The Effect of Server Energy Proportionality on Data Center Power Oversubscription

Sulav Malla, Ken Christensen

Department of Computer Science and Engineering, University of South Florida, Tampa, United States

Abstract

Modern data centers improve resource utilization with power oversubscription. The power hierarchy in a data center is oversubscribed by installing more servers than allowed by the power budget based on server peak power consumption. Power oversubscription is possible due to the statistically low likelihood of simultaneous peak power operation of multiple servers. As future servers become more energy proportional, the opportunity for greater power oversubscription increases. The challenge is to quantify the level of oversubscription that can be attained. In this paper, we quantify the level of oversubscription possible for a given acceptable probability of power overload for servers characterized by energy proportionality metric and workload distribution. We develop a theoretical framework to characterize and predict the relationship between server energy proportionality and power oversubscription. We verify our framework through an extensive empirical study using publicly available SPECpower benchmark data for over 500 server models and publicly available Google cluster utilization data. Using our framework, a data center operator can predict the additional power oversubscription possible when replacing existing servers with a newer model of more energy proportional servers.

Keywords: Data center, Energy proportionality, Power oversubscription

1. Introduction

The rapid growth of cloud services and the trend towards server-side computing has resulted in a demand for more data centers. Building a new data center or expanding an existing one can be expensive. Construction cost increases almost linearly with the power provisioned and can range from $10 - $20 per Watt [1]. To make matters worse, the provisioned power is not fully utilized. Although a server may consume peak power from time to time, a group of servers is less likely to reach peak aggregate power due to statistical multiplexing of individual server power. Fan et al. [2] reported that, over the course of six months, a group of 5,000 servers under study at Google never exceeded 72% of their aggregate peak power.

The fact that data centers are expensive to build and their power infrastructures are under-utilized provides motivation for *power oversubscription*. Power oversubscription of a data center refers to deploying more servers than allowed by the *power limit*. Power limit is a fixed quantity which can be physical, defined by circuit breaker limits, or contractual, defined by a contract between the utility and the data center operator. The benefit of power oversubscription is that we save on building cost with better utilization of power resources that would have otherwise gone to waste. Data center power oversubscription is a common practice today. For multi-tenant data centers, where the operator leases data center space (including power, cooling, and security) to different tenants, the standard is to oversubscribe power by 20% [3] which results in 20% more revenue at no additional cost. Major IT companies (for example, Facebook [4]) who own large data centers have also reported to have oversubscribed their data center power to save data center infrastructure cost. A risk associated with data center power oversubscription is that aggregate power could exceed the power limit due to simultaneous peaking of servers resulting into a *power overload*. Various control mechanisms have been proposed to control or avoid power overload by capping power in such aggressively provisioned data centers [5, 6, 4, 7, 8]. Such control mechanisms allow us to harness the benefits of data center power oversubscription in a controlled and manageable way.

The data center community has advocated for energy proportional servers for over a decade [9]. Ideal energy proportional servers consume almost no power when idle and their power consumption increases linearly with its utilization. An attractive property of such servers is that it has an uniform energy efficiency over the entire server utilization range, so no matter what the workload looks like, the server always operates at its peak energy efficiency region. Ideal energy proportionality had been a design goal for various server components (CPU, memory, disk, etc.). It has been found that servers are getting more energy proportional over the years with some newer model servers being very close to ideal energy proportional [10, 11]. A server in 2007 that consumed more than 60% of peak power when idle, has improved to consume only a little over 10% of peak power when idle, in 2018 [1].

When servers get more energy proportional, their power consumption pattern changes, which in turn affects the aggregate data center power consumption. In the future, as more energy proportional servers replace existing less energy proportional servers, we would like to know how this will impact opportunities to oversubscribe data center power infrastructure. To the best of our knowledge, the effect of increasing server energy proportionality on opportunities for data center power

oversubscription has not been studied or quantified before. This is the focus of this paper. One key question we explore is: *How does increasing server energy proportionality affect opportunities to oversubscribe data center power infrastructure?* Two major contributions of this paper are:

- We show how increasing server energy proportionality opens up the opportunity for more power oversubscription by modelling the relationship between server energy proportionality and possible power oversubscription for a fixed probability of power overloading.
- We validate our theoretical framework using real world data center server utilization data from a Google cluster and power/performance characteristics data for various server models from SPECpower benchmark.

We hope that data center operators will be able to estimate the data center power oversubscription possible for their particular scenario using our proposed framework.

The rest of this paper is organized as follows. In Section 2, we provide the preliminary concepts. Section 3 is where we present the mathematical framework followed by evaluation in Section 4. We discuss the related work in Section 5 and finally conclude with possible future work in Section 6.

2. Background

In this section, we describe the data center power hierarchy, power oversubscription, and energy proportionality of servers.

2.1. Data center power hierarchy

The power infrastructure of a data center has a hierarchical structure as shown in Fig. 1. Power from the utility is distributed to the data center site and the voltage is stepped down (typically, 480 V in the U.S.) for on-site distribution. The utility is the primary source of power to the data center while diesel generators provide backup power during a utility power outage. An Automatic Transfer Switch (ATS) selects utility power by default and automatically switches to the backup diesel generator power in case of a utility power failure. Power then goes to central Uninterruptible Power Supply (UPS) which removes power spikes/sags from the input and also performs power factor corrections on the output side. Additionally, UPS have some form of energy storage devices (such as, batteries) to provide immediate transition power in case of utility power failure, since backup diesel generators take several seconds to carry the full load. The UPS provides a reliable and regulated power ready for distribution inside the data center floor.

Power from the UPS goes to multiple Power Distribution Units (PDU) spread across the data center floor. PDUs perform the final voltage step down appropriate for individual IT equipment (typically, 110 V in the U.S.). PDUs distribute incoming power through several smaller power lines that go to individual server racks/cabinets. Power lines coming into a server rack from a PDU have circuit breakers to prevent power overdraw or short circuits from travelling up the power hierarchy, thus,

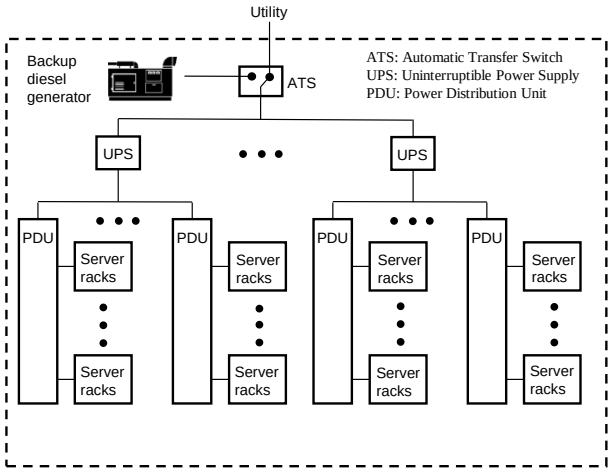


Figure 1: Power delivery infrastructure of a typical data center [12].

avoiding cascading failure of power equipment (UPS, PDU). Data center power infrastructure can additionally have redundant UPS and PDUs (N+1 or N+2 redundancy) or entirely independent power feeds (2N redundancy) for higher availability but at a higher cost.

2.2. Power oversubscription

Power at a particular level in the data center power hierarchy is said to have been oversubscribed if the power limit at that particular level can be exceeded by the aggregate power consumption from the level below. For example, deploying more servers in a rack than allowed by the circuit breaker limit (rack level power oversubscription), distributing power lines to more server racks than allowed by the PDU maximum power rating (PDU level power oversubscription), or having more PDUs than allowed by the UPS power rating (UPS level power oversubscription). Note that power can be oversubscribed at a single particular level or at all levels of the power hierarchy. While power oversubscription increases the utilization of the power hierarchy, there can be power overload instances when the aggregate power consumption exceeds the power limit. A sustained power overload can trip circuit breakers or permanently damage power equipment leading to service disruptions. Power overload at the data center level may result in fines from the utility as they impose a contractual power limit [12].

