

# Towards Realizing a Low Cost and Highly Available Datacenter Power Infrastructure

Sriram Govindan\*, Di Wang\*, Lydia Chen†, Anand Sivasubramaniam\*, and Bhuvan Urgaonkar\*

\*The Pennsylvania State University. †IBM Research Zurich

{sgovinda, diw5108}@cse.psu.edu, yic@zurich.ibm.com, {anand, bhuvan}@cse.psu.edu

**Abstract.** Realizing highly available datacenter power infrastructure is an extremely expensive proposition with costs more than doubling as we move from three 9's (Tier-1) to six 9's (Tier-4) of availability. Existing approaches only consider the cost/availability trade-off for a restricted set of power infrastructure configurations, relying mainly on component redundancy. A number of additional knobs such as centralized vs. distributed component placement and power-feed interconnect topology also exist, whose impact has only been studied in limited forms. In this paper, we develop detailed datacenter availability models using Continuous-time Markov Chains and Reliability Block Diagrams to quantify the cost-availability trade-off offered by these power infrastructure knobs.

## 1. INTRODUCTION AND MOTIVATION

It is now widely recognized that power consumption of datacenters is a serious and growing problem from the cost, scalability and eco-footprint viewpoints. EPA has projected the electricity cost of powering the nation's datacenters at \$7.4 billion for 2011. Over and beyond the electricity consumption, power also plays a dominant role in capital expenditures for provisioning the infrastructure to sustain the peak draw by the datacenter. For instance, provisioning the power infrastructure for a 10 MW datacenter costs around \$150 Million [3, 6] - in fact amortizing this monthly overshadows the electricity bill. A root cause for this high cost in the power infrastructure is the necessity of providing redundancy in case of any failures in order to ensure uninterrupted operation of the IT equipment (IT equipment include servers, storage and network devices). Table 1 illustrates the power infrastructure cost (shown on a per rack basis) increase as we progressively move from a basic Tier-1 datacenter

(with little/no redundancy) to a highly redundant Tier-4 datacenter, where the cost more than doubles (source: [2]). The goal of this paper is to *understand and analyze the cost ramifications of power infrastructure availability*, and use this understanding to answer the question “*can we attain the availability of a higher Tier datacenter at a substantially lower cost?*”

Before getting to the IT equipment, power flows through several infrastructural components each serving a different purpose. These components include sources (e.g. diesel generators), UPS batteries/units, and transformers. Traditional approach of replicating these expensive components to provide high availability amplifies the cost substantially. It is not clear whether this is the most cost-effective way of realizing the required availability target. Instead, it is important to be able to systematically define and analyze different power infrastructure configurations to quantify the cost-availability tradeoffs.

| Tier # | Availability | Cost/Rack |
|--------|--------------|-----------|
| Tier-1 | 0.999200     | \$18000   |
| Tier-2 | 0.999300     | \$24000   |
| Tier-3 | 0.999989     | \$30000   |
| Tier-4 | 0.999999     | \$42000   |

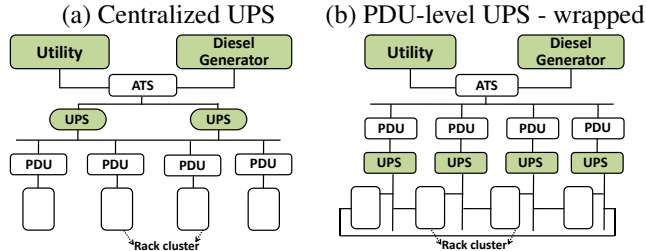
**Table 1: Datacenter power infrastructure cost more than doubles while transitioning from a low availability datacenter to a highly available datacenter.**

Apart from the conventional approach of introducing redundancy in the components, there are 3 main mechanisms/knobs for configuring the power infrastructure, each of which has an impact on the cost, complexity and resulting availability. The first consideration is the issue of where in the power infrastructure hierarchy to place each of these components. For instance, most existing power infrastructure uses centralized UPS units, while there are datacenters (such as those at Google [5]) which choose to place UPS units at each server instead. The second consideration is related to the capacity of these components - a few of high capacity or many with lower capacity? Finally, the connectivity between successive stages of the hierarchy can also have a crucial

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

*HotPower'11*, October 23, 2011, Cascais, Portugal.

