A TRANSFORMATIVE APPROACH TO POWER OPTIMIZATION IN DATA CENTERS

by © Virtual Power Systems, Inc.

TABLE OF CONTENTS

1 INTRODUCTION 4

2 DATA CENTER POWER PROVISIONING 6

2.1 AVAILABILITY CONSIDERATIONS 7

2.2 CAPACITY CONSIDERATIONS IN FLUCTUATING IT LOAD 8

3 INTRODUCTION TO ICE – INTELLIGENT CONTROL OF ENERGY® 10

4 ICE MECHANISMS 12

4.1 ICE PEAK SHAVING. USE OF ALTERNATE SOURCE OF ENERGY 13

4.2 DYNAMIC REDUNDANCY 15

5 TOTAL COST OF OWNERSHIP [TCO] CONSIDERATIONS 18

6 CONCLUSION 20

REFERENCES 21

FIGURE INDEX

FIGURE 1: DATA CENTER MONTHLY COSTS SOURCE: JAMES HAMILTON'S BLOG 4
FIGURE 2: SIMPLIFIED DATA CENTER POWER DISTRIBUTION HIERARCHY 6
FIGURE 3: 2N CONFIGURATION POWER SETUP IN A TIER 3/4 DATA CENTER 7
FIGURE 4: LARGE COLOCATION DATA CENTER 9
FIGURE 5: ICE SYSTEM PLATFORM 10
FIGURE 6: ICE ABSTRACTION FOR SOFTWARE DEFINED DATA CENTER 11
FIGURE 7: CONCEPT OF PEAK SHAVING 14
FIGURE 8: ICE SOFTWARE AND HARDWARE SETUP 12
FIGURE 9: CHARACTERIZATION OF PEAKS 13
FIGURE 10: 2N REDUNDANT DATA CENTER CAN NOW HAVE EXTRA CAPACITY FOR LOWER PRIORITY RACKS 15
FIGURE 11: REMEDIATION DURING FAILURE, IMPACTING ONLY LOWER SLA RACKS 16
FIGURE 12: POWER UTILIZATION IN A RACK OF A LARGE COLO 18

1.1 INTRODUCTION

Recently, the IT industry has made tremendous progress in delivering services with ease and scale. Cloud computing and Infrastructure as a Service can provision IT infrastructure at the blink of an eye. Software Defined data centers allow self-serviced users to deploy services and workload in seconds. These advancements, although revolutionary, are putting tremendous pressure on data centers to scale. Often the most constrained commodity in data centers is power.

Existing data centers exhaust power before running out of space but adding power within a data center is both time consuming and expensive. According to James Hamilton of AWS [1], Power and Cooling Infrastructure is second in cost only to servers, with power itself coming in a close third. Accounting for the time spent and the opportunity cost, readying infrastructure for power delivery only increases the cost of power.

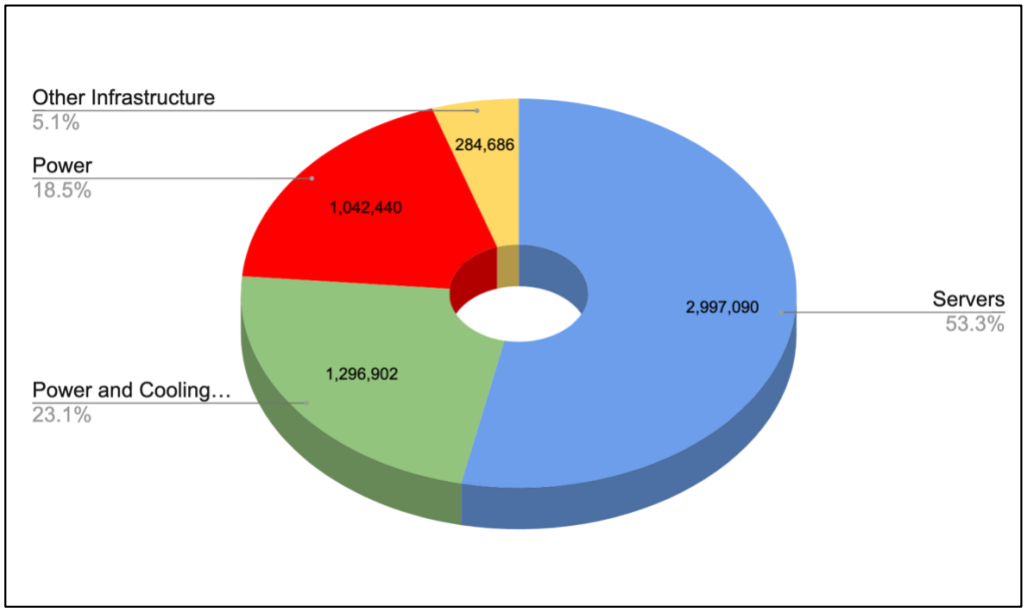

Figure 1: Data center Monthly Costs Source: James Hamilton's Blog

Timely provisioning of power for new IT workloads or racks is a challenge in itself, as is finding available power for expansion or new data center projects from the grid. Many utility locations may not be in a position to increase power capacity.

The industry is not standing still in the face of this power problem. Server designs are far more power efficient, with lower idle power consumption, and higher compute density. Additionally, Virtualization, or Software Defined Compute (SDC), is now a default technology in IT and allows for more efficient utilization of IT infrastructure. Data center facilities have also seen major improvements in power efficiencies, with the Mega data centers now operating at sub 1.2 PUE.

Such progress has helped drive efficiency in power utilization but has not been able to improve the power capacity utilization levels across the data center. The power infrastructure is still over-provisioned and designed to serve peak loads. Current architectures entail rigid and static allocation of power and power back-ups, locking and undermining power capacity and thus lowering power utilization.