Power overload events can be controlled or entirely prevented through *power capping*, the processes of limiting the power consumption of a server or a group of servers. For example, Intel's Running Average Power Limit (RAPL) [13] interface allows a server's average power consumption (inside a time window) to be capped. Individual server power capping can be utilized to achieve power capping at higher levels in the power hierarchy. However, frequent server power capping has an adverse effect on system performance as the servers are throttled to run at a lower frequency. The trade-off is to oversubscribe power by an amount such that *probability of overloading* is within an acceptable threshold. Such power oversubscription where the probability of overloading, the perfor-

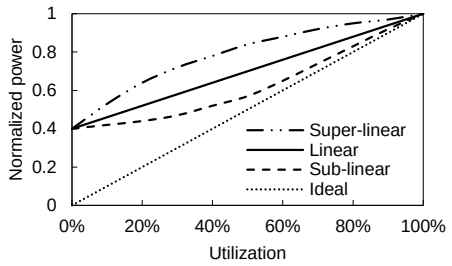


Figure 2: Normalized power-utilization curves of different types of energy proportional servers.

mance penalty, and the associated risk is within a small acceptable range (defined by the data center operator) is considered as the "safe" power oversubscription. We want a probability of overloading to be as small as possible, for example, a threshold of 50% (power overload half of the time) may be unacceptable, but 0.1% (about 86 seconds a day) may be acceptable, depending upon the data center.

2.3. Server energy proportionality and metrics

Energy proportional servers consume power proportional to their utilization. Power consumption of a server at different utilization levels can be represented as a *power-utilization curve*. An ideal energy proportional server will have a power-utilization curve as a straight line joining zero to peak power. Fig. 2 shows a normalized power-utilization curve for different types of energy proportional servers. The solid line represents a linearly energy proportional server, dashed line represents a sub-linearly energy proportional server, and dot-dashed line represents a super-linearly energy proportional server. We also have the ideal energy proportional server shown as dotted line for reference.

Numerous metrics have been proposed to measure and compare energy proportionality of servers [14] with Dynamic Range (DR) being one of the simplest metric. The DR of a server is given as the difference between peak and idle power of a server, normalized with respect to its peak power [15]

$$\text{DR} = \frac{\text{Peak power} - \text{Idle power}}{\text{Peak power}}.$$

DR ranges from 0 to 1 with higher value meaning greater energy proportionality. However, DR cannot distinguish between linear, sub-linear, and super-linear servers if they have the same idle and peak power. One way of measuring linearity of servers is through the Linear Deviation (LD) metric [16] given as

$$\text{LD} = \frac{\text{Actual power curve area}}{\text{Linear power curve area}} - 1$$

where "Linear power curve area" is the area under the line joining idle and peak power of an actual power-utilization curve. For example, the LD of the super-linear server in Fig. 2 would be the ratio of area under the super-linear curve and area under the linear curve, minus 1, which in this case would evaluate to a positive value. Super-linear servers will have LD greater than 0 while sub-linear servers will have LD less than 0. A server that is linearly energy proportional will have LD equal to 0.

Table 1: List of symbols used.

Symbol	Meaning
μ	average utilization of a server
σ	standard deviation of utilization of a server
DR	dynamic range of a server
EP	energy proportionality of a server
$F(P)$	CDF of aggregate power
n	number of servers in a data center
S	safe power oversubscription level
$p(u)$	power consumption of a server at utilization, u
P_{limit}	data center power limit/budget
P_{max}	data center maximum possible power
$Pr(v)$	probability of overloading

A metric that captures both energy proportionality as well as linearity into a single value is the EP metric [10] given as

$$\text{EP} = 1 - \frac{\text{Actual power curve area} - \text{Ideal power curve area}}{\text{Ideal power curve area}}$$

where "Ideal power curve area" refers to the area under the power-utilization curve for an ideal energy proportional server. EP ranges from 0 to 2 with higher value meaning greater energy proportionality and an ideal energy proportional sever will have EP equal to 1. Unlike DR which only considers the power values at peak and idle utilization, EP takes the entire utilization range into account by calculating the area under the normalized power-utilization curve. For example, in Fig. 2, the three servers with different linearity property will have same DR, but EP of sub-linear server will be greater than linear server which will be greater than super-linear server. For servers with linear power-utilization curve, DR and EP evaluate to the same value.

3. Formulation

In this section we describe how to estimate the safe oversubscription level for different energy proportional servers at a given probability of overloading. We describe the theory and in the next section, validate with real world data.

3.1. Individual and aggregate server power

Let us consider a data center with n identical servers, and denote the aggregate power of the data center, P, as the sum of individual server power

$$P = \sum_{i=1}^{n} p(u_i) \tag{1}$$

where $p(u)$ is the power consumption of a server at utilization $0 \leq u \leq 1$. Let us denote the maximum possible aggregate power, the sum of individual server peak power, as P_{max} and

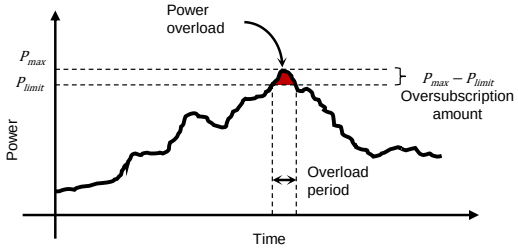


Figure 3: Variation of aggregate power over time in an oversubscribed data center.

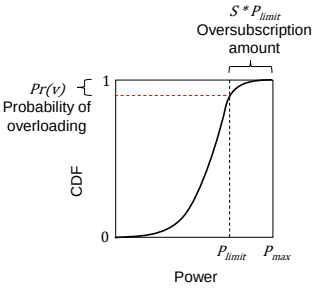


Figure 4: CDF of aggregate power consumption.

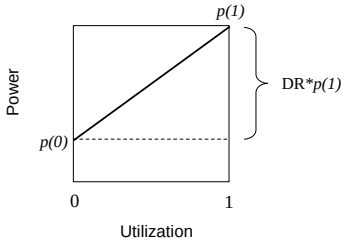


Figure 5: Linear power-utilization curve for a server.

the power limit/budget of the data center as P_{limit}. Assuming servers consume their peak power at peak utilization, $u = 1$, we have $P_{max} = n \cdot p(1)$. Note that power limit (or budget) is a fixed quantity for a data center determined by the capacity of the power equipment (circuit breakers, PDU, and UPS). This may or may not be equal to the maximum possible aggregated power use, P_{max}.

3.2. Probability of overloading and safe oversubscription

We have the case of power oversubscription whenever the maximum possible aggregate power is greater than the power limit, $P_{max} > P_{limit}$, and the amount of power oversubscription is given as

$$S = \frac{P_{max} - P_{limit}}{P_{limit}} \tag{2}$$

where S is the safe power oversubscription level (generally expressed as a percentage). An example of aggregate power variation in an oversubscribed data center can be seen in Fig. 3. The variation is due to servers having different power consumption at different utilization levels. Although rare, server power consumption may peak simultaneously causing power overloads, a situation where the aggregate power consumption is greater than the power limit as shown by the red shaded area in Fig. 3. Safety mechanisms must be in place to prevent extended power overload situations which might trip circuit breakers and cause service outage.

Whenever there is oversubscription, there is a certain probability of overloading the power infrastructure associated with it (the fraction of time that power overload occurred over the total observation time), denoted as $Pr(v)$. If we know the cumulative density function (CDF) of data center aggregate power, $F(P)$, we can get the probability of overloading as

$$\begin{aligned} Pr(v) &= Pr(P > P_{limit}) \\ &= 1 - F(P_{limit}). \end{aligned} \tag{3}$$

$Pr(v)$ will depend on the power limit of the data center. Combining Eq. (2) and Eq. (3) we get the relationship between probability of overloading and safe power oversubscription level as

$$Pr(v) = 1 - F\left(\frac{P_{max}}{1+S}\right). \tag{4}$$

This shows that, for a given set of servers, the relationship between $Pr(v)$ and S is determined by the CDF of aggregate power. Fig. 4 shows an example CDF of data center aggregate power. As the amount of power oversubscription is increased (power limit P_{limit} is decreased), probability of overloading ($Pr(v)$) increases. $S = 0$ implies $Pr(v) = 0$ and as $S \to \infty, Pr(v) \to 1$. We can fix the probability of overloading to a certain acceptable threshold, for example $Pr(v) = 0.001$ (0.1% probability of overloading), and find the corresponding safe power oversubscription level (S).