Copyright 2011 ACM 978-1-4503-0981-3/11/10 ...\$5.00.



**Figure 1:** (a) Centralized UPS with 1+1 redundancy. PDUs are connected to their rack clusters using the one-one power-feed topology (b) Distributed UPS at PDU-level. UPS to Rack cluster is connected using wrapped topology.

effect on cost-availability trade-off (e.g. [9]). Further, each of these knobs is not independent, and can interact in complex ways with the others, having a further consequence on availability. *What are the power infrastructure parameters that impact availability? By how much? At what cost? Can we come up with a power infrastructure blue-print to meet the availability targets?*

It is very important to be able to quantify the availability of a datacenter leveraging these knobs to meet the requirements in a cost-effective manner. To our knowledge, there has been no prior work on developing a set of tools to comprehensively evaluate this design space quantitatively, even though there have been a few isolated studies pointing out reliability issues with specific configurations [2, 8, 9]. In this paper, we develop detailed Markov-chain and Reliability Block Diagram (RBD) based availability models to systematically evaluate the cost-availability trade-off involved in constructing a datacenter power infrastructure using these rich set of knobs.

## 2. MODELING THE AVAILABILITY OF POWER INFRASTRUCTURE

### 2.1 Datacenter Power Hierarchy

As shown in Figure 1(a), power enters the datacenter through a utility substation which serves as its primary power source. Datacenters also employ a Diesel Generator unit (DG) which acts as a secondary backup power source upon a utility failure. An Automatic Transfer Switch (ATS) is employed to automatically switch between these two sources. Upon a utility failure, it takes about 10-20 seconds (the *startup time*) for the DG to get activated, before it can supply power. Uninterrupted Power Supply (UPS) units are typically employed to bridge this time gap between utility failure and DG activation. UPS batteries typically have a runtime (reserve charge level) of about 10 minutes to power the datacenter. Datacenters typically employ *double-conversion* UPSes, which have zero transfer-time (unlike standby UPS) to batteries upon a utility failure. Since the UPS units are always involved in the double-conversion pro-

cess (even when they are not used to power the datacenter), their failure will render the whole datacenter unavailable. Power from the UPS units is fed to several Power Distribution Units (PDUs) which route power to several racks that host IT equipment. We refer to the set of racks associated with a given PDU as a *rack cluster* in Figure 1. The *power infrastructure* with all these components is often viewed as a hierarchy of different levels, e.g., in Figure 1(a), utility and DG form the top-most level, ATS forms the next level, the 2 UPS units form the third level, the 4 PDUs form the fourth level and finally, the rack clusters form the last level. Failure of one or more of these power infrastructure components may result in power unavailability to the IT equipment. A number of factors impact the availability of power infrastructure components, which we discuss below,

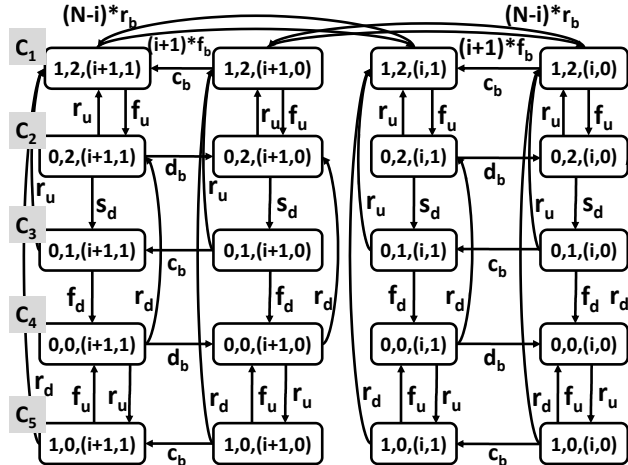
*Component Redundancy:* Redundancy in the power infrastructure components is typically incorporated to tolerate one or more failures. Let  $N$  denote the number of components that are required at a particular level of the power hierarchy to sustain the overall datacenter power load. Then  $N+M$  denotes the redundancy configuration where  $M$  component failures can be tolerated out of a total of  $N+M$  components.