The problem of underutilization of capacity is not a new phenomenon in data centers and has been successfully addressed in other non-power areas. Virtualization has increased the productivity in server capacity and hence prevented server sprawl. The Software definition of infrastructure is a key arena in the search to unlock lost capacity. Power infrastructure, which is dominated by hardware and hardware vendors, has failed to make the same strides in **Software Defined Power**® as it has in hardware.

A fortuitous outcome of capacity repatriation with virtualization or software defined infrastructure has been an increased level of automation and intelligence to infrastructure management. This mask the increasing complexity of systems and architecture and enables more efficient scaling and operations.

This paper defines a comprehensive approach to employing Software Defined Power to resolve the power capacity problem across data centers. It begins with a description of current 2N Redundant data centers that highlights where capacity is constrained. After this is an overview of **ICE** (Intelligent Control of Energy) hardware and software solutions followed by the use cases that demonstrate how locked capacity can be released using said solutions. We believe that the methods described will optimize power capacity use by IT equipment when weighed empirically to account for service levels and high availability requirements.

2 DATA CENTER POWER PROVISIONING

To start our analysis, we are going to describe the level set on the basics of power infrastructure topology and the three control points where power management decisions are made:

1. Automatic Transfer Switch (ATS) in the Gray Space
2. Power Distribution Units
3. Rack level

The figure to the right shows the topology of a typical data center. Power management decisions are made at the Automatic Transfer Switch (ATS) and the rack decisions are based on either availability or capacity. Availability decisions are tied to redundancy, which safeguards against failure scenarios while capacity decisions are for planning purposes.

Operators collect power consumption data for capacity planning to feed IT demand. We will now dig deeper into availability and resiliency of the existing capacity to better understand the problem at hand.

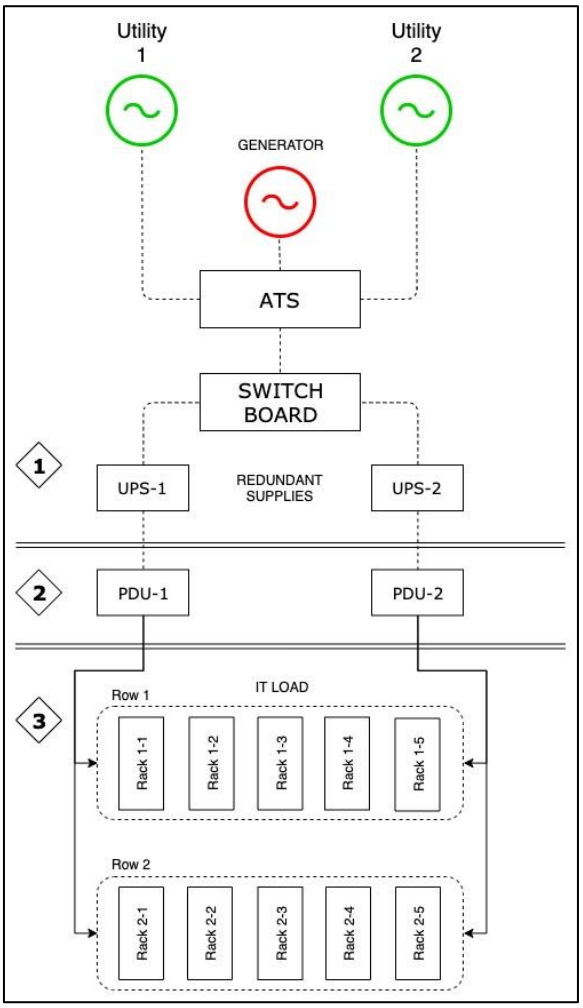

Figure 2: Simplified data center Power Distribution Hierarchy

2.1 AVAILABILITY CONSIDERATIONS

As illustrated in Figure 3 below, redundancy is built across the 2N data center. The figure depicts two completely independent feeds to the IT servers (Utility 1 and Utility 2). Each is made up of utility power feeds, followed by a generator feeding an UPS (Uninterruptible Power Supply) feeding PDUs (Power Distribution Units), then continuing all the way to dual corded servers. Because there are two of these pathways, the entire power line is redundant all the way to the IT workloads running on the server.

