Power Availability Provisioning in Large Data Centers

Sriram Sankar
Microsoft Corporation
One Microsoft Way,
Redmond, WA 98052
sriram.sankar@microsoft.com

David Gauthier
Microsoft Corporation
One Microsoft Way,
Redmond, WA 98052
davidgau@microsoft.com

Sudhanva Gurumurthi
Department of Computer Science
University of Virginia
Charlottesville, VA 22904
gurumurthi@virginia.edu

ABSTRACT

Enterprise data centers are provisioned with conservative redundancies built into their power infrastructures to handle failures. Conservative over-provisioning of power capacity for availability reasons results in significant capital investment for large enterprises because this capacity is designed for failure conditions that do not happen often. On the other hand, under-provisioning this capacity runs the risk of affecting the performance of the data center when failures do happen, through either service unavailability or degraded service performance. Hence, there are interactions and tradeoffs between power capacity utilization, power redundancy, and data center performance that is often overlooked. Our work proposes a provisioning methodology for the power delivery infrastructure called *power availability provisioning* that addresses this challenge. We provide observations on power infrastructure design based on industry experience operating large data centers. We characterize power availability events, motivate the need for workload-driven power availability provisioning, and describe a methodology to estimate performance impact due to power availability events. We then present an unconventional redundancy technique (N-M redundancy) that proposes reducing redundant power equipment, leveraging observations from our study.

Categories and Subject Descriptors

C.4. Performance of Systems, C.5. Computer System Implementation

General Terms

Management, Measurement, Performance, Design, Reliability.

Keywords

Data Center, Availability, Power Infrastructure, Large Scale Systems

CF'14, May 20–22, 2014, Cagliari, Italy

1. INTRODUCTION

Large enterprises continue to build data centers in order to respond to the growing demand in online and cloud services. Power infrastructure costs alone amount to $10-20 million per megawatt of data center capacity [12]. Utilizing the allocated power capacity efficiently results in full use of capital that is invested. Given the importance of power infrastructure for continuously running a data center, data center operators typically tend to be conservative when designing the data center. While power-capacity provisioning has received considerable focus [6] [13], availability plays a key role in overall infrastructure provisioning. For instance, data center operators usually provision additional power capacity for enabling concurrent maintenance during times when certain parts of the infrastructure need to be taken down for maintenance. They also provide power back-up systems that are redundant (systems that have twice the number of generators are termed 2N, while systems that have one additional generator are termed as N+1). While the generators themselves are back-up power source for the primary power source, conservative design can lead to over-provisioning of the back-up sources as well.

When provisioning the power capacity of a data center, the availability requirements are factored into the design itself. The "number of nines" is a commonly used term that denotes the overall availability of the data center [10]. A data center with 99.9 availability denotes that out of 8,760 hours in a year, the data center can be unavailable for close to 8.76 hours. A 99.999 data center facility can be expected to be unavailable for 5.256 minutes out of a year. Typically, the "number of nines" availability metric is computed with component design mean time between failures (MTBF) and mean time to repair (MTTR) numbers [3]. A data center with a higher availability requirement tends to be more expensive to design and construct because it requires redundant components needed to provide that availability. The Uptime Institute, a widely recognized research and consulting organization, has published a tier mechanism [24] that classifies data centers into various classes based on availability. Tiers fall into four types starting with the basic reliability option (Tier 1), where there is no redundancy or concurrent maintenance, and proceeding to Tier 4 data centers, which feature fault tolerance and concurrent maintenance. Figure 1 shows the construction cost and corresponding total expected downtime of different tiers of data centers. A Tier 4 data center, with the highest possible availability among the tiers, costs almost 2.5 times more to build than a Tier 1 facility.

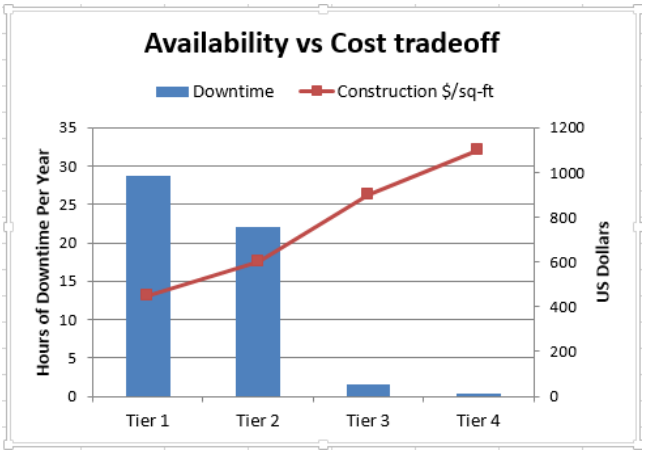

Figure 1: Expected Downtime associated with each tier and corresponding cost to construct the data center per square foot (lower downtime implies higher availability)

However, a data center's number of nines alone is not a good operability metric for power provisioning or management because it fails to capture the impact of availability on data center performance. For instance, the number of nines is a summary metric that abstracts frequency or duration of events: 8 hours of unavailability could have one event totaling 8 hours or 100 events of 0.08 hours each. Hence, this metric is inadequate to capture the actual distribution of the power availability events in the data center, which is important for making power provisioning decisions.

In this paper, we make the following key contributions:

- Using observations from industry experience operating large facilities, we present insights into power events in real data centers. We show that a significant portion of observed power events are of short duration (within 21 seconds) and typically have varied inter-arrival times (several days between each event).
- We present a methodology to capture the impact of availability events on the data center design. We argue that power utilization and power availability models should be constructed and be used as workloads to guide power availability provisioning.
- We present a method for estimating data center performance in the presence of power events by using artificial power caps on two real data center workloads running on actual servers.
- We present a novel redundancy technique, N-M redundancy, which is a reduced redundancy provisioning method that contrasts with the additive redundancy policies currently used (2N or N+1). We also show the validity and cost benefits of N-M redundancy.