*UPS Placement:* Placement of UPS units in the power hierarchy can have a significant impact on datacenter availability. Centralized UPS units are connected using a parallel bus and are placed above the PDUs as shown in Figure 1(a). In this paper, we also consider a variety of distributed UPS placements. Figure 1(b) shows a PDU-level distributed UPS placement. Similarly rack-level and server-level placements employ UPS units at each rack and each server, respectively.

*Power-feed topology:* The connectivity configuration between components placed in two consecutive levels of the power hierarchy impacts availability. In general, denser connectivity, accompanied by larger component capacity results in improved availability. In this paper, we consider four power-feed topologies, (i) *one-one* (between PDU and rack cluster in Figure 1(a)), (ii) *wrapped* (between UPS and rack cluster in Figure 1(b)), (iii) *serpentine*, and, (iv) *fully-connected*. We refer the reader to [9] for more details about these topologies.

### 2.2 Markov Availability Model

The key non-trivial aspect of power infrastructure availability modeling arises from certain idiosyncrasies of the interactions between utility, DG, and UPS. Apart from the steady-state failures associated with power infrastructure components (see Table 2), the UPS has a second form of failure which may happen when its battery becomes completely discharged. This can happen if the UPS gets completely discharged when powering IT in the following two scenarios: (i) both utility and DG



**Figure 2: Continuous-time Markov Chain captures the interaction between Utility, DG and UPS units. Shown is the transition between  $(i+1)$  active UPS units and  $i$  active UPS units. The failure and recovery rates of UPS units are presented only for the states in the top row for clarity, but those transitions exist in all the lower rows as well.**

have failed or (ii) utility has failed and DG is in the process of starting up. It can be seen that these special failure scenarios related to UPS discharge are conditional on the utility being unavailable and the DG either failing or taking too long to start up. Additionally, the amount of time the UPS can power IT available under these scenarios depends on the amount of charge in its battery, indicated by the battery runtime (in minutes). This suggests that a modeling technique to capture the impact of these interactions on overall availability should “remember” utility/DG failures and UPS charge level. Continuous-time Markov Chains (MC) fit this requirement and have been used in some existing research [2].

We consider a continuous-time MC-based model for a power infrastructure with one utility, one DG unit, and  $N$  identical UPS units  $b_1 \dots b_N$  ( $b$  for battery). We assume exponential failure and recovery processes for utility, DG, and UPS units with rates  $\{f_u, r_u\}$ ,  $\{f_d, r_d\}$ , and  $\{f_b, r_b\}$ , respectively. We also assume exponentially distributed rates  $s_d$  for DG startup,  $d_b$  for UPS battery discharging, and  $c_b$  for UPS battery charging.

The states within our MC are 3-tuples of the form  $\{u, d, b\}$ , with  $u, d, b$  representing states of the utility, DG, and UPS units, respectively.  $u \in \{0, 1\}$ : 0 means “utility failed” and 1 means “utility available and powering.”  $d \in \{0, 1, 2\}$ : 0 means “DG failed,” 1 means “DG available and powering,” and 2 means “DG available but not powering.” Finally,  $b \in \{(n, 0/1)\}$  denotes the state of the UPS units and their batteries:  $0 \leq n \leq N$  denotes the number of available UPS units, while 0 and 1 represent whether these  $n$  units are fully discharged

| Component | Reliability parameters            |
|-----------|-----------------------------------|
| Utility   | $f_u=3.89E-03, r_u=30.48$         |
| DG        | $f_d=1.03E-04, r_d=0.25$          |
| UPS       | $f_b=3.64E-05, r_b=0.12$          |
| ATS       | $f_{ats}=9.79E-06, r_{ats}=0.17$  |
| PDU       | $f_{pdu}=1.80E-06, r_{pdu}=0.016$ |