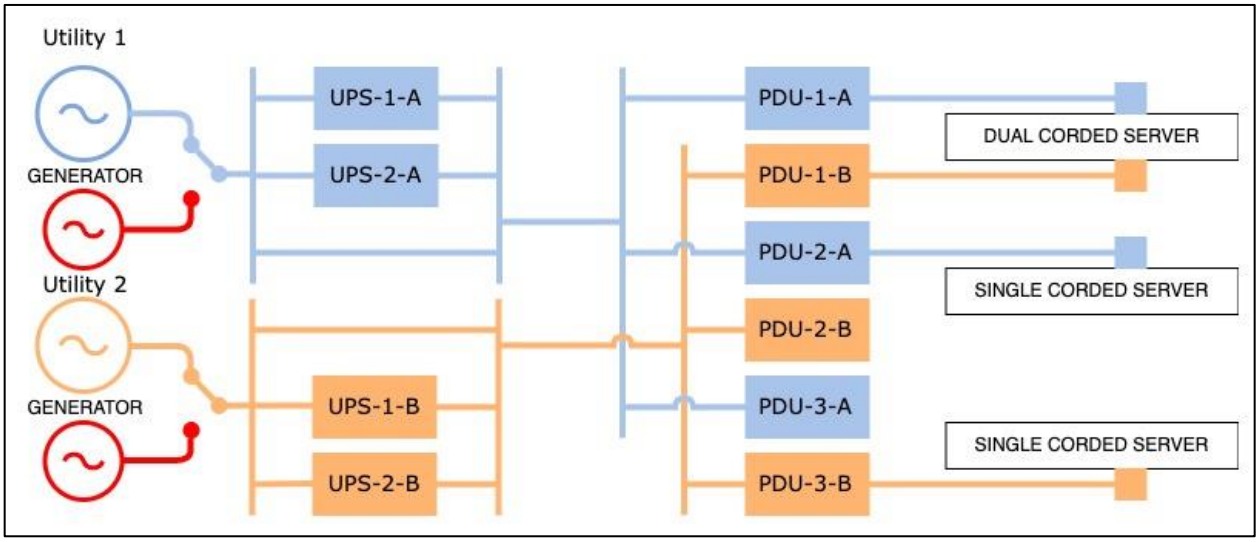

Figure 3: 2N Configuration Power Setup in a Tier 3/4 data center

These data centers are built on a 100% redundancy premise to deliver a highly available Mission Critical infrastructure. Though they are successful in doing so, they operate on the imprecise premise that all workloads are created equal and that they are all Mission Critical. Even in Tier 3-4 data centers, many workloads are not considered Mission Critical. For example, 20-30% of servers could be Dev/Test workloads, which many would not consider Mission Critical. If X% of your workloads under 2N redundancy are not Mission Critical, you have blocked X/2% of power capacity as unnecessarily unusable. It appears that there is no easy choice in this regard today, as the rigidity of power infrastructure does not allow for any other alternatives. But as we will see later on, configuring power through Software can unlock this constraint and make it an elastic resource.

2.2 CAPACITY CONSIDERATIONS IN FLUCTUATING IT LOAD

Measuring, predicting, and addressing the demand consumption is a big source of trouble in data center planning. The root of the issue lies in the variability of power consumption in IT equipment. Let us illustrate with a simple linear CPU utilization model, as this is known to be a fairly good estimator of power usage.

$$P_{server} = P_{idle} + \mu (P_{full} - P_{idle})$$

Where P_{idle} is the server power consumed at idle, and μ represents the CPU utilization. If idle power decreases as desired, the power spread on an individual server could easily exceed 600W. This leads to a larger spread at the Rack level.

Workloads add another dimension to this variability, something that is showcased by some excellent case studies. One of the most notable and pioneering came from Google and was published a whitepaper called 'Power Provisioning for a Warehouse-sized Computer' [2]. Below is a table from the paper that shows the ratio of average power to observed peak power for different workloads and mixes of workloads.

WORKLOAD	AVG OBSERVED PEAK
WEB SEARCH	**72.7 %**
WEB MAIL	**89.9 %**
MAP REDUCE	**77.2 %**
MIX	**84.4 %**
REAL DC	**82.8 %**

The best-case scenario in a mix of heterogeneous workloads in a real data center creates a 17% difference between average power consumed and peak power observed.

While average power largely determines the power bill, peak power guides the deployment of infrastructure and capacity in the data center. So even one of the best run data centers in the world could be stranding 17% of its power capacity.

Compounding the problem of the peaks is a fear of the peaks, causing outages, that leads data centers to add safety buffers on top of them to prevent breakers from tripping or from breaching UPS capacity.

Figure 4 shows a rack at a Large CoLo provider. The mean load was only 3.36kW in racks, with 4.8kW derated capacity (80% derated capacity of 6kW rack), yet this particular data center was considered out of capacity. Peaks and Peak safety buffers clearly played a role in this. It is due to these reasons; it is not surprising that average utilization in data centers worldwide is less than 40%. Utilization shrinks even further when including the redundancy factor that was discussed in the previous section.

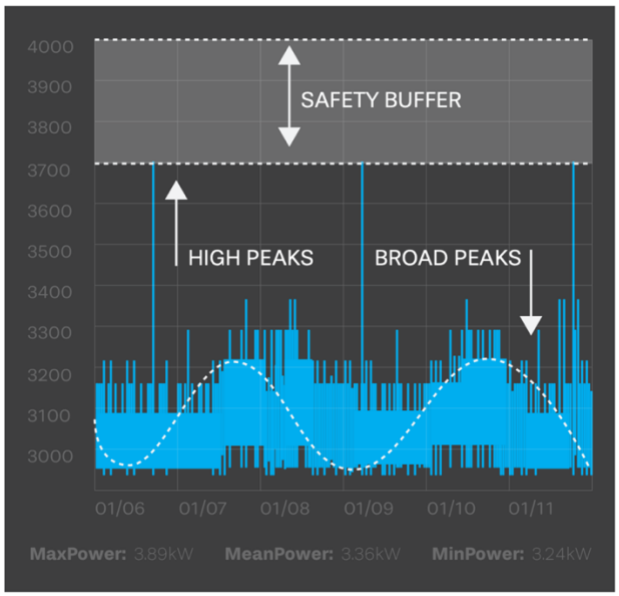

Figure 4: Large Colocation data center

Hope that the explanations given above have adequately laid out the rationale behind James Hamilton's observation that Power infrastructure is the second largest expense in the data center. In fact, we will realize later in this whitepaper that over provisioning and design of power infrastructure is the single biggest factor in the Total Cost of Ownership in a data center today.

3 INTRODUCTION TO ICE – INTELLIGENT CONTROL OF ENERGY®

Before diving into how we solve the power over provisioning problem, we will provide a brief introduction to our ICE product line.