The rest of the paper is structured as follows: Section 2 describes power infrastructure design for data centers. Section 3 describes power event traces and presents power availability characterization from real data centers. Section 4 applies the N-M redundancy technique to a data center using performance models. Section 5 provides related work, and Section 6 presents future work and conclusion.

2. POWER INFRASTRUCTURE

This section presents background on the power delivery infrastructure in data centers, describes the common availability related metrics, and presents various types of power availability events that occur in data centers.

2.1 Power Delivery Infrastructure in Data Centers

The power delivery infrastructure in a data center consists of a primary utility that supplies power to the entire facility (in some scenarios, more than one utility supplies power for increased redundancy). Once power is delivered from the utility, it goes through a utility substation, and then through a medium-voltage (MV) distribution. In some cases, smaller data centers take power at low-voltage (LV) levels. The power distribution includes transformers to transform to lower voltage levels (from distribution levels to operability levels) and circuit breakers to guard against unintended power scenarios. In addition to the utility distribution, the power infrastructure also includes back-up (that allows for continuous operation when the utility provider encounters a failure) and reserve distributions (that allow for concurrent maintenance). Uninterruptible power supply (UPS) systems enable the transfer of power from the utility to generators in case of a power event. The UPS systems have batteries that store power for a pre-designed amount of time and act as an intermediary power source for the IT equipment (servers and network devices), in order to provide time for the generators to power up and provide the continuous back-up supply. The generators themselves are rated for the power capacity that they need to support and typically have a fuel tank that supplies the fuel needed for producing power to the facility.

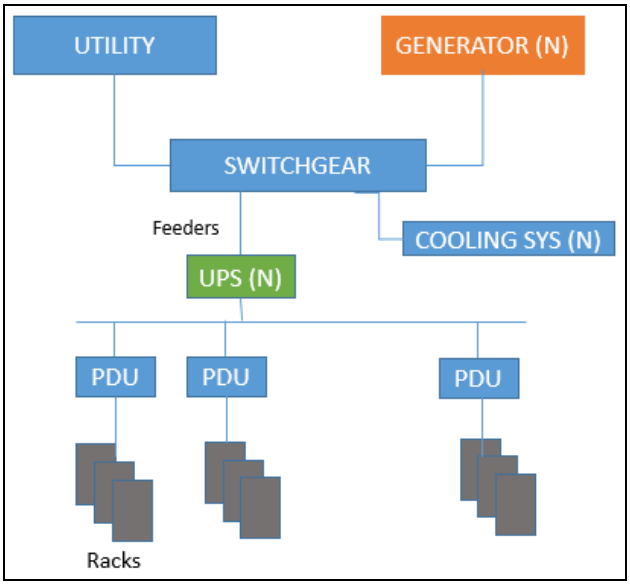

Figure 2: DC Power Infrastructure with Generators, but no redundancy

Figure 2 shows a power-distribution hierarchy with different high-level components, including utility, generator, switchgear (that allows for transfer and switching between sources), UPS, and power-distribution units (PDUs). The racks hold server units that

are powered through power whips from the PDU. Typically, the servers present in the racks consume the highest proportion of power delivered; however, the cooling equipment and power distribution equipment also consume a proportion of total power supplied to the data center from the utility. This proportion is commonly measured using PUE (Power Usage Effectiveness). PUE refers to the fraction of power consumed by the entire facility, including cooling, divided by power consumed by IT equipment alone. A PUE closer to 1 denotes a very efficient data center facility (1.25 PUE is typical of traditional data centers). A lower PUE denotes that the bulk of power spent in the data centers are actually consumed by the IT equipment (server and network equipment, including storage, compute devices). PUE is still a measure of power consumption effectiveness, and does not reflect power availability impact.

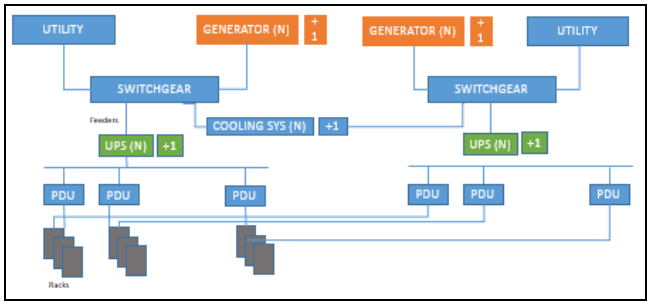

Figure 3: Power Infrastructure with dual utility, 2N distributions, N+1 component in each distribution

Figure 3 shows a highly redundant distribution that has two utilities feeding the data center. If one utility goes down, the expectation is that the other utility takes over. If the other utility also fails, there are generators that take over. This example illustrates a highly redundant distribution path. In some implementations, the generators are shared between the two utility paths. The UPS are N+1 and are present on both distribution paths, and they also can be shared between distribution paths in certain implementations.

2.2 Power Availability Metrics

2.2.1 MTBF versus MTTR:

During design, it is common to look at components that fail often and incorporate redundancies in those components. However, with the shift to software resiliency and cloud services [9], applications are designed to tolerate failures because they have redundant copies and application migration built into the system. From the perspective of a data center operator, it is important to repair the failure so that applications would not be affected for a long duration. Consider the case of a cloud application designer who needs to write an application based on expected availability. The application developer needs to know the time it might take for a data center to come back up so s/he can decide whether to migrate services to another location or not. In such cases, the focus is on reducing the repair time, especially for components that are higher in the distribution hierarchy. Table 1 shows some of the major components in a data center and their MTBF and MTTR. If the MTBF is a large number, then the component has a high reliability. If the MTTR is a large number, then the component takes a long time to repair. The numbers are derived from the IEEE Gold Book [14] based on the IEEE Standard 493, which provides failure data for industrial and commercial power systems based on surveys of industrial plants. These numbers are a publicly available source; the actual MTBF and MTTR numbers can vary for real data center deployments.