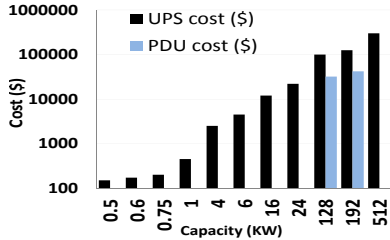
**Table 2: Failure rates ( $f_x$ ) and recovery rates ( $r_x$ ) of power infrastructure components. The rates presented indicate failure/recovery events per hour.**

or charged, respectively. For clarity, we present our Markov states only for two battery charge states, either fully charged or fully discharged, though our actual model considers discrete charge states (one state for every one minute of battery runtime). Figure 2 presents the transitions among states corresponding to  $(i+1)$  and  $i$  available UPS units ( $0 \leq i \leq N-1$ ). States in a given column all possess the same number of available UPS units and battery charge (with utility and DG states varying), while states within a given row all possess the same utility and DG states (with UPS states varying). Consequently, transitions among states within a given column capture failure/recovery of utility and DG, whereas those among states within a given row capture events pertaining to failure/recovery and charge/discharge of UPS units. The transitions between the states are self-explanatory and details are omitted for space.

We combine the availability of *Utility-DG-UPS* units obtained using the above Markov model with that of the PDU and ATS units using simple Reliability Block Diagrams (RBDs). We obtain failure and recovery rates of the power infrastructure components from the IEEE Gold-book [4] and present those in Table 2.

### 3. EVALUATION

We consider a 4MW datacenter with 32 PDUs, 256 racks and 8192 servers for our evaluation. We vary the number and capacity of UPS units depending on their placement within the power hierarchy. We also vary the capacity of UPS and PDU units depending on the over-provisioning capacity associated with the power-feed topology [9]. Only selective UPS models allow for higher capacity than 512KW - those that exist offer only 1MW and their cost numbers are not known for us to make useful comparison [1]. Consequently, we assume any larger UPS capacity to be obtained using multiple 512KW UPS units connected to the same parallel bus. Since the UPS and PDU subsystem constitutes to a significant portion of overall power infrastructure costs [2, 7], we only consider the cost of these two components for our evaluation (cost numbers obtained from APC Website are presented in Figure 3). We use the prevalent notation of representing availability as the number of leading nines - e.g., 0.999193 would simply be referred to as three 9’s of availability.



**Figure 3:** Cost of UPS and PDU units for different capacities that we explore (y-axis is in log scale). Cost per unit capacity (\$/W) is much lower for server-level (0.5 KW) UPS units compared to centralized UPSes (512 KW).

| Configuration            | Cost (Million \$) | # of 9's of availability |
|--------------------------|-------------------|--------------------------|
| "Centr N, 1-1 PDU"       | 3.42              | 2                        |
| "Centr N+1, 1-1 PDU"     | 3.72              | 2                        |
| "Centr N+1, wrapped PDU" | 4.04              | 5                        |
| "Centr N+2, wrapped PDU" | 4.34              | 6                        |
| "Centr 2N, wrapped PDU"  | 5.82              | 6                        |

**Table 3:** Cost and availability of different centralized UPS configurations. While the \$ per 9 for scaling from two 9's to five 9's is just \$100000, the incremental cost for scaling from five 9's to six 9's becomes \$300000.

### 3.1 Availability/Cost for Centralized UPS

In this section, we discuss the cost-availability trade-offs associated with different centralized UPS configurations by varying two knobs. The first knob is the number of UPS units connected to the parallel bus. The second knob is the power-feed topology (we only need to consider the topology connecting PDUs to rack clusters since the centralized UPS units are connected to PDUs via a parallel bus). Table 3 present our results. The configuration shown as "Centr. N, 1-1 PDU" represent Tier-1 datacenters [10]<sup>1</sup> with no redundancy in its UPS/PDU levels. Its UPS level consists of 8 units of 512KW each for a total of 4MW and the topology connecting PDU and rack cluster is one-one. This configuration offers only two 9's of availability, since it requires all 32 PDUs and 8 UPS units to be available.