3.3. Effect of DR on server power

If we assume the power-utilization curve of a server, $p(u)$, to be linear as shown in Fig. 5, the power consumption of a server at utilization u can be expressed in terms of its DR

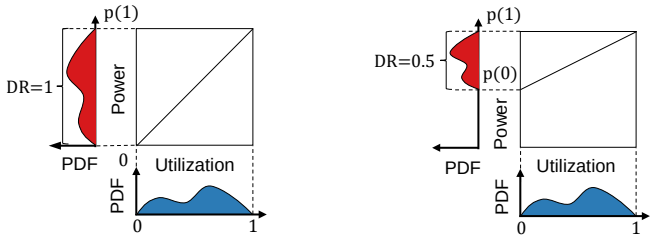
$$\text{DR} = \frac{p(1) - p(0)}{p(1)} \tag{5}$$

and peak power as

$$\begin{aligned} p(u) &= p(0) + u[p(1) - p(0)] \\ &= p(1)[1 - \text{DR}] + u \cdot p(1) \cdot \text{DR} \\ &= p(1)[1 - \text{DR} + u \cdot \text{DR}]. \end{aligned} \tag{6}$$

Since, DR and EP for a linear power-utilization curve are equal, we can use them interchangeably.

A server in a data center will have a time varying utilization due to a time varying workload. Variation in server utilization can be represented by a probability density function (PDF) as shown with blue shaded area in Fig. 6. This variation will also be seen in the server power consumption. From Eq. (6) we can observe that when DR $= 1$, the server power consumption will be in the range $0 \le p(u) \le p(1)$ as $0 \le u \le 1$. Power consumption of a server at a particular utilization increases as the value of DR decreases, since the server idle power, $p(0)$, increases, causing the power consumption to be in



(a) Fully energy proportional server

(b) Partially energy proportional server

Figure 6: DR of server changes the distribution of power consumption of the server.

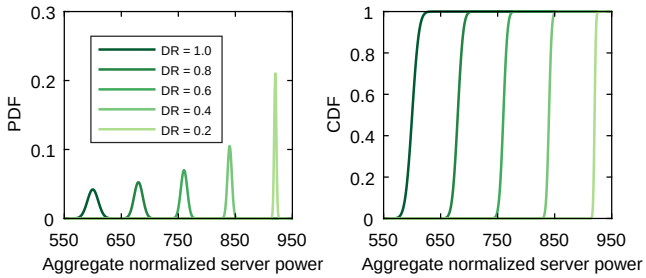


Figure 7: PDF and CDF of aggregate normalized power consumption of 1000 servers with $\mu = 0.6$ and $\sigma = 0.3$ for varying DR.

the range $p(0) \leq p(u) \leq p(1)$ and ultimately having a constant peak power consumption, p(1), irrespective of utilization when DR = 0. Hence, the DR value of a server affects the server power distribution even when the utilization distribution of the server does not change as shown in the red shaded area in Fig. 6. For an ideal energy proportional server, the distribution of power consumption is the same as the distribution of utilization as shown in Fig. 6a. However, when a server is not fully energy proportional, the distribution of power consumption becomes a scaled and shifted version of the utilization distribution, scaled by DR and shifted by idle power, as shown in Fig. 6b.

3.4. Effect of DR on data center power

The DR of the server affects its power consumption and therefore, it will also have an effect on the aggregate power. We can come up with the relationship by combining Eq. (1) and Eq. (6)

$$\begin{aligned} P &= \sum_{i=1}^{n} p(u_i) \\ &= \sum_{i=1}^{n} p(1)[1 - \mathrm{DR} + u_i \cdot \mathrm{DR}] \\ &= n \cdot p(1)[1 - \mathrm{DR}] + p(1) \cdot \mathrm{DR} \sum_{i=1}^{n} u_i \\ &= \underbrace{P_{max}[1 - \mathrm{DR}]}_{\text{idle aggregate power}} + \underbrace{p(1) \cdot \mathrm{DR} \sum_{i=1}^{n} u_i}_{\text{varying aggregate power}} . \end{aligned} \tag{7}$$

From Eq. (7) we observe that if all the servers in the data center have DR = 1, the aggregate power consumption takes values in the range $0 \leq P \leq P_{max}$. Now, if we replace the servers with ones having lower DR, the range of aggregate power gets narrower to $P_{max}[1 - \mathrm{DR}] \leq P \leq P_{max}$. Furthermore, the aggregate power variation is scaled by DR, that is, as the server DR decreases, the aggregate power variation also decreases.

We take a hypothetical working example to illustrate our point. Let us assume that utilization of each server in a data center is independent and identically distributed (i.i.d.) with mean, μ and standard deviation σ (we will use a more realistic scenario in our evaluation section). That is, the server utilization varies in the range $[0, 1]$ with the mean, μ, standard deviation, σ, and this is true for all servers (the i.i.d. assumption). For a server with linear power-utilization curve, the normalized power use of the server will vary between its idle power, $p(0) = 1 - \mathrm{DR}$, and peak power, $p(1) = 1$. Furthermore, the normalized server power distribution is shifted by the idle power, $1 - \mathrm{DR}$, and scaled by DR, as explained above in section 3.3, with the mean $1 - \mathrm{DR} + \mathrm{DR} \cdot \mu$ and standard deviation $\mathrm{DR} \cdot \sigma$. Since the normalized power of a server is also bounded in the range [0, 1], the maximum variance possible is 0.25 (standard deviation of 0.5), as variance of a bounded random variable in the range $[a, b]$ is given by the inequality $\sigma^2 \leq \frac{(b-a)^2}{4}$ [17]. The finite variance of server power consumption allows us to use the central limit theorem. If we aggregate n such normalized server power consumption, according to the central limit theorem, the sum of normalized server power consumption will tend to a Gaussian distribution with mean of $n(1 - \mathrm{DR} + \mathrm{DR} \cdot \mu)$ and standard deviation of $\sqrt{n} \cdot \mathrm{DR} \cdot \sigma$. Fig. 7 shows the Gaussian PDF and CDF of aggregate normalized server power when we have $n = 1000$ servers and the utilization distribution has $\mu = 0.6$ and $\sigma = 0.3$, for various DR values. We can see that as the DR of the server decreases, the PDF and CDF become narrower and shift to the right, suggesting that, as the server energy proportionality decreases, the average aggregate power consumption increases while its variance decreases.

The CDF of aggregate power will scale with the DR of servers. If we denote the CDF of aggregate power with fully energy proportional (DR = 1) servers as $F_1(P)$, we can get the CDF of aggregate power when servers of a given DR is used

$$F_{\mathrm{DR}}(P) = \begin{cases} 0 & \text{if } P \leq P_{max}[1 - \mathrm{DR}] \\ F_1\left(\frac{P - P_{max}[1-\mathrm{DR}]}{\mathrm{DR}}\right) & \text{if } P > P_{max}[1 - \mathrm{DR}]. \end{cases}$$

where $F_{\mathrm{DR}}(P)$ is the CDF of aggregate power when servers of energy proportionality DR are used. This CDF is a scaled and shifted version of $F_1(P)$ as shown in Fig. 7, scaled by DR and shifted by $P_{max}[1 - \mathrm{DR}]$. The shape of CDF is maintained as we assumed a linear power-utilization curve for the servers.

3.5. Effect of DR on safe power oversubscription

Rewriting Eq. (4) to account for a change in the server DR we have

$$Pr_{\mathrm{DR}}(v) = 1 - F_{\mathrm{DR}}\left(\frac{P_{max}}{1 + S}\right). \tag{8}$$

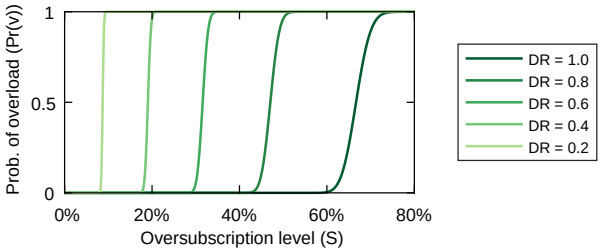


Figure 8: Probability of overloading as oversubscription amount increases for servers with different DR.

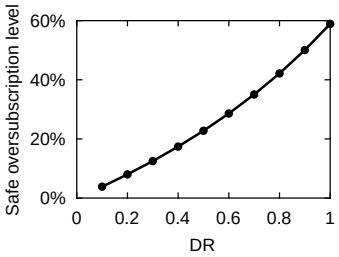


Figure 9: Relationship of safe oversubscription amount and server DR when probability of overloading is fixed at 0.1%.