Table 1: Example MTBF hours and MTTR hours of power infrastructure components

Component	MTBF (hours)	MTTR (hours)
Dual Utility	27,712	0.56
Switchgear	5,156,470	2.40
Feeder	3,718,331	15.70
Transformer	7,895,436	5.00
Generator	13,839	13.62
UPS	499,774	6.09
PDU	55,800,000	3.10

From Table 1, we observe that the feeder MTBF is higher than the UPS MTBF by almost an order of magnitude; however, it takes longer to fix a feeder that feeds several distributions underneath it, whereas it is easier to replace a UPS component. Another important factor in power availability design is the hierarchy of the power distribution. A failure of a component at a higher level in the hierarchy can impact a larger number of IP equipment, whereas a failure in a component lower in the hierarchy has less impact and provides greater fault isolation. For this very reason, the common practice is to use distributed UPS [17] at a rack level rather than a conventional centralized UPS at the low-voltage distribution level.

Observation 1: Redundancy decisions need to factor in the cost of repair (MTTR) and not just the MTBF of the components.

2.2.2 Number of Nines:

Typically, the number of nines is calculated by taking a power infrastructure layout (e.g., those shown in Figures 2 and 3) and applying series or parallel calculation of individual component availability [3], much like circuit-resistance calculations. This methodology of constructing a reliability block diagram (RBD) is used commonly in the industry, and the number of nines is a popular term for denoting availability. While this term is useful for understanding expected availability, there is no guarantee that the actual data center performs to the design constraint. The operational behavior of the data center could be significantly different, and it could experience downtimes longer or shorter than the number of nines for which it is designed. The number of nines does not provide any information on the frequency or duration of events. Hence, it is imperative to talk about availability in terms of workloads and data center behavior. The rest of this paper attempts to provide a framework to define better such behaviors and the concept of availability.

Observation 2: The number of nines is a design-time metric that is insufficient for measuring the actual operational performance of a data center.

2.3 Power Availability Events:

To understand the impact of power events, it is necessary to understand the different types of possible events. The Information Technology Industry Council (ITIC) provides a curve that defines power-event timeliness and voltage levels for tolerances; the ITIC curve broadly captures the impact power events can have on IT equipment. There are seven types of power events: transients, interruptions, sags (under-voltage), swells (over-voltage), waveform disturbances, voltage fluctuations, and frequency variations [8][23]. Based on the event and its impact on IT equipment, the regions of operation on the ITIC curve are divided into three broad categories: an area of the curve in which there is no interruption in function, an area that is the prohibited region (damage to equipment), and an area in which functionality is affected but no damage is done to the equipment. Figure 4 presents the ITIC curve and shows the different regions of interest. The y-axis in the figure shows the percent of nominal voltage, where voltage more than 100% nominal leads to swells and voltages lower than 100% lead to sags. The x-axis shows the duration for which the operation regions are defined. For instance, for percent of nominal voltage exceeding 200% and over 0.001 seconds puts the equipment in Prohibited Region, whereas for any duration lower than that, the equipment still functions as expected without damage.

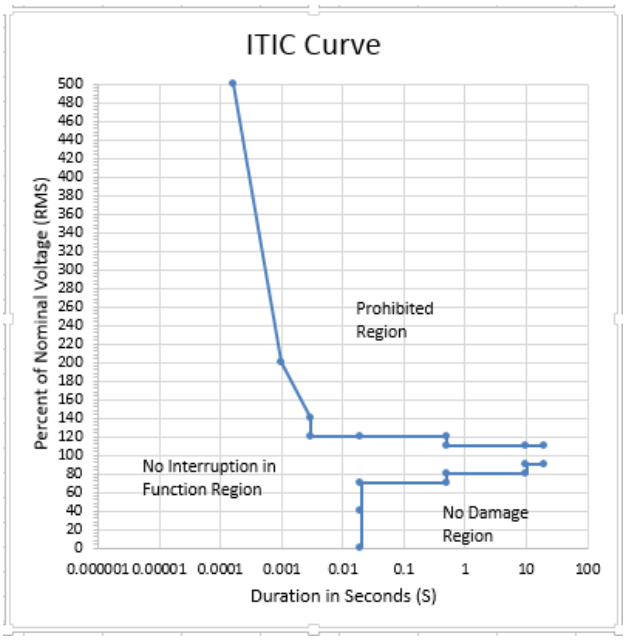

Figure 4: ITIC Curve showing operation region as a function of nominal voltage and duration of event (Prohibited Region – equipment suffers damage when operating beyond rated voltage for extended periods of time; No Damage Region – equipment might not function, but there is no damage).

Data center operators are concerned with the prohibited region, and thus design circuits with surge protectors and circuit breakers to prevent damage to expensive equipment. The "no damage region" is also of interest because this is a period during which there is no damage to the equipment, nevertheless this affects the data center by causing downtime. We design redundancy into distribution to protect against such events. Among the events in this region, sags up to 80% of the nominal voltage are rated to be tolerated for up to 10 seconds, while sags up to 70% of the nominal voltage are tolerated for only 0.5 seconds, beyond which the IT equipment either is turned off or switched over to some form of back-up power. If there is no back-up power source, lower under-voltage levels that occur for more than 20 ms result in IT equipment being powered off with no tolerance limit.

Table 2: Typical ride-through timing for energy storage components in the data center power infrastructure

Component	Energy Storage	Duration of Ride-through
Power Supply	Capacitor	20 ms for 50 Hz (1 cycle)
UPS	Batteries	More than 1 minute (2-5 minutes)
Generator	Diesel/Fuel	24-48 hours

While the ITIC curve specifies operating tolerances, IT equipment still needs to function in the presence of power events. Hence, data center power distribution includes energy storage components that can provide energy for a specific duration for which it is provisioned. Table 2 shows typical times for which different power equipment in the distribution is provisioned. Power supplies that are close to the server can ride through power events that are less than 20 ms (for 50 Hz) within 80-120% of nominal voltage limits. UPS devices are provisioned to handle events longer than that (on the order of multiple seconds). However, they are designed to be short-term battery storage, enough to transition load from main utility to generator back-up, and provide only transient energy to IT equipment while the generator starts up. Generators themselves can continue to operate for a long duration; however, they have a limited fuel supply (24-48 hours) that must be replenished. Each layer of redundant power is expensive to provision, and the objective of this paper is to provision the redundant power equipment efficiently.

