SoH – lim. comm.	30 (19%)	23.1 (18%)	14.3 (16%)	6.6 (13%)	2.5 (10%)
Max-SoH – m-lim. comm.	30 (19%)	18.1 (17%)	11.2 (15%)	4.8 (12%)	2.5 (10%)

Table VIII shows energy savings per week due to various peak power shaving strategies scaled to a datacenter of peak capacity 10MW, along with peak power shaving percentages for each configuration based on the smallest power cap we can guarantee. Google's 10MW, 45 container datacenter, with 40000 servers [10] is an example of such a deployment. The best peak power shaving can be achieved with a centralized controller – as much as 19% of the peak power of the entire datacenter, equivalent to 30MWh/week of the 10MW datacenter, or 9380 more servers with no additional peak power cost. Although we have the same total battery capacity in all of the configurations, the power shaving capability decreases significantly with lower level controllers because of their limited knowledge of the total power demand. They shave up to 50% less power and 92% less energy compared to the best solution. In contrast, we observe that our PDU level controllers with communication can shave 18% of the peak power and 23MWh energy, within 6% and 23% of the centralized solution.

Table IX shows the average battery lifetime, normalized to the case with the individual server level controllers. Local policies perform poorly regardless of their battery selection algorithm as they are unaware of batteries in other groups. Changing the group size does not affect performance of the local policies, except for Max-SoH, which reduces to Max-SoH-global when there is only one group. The centralized controller gives the best results, performing 2x better than the local policies by processing the data from all the batteries. The performance of policies with limited communication depends on the group size and communication span. Increasing span with 5 PDU level controllers using limited communication by one group results in up to 20% longer battery lifetime, within 3.3% of the centralized solution. Thus, our distributed controllers well approximate the performance of the centralized controller in terms of both power shaving and battery lifetime, showing that intelligent control and good characterization of datacenter's physical infrastructure can dramatically improve the overall system efficiency.

TABLE IX. NORMALIZED AVERAGE BATTERY LIFETIME

Policies	**Datacenter partitioning**				
	1 cont.	**5 PDUs**	**10 PDUs**	**50 Racks**	**1000 Servers**
Random	1.02	1.03	1.04	1.04	1.00
LRU	1.07	1.07	1.07	1.07	1.00
Max-SoH-local	1.97	1.07	1.07	1.07	1.00
Max-SoH – glob.	1.97	1.97	1.97	1.97	1.97
Max-SoH – lim. comm.	1.97	1.91	1.76	1.77	1.73
Max-SoH - more lim. comm.	1.97	1.59	1.59	1.51	1.48

C. Communication overhead analysis

In this architecture, each group controller polls the servers in its group using the datacenter network to collect server power consumption and battery statistics. The controller then delivers the battery selection decision to the servers. Intra-rack communication is extremely fast, but relaying messages through multiple switches introduces far more delays. Assuming a common a fat-tree topology, we model the links in the network with 10 Gbps capacity, which can transmit a 1K package at 1us. We evaluate an ideal network, without queuing delay, a network with normal level congestion where a single message transmission delay in a switch is 50us and a network with a high level congestion reaching 350us delay [21]. In this experiment, container level models global communication.

Figure 10 shows the results of the communication analysis. The vertical axes are on a log scale. The total delay increases exponentially with higher level controllers because of the increasing number of out-of-rack communication signals, going over several hops. Rack level controller gives the best

results with only tens of ms total delay even in the presence of high congestion. However, it has 32% less power shaving and 11% shorter battery lifetime compared to the centralized solution. In contrast, the container level controller may have seconds of delay, 100x more than the rack level in high congestion. With 5 PDU controllers there is a 10x decrease in the total communication delay relative to the global solution while being within 6% and 3.3% of the centralized controller in terms of peak power shaving and battery lifetime. Clearly this is a great replacement for the centralized control for peak power shaving with batteries.

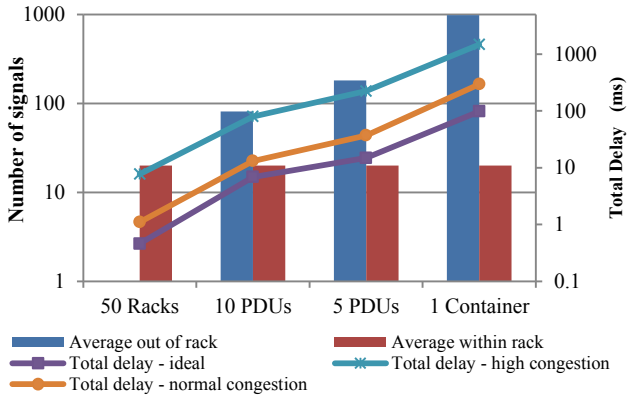


Figure 10. Communication overhead

VI. CONCLUSION

Peak shaving with batteries has gained significant importance because of its ease of applicability and no performance overhead. Previous work does not model the physical characteristics of the batteries and therefore overestimates the benefits by as much as 2.4x longer battery lifetime and up to 113% longer peak power shaving duration. We propose a distributed control mechanism to manage the physical properties of the batteries. Our mechanism removes the single point of failure of the traditional centralized control and reduces its communication overhead by 10x while being within 6% and 3.3% of its peak power shaving and battery lifetime, respectively. This power shaving leads to 23.1 MWh energy shaving when scaled to a typical 10MW datacenter [10]. This work illustrates the benefits of correctly modeling and tracking the physical phenomena (batteries). Thus, designing an appropriate infrastructure to manage the batteries is critical for obtaining great results.