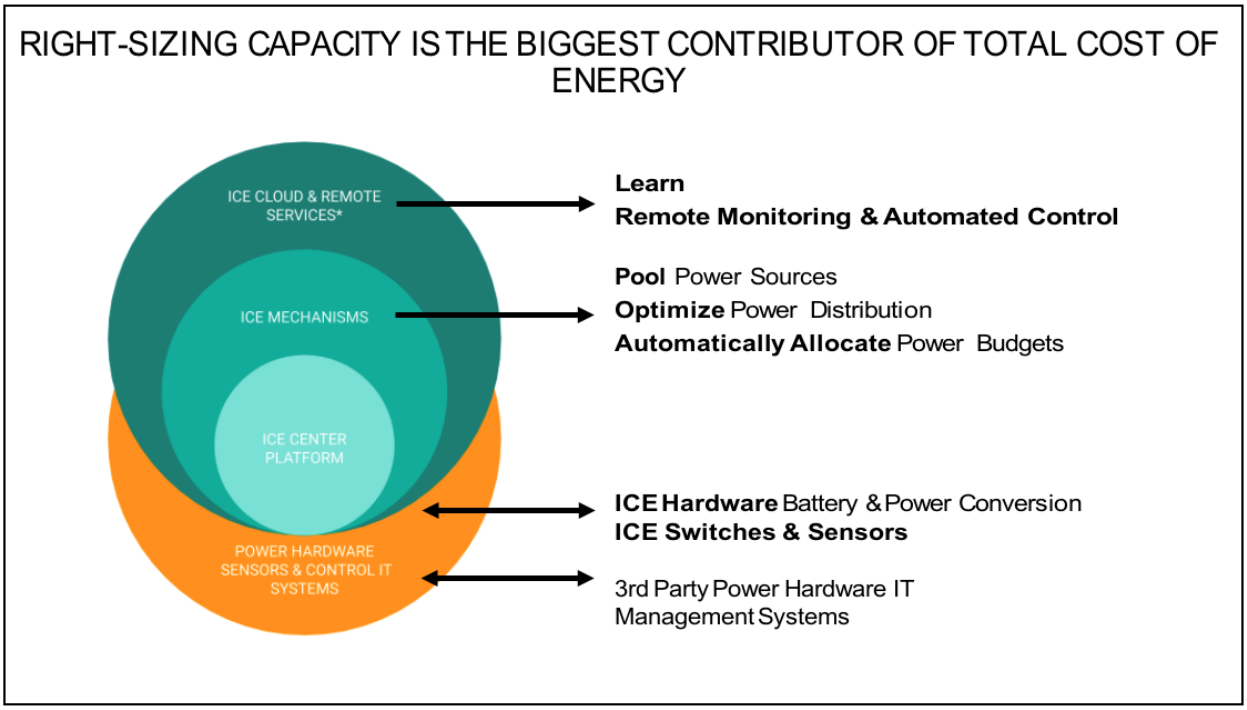

Figure 5: ICE System Platform

VPS Intelligent Control of Energy (ICE®) solutions are a combination of hardware and software modules that can be placed on the three power decision control points (Automatic Transfer Switch, Static Transfer Switch / Power Distribution Units, Rack) and help make dynamic, on-demand decisions on capacity and sourcing.

Capacity management in the power system is managed by Virtual Power Systems software stack called ICE (Intelligent Control of Energy®). The ICE software consists of an operating system, the ICE Center, that collects telemetry data from ICE devices and other infrastructure hardware and makes this data presentable for real time control operations in the user interface ICE Console. The ICE Console sit on top of the ICE OS and perform a variety of functions.

ICE OS and Applications can be provisioned on either a commodity server or a virtual machine. The ICE compatible hardware are third party control modules, like Batteries or Switches, which are optimized for and perform to appropriate function calls from the ICE software. The ICE hardware modules interface via an Ethernet port using SNMP or Mod Bus protocols. These modules are designed for real time control system operations, reporting sub second data and executing control operations within allocated line cycles.

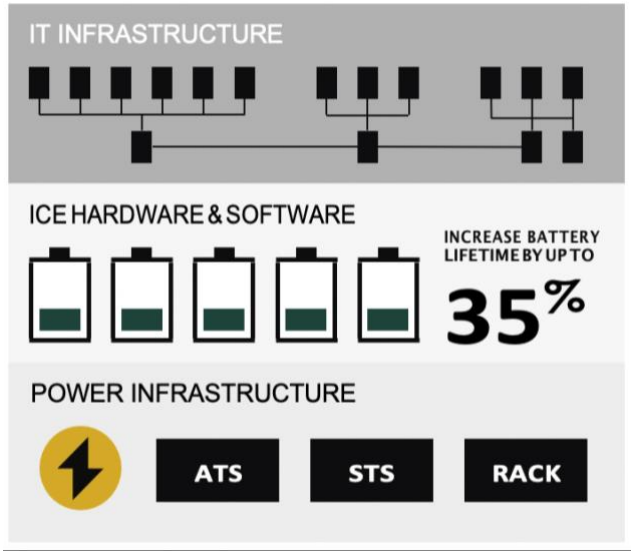

Figure 6: ICE Abstraction for Software Defined data center

Conceptually, ICE Hardware and Software introduce a hypervisor like abstraction between the three power control points and IT infrastructure to produce power pooling and elasticity. This dynamically distributes and allocates power, increasing capacity utilization of the infrastructure. We will go over some specific ICE mechanisms in the following sections to see how exactly ICE addresses the capacity and sourcing decisions described earlier.

4 ICE MECHANISMS

The figure below shows a conceptual ICE Center setup between the PDU and the servers. All ICE compatible hardware devices, including the one in the figure below, are monitored and controlled through the ICE Center, connected via Ethernet.