Observation 3: Sags up to 80% of nominal voltage are tolerated for 10 seconds, while sags up to 70% are tolerated for 0.5 seconds. This implies that safeguards need to be put in place to handle power events that fall outside the acceptable operating ranges.

3. POWER WORKLOADS

To evaluate the trade-offs among power usage, power availability, and data center performance, we need to characterize and model power utilization, power availability and performance. In this section, we identify the set of workloads needed to determine an efficient methodology for power availability provisioning.

3.1 Power Utilization Traces

The power infrastructure carries a maximum load budget that is a design-time constraint. All the equipment, including the substation, transformers, and circuits, is sized based on this maximum supportable capacity. If actual power utilization is lower than the maximum rated capacity of the data center, then the power infrastructure could afford to ride through a few power events. For instance, in a bank of ten generators rated for 20 MW, if one of them fails, the resultant capacity is 18 MW. If the entire

facility load during the time of failure is less than 18 MW, then the generators should be able to ride through this failure event. Hence, it is important to understand the workload power utilization. While it is desirable to use the power capacity of the data center completely, it is not always the case. The likelihood of a power availability event happening at the same time as a power peak must be evaluated. To understand power-spike behavior that would be valuable for UPS sizing, we need to consider actual power traces from real data centers. A recent study from Microsoft data center power traces [26] concludes that power consumption has potential for statistical multiplexing (the occurrence of peaks at the same time is fairly minimal, and we can multiplex different component loads together such that the sum of the total is smoothed out) as we move up the hierarchy. The same study also shows that there is significant self-similarity in the power traces using Hurst parameter analysis. This implies that power utilization traces can be regenerated for studies by observing shorter duration workloads.

Observation 4: Data center power utilization has good statistical multiplexing as we move up the hierarchy. Peaks rarely occur at the same time across multiple clusters of machines. Traces show self-similarity, which enables us to construct models and regenerate workloads.

3.2 Power Availability Characterization

While there have been several previous works on power-usage characterization, there is very little work on power availability characterization. In this section, we observe data from two functioning data centers: data center 1 (DC1) during a period of one year, and data center 2 (DC2) during a period of five months. We anonymize the data center location and actual date and provide normalized data to protect proprietary information, and characterize the duration of power events and the inter-arrival times.

Duration of power events: Figure 5 shows the cumulative distribution function (CDF) of the duration of all power events recorded at DC1 and DC2 during the period of study. If all of the events are shorter than a full AC power cycle (20 ms at 50Hz or 16.7 ms/cycle at 60 Hz), then a power supply would be able to ride-through that event, as shown in Table 2.

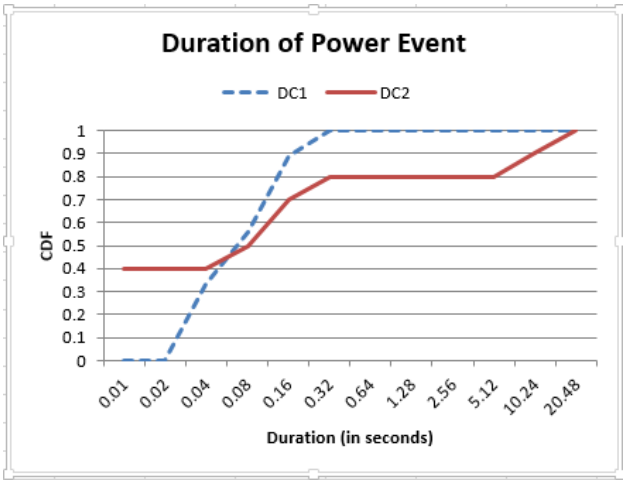

Figure 5: CDF of duration of power events from the two data centers

However, as we can see, almost all events at DC1 were longer than 20 ms and more than 60% of events at DC2 were longer than 20 ms. Hence, a UPS ride-through is a necessity here. The last bucket (20.48 seconds) includes all events that are greater than that value as well, in the interest of capturing the significant portion of the population. It is interesting to note that almost all events are less than 21 seconds duration (except for a small percentage at DC2), which is well within the tolerance limit of UPS as shown in Table 2. We need generators only to provide a continuous power source, and UPS can act as a temporary power source. This data shows that, in a large majority of cases, even before the generators start up, the main utility line is back on; hence, the generators would not have been used at all. However, it is very hard to predict duration of all possible power events during operation, so generators get started assuming that there would be long power event.

Inter-arrival of power events: Figure 6 shows the cumulative distribution function of the inter-arrival time for power events. Of interest is that the CDF is not skewed, favoring short or long durations. Instead, it is spread equally across the measured values, thereby reflecting a varying pattern of occurrence. About 40% of the power events occur within 10 days of each other, whereas the maximum inter-arrival time is within 100 days (40 days for DC2 and 100 days for DC1). Also from the data we analyzed, we found that no two power events happened within seconds or minutes of each other. To avoid such an occurrence, data center operators typically run on back-up power (generators) for a period longer than necessary, even after the main utility is back on, to make sure that the power delivery is stable.

Observation 5: Real power events in data centers are of short duration in the sample we characterized. UPS systems would be sufficient to ride-through the duration of these power events. Generators would have seen minimal usage for such events.