ACKNOWLEDGEMENTS

This work was sponsored in part by Google, NSF ERC CIAN (grant number 812072), NSF IRNC TransLight–StarLight (grant number 962997), and CNS. The authors also acknowledge the support of the Multiscale Systems Center (MuSyC), one of six centers under the Focus Center Research Program (FCRP), a Semiconductor Research Corporation program (SRC).

REFERENCES

[1] S. Govindan, D. Wang, A. Sivasubramaniam and B. Urgaonkar, "Leveraging stored energy for handling power emergencies in aggressively provisioned datacenters," in *ASPLOS*, 2012.

[2] D. Palasamudram, R. Sitaraman, B. Urgaonkar and R. Urgaonkar, "Using Batteries to Reduce the Power Costs of Internet-scale Distributed Networks," in *ACM Symp. on Cloud Computing*, 2012.

[3] V. Kontorinis, L. Zhang, B. Aksanli, J. Sampson, H. Houman, E. Pettis, D. Tullsen and T. Rosing, "Managing Distributed UPS Energy for Effective Power Capping in Data Centers," in *ISCA*, 2012.

[4] S. Govindan, A. Sivasubramaniam, and B. Urgaonkar, "Benefits and Limitations of Tapping into Stored Energy For Datacenters," in *ISCA*, 2011.

[5] X. Fan, W. Weber and L. Barosso, "Power provisioning for a warehouse-sized computer," in *ISCA*, 2007.

[6] D. Meisner, C. Sadler, L. Barroso, W. Weber and T. Wenisch, "Power management of online data-intensive services," in *ISCA*, 2011.

[7] R. Nathuji and K. Schwan, "Vpm tokens: virtual machine-aware power budgeting in datacenters," in *HPDC*, 2008.

[8] N. Buchbinder, N. Jain and I. Menache, "Online job-migration for reducing the electricity bill in the cloud," in *Networking*, 2011.

[9] L. Rao, X. Liu, L. Xie and W. Liu, "Minimizing Electricity Cost: Optimization of Distributed Internet Data Centers in a Multi-Electricity-Market Environment," in *INFOCOM*, 2010.

[10] Google, "Google Summit," 2009. http://www.google.com/corporate/datacenter/events/dc-summit-2009.html.

[11] Facebook, "Hacking conventional computing infrastructure," 2011. http://opencompute.org/.

[12] D. Wang, C. Ren, A. Sivasubramaniam, B. Urgaonkar and H. Fathy, "Energy Storage in Datacenters: What, Where, and How much?," in *SIGMETRICS*, 2012.

[13] S. Drouilhet and B. Johnson, "A Battery Life Prediction Method for Hybrid Power Applications," in *AIAA Aerospace Sciences Meeting and Exhibit*.

[14] SmartGauge, "Peukert's law equation and its explanation," 2011. [Online]. Available: http://www.smartgauge.co.uk/peukert.html.

[15] PowerSonic, "Technical manual of LA batteries," http://www.power-sonic.com/technical.php.

[16] L. Benini, G. Castelli, A. Macii, E. Macii, M. Poncino and R. Scarsi, "Discrete-time battery models for system-level low-power design," in *IEEE Transactions on VLSI Systems*, 2001.

[17] D. Rakhmatov, S. Vrudhula and D. A. Wallach, "Battery lifetime prediction for energy-aware computing," in *ISLPED*, 2002.

[18] K. Soon Ng, C.-S. Moo, Y.-P. Chen and Y.-C. Hsieh, "Enhanced coulomb counting method for estimating state-of-charge and state-of-health of lithium-ion batteries," *Applied Energy,* vol. 86, no. 9, 2009.

[19] F. Harvey, "Table with Peukert's exponent for different battery models," 2001. http://www.electricmotorsport.com/store/ems_ev_parts_batteries.php.

[20] B. Saha and K. Goebel, "Battery Data Set, NASA Ames Prognostics Data Repository," 2007. http://ti.arc.nasa.gov/tech/dash/pcoe/prognostic-data-repository/.

[21] "Priority flow control: Build reliable layer 2 infrastructure," http://www.cisco.com/en/US/prod/collateral/switches/ps9441/ps9670/white_paper_c11-542809.pdf.

[22] "RUBiS," http://rubis.ow2.org/.

[23] Apache, http://incubator.apache.org/olio/.

[24] "Hadoop," http://hadoop.apache.org/.

[25] Google, "Google Transparency Report," http://www.google.com/transparencyreport/traffic.

[26] Y. Chen, A. Ganapathi, R. Griffith and R. Katz, "The case for evaluating MapReduce performance using workload suites," In Technical Report No. UCB/EECS-2011-21, 2011.

[27] Windsun, "Lead-acid batteries: Lifetime vs Depth of discharge," 2009. http://www.windsun.com/Batteries/Battery_FAQ.htm.

[28] M. Swierczynski, R. Teodorescu and P. Rodriguez, "Lifetime investigations of a lithium iron phosphate (LFP) battery system connected to a wind turbine for forecast improvement and output power gradient reduction," in *In BatCon'08.*, 2008.