Using our first knob of adding redundancy at the UPS level, we obtain the configuration "Centr. N+1, 1-1 PDU" with (8+1) UPS units of 512KW each. The availability is still bottlenecked by the PDU level with just two nines since it requires all 32 PDUs to be available. Our second knob helps address this. Using the wrapped topology between PDU and rack cluster (indicated as "Centr. N+1, wrapped PDU" in Table 3), the availability increases from two to five 9's. This configuration corresponds to that employed in many of today's Tier-2 data centers. Next, we consider how the availability can be improved beyond five 9's. For this, we investigate

<sup>1</sup>Note that the availability numbers we report for the different Tier configurations are specific to our datacenter size and can vary widely across different datacenter sizes.

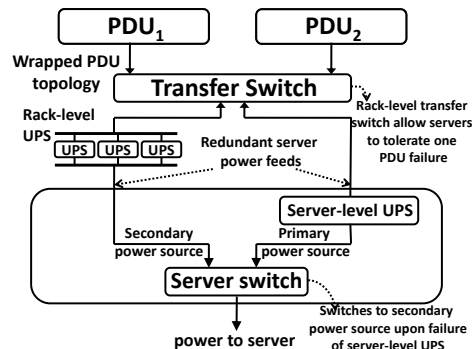
| Configuration           | Cost (Million \$) | # of 9's of availability |
|-------------------------|-------------------|--------------------------|
| "Centr. Tier-2"         | 4.04              | 5                        |
| "Dist. Server-level"    | 2.57              | 0                        |
| "Dist. 2N Server-level" | 3.8               | 3                        |
| "Dist. Rack-level"      | 4.40              | 1                        |
| "Dist. PDU-level"       | 4.50              | 2                        |
| "Hybrid"                | 2.68              | 6                        |

**Table 4:** Cost and availability for different distributed UPS placement configurations. Our hybrid scheme that employs a combination of server-level UPSes and three extra UPSes per rack achieves the availability of six 9's at 33% lower cost than centralized.

the effect of using the first knob to increase UPS redundancy to N+2 and 2N, while keeping the wrapped topology at PDU level. We find that both N+2 and 2N achieve six 9's. The table also suggests any redundancy beyond N+2 becomes unnecessary. It is interesting to note that while increasing availability from (Tier-1) two 9's to (Tier-2) five 9's incurs only a small incremental cost (\$100000 per 9 between two and five 9's), further improvements involve significant investments (\$300000 between five and six 9's). We also find that providing denser interconnect at the PDU level (serpentine and fully-connected) do not result in any further improvement in availability and therefore we assume wrapped at the PDU-level throughout this section.

**Key insights:** (i) Wrapped PDU suffices (achieves five to six 9's); (ii) N+1 is good enough for centralized UPS (five 9's at a small additional cost);

### 3.2 Availability/Cost for Distributed UPS



**Figure 4:** Illustration of our hybrid UPS placement. It has one UPS per server and a rack-level UPS module which is shown to have three UPS units connected to a parallel bus. This hybrid configuration can tolerate failure of at most three server-level UPSes within each rack.

In this section, we study availability offered by distributed UPS placements and compare it with that of centralized N+1 (Tier-2) configuration discussed above, denoted as 'Centr. Tier-2' in Table 4. We see that as we move from centralized to PDU-level to rack-level to server-level, the availability decreases. This is due

to increase in number of UPS components - 8+1 (centralized), 32 (PDU-level), 256 (rack-level), and 8192 (server-level) - with accompanying increase in probability of at least one UPS unit failing. This is evident for the configuration with 8192 UPS units (labeled "Dist. server-level") which has only zero 9's, due to a relatively high probability (about 0.9) of at least one UPS unit being down at a given time. The configuration with 2 UPSes per server (labeled "Dist. 2N server-level") increases the availability to three 9's. Table 4 shows that distributed server-level placement (even with 2N redundancy) is much cheaper (36% lower cost) than Tier-2 centralized but has poor availability. The table also shows that PDU-level and rack-level UPS placements are undesirable from both cost and availability dimensions. We also find that incorporating power-feed connectivity (wrapped and others) at the UPS level, though increase availability, does so with significant cost additions, making them less likely to be adopted in practice.