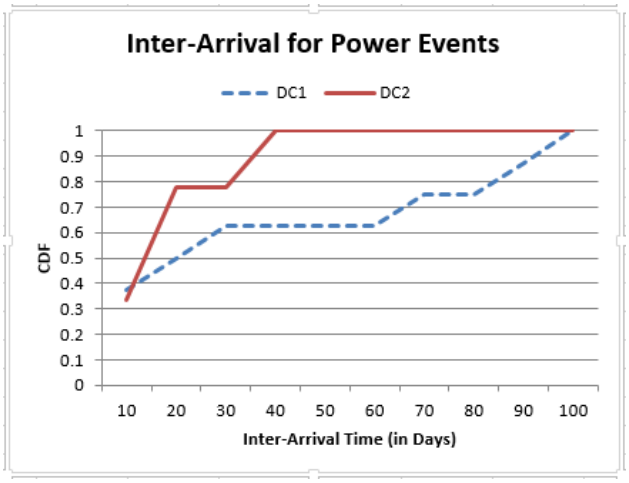

Figure 6: Inter-arrival time CDF for power events from two data centers

3.3 Putting it all together

The earlier sections provided data on power events, and observed two important qualifications: power capacity need not be utilized

to its maximum limit at all times, and power availability events are of short duration. To simulate power behavior, we need to model both the spatial and temporal characteristics of power utilization and power availability. There are multiple approaches to regenerate power workloads, including replaying the trace of the workload, creating synthetic traces based on real workload properties, and using hierarchical state-based models that can represent different levels of detail. The hierarchical model has been used previously to generate storage and network traces with high fidelity and efficiency [4][5]. To evaluate operational behavior of data centers, we need a method to regenerate workloads and simulate behavior. We use the duration and inter-arrival time distributions from this section in Section 4.3 to estimate the total performance impact of availability events.

4. POWER AVAILABILITY PROVISIONING FOR DATA CENTERS

A simple way of saving cost on redundancy is to reduce the capacity that offers redundancy in the infrastructure. In this section, we capture the impact of power events on performance SLAs and propose a new methodology to reduce the cost that is spent on provisioning generator capacity, using our observations from earlier sections. Generators are power sources that are used only when the main utility line has a failure. We propose a mode in which, during these times, the provisioned capacity of the generators is less than the overall capacity of the data center. We call this mode N-M redundancy, since compared to traditional provisioning of N, N+1 or 2N, we actually reduce the total capacity of generators. We use M to denote the total capacity that could be reduced in percentage, compared to overall capacity. Figure 7 explains this concept visually for an example data center provisioned for 10 MW. We assume that each generator is a 1-MW diesel generator and the UPS capacity is N+1 (UPS are also constructed in 1 MW blocks).

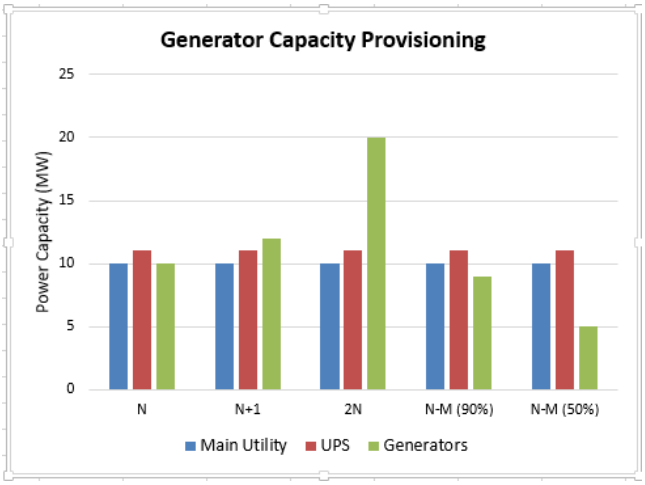

Figure 7: Generator capacity provisioning for different redundancy levels (lower capacity implies cost savings)

To provision reduced generator capacity when compared to total capacity that is provided by the main utility, we still have to be able to throttle the load to the maximum allowed capacity of the generator block when the load actually runs on the generator power source. This happens only during outage scenarios. In Section 3.2, we noted that the duration of power events that we characterized are shorter than 20 seconds and the generator system would not kick in. However, we cannot claim that all events will be of lesser duration similar to our observations in data centers. Different data centers might have different behavior. Hence, we also need to provide a methodology for those cases when the power events are longer than 20 seconds in duration.

As shown in Table 2, all events that are greater than 20 ms cannot be tolerated by the power supply in the server. One way to address the limitations of the power supply is to augment it with a supercapacitor [2]. However, such supercapacitors are expensive and are not widely available. Hence, the power events are handled by the UPS system, while the generator starts up. One valid question is whether the UPS itself could be reduced in capacity. For the UPS to be under-sized, power throttling needs to be applied within one AC cycle (i.e., the limit of power supply). If this does not happen, then the power consumed by the IT load could overrun the provisioned capacity of the UPS, cause circuit breakers to trip, and potentially bring down the entire data center. UPS can be under-provisioned only in the case when power utilization traces (from Section 3.1) can show that for all times of measurement, power usage was below a certain threshold, and can never exceed that threshold from an application-level guarantee. The typical approach to reducing the power usage of servers is to use power capping. However, power capping [11] is not sufficiently fast to allow the UPS capacity to be reduced. The typical latency to enact a power capping command through an external controller using server power capping is in the order of 110-350 ms [1], and hence it violates the 20 ms limit of the power supply by an order of magnitude. Based on UPS capabilities and data center power demands, for the application considered in this section, we apply our methodology only to reduce the sizing of generators. Generators take a certain amount of time to start up while the UPS provides a capacity buffer. The time provided by the UPS buffer is sufficient to implement power capping. Hence, we translate outage events into power-capping events for the duration of the outage. If the outage event happens during the time that the power utilization itself is lower, then the power capping will not have any impact on performance. However, if the power utilization is greater than the provisioned generator capacity (N-M case), then power capping will affect the performance SLA. The next section provides a model based on an actual application benchmark to characterize the relationship.

4.1 Performance Model

Since underprovisioning the generator capacity leads to less power delivery in the case of an outage, measures have to be taken to reduce the total power draw of the IT equipment to match the generator capacity. Such measures can have a detrimental impact on the performance of the applications running on the servers. We now evaluate the impact on performance due to power availability events.