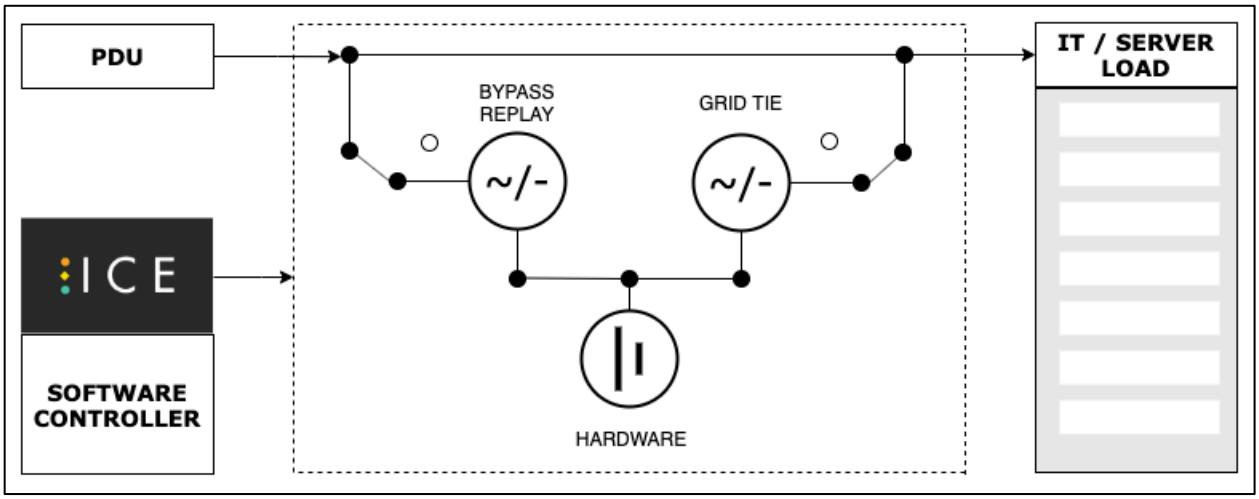

Figure 7: ICE Software and Hardware Setup

Another important observation is that the ICE software works in a scaled out distributed mode, treating the ICE compatible hardware, like batteries, as a pool of resources across the data center as well as separate units. Treating racks as a group (like a row or a room or data center), rather than controlling individual racks, enables an efficient means of power optimization. Acting as a unit, the ICE hardware allows exceeding, for short periods, the rack or branch circuit breaker capacity providing peak assurance beyond breaker limits without tripping any breaker. This group policy control is referred to as ICE Peak Shaving and acts as an application running on the controller.

4.1 ICE PEAK SHAVING. USE OF ALTERNATE SOURCE OF ENERGY

Battery based peak shaving is a powerful tool because it reacts very quickly to peaks according to objective functions or goals set up within the ICE console. This quick reaction time in peak shaving, as we will see later in this section, also enables the use of other hardware devices to address longer and broader peaks.

The figure shows two kinds of peaks: thin high peaks and thick broad peaks. For more technically oriented readers, the researchers at Penn State (Data Center cost optimization via Workload modulation) [3] characterize the peaks in terms of:

1. The peak to average ratio (PAR), which captures the peak shaving potential of the workload.
2. A function of peak width and frequency (Px), defined as the percentage of the number of time slots in which the power demand value is larger than x%.

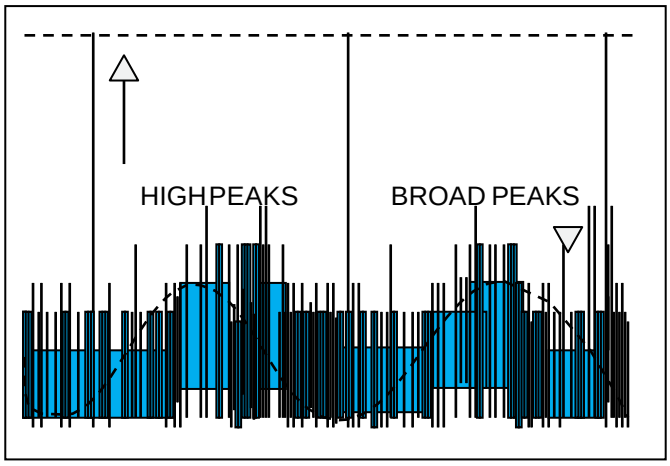

Figure 8: Characterization of Peaks

ICE compatible batteries are optimized to remove high, thin, spaced peaks characterized by high PAR values and low Px (thin spaced). The characteristics of these peaks allow batteries to recharge while peak shaving. The thick / broad peaks (low PAR and high Px) have to resort to alternate energy sources, such as Genset or higher capacity batteries, for peak shaving.

Although Genset has long been used for peak shaving, it is only useful for static time-based peak shaving, like what happens with pricing tiers. It is impractical to use Gensets for dynamic utilization-oriented peaks, due to the ad hoc nature of such peaks. Gensets take time to start and may not be able to quickly stop and re-start. ICE controls solve that problem by using Gensets in conjunction with ICE compatible batteries. ICE controlled batteries are able to shave high thin peaks and remove false positives for broad peaks on the Gensets. They pass the Peak Shaving to the Gensets only when it is sure of the occurrence of broad peaks. Using alternate sources like Gensets in conjunction with fast acting batteries further enhances the effectiveness of peak shaving by going deeper into the peaks, hence releasing further chunks of power capacity to be used for powering IT equipment.

As already discussed, there is variability between average and peak power draw. Deploying power infrastructure for peak loads leads to locked unusable capacity, for those infrequently occurring peaks. ICE works on the simple yet elegant idea of using stored energy in the form batteries to eliminate the peak draw from infrastructure and hence unlock the capacity for IT use.