Continuing with our working example, we have P_{max} fixed at 1000 while we can vary the power oversubscription amount by varying P_{limit}. The corresponding probability of overloading as given by Eq. (8) is shown in Fig. 8 as server DR changes. We can make two main observations from Fig. 8:

- As the DR of a server increases, we are able to oversubscribe more for the same probability of overloading
- The increase in probability of overload is abrupt (more sensitive) for lower DR while it is gradual for higher DR.

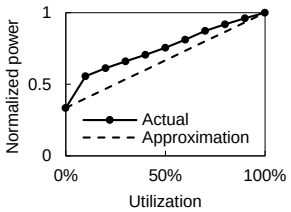
Rearranging Eq. (8), we can get the expression for safe power oversubscription level for a fixed probability of overloading as

$$S = \frac{P_{max}}{F^{-1}_{\mathrm{DR}}\left(1 - Pr_{\mathrm{DR}}(v)\right)} - 1. \tag{9}$$

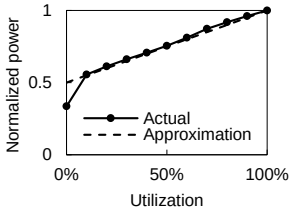
Setting the acceptable threshold for probability of overloading to 0.1% ($Pr(v) = 0.001$), we can find the corresponding safe power oversubscription level (S) for various server DR. Fig. 9 shows how safe oversubscription level varies with DR of the server when $Pr(v)$ is fixed at 0.001. As servers get more energy proportional, we are able to oversubscribe more and the relation is better than linear. Specifically, doubling the DR of the servers from 0.4 to 0.8 can increase power oversubscription from 17% to 42% (more than double) for the same probability of overloading.

3.6. Servers with non linear power-utilization curve

In all of the derivations, we have assumed that the power-utilization curve of the server is linear. However, a real server may not have a linear power-utilization curve. A major implication is that the shape of server power distribution would be different than the shape of utilization distribution due to non-linearity (unlike what we have in Fig. 6). For the non-linear power-utilization curve, derivations in Eq. (7) and Eq. (9) will not hold exactly but will be a linear approximation with the DR metric. We can have a better linear approximation to a real server with the EP metric,



(a) Linear approximation by DR metric



(b) Linear approximation by EP metric

Figure 10: Linear approximation for a power-utilization curve of a real server by (a) DR metric and (b) same server by EP metric. The solid line represents the actual power power-utilization curve and the dashed line is the linear approximation.

$$\begin{aligned}\mathrm{EP} &= 1 - \frac{\text{Actual power curve area} - \text{Ideal power curve area}}{\text{Ideal power curve area}} \\ &= 1 - \frac{\int_0^1 p(u)\,\mathrm{d}u - \frac{p(1)}{2}}{\frac{p(1)}{2}} \\ &= 2 - \frac{2\int_0^1 p(u)\,\mathrm{d}u}{p(1)}.\end{aligned} \tag{10}$$

The EP metric in Eq. (10) and DR metric in Eq. (5) are equal for a linear power-utilization curve and can be used interchangeably in Eq. (9) as we have assumed a linear power-utilization curve when deriving it

$$S = \frac{P_{max}}{F^{-1}_{\mathrm{DR}}\left(1 - Pr_{\mathrm{DR}}(v)\right)} - 1 = \frac{P_{max}}{F^{-1}_{\mathrm{EP}}\left(1 - Pr_{\mathrm{EP}}(v)\right)} - 1. \tag{11}$$

However, for a non-linear power-utilization curve, EP and DR approximate different linear servers. Fig. 10 shows a non-linear power-utilization curve of an actual server as a solid line. If we calculate the DR for this server and use it in our analysis, it would be like approximating the server with the dotted line shown in Fig. 10a. This line simply connects the end points of the actual curve (for both to have the same DR). Similarly, if we calculate the EP for the same server instead of DR, it would be like approximating the server with the dotted line shown in Fig. 10b. The area under the line would be the same as the area under the actual curve (for both to have the same EP). The linear approximation by EP is better than the approximation by DR as seen in Fig. 10. This is because the area under the power-utilization curve of a server more accurately depicts the power it will consume [18].

Notice that the probability of overloading given by Eq. (8), from which we derive Eq. (9) and further extend it in Eq. (11), are long run probabilities. However, actual power overload events are instantaneous occurrences. What happens during a power overload (when we are past the power limit) de-

pends on the physical and electrical characteristics of the circuit breaker. Tripping of a circuit breaker depends primarily on two factors 1) the magnitude of the power overload and 2) the duration of the power overload. For example, Wu et al. [4] experimented with tripping characteristics of circuit breakers at different power overload level and found that even when the power drawn was twice the rated power, it still took about 30 seconds to trip the circuit breaker. Furthermore, circuit breakers could sustain a 10% power overload for more than 15 minutes [4, 19].

3.7. Taking server performance into account

We have not taken server performance into account up to this point. An implicit assumption that we have made is that the servers have the same level of performance while only differing in energy proportionality. In practice, performance of different servers are not the same. Server performance can be characterized by its throughput (operations per second) at various utilization levels. Server throughput generally increases linearly with utilization [18, 20]. As a result, the throughput-utilization curve is a line from zero to peak throughput, and thus, server performance can be compared using their peak throughput (throughput when a server is 100% utilized). For example, a server that is twice as fast will have twice the peak throughput.

We define workload as the offered load (operations per second) to the server. Serving the workload causes a server to run at a particular utilization depending upon the peak throughput of the server. Same amount of workload may cause two different servers to operate at two different utilization level. For example, for the same amount of workload, a server might be at 70% utilization but a server that is twice as fast might only be 35% utilized. Hence, if two servers had the same power-utilization curve but different peak throughput, the faster server would be at a lower utilization and thus have a lower power consumption level leading to more opportunity for power oversubscription. For a server with a linear power-utilization curve, the power consumption for a given workload would be inversely proportional to the peak throughput of the server. Therefore, the server with higher peak throughput (and same power-utilization curve) can be oversubscribed more, the relation being linear.

We note that performance impacting proactive control mechanisms like throttling of servers or workload scheduling (admission control) may be in place to prevent power overload. Since power overload event are undesirable, they should be rare even when such control mechanism is in place. If a power overload event occurs and control is triggered, the performance will be impacted which may even lead to service level agreement (SLA) violations. Our results can inform on how often a control mechanism will trigger (and thus also predict performance impacts). It is up to the data center operator to decide what level of performance degradation is acceptable and oversubscribe the data center power hierarchy accordingly.

4. Experimental validation

In this section we validate our theoretical results using a real world data center trace. For this we need server utilization data from a real data center as well as power-utilization data for real servers with varying energy proportionality.

4.1. Server utilization data from Google cluster

Power or resource usage of real world data centers are generally not publicly available due to privacy concerns. However, one such usage trace from a Google data center [21] has been publicly released after obfuscating the data to prevent leaking of sensitive information. This dataset contains 6 tables with various information about a cluster of about 12.5 thousand servers for a period of 29 days, from May 1, 2011 to May 30, 2011. Many similar recent studies have used this dataset for evaluation and is considered representative of real world data center workload [22]. Our interest is in the "task_usage" table, which contains task resource usage information (resource refers to CPU, memory, or disk) for every 5-minute measurement interval, and the "machine_events" table, which contains server resource information along with the time it was added, updated, or removed from the cluster. While server CPU utilization is not directly provided in the data, we can derive this information for every 5-minute measurement interval from the "task_usage" table by summing the CPU utilization of all the tasks running on a particular server during the measurement interval. This is repeated for all servers in the cluster. Following are some specific cases and how we handled them.

- If no record exists for a server in a measurement interval, this means that no task was assigned to that server and the CPU utilization of that server is assumed to be zero.
- There may be multiple records for a server in a measurement interval (for different tasks), we sum the CPU utilization of these records to get the CPU utilization of the server. Some records have intervals less than the measurement interval (less than 5 minutes). To account for this, we weight the CPU utilization according to the interval of the record. For example, our measurement intervals are fixed at 5 minutes but if a record indicates an interval of 2 minutes and the CPU utilization is 0.2, then we weight it as $0.2 \times ({}^{2}/_{5})$.
- Out of all the task utilization records, 583 records (less than 0.00005%) have CPU utilization more than 1, with one measurement as high as 145.8. This may be due to some measurement error. We truncate such values to 1.