4.1.1 Performance Workloads:

For evaluating the performance impact of power limits, we consider two large scale online services workloads: WebSearch and Cosmos. Both these workloads are cloud-scale workloads,

typically spread over several hundreds to thousands of servers in a data center [18]. These workloads are also load-balanced by a cloud scheduler, and hence performance deviations are tolerable as long as the impacts do not violate SLAs designed into these applications. These workloads are also designed to be fault tolerant within specific threshold limits. Note that we consider a class of large-scale online services in this study, and use real application workloads to validate our approach. Applications that are delay sensitive might exhibit different SLA requirements, and should be characterized through a similar methodology.

WebSearch: In large scale web search, queries are distributed to many nodes as they arrive, by a top level aggregator, in the case when the aggregator is not able to serve the request from its cache. After examining its subset of index, each node returns the most relevant pages and their dynamic page rank. The content served to the user is then accessed through the sorted indices. The performance requirements of the WebSearch application is defined by QoS (Quality of Service), throughput and latency. Given a target QoS, we measure the sustainable queries served per second (QPS). The sustainable QPS also satisfies the latency constraints of the application.

Cosmos: Cosmos is a data storage and large scale data analysis application [15]. A large set of machines are available for scheduling a pool of jobs that gets executed in parallel. This workload is representative of MapReduce and Hadoop, parallel databases and other application instances processing a large set of data. Cosmos is typically CPU intensive, and has high utilization on its servers. Execution time of a job is a direct measurement of its performance, since the faster a job gets executed, several more jobs can be scheduled from the pool.

4.1.2 Power Capping:

In order to emulate how the workload behaves under different power loads, we employed server level power-capping [11]. Power capping is a methodology by which the server is limited to run at a pre-defined power level [7][18][22]. Power capping could be implemented through hardware mechanisms on the server, or through the OS. Server vendors and data center solutions providers have started to offer power capping solutions [11]. Power capping using online self-tuning methodology [22] and feedback control algorithms [27] has been studied for individual servers. For the purposes of our experiments in this work, we chose to implement system level power capping mechanism that measures total server power and caps the server to a pre-described maximum limit in operation.

4.1.3 Results:

We run our workload at different power caps to estimate the performance at different power levels. We can then use this as a proxy to determine the overall performance during the times when the power capacity is diminished due to failure events. The % Cap is the power cap at a level that is 90% - 60% of the peak power that can be consumed by the server. We limit our experiments upto 60% cap since cap limits below that are not guaranteed in accuracy (a setting below 60% might not achieve power cap equivalent to the set value). Figure 8 shows the normalized sustained QPS measured at different power levels. If the data center runs at a reduced power level of 80%, we can use the data from Figure 8 to compute the corresponding QPS to be 83%. The figure shows that using a 90% power cap has zero impact on the performance SLAs of this workload. Hence, reduction in data center power capacity by 10% would not affect this particular workload materially, and it is a straightforward optimization that we could make (N-M, where M=10%). This methodology assumes that the workload of interest has been profiled as shown and the performance benchmarked at different power levels before identifying what levels of degradation of power capacity are tolerable by this workload.

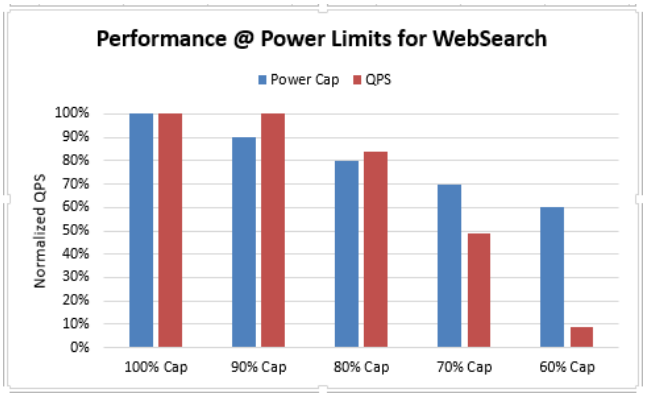

Figure 8: Performance model for WebSearch (QPS = queries per second, measured at power caps from 100% to 60%)

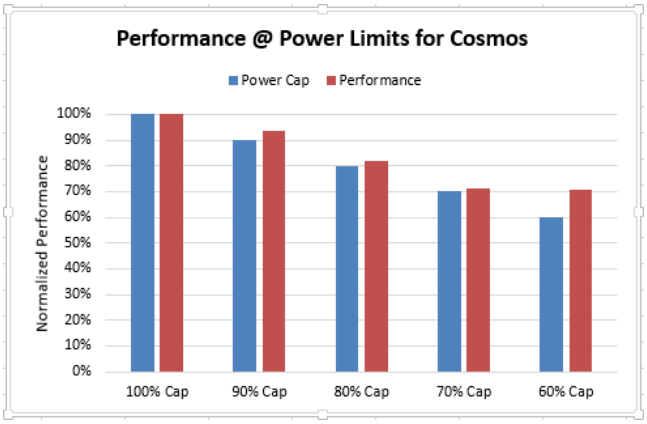

Figure 9: Performance Model for Cosmos (Performance is normalized from execution time for Cosmos jobs)

Figure 9 shows a similar graph for Cosmos workload, where the performance is a normalized measure of the total execution time. When compared to the WebSearch workload, the Cosmos workload shows a drop in performance even for the 90% cap case. This is due to the fact that Cosmos is a map-reduce like workload, with heavy CPU load. Hence even small power caps impacts CPU utilization, and hence the execution time of the job. However, the drop in performance is more proportional as the power cap is reduced. For instance, the Cosmos workload has a performance of 71% at 70% power cap, whereas the WebSearch workload has a normalized performance of only 49% at the same 70% power cap limit. This shows that WebSearch is highly sensitive to lower power cap limits compared to Cosmos workload. We use both these performance models for computing the impact to performance for these respective workloads in the following section.