The operations of the ICE block are analogous to those of a hybrid car in that the hybrid car charges its batteries when there is excess power in the car and discharges this battery power when needed. They are different in that ICE must have power available when required whereas hybrid cars can rely on gas when their batteries run out.

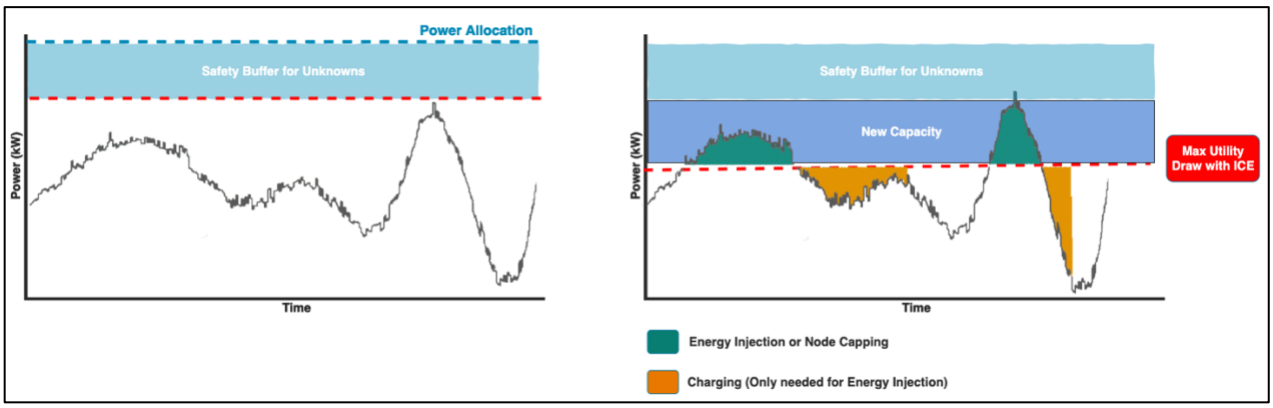

Figure 9: Concept of Peak Shaving

Figure 7 shows an example of this specific ICE Mechanism. The left image shows the power draw before ICE deployment. The right image shows the power draw post ICE deployment, and displays a drop in Utility power draw, when the ICE Peak shaving mechanism is activated. This generates an increase in capacity - enough to add another rack to this pod, without any new addition of traditional power equipment. The right-side image also shows battery draw, which is equivalent to the peaks shaved.

The battery supplements the current required for the peaks and recharges when the power curve dips. This is the fundamental concept behind how ICE solution flattens the power curve, bringing it closer to the mean. Thereby allowing deployment of more IT, within the same power footprint.

4.2 ICE DYNAMIC POWER

As stated earlier, current Tier 3/4 data centers assume that all workloads are Mission Critical all the time, which is clearly not the case. We are proposing to trade electrical redundancy on lower SLA racks to provision more Racks or IT equipment. This would create pockets of lower redundancy racks in an otherwise 2N redundant data center.

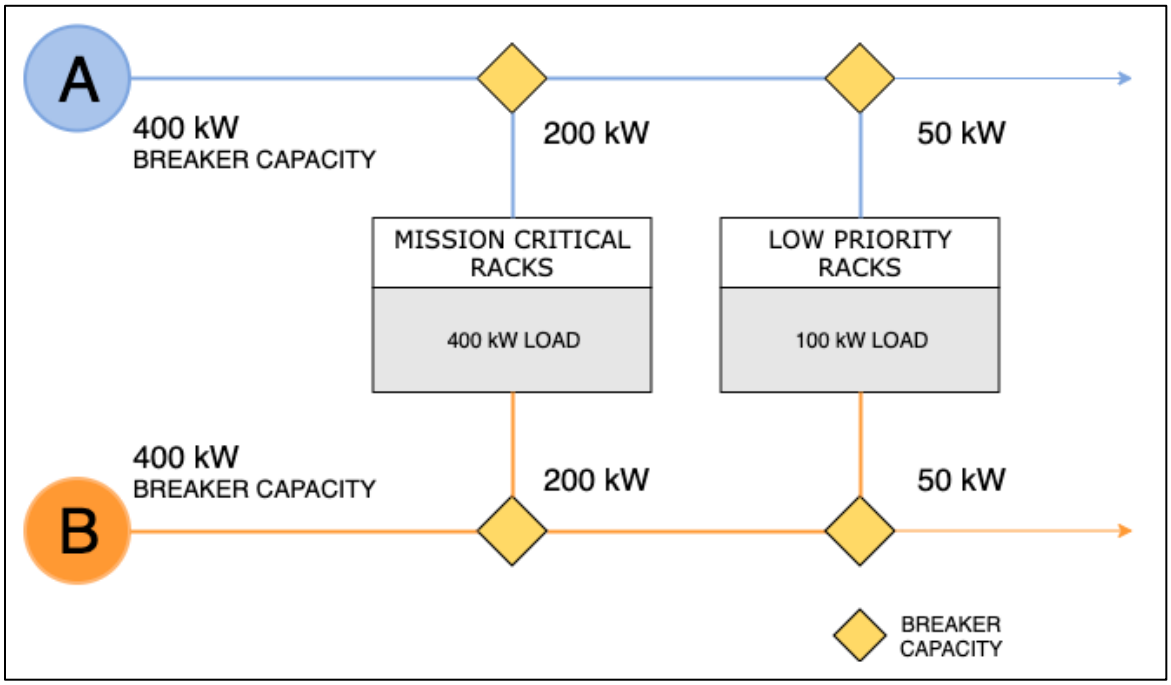

Figure 10: 2N Redundant data center can now have extra capacity for Lower Priority Racks

The figure 10 above shows a 2N redundant setup with ICE-enabled Dynamic Redundancy. We have oversubscribed the power consumptions by adding lower priority racks. In a 2N redundancy, both power feeds, A and B, feed to a 400kW peak demand (200kW on each feed in the figure). If one feed fails, the other takes over for the full 400kW peak load. In Figure 10, an additional 100kW is added as lower priority racks, making it a peak demand of 500kW (400kW Mission Critical load and added 100kW of lower SLA rack). There is no problem with this setup when both Feed A and Feed B are active. However, during a failure scenario, of either Feed A or Feed B, will leave 100kW short in traditional setup without ICE.