We constructed a time series of CPU utilization for each of the 12,583 unique servers. Each time series has 8,351 values corresponding to 5-minute measurement intervals throughout the 29 day period. Servers are identified using a unique machine ID in the dataset. We plot the CPU utilization time series and histogram of a few selected servers in Fig. 11. Daily and weekly patterns are visible in some of them. The CPU utilization for some servers never goes above 0.25 (as in the case of machine ID 3696086053) or 0.5 (as in the case of machine ID 400501120 and 38649400). This is because the CPU utilization is normalized with respect to the largest CPU capacity

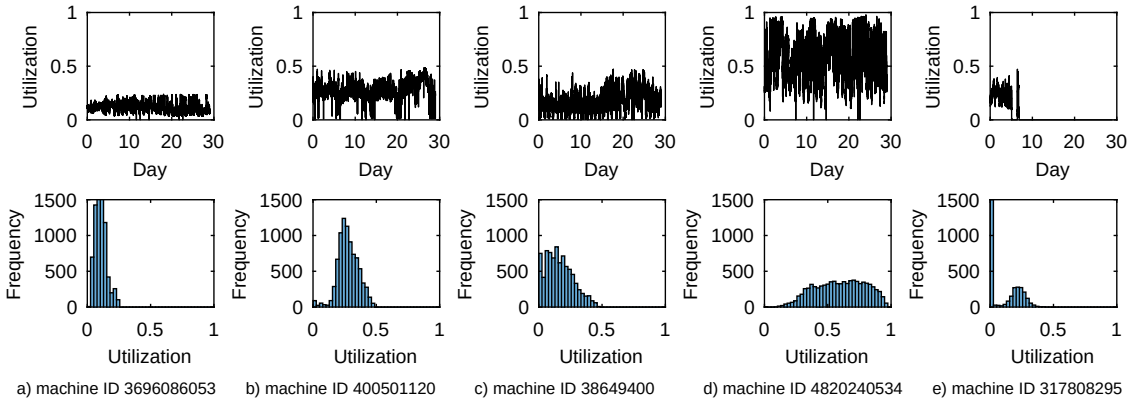


Figure 11: CPU utilization time series and histogram of five selected servers. a) CPU utilization is under 0.25; daily pattern visible b) CPU utilization is under 0.5; weekly pattern visible c) CPU utilization is under 0.5; daily pattern visible d) CPU utilization is under 1; high utilization e) CPU utilization ends after the first week.

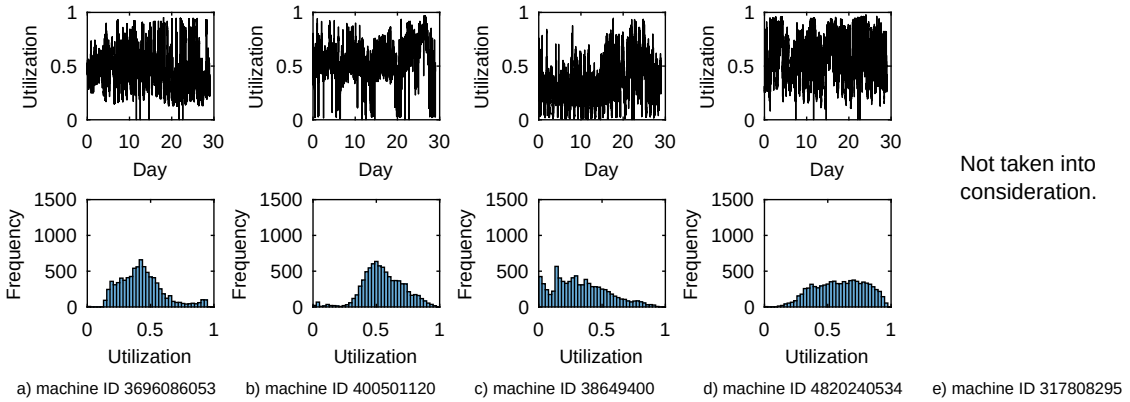


Figure 12: CPU utilization time series and histogram of same servers in Figure 11 after scaling or not taking into consideration. a) scaled by 4 b) scaled by 2 c) scaled by 2 d) no scaling required e) not taken into consideration as this server is removed from cluster during the trace period.

[23]. CPU capacity of each server can be found in the "machine_events" table which we used to scale the CPU utilization for that server. For example, the CPU capacity for server with machine ID 400501120 is 0.5, shown in Fig. 11b, so we multiply its CPU utilization by 2. We get the actual CPU utilization of servers after scaling. CPU utilization for server with machine ID 317808295, shown in Fig. 11e, ends abruptly after a week as this server was removed from the cluster. We do not take such servers, which get removed or which were later added to the cluster, into further consideration to end up with 7,171 server traces. These steps were taken to ensure that we don't overestimate the oversubscription possible. The CPU utilization time series and histogram of this set of servers, after filtering and scaling, is shown in Fig. 12.

4.2. Power-utilization data from SPECpower benchmark

Standard Performance Evaluation Corporation (SPEC) has developed an energy efficiency benchmark called SPECpower_ssj2008 [20] to measure and compare the energy efficiency of servers. The benchmark loads a server with a server-side Java (ssj) graduated workload from idle to 100% utilization at steps of 10% utilization and measures the throughput (in ssj operations per second) and power (in Watts) at these 11 utilization levels. Various hardware vendors test their servers using this benchmark and report it to SPEC. The self-reported results can be downloaded from the SPEC website [20]. There are 594 results published thus far on the website out of which 40 are non-complaint. We use the 554 complaint results for our experiment. In addition to power and throughput values at different utilization level, this data contains various information about the server and test conditions, such as, its technical specifications, hardware availability date, publication date, software settings for the benchmark, etc.

In Fig. 13, we plot the normalized power versus utilization curves for all the 554 servers from the SPECpower data. Each faint line in Fig. 13 represents a SPECpower server and we can observe servers having varying energy proportionality, with DR ranging from 0.21 to as high as 0.91. Moreover, these servers have varying hardware availability dates with some models from 2004 while some very recently released in 2018. To see a trend in energy proportionality over the past 15 years, we divide the servers into three 5-year intervals and calculate the average normalized power over utilization for each range. This is represented by the three thick lines with different markers in Fig.

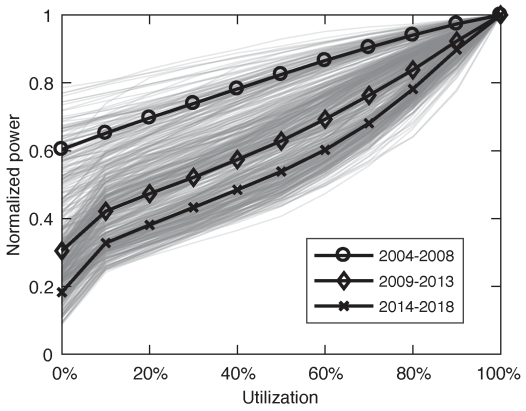


Figure 13: Normalized power versus utilization curves for all 554 servers from SPECpower data along with average for three 5-year intervals.

Table 2: Characteristics of the three selected SPEC servers.

System	Year	DR	EP	LD
HP G5	2006	0.33	0.37	-0.02
HP G6	2009	0.53	0.53	0.01
IBM M3	2010	0.73	0.65	0.07

13. We can see that servers are getting more energy proportional over the years.The improvement in energy proportionality seems to have slowed down for the present year interval.

4.3. Safe power oversubscription prediction

For each of the 554 servers in the SPECpower data, power consumption by the server at a particular utilization is known (we linearly interpolate values for intermediate utilization). We can calculate the power consumption of the Google cluster for a particular type of SPECpower server model by taking the sum of power consumption of all the (homogeneous) servers in the cluster. This aggregate cluster power consumption depends on the power-utilization characteristics of the server that was deployed. For now, we assume that all SPECpower servers have same performance (ignore throughput-utilization characteristics) implying same utilization for the Google cluster workload.

We select three servers for illustration purposes, 1) HP ProLiant DL380 G5, 2) HP ProLiant DL385 G6, and 3) IBM System x3400 M3, which we will refer to as HP G5, HP G6, and IBM M3 for brevity, respectively. All three servers have the same peak power of 258 Watts but differ in their energy proportionality. Fig. 14a shows the power-utilization curve and Table 2 lists the characteristics for the three selected servers. IBM M3 is more energy proportional than HP G6, which in turn is more energy proportional than HP G5 as indicated by the DR and EP values. Equal DR and EP values as well as LD being close to zero signifies that HP G6's power-utilization curve is most close to being linear.