4.2 N-M Generator Capacity Sizing

The previous section showed that a 10% reduction in generator capacity would not affect the performance SLA for WebSearch when the load runs on the generator during a power event. This section determines whether we can reduce the generator capacity further and evaluate the impact of such a design on application performance. To estimate the impact of generator sizing, we assume a hypothetical UPS sizing that is just sufficient to implement power capping at the high-end mark (350 ms) [1]. We assume that the UPS transitions to the generator successfully within this time, and that the power capping is implemented according to the generator sizing (90%, 80%, 70%, and 60%). These assumed UPS sizes are not common in actual data center provisioning, and we use the aggressive time limit (350 ms) specifically to illustrate our methodology. We use the duration distribution from Figure 5 to calculate event durations that would result in generators supplying power to the facility for that duration. We also use the inter-arrival time distribution obtained from Figure 6 to generate a time-series of those events to simulate power event frequency during a year. We then use our performance model from Figure 8 and Figure 9, to compute the impact of power events at various levels of generator capacity reduction (ranging from 90% to 60%) during the entire year.

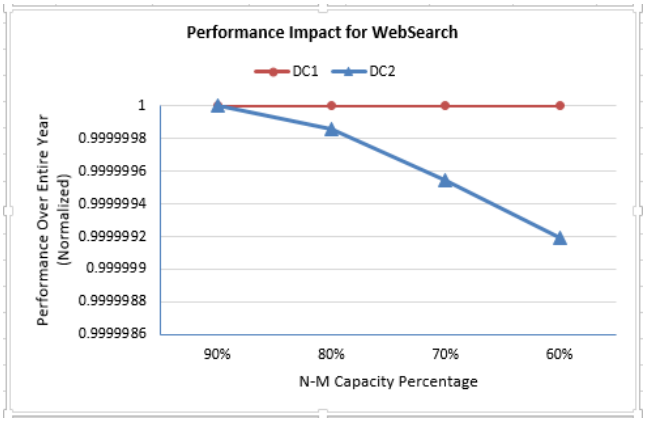

Figure 10: Performance Impact of Power Events with N-M generator capacity for WebSearch

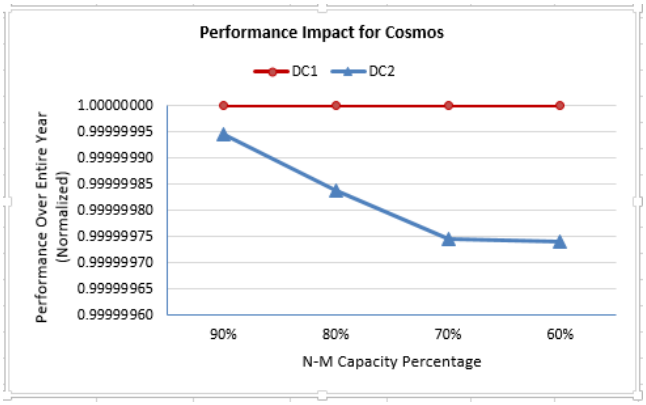

Figure 11: Performance Impact of Power Events with N-M generator capacity for Cosmos

We represent the overall impact on performance as a percentage of the data center's entire performance for a full year in Figure 10 for WebSearch and Figure 11 for Cosmos (The y-axis shows the deviation but does not start from zero). As we can see, there is minimal impact given our power-event distribution and performance model for WebSearch. For 90% capacity provisioning, there is no impact. For 60% capacity provisioning, the impact is only 0.00008% of overall performance during a year for DC2. Note that these results would be different with a different distribution of power availability events. There would be no impact at DC1 because all of its events were less than 350 ms and were absorbed by the UPS energy storage capacity.

Figure 11 presents performance for Cosmos, however even at a N-M redundancy level of 90%, this workload suffers some performance penalty. For the 60% power cap case, the penalty suffered by Cosmos workload is almost 3 times that of WebSearch workload. However given the nature of the highly parallel batch jobs that run on Cosmos machines, it is more tolerant towards performance differences. In both these figures, we observe that performance penalties are minimal, motivating our methodology that workload profiles should be considered along with application service-level agreements (SLAs) and appropriate provisioning should be adopted for specific power workload profiles in the data center.

Cost Impact: While the performance impact seems low from Figure 10 and Figure 11, the impact to total cost of operation of the data center could be significant. Because the data center runs online services, the impact of downtime to the service is significant. In addition, the cost of not using the servers for even a short duration is equivalent to wasting the capital that was invested in building the data center for the same duration. It is important to understand this consequence when looking at the cost savings from N-M generator capacity proposal. However, if an application is able to tolerate short durations of reduced performance, then it could leverage the benefit of having reduced generator capacity. Assuming that generators contribute to 20% of overall power infrastructure cost [10], the total cost savings at 60% N-M capacity provisioning for a 10-MW data center is close to $8 million. In the 90% case, the total savings would be close to $2 million. Given these considerable savings, N-M redundancy option should be a serious consideration if the performance impact due to power events is within tolerable limits for the workload.

5. RELATED WORK

In contrast to power consumption management, provisioning is an activity that is performed typically before the servers are installed in the data center. There are several power management works that deal with shutting down servers, hibernating using power states, DVFS, and using intelligent workload migration techniques [19]. In the realm of power provisioning itself, previous work [6][7][10][20][22] focused mainly on power capacity provisioning and rarely looked at power availability as a function of adding additional capacity to the data center. For instance, CPU utilization correlation to power utilization was shown by Fan et al. [6], which also presented a provisioning methodology to allocate power capacity based on workload power usage. Power management at the level of blade ensembles was

discussed by Ranganathan et al [20]. Recently, researchers have started to look at software techniques to provision data center infrastructures [25], which leverage battery storage and application throttling techniques to implement provisioning. However, there is very little work in understanding the relationships among availability, power capacity, and data center performance. To the best of our knowledge, our paper is one of the first to propose a workload-driven power availability provisioning methodology to reduce the cost incurred in power infrastructures that also are provisioned for availability.