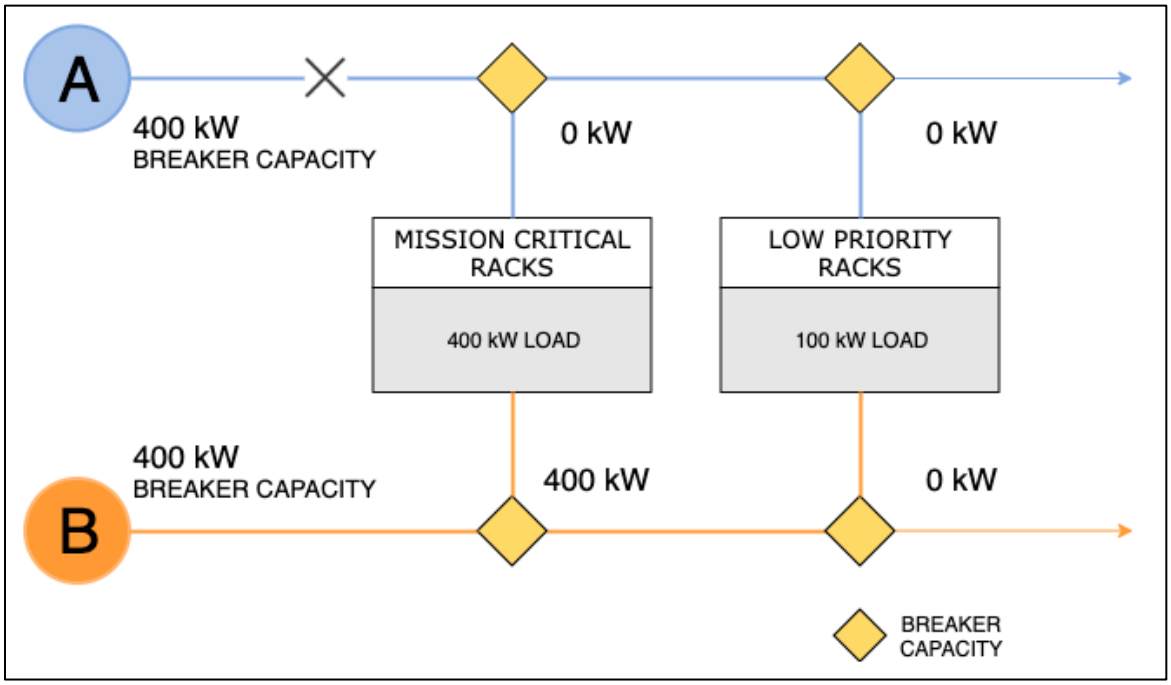

Figure 11: Remediation During Failure, Impacting Only Lower Sla Racks

Figure 11 shows the scenario when Feed A fails, putting the complete 500kW load on Feed B. ICE manages this failure scenario in two ways. Lowest priority racks are instantly turned off, using an ICE enabled switch at the racks. Keeping load on Feed B to 400kW, within the breaker capacity.

ICE batteries can also be used to sustain Low Priority Racks for a configurable amount of time, before shutting them off. This allows us time for a graceful shutdown of racks or other means, like workload migration, to transition the workload to safe racks.

As an alternative to shutting down racks, ICE is able to create policies at the granularity of a PDU outlets or servers. In that case, low priority servers can be turned off with the use of PDU outlet level controls or server BMC controls. The server BMC controls can also be leveraged to run the servers at a lower power state, with the technology offered on X86 CPUs called power capping. With power capping, the turning off of servers or racks may only be needed as means of last resort

This expansion of the data center capacity would not have been possible without ICE, as the failure options only materialize thanks to the abstraction and capacity buffer provided by ICE.

Such redundancy can be dynamically configured with the help of ICE Software, thereby changing any racks to 1N or 2N depending on the changing priority of time. Enabling redundancy to be dynamically configured by software, as opposed to statically configured through single or double corded racks creates resiliency which has never been previously available.

Not only is this feature interesting for Enterprise, it is a very profitable Use Case for CoLo providers. With Dynamic Redundancy, CoLo providers can now offer flexible service and additional low priority services, on top of fully subscribed 2N services.

5 TOTAL COST OF OWNERSHIP (TCO) CONSIDERATIONS

Projecting and measuring the TCO is key in making decisions about deploying infrastructure and becomes even more relevant when considering a transformational technology like ICE. Research shows that the typical data center is only utilized to 30% of its capacity. We have already discussed redundancy and peaks, which are the main causes of this underutilization.

The figure below shows that ICE can bring utilization all the way to 80% or higher using ICE mechanisms. This is significant given the study from Schneider, in the whitepaper on 'Determining Total Cost of Ownership for data center' [4], finding that the single largest cost driver of data center TCO to be the unabsorbed overhead cost of underutilized infrastructure.