The aggregate power consumption of the Google cluster, if any one of the servers were homogeneously deployed, is shown in Fig. 14b, with corresponding CDF in Fig. 14c. Here, P_{max} (maximum aggregate cluster power possible) is the sum of individual server peak power, about 1.85 MW (258 Watts $\times$ 7,171 servers) in these cases as all three servers have the same peak power consumption. We see that the aggregate cluster power does not reach P_{max} in any of the three scenarios, as all servers in the cluster do not peak (are fully utilized) simultaneously. This represents the opportunity for power oversubscription in real data centers. Furthermore, as the servers get more energy proportional, aggregate power consumption decreases even though all three servers have the same peak power rating, providing more opportunity for oversubscription for the same probability of overloading. Fixing the probability of overloading at 0.001 (0.1%), we can find the power limit, P_{limit}, as the aggregate cluster power at which CDF reaches 0.999. Corresponding safe oversubscription level can be calculated for each scenario. Repeating this procedure, we find the safe oversubscription level for each of the 554 SPECpower servers.

Plotting safe oversubscription level against DR of the server, we get the scatter plot as shown in Fig. 15. Each point represents one of the 554 servers. The three selected servers discussed earlier are marked and labelled separately for reference. The dotted line represents the mathematically calculated safe oversubscription level for varying DR according to Eq. (11) derived in the previous section. The empirically calculated points are dispersed around the mathematically predicted line with mean absolute percentage error (MAPE) of about 21.17% since our mathematical derivation assumed linear power-utilization curves. However, the points show a general trend of increasing safe oversubscription possible as the DR of servers increases. We have colored the points according to the LD of servers which measures their deviation from linearity. Servers which have LD close to zero fall very close to the predicted line while those having positive LD fall below the line and vice versa. Furthermore, the departure of the empirical value from the mathematically calculated value is proportional to the magnitude of LD.

LD less than zero implies that the power-utilization curve of the server is sub-linear, meaning, it would consume lower power than a linear power-utilization curve server with same DR. This results in more oversubscription possible than predicted by liner approximation. In other words, our mathematically predicted values are conservative estimates of actual possible power oversubscription possible with such servers. This is reflected in Fig. 15 where blue points (negative LD) are above the predicted line, that is, more oversubscription is possible than mathematically predicted. The situation is opposite for servers with positive LD. Our prediction is an over estimate in such case as evident by red points being below the line.

Plotting a similar scatter plot but with the EP metric of servers, we have a scatter plot as shown in Fig. 16. Here the points are more close to the predicted line with mean absolute percentage error (MAPE) being around 10.97% compared to 21.17% MAPE for the case with DR. This is due to fact that DR only looks at end points and is independent to the shape of the power-utilization curve while EP takes the entire area into consideration. Hence, the EP metric serves as a better linear approximation to predict the safe power oversubscription level compared to the DR metric. Once again, we observe that more

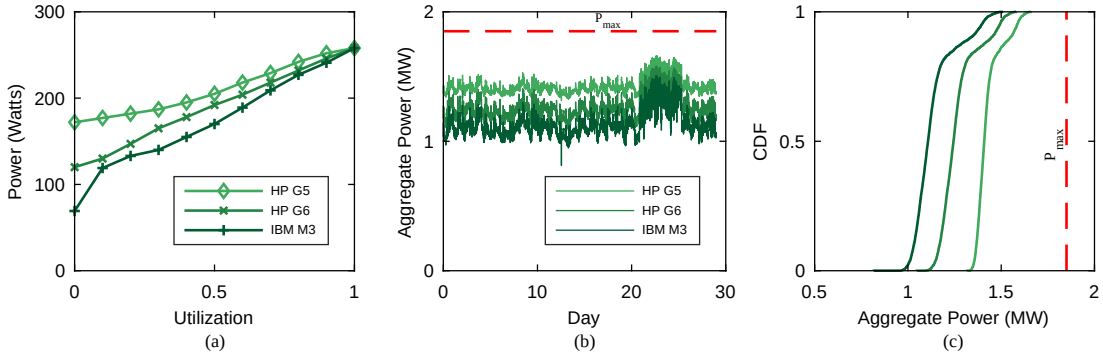


Figure 14: (a) Power-utilization curve, (b) aggregate cluster power consumption over 29 days, and (c) corresponding CDF of aggregate cluster power consumption over 29 days, for the three selected SPEC servers.

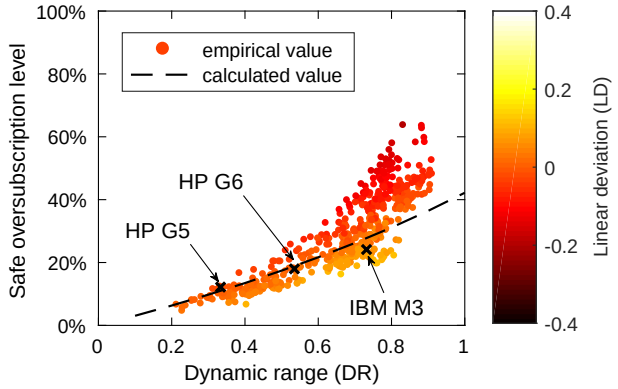


Figure 15: Predicted and actual safe oversubscription level at different DR of servers at 0.1% probability of overloading.

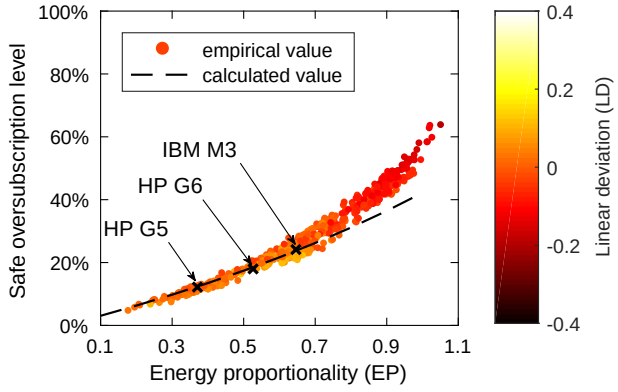


Figure 16: Predicted and actual safe oversubscription level at different EP of servers at 0.1% probability of overloading.

oversubscription is possible as servers get more energy proportional and the relation is better than linear, same as we observed before in Fig. 9 for the i.i.d. case in the previous section.

4.4. Taking server performance into account

The SPECpower benchmanrk measures server throughput in terms of ssj operations per second (ssj_ops). However, the SPECpower dataset contains a wide range of servers with peak throughput ranging from 0.026 million ssj_ops to 70.6 million ssj_ops. This is a difference of three orders of magnitude in terms of throughput and we cannot have a meaningful comparison of such contrasting servers as a workload that drives the slowest server to 100% utilization could be negligibly small to the fastest server. Hence, we select 119 servers (about one fifth of total) with peak throughput between 1 million ssj_ops to 2 million ssj_ops. Note that the fastest server is already twice as fast as the slowest server in this selected group of servers. Similarly, we multiply the Google server utilization data by 1 million ssj_ops to get the workload trace for further evaluation, such that all selected servers are able to serve the offered ssj_ops (that is, even the slowest server can handle the workload without reaching 100% utilization).

Taking the performance characteristics of SPECpower servers into account will result in varying utilization distribution of servers according to their peak throughput for the same workload. This implies that, in addition to energy proportionality, peak throughput of the server also affects the safe oversubscription level. As the server peak throughput gets higher, the corresponding utilization, and thus power, are proportionally lower for a given workload, leading to more opportunity for power oversubscription. Calculating and plotting the safe oversubscription level (at 0.01% probability of overloading) for each server against EP and peak throughput, we get a 3D scatter plot as shown in Fig. 17 where each point represents one of the 119 servers. We also mathematically calculate the predicted safe oversubscription level for varying EP according to Eq. (11), but now repeatedly for various peak throughput in the range 1 million ssj_ops to 2 million ssj_ops. Hence, we end up with a surface rather than a line as shown in Fig. 17. We can observe from this surface that safe oversubscription level increases with, both, increasing peak throughput as well as increasing EP. The empirically calculated points are very close to the mathematically calculated surface with deviation caused by non-linearity of power-utilization curve of the server. Mean absolute percentage error (MAPE) between the predicted surface and the actual points comes out to be about 10% as in the previous case where we ignored server performance.