6. FUTURE WORK & CONCLUSION

In this paper, we show that typical availability metrics are not sufficient to predict data center behavior for power availability events. We then provide a technique to measure performance impact due to power availability events and propose a novel redundancy mechanism (N-M redundancy) for generator capacity sizing. Future work involves characterizing power events for a multi-year or longer duration to detect workload patterns that could be exploited for additional savings and developing an extended simulator [18]. We also plan to expand the scope of power availability provisioning to other areas, including mechanical infrastructures, server equipment, and networks. We believe that a workload-driven provisioning methodology across all dimensions (power, mechanical, network, and servers) is key to reducing capital cost and operational expenditure on data center infrastructure.

7. REFERENCES

[1] Bhattacharya, A. A., Culler, D., Kansal, A., Govindan, S., & Sankar, S. (2013). The need for speed and stability in data center power capping. Sustainable Computing: Informatics and Systems.

[2] Casadei, D., Grandi, G., & Rossi, C. (2002). A supercapacitor-based power conditioning system for power quality improvement and uninterruptible power supply. In Industrial Electronics, 2002. ISIE 2002.

[3] Čepin, M. - Assessment of Power System Reliability, 2011 – Springer

[4] Delimitrou, C., Sankar, S., Kansal, A., and Kozyrakis, C.. 2012. ECHO: Recreating network traffic maps for datacenters with tens of thousands of servers. In Proceedings of the 2012 IEEE International Symposium on Workload Characterization (IISWC) (IISWC '12). IEEE Computer Society, Washington, DC, USA, 14-24.

[5] Delimitrou, C., Sankar, S., Vaid, K., and Kozyrakis, C.. Decoupling datacenter studies from access to large-scale applications: A modeling approach for storage workloads. In the Proceedings of the 2011 IEEE International Symposium on Workload Characterization (IISWC), November 2011, Austin, TX.

[6] Fan, X., Weber, W., and L., A. Barroso. Power provisioning for a warehouse-sized computer. In ISCA 2007 (San Diego, CA, USA 2007).

[7] Felter, W., Rajamani, K., Keller, T. and Rusu, C. A performance-conserving approach for reducing peak powerconsumption in server systems. In ICS '05: Proceedings of the 19th annual international conference on Supercomputing, pages 293–302, 2005.

[8] Fortenbery, B. Power quality in internet data centers, PQ TechWatch, Nov 2007

[9] Gauthier, D. Software reigns in Microsoft's cloud-scale data centers. Feb 2011, Global Foundation Services, Microsoft

[10] Govindan, S., Wang, D., Chen, L., Sivasubramaniam, A., & Urgaonkar, B. (2011, October). Towards realizing a low cost and highly available datacenter power infrastructure. In Proceedings of the 4th Workshop on Power-Aware Computing and Systems (p. 7). ACM.

[11] HP power capping and HP dynamic power capping for ProLiant servers. Technology brief, 2nd edition. http://h20000.www2.hp.com/bc/docs/support/SupportManual/c01549455/c01549455.pdf.

[12] Hamilton, J. Cost of power in large-scale data centers. http://perspectives.mvdirona.com (November 2008).

[13] Hoelzle, U. , Barroso, L. A., The datacenter as a computer: An introduction to the design of warehouse-scale machines, Morgan and Claypool Publishers, 2009

[14] IEEE Std 493-1997: IEEE Recommended Practice for the Design of Reliable Industrial and Commercial Power Systems

[15] Isard, M., Budiu, M., Yu, Y., Birrell, A., & Fetterly, D. (2007). Dryad: distributed data-parallel programs from sequential building blocks. ACM SIGOPS Operating Systems Review, 41(3), 59-72.

[16] ITI (CBEMA) Curve Application Note, Technical Committee 3, Information Technology Industry Council

[17] Kontorinis, V., Zhang, L. E., Aksanli, B., Sampson, J., Homayoun, H., Pettis, E., Tullsen, D. T., and Rosing, T. S. 2012. Managing distributed ups energy for effective power capping in data centers. In Proceedings of the 39th Annual International Symposium on Computer Architecture (ISCA '12). IEEE Computer Society, Washington, DC, USA, 488-499.

[18] Kozyrakis, C., Kansal, A., Sankar, S., & Vaid, K. (2010). Server engineering insights for large-scale online services. *IEEE micro*, *30*(4), 8-19.

[19] Lu, Y. H., & De Micheli, G. (2001). Comparing system level power management policies. Design & Test of Computers, IEEE, 18(2), 10-19.

[20] Ranganathan, P., Leech, P., Irwin, D., and Chase, J. 2006. Ensemble-level power management for dense blade servers. SIGARCH Comput. Archit. News 34, 2 (May. 2006), 66-77.

[21] Sankar, S., Kansal, A., & Liu, J. Towards a holistic data center simulator. MODSIM 2013

[22] Saravana, M. and Govidan, S. Using on-line power modeling for server power capping. In Workshop on Energy-Efficient Design 2009 (2009), University of Texas and IBM.

[23] Seymour, J. (2005), The seven types of power problems, American Power Conversion, Schneider Electric White Paper 18

[24] Turner IV, W. P., PE, J., Seader, P. E., & Brill, K. J. (2006). Tier classification define site infrastructure performance. Uptime Institute, 17.

[25] Wang, D., Govindan, S., Sivasubramaniam, A., Kansal, A., Liu, J., & Khessib, B. (2013). Underprovisioning Backup Power Infrastructure for Data centers. Technical Report CSE-13-012, The Pennsylvania State University.

[26] Wang, D., Ren, C., Govindan, S., Sivasubramaniam, A., Urgaonkar, B., Kansal, A., and Vaid, K.. 2013. ACE: abstracting, characterizing and exploiting peaks and valleys in datacenter power consumption. SIGMETRICS Perform. Eval. Rev. 41, 1 (June 2013), 333-334

[27] Wang, Z., Zhu, X., McCarthy, C., Ranganathan, P., & Talwar, V. (2008, June). Feedback control algorithms for power management of servers. In Third International Workshop on Feedback Control Implementation and Design in Computing Systems and Networks (FeBid), Annapolis, MD.