Based on the insights gained from the above analysis, we now propose "hybrid" placement schemes that combine the high availability offered by centralized UPS placement with the cheaper cost offered by server-level UPS placement. These hybrid schemes, in addition to the UPS unit at each server (as in "Dist. server-level"), add one or more UPS units per-rack with capacity same as that of server-level UPSes. We find that placing three such additional UPS units per-rack exceeds the availability of 'Centr. Tier-2' configuration (see Table 4). Our proposed hybrid configuration is shown in Figure 4. Each server has one of its dual power feeds connected to its local UPS unit, and the other connected to rack-level UPS units through a parallel bus. During normal operation, each server draws power through its local UPS. Upon a failure of its local UPS, a server starts to draw power from the rack-level UPS units. Both the server-level and rack-level UPS units are connected to both the PDUs (wrapped PDU shown) through a rack transfer switch. Other desirable levels of availability may then be obtained by varying the number of these redundant rack-level UPS units.

**Key insights:** Hybrid schemes that combine the high availability of centralized and lower cost of distributed server-level allow significantly better availability/cost trade-offs than existing configurations.

#### 4. DISCUSSION AND FUTURE WORK

We have developed a systematic Markov/RBD based availability model to evaluate the cost-availability trade-off associated with different datacenter power hierarchy configurations. We have shown that a hybrid technique combining server-level UPS with rack-level UPS units can achieve availability as high as current centralized placement at just two-thirds of its cost. There are several interesting directions for future work,

Most existing work on datacenter availability focuses solely on how likely a datacenter's power infrastructure is to support its *entire* IT equipment. But in general, different power infrastructure-induced failure scenarios can render varying fractions of the overall IT equipment unavailable. For example, while failure of centralized UPS unit may result in unavailability of the entire IT equipment, failure of distributed server/rack/PDU-level UPS units result only in partial IT equipment failure. In fact, datacenters willing to tolerate few IT equipment failures can be constructed with much lower cost. Using our availability model, we find that server-level UPS placement ("Dist. Server-level") which is much cheaper than conventional centralized placement achieve 6 nines of availability for handling 99% of the IT load, (compare it with 0 nines at 100% IT load in Table 4). Interesting workload placement strategies can be developed to leverage of such fractional IT availability, (i) PDU failures in Figure 1(b) need not necessarily result in IT unavailability since the distributed UPS units can be leveraged to migrate (live migration takes only 1-2 minutes) the workload to the active PDUs. (ii) Load-balancers can be tuned to direct client requests to active part of the power infrastructure. Capturing such effects to have a workload and migration policy aware availability model is an important future research direction.

We would also like to study in detail the feasibility of distributed and hybrid server/rack-level UPS placements, especially on raised-floor real estate and associated cooling inefficiencies as part of our future work.

#### 5. REFERENCES

- [1] APC UPS. <http://www.apc.com/products/>.
- [2] APC White paper 75: Comparing UPS System Design Configurations, 2008.
- [3] L. A. Barroso and U. Holzle. *The Datacenter as a Computer: Design of Warehouse-Scale Machines*. Morgan and Claypool Publishers, 2009.
- [4] Gold Book, IEEE Recommended Practice for the Design of Reliable Industrial and Commercial Power Systems, 1998.
- [5] Google Servers-level Batteries. [news.cnet.com/8301-1001\\_3-10209580-92.html](http://news.cnet.com/8301-1001_3-10209580-92.html).
- [6] J. Hamilton. Internet-scale Service Infrastructure Efficiency, ISCA Keynote, 2009.
- [7] Liebert White paper: Choosing The Right UPS for Small and Midsize Datacenters, 2004.
- [8] M. Marwah, P. Maciel, A. Shah, R. Sharma, T. Christian, V. Almeida, C. Araújo, E. Souza, G. Callou, B. Silva, S. Galdino, and J. Pires. Quantifying the sustainability impact of data center availability. *SIGMETRICS Perform. Eval. Rev.*, 2010.
- [9] S. Pelley, D. Meisner, P. Zandevakili, T. F. Wenisch, and J. Underwood. Power Routing: Dynamic Power Provisioning in the Data Center. In *Proceedings of the Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*, 2010.
- [10] K. G. B. W. P. Turner, J. H. Seader. Tier Classifications Define Site Infrastructure Performance, 2008.