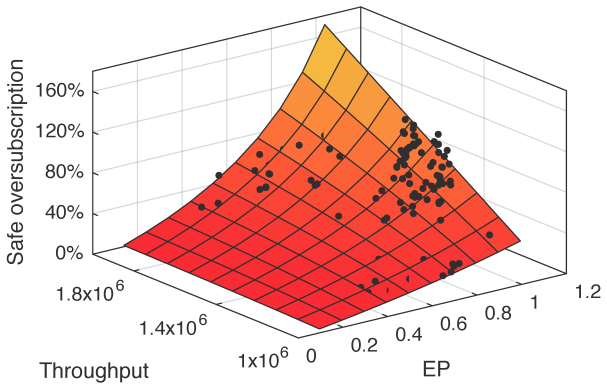


Figure 17: Predicted and actual safe oversubscription level at different EP and peak throughput of servers at 0.1% probability of overloading.

5. Related work

Many extensive surveys [24, 25, 26, 27, 28] have reviewed the numerous works in the area of power management and energy efficiency of data centers. However, the focus on safe power oversubscription of data centers has been relatively limited. The concept of oversubscribing or overbooking resources can be found within various aspects of a data center, for example, virtual machines oversubscribing physical servers [29] or servers oversubscribing network bandwidth [30], and seems only natural to be extended to oversubscribing the power infrastructure as it increases profit as well as resource utilization.

Determining the optimum number of servers that can be deployed within a given power limit is non-trivial [1]. If too few a servers are deployed, the expensive data center power infrastructure is left severely under utilized. While, if too many servers are deployed, there is a risk of frequent power overload. Femal and Freeh [31, 32] and Ranganathan et al. [33] were among the first to demonstrate power oversubscription in data centers using prototypes with a few servers. They used dynamic voltage-frequency scaling of CPU to control individual server power consumption and avoid simultaneous peaking of servers. Fan et al. [2] analyzed power profile of a real production data center at Google and observed that dynamic range of power usage decreased as you go up the power hierarchy. They found that there was more stranded power left never utilized at the cluster level (over one-fourth of the power limit) than at the rack level suggesting that plenty of opportunity exists in practice for safe power oversubscription of data centers.

Power overload is always a risk in oversubscribed data centers and control mechanisms, acting as a safety net, have been proposed to manage aggregate peak power [34, 35, 5, 6, 36, 7, 8]. Wang et al. [35] proposed a hierarchical power capping architecture based on control theory to cap power at different levels in an oversubscribed data center. Bhattacharya et al. [36] considered admission control of workload, in addition to dynamic voltage-frequency scaling, as a control knob to cap individual server power. Li et al. [8] extended these techniques to data centers with redundant power infrastructure. Power capping requires throttling of individual servers which results in performance degradation. An orthogonal approach to handle short power overloads is to use UPS batteries for the required excess power as proposed in [34, 5, 6]. Govindan et al. [5] studied the feasibility of using distributed UPS batteries during power overloads. There are also works studying power oversubscription in specific types of data centers, such as, High Performance Computing (HPC) clusters [37, 38] and Multi-Tenant Data Centers (MTDC) [39, 12]. However, these works do not study the effect of a server's power and performance characteristics on the amount of power oversubscription possible.

With most of the studies evaluated through simulations or on a small prototype, very little published work exists on power oversubscription of production systems. Wu et al. [4] first reported that Facebook had been using an in-house power management system called Dynamo to oversubscribe their data centers since 2013. Dynamo consists of a light agent running on each server able to measure and cap its power while higher level controllers monitor agents and make power capping decisions. Similarly, Sakamoto et al. [7] study oversubscription of a production HPC system at Kyushu University containing 965 compute nodes (with 23,160 cores and 120 TB memory). They extend the SLURM [40] resource manager to incorporate power-awareness and oversubscribe power to increase power utilization. Both [4] and [7] use Intel's RAPL (Running Average Power Limit) [41] interface to cap power of individual servers.

Barroso et al. [9] advocated for the need of energy proportional computing to make data centers more energy efficient. At that time, servers were consuming more than 60% of their peak power even when completely idle. Using real production data center workload, Fan et al. [2] projected that well over 30% energy saving was possible if the servers were replaced by more energy proportional ones. Dynamically provisioning servers according to the workload have also been studied [42]. Cluster management techniques dynamically scale the number of active servers to make the cluster energy proportional even though the underlying individual servers are highly energy disproportional. Wong et al. [43] argued that such complex cluster management techniques might not be needed for energy saving as underlying servers become more energy proportional. Over the years, energy proportionality of servers has steadily improved [10, 14, 11] and has helped data centers become more energy efficient. However, the effect of energy proportional servers on opportunities for data center power oversubscription has not been studied and our paper is a step in that direction. To the best of our knowledge, our work is the first to characterize power oversubscription of data center in terms of server's power and performance characteristics.

6. Summary and future work

In this work, we formulated the relationship between safe power oversubscription of a data center and the energy proportionality of the servers deployed within it. Using real world cluster utilization data and power-utilization data for different server models, we showed how our framework based on

a linear power-utilization approximation can successfully predict the safe power oversubscription for different energy proportional servers within a 10% error on average. Prediction error is caused by non-linear server power characteristics, with higher error resulting from more aggressive violation of our linear assumption and the direction of error is known. We found, through both a synthetic i.i.d scenario as well as real world data, that although the exact value of safe oversubscription possible will depend upon the aggregate power distribution, the safe oversubscription level increases better than linearly with increasing server energy proportionality for a fix probability of overloading. We also found that EP is a better server metric than DR for prediction, implying that safe power oversubscription depends upon the entire power-utilization curve of the server rather than just its idle and peak power. Furthermore, a server with sub-linear power-utilization curve could be oversubscribed more than a server with super-linear power-utilization curve.

In the future, we would like to define the acceptable probability of power overloading for an oversubscribed data center more precisely. A 0.01% probability of overloading implies about 43 minutes of power overload every month. However, if that 43 minutes of power overload occurred continuously on a single day of the month, it could be more severe than, for example, a 10 minute power overload every week. We would like to characterize power overload in a way that distinguishes between these two scenarios and study the effect of server energy proportionality on it.

Acknowledgement

We would like to thank the anonymous reviewers for their valuable feedback in improving this paper.

References

[1] L. A. Barroso, U. Hölzle, P. Ranganathan, The datacenter as a computer: Designing warehouse-scale machines, Synthesis Lectures on Computer Architecture 13 (3) (2018) i–189.

[2] X. Fan, W.-D. Weber, L. A. Barroso, Power provisioning for a warehouse-sized computer, in: Proceedings of the 34th Annual International Symposium on Computer Architecture, ISCA '07, ACM, 2007, pp. 13–23.

[3] M. A. Islam, L. Yang, K. Ranganath, S. Ren, Why some like it loud: Timing power attacks in multi-tenant data centers using an acoustic side channel, Proc. ACM Meas. Anal. Comput. Syst. 2 (1) (2018) 6:1–6:33.

[4] Q. Wu, Q. Deng, L. Ganesh, C.-H. Hsu, Y. Jin, S. Kumar, B. Li, J. Meza, Y. J. Song, Dynamo: Facebook's data center-wide power management system, in: 2016 ACM/IEEE 43rd Annual International Symposium on Computer Architecture (ISCA), 2016, pp. 469–480.

[5] S. Govindan, D. Wang, A. Sivasubramaniam, B. Urgaonkar, Leveraging stored energy for handling power emergencies in aggressively provisioned datacenters, SIGARCH Comput. Archit. News 40 (1) (2012) 75–86.

[6] V. Kontorinis, L. E. Zhang, B. Aksanli, J. Sampson, H. Homayoun, E. Pettis, D. M. Tullsen, T. S. Rosing, Managing distributed ups energy for effective power capping in data centers, in: 2012 39th Annual International Symposium on Computer Architecture (ISCA), 2012, pp. 488–499.

[7] R. Sakamoto, T. Cao, M. Kondo, K. Inoue, M. Ueda, T. Patki, D. Ellsworth, B. Rountree, M. Schulz, Production hardware overprovisioning: Real-world performance optimization using an extensible power-aware resource management framework, in: 2017 IEEE International Parallel and Distributed Processing Symposium (IPDPS), 2017, pp. 957–966.