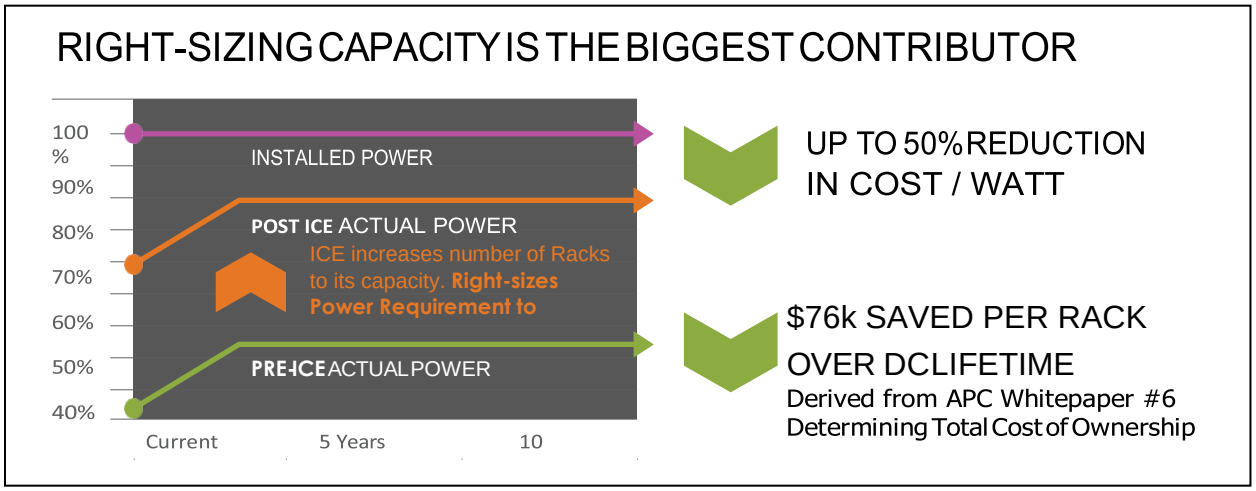

Figure 12: Power Utilization in a Rack Of A Large Colo

The most obvious TCO metric that ICE impacts is $ / Watt, where $ is measured in dollars spent on Electrical equipment. On average, data centers spend $7 on capacity for every Watt added. ICE implementation increases the Watts used in IT equipment by using Electrical capacity to what is provisioned. As such, it minimizes the $ / Watt, reducing it to as much as 50%.

Although $ / Watt is a widely used metric, it fails to impress IT folks because of its lack of comprehension of useful work done by IT equipment. Schneider's whitepaper overcomes that with the introduction of the metric relating TCO to the useful work performed by a 'rack'.

Since the only unit of work in a data center is IT equipment hosted in a rack, expressing facility infrastructure in terms of ‘Racks’ is gaining wide acceptance. The Total Cost of Ownership of a Rack in a data center is approximately $120k over the data center depreciation lifecycle (usually 10 years). The ‘Right-sized use of the systems to the actual requirement’ was found to be the biggest TCO factor, saving $76,400 per rack over 10 years. Assuming that there are 15 servers in each rack, this amounts to $2000 per server. Since the average server costs $4000, this corresponds to 50% of the server’s cost, a substantial savings delivered in the area that Schneider rightly identifies as the biggest TCO factor.

6 CONCLUSION

We hope that this paper has thoroughly described how VPS' Intelligent Control of Energy (ICE®) solutions provide a flexible Software defined solution to the problem of stranded power capacity in data centers. Below is a brief summary of its main points.

Power Infrastructure today is over provisioned, as many data centers are running out of space before running out of power. This is due to the fact that power capacity is locked and unusable due to Peaks in power consumption and Redundancy for High Availability. A significant part of such power capacity can be unlocked through Software defined Power.

ICE unlocks unused power capacity and increases power utilization levels. This allows existing data centers to favor racks with higher power density in their existing power footprint. New data centers can build smaller power footprints with existing workload plans. In this paper we reviewed two mechanism that are currently production-ready within the ICE Console.

1. By the use of Stored Energy (Battery / Genset).
2. Controlling Redundancy only as needed. Redundancy control can be at rack or server or VM (workload) level, depending on the granularity of operational control required and feasible.

Software plays a critical role in managing both of these approaches. The optimal use of power capacity, or right sizing the provisioned capacity, delivers significant reduction in Total Cost of Ownership. In essence, ICE uses a systematic approach to overcome the limitations imposed by uncertainty about the current and the future operations, likely failure scenarios, and lack of precise controls. It delivers the most cost optimized solution for the data center through a variety of control knobs.

REFERENCES

[1] Hamilton J. (2008, November). Cost of Power in Large-Scale data centers. Retrieved from https://perspectives.mvdirona.com/2008/11/cost-of-power-in-large-scale-data-centers/

[2] Xiaobo Fan, Wolf-Dietrich Weber, and Luiz Andre Barroso. 2007. Power provisioning for a warehouse-sized computer. SIGARCH Comput. Archit. News 35, 2 (June 2007), 13-23. DOI: https://doi.org/10.1145/1273440.1250665

[3] Wang, C., Urgaonkar, B., Wang, Q., Kesidis, G., & Sivasubramaniam, A. (2013). Data center power cost optimization via workload modulation. In *Proceedings - 2013 IEEE/ACM 6th International Conference on Utility and Cloud Computing, UCC 2013* (pp. 260-263). [6809409] (Proceedings - 2013 IEEE/ACM 6th International Conference on Utility and Cloud Computing, UCC 2013). IEEE Computer Society. https://doi.org/10.1109/UCC.2013.52

[4] Rasmussen N. (2019, July). Determining Total Cost of Ownership for data center and Network Room Infrastructure. Retrieved from http://www.schneider-electric.com/en/download/document/APC_CMRP-5T9PQG_EN/