[8] Y. Li, C. R. Lefurgy, K. Rajamani, M. S. Allen-Ware, G. J. Silva, D. D. Heimsoth, S. Ghose, O. Mutlu, A scalable priority-aware approach to managing data center server power, in: 2019 IEEE International Symposium on High Performance Computer Architecture (HPCA), 2019, pp. 1–14, (to appear).

[9] L. A. Barroso, U. Hölzle, The case for energy-proportional computing, Computer 40 (12) (2007) 33–37.

[10] F. Ryckbosch, S. Polfliet, L. Eeckhout, Trends in server energy proportionality, Computer 44 (9) (2011) 69–72.

[11] C. Jiang, Y. Wang, D. Ou, B. Luo, W. Shi, Energy proportional servers: Where are we in 2016?, in: 2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS), 2017, pp. 1649–1660.

[12] S. Malla, K. Christensen, Reducing power use and enabling oversubscription in multi-tenant data centers using local price, in: 2017 IEEE International Conference on Autonomic Computing (ICAC), 2017, pp. 161–166.

[13] B. Rountree, D. H. Ahn, B. R. De Supinski, D. K. Lowenthal, M. Schulz, Beyond dvfs: A first look at performance under a hardware-enforced power bound, in: 2012 IEEE 26th International Parallel and Distributed Processing Symposium Workshops PhD Forum, 2012, pp. 947–953.

[14] C.-H. Hsu, S. W. Poole, Measuring server energy proportionality, in: Proceedings of the 6th ACM/SPEC International Conference on Performance Engineering, ICPE '15, 2015, pp. 235–240.

[15] A. Shehabi, S. Smith, D. Sartor, R. Brown, M. Herrlin, J. Koomey, E. Masanet, N. Horner, I. Azevedo, W. Lintner, United states data center energy usage report (2016).

[16] D. Wong, M. Annavaram, Knightshift: Scaling the energy proportionality wall through server-level heterogeneity, in: Proceedings of the 2012 45th Annual IEEE/ACM International Symposium on Microarchitecture, MICRO-45, 2012, pp. 119–130.

[17] R. Sharma, M. Gupta, G. Kapoor, Some better bounds on the variance with applications, Journal of Mathematical Inequalities 4 (3) (2010) 355–363.

[18] S. Malla, K. Christensen, Choosing the best server for a data center: The importance of workload weighting, in: IEEE International Performance Computing and Communications Conference (IPCCC), 2018, pp. 1–8, (to appear).

[19] X. Fu, X. Wang, C. Lefurgy, How much power oversubscription is safe and allowed in data centers, in: Proceedings of the 8th ACM International Conference on Autonomic Computing, ICAC '11, 2011, pp. 21–30.

[20] Standard Performance Evaluation Corporation, SPECpower_ssj 2008, (accessed 3 December 2018).
URL `https://www.spec.org/power_ssj2008/`

[21] J. Wilkes, More Google cluster data, Google research blog, posted at `http://googleresearch.blogspot.com/2011/11/more-google-cluster-data.html`. (Nov. 2011).

[22] M. Carvalho, D. A. Menascé, F. Brasileiro, Capacity planning for IaaS cloud providers offering multiple service classes, Future Generation Computer Systems 77 (2017) 97–111.

[23] C. Reiss, J. Wilkes, J. L. Hellerstein, Google cluster-usage traces: format + schema, Tech. rep., Google Inc., Mountain View, CA, USA, revised 2014-11-17 for version 2.1. (2011).

[24] A. Beloglazov, R. Buyya, Y. C. Lee, A. Zomaya, A taxonomy and survey of energy-efficient data centers and cloud computing systems, Advances in Computers 82 (2) (2011) 47–111.

[25] K. Bilal, S. U. R. Malik, O. Khalid, A. Hameed, E. Alvarez, V. Wijaysekara, R. Irfan, S. Shrestha, D. Dwivedy, M. Ali, U. S. Khan, A. Abbas, N. Jalil, S. U. Khan, A taxonomy and survey on green data center networks, Future Generation Computer Systems 36 (2014) 189–208.

[26] A.-C. Orgerie, M. D. d. Assuncao, L. Lefevre, A survey on techniques for improving the energy efficiency of large-scale distributed systems, ACM Comput. Surv. 46 (4) (2014) 47:1–47:31.

[27] M. Zakarya, L. Gillam, Energy efficient computing, clusters, grids and clouds: A taxonomy and survey, Sustainable Computing: Informatics and Systems 14 (Supplement C) (2017) 13–33.

[28] S. Malla, K. Christensen, A survey on power management techniques for oversubscription of multi-tenant data centers, ACM Comput. Surv.(to appear).

[29] F. Lopez-Pires, B. Baran, L. Benitez, S. Zalimben, A. Amarilla, Virtual machine placement for elastic infrastructures in overbooked cloud computing datacenters under uncertainty, Future Generation Computer Systems 79 (2018) 830–848.

[30] J. Cao, Z. Ma, J. Xie, X. Zhu, F. Dong, B. Liu, Towards tenant demand-aware bandwidth allocation strategy in cloud datacenter, Future Generation Computer Systems(in press, 6 July 2017).
[31] M. E. Femal, V. W. Freeh, Safe overprovisioning: Using power limits to increase aggregate throughput, in: Power-Aware Computer Systems, Springer Berlin Heidelberg, 2005, pp. 150–164.
[32] M. E. Femal, V. W. Freeh, Boosting data center performance through non-uniform power allocation, in: Second International Conference on Autonomic Computing (ICAC'05), 2005, pp. 250–261.
[33] P. Ranganathan, P. Leech, D. Irwin, J. Chase, Ensemble-level power management for dense blade servers, in: Proceedings of the 33rd Annual International Symposium on Computer Architecture, ISCA '06, 2006, pp. 66–77.
[34] S. Govindan, A. Sivasubramaniam, B. Urgaonkar, Benefits and limitations of tapping into stored energy for datacenters, in: 2011 38th Annual International Symposium on Computer Architecture (ISCA), 2011, pp. 341–351.
[35] X. Wang, M. Chen, C. Lefurgy, T. W. Keller, Ship: A scalable hierarchical power control architecture for large-scale data centers, IEEE Transactions on Parallel and Distributed Systems 23 (1) (2012) 168–176.
[36] A. A. Bhattacharya, D. Culler, A. Kansal, S. Govindan, S. Sankar, The need for speed and stability in data center power capping, Sustainable Computing: Informatics and Systems 3 (3) (2013) 183–193.
[37] T. Patki, D. K. Lowenthal, B. Rountree, M. Schulz, B. R. de Supinski, Exploring hardware overprovisioning in power-constrained, high performance computing, in: Proceedings of the 27th International ACM Conference on International Conference on Supercomputing, ICS '13, 2013, pp. 173–182.
[38] O. Sarood, A. Langer, A. Gupta, L. Kale, Maximizing throughput of overprovisioned hpc data centers under a strict power budget, in: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, SC '14, 2014, pp. 807–818.
[39] M. A. Islam, X. Ren, S. Ren, A. Wierman, X. Wang, A market approach for handling power emergencies in multi-tenant data center, in: 2016 IEEE International Symposium on High Performance Computer Architecture (HPCA), 2016, pp. 432–443.
[40] A. B. Yoo, M. A. Jette, M. Grondona, Slurm: Simple linux utility for resource management, in: Job Scheduling Strategies for Parallel Processing, Springer Berlin Heidelberg, 2003, pp. 44–60.
[41] H. David, E. Gorbatov, U. R. Hanebutte, R. Khanna, C. Le, Rapl: Memory power estimation and capping, in: 2010 ACM/IEEE International Symposium on Low-Power Electronics and Design (ISLPED), 2010, pp. 189–194.
[42] A. Gandhi, M. Harchol-Balter, R. Raghunathan, M. A. Kozuch, Autoscale: Dynamic, robust capacity management for multi-tier data centers, ACM Trans. Comput. Syst. 30 (4) (2012) 14:1–14:26.
[43] D. Wong, M. Annavaram, Implications of high energy proportional servers on cluster-wide energy proportionality, in: 2014 IEEE 20th International Symposium on High Performance Computer Architecture (HPCA), 2014, pp. 142–153.