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, power-feed interconnect topology and component capacity over-provisioning also exist, whose impact has only been studied in limited forms. In this paper, we provide a systematic approach to understand the cost/availability trade-off offered by these configuration parameters as a function of supported IT load. We develop detailed datacenter availability models using Continuous-time Markov Chains and Reliability Block Diagrams to quantify the relative impact of these parameters on availability. Using real-world component availability data to parametrize these models, we offer a number of interesting insights into developing cost-effective yet highly available power infrastructure. As two salient examples, we find (i) although centralized UPS placement offers high availability, it does so with significant cost, and (ii) distributed server-level UPS placement is much more cost-effective but does not offer meaningful availability for operating the datacenter at full load. Based on these insights, we propose a novel hybrid strategy that combines the server-level UPS placement with a rack-level UPS, achieving as good availability as existing centralized techniques, at just two-thirds of its cost.

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 [11] has projected the electricity cost of powering the nation’s datacenters at $7.4 billion for 2011, with a peak load imposed by them on the grid at 12 Gigawatts, subsuming the generation capacity of over two dozen baseload power plants. 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, 17] - 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 redundancy) to a very stringent Tier-4 datacenter that provides complete redundancy right from the utility lines, 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?”
<table>
<thead>
<tr>
<th>Tier #</th>
<th>Availability</th>
<th>Cost/Rack</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tier-1</td>
<td>0.999200</td>
<td>$18000</td>
</tr>
<tr>
<td>Tier-2</td>
<td>0.999300</td>
<td>$24000</td>
</tr>
<tr>
<td>Tier-3</td>
<td>0.999989</td>
<td>$30000</td>
</tr>
<tr>
<td>Tier-4</td>
<td>0.999999</td>
<td>$42000</td>
</tr>
</tbody>
</table>

Table 1: Datacenter power infrastructure cost more than doubles while transitioning from a low availability datacenter to a highly available datacenter (Source: [2]).

The IT equipment in the datacenter may be providing mission-critical services, and housing invaluable information, with any service disruption having serious economic, societal and security consequences. There is a considerable body of work on building robust and trustworthy IT systems in the datacenter. Equally important is the robustness of the power infrastructure in order to ensure continuous operation. The availability\(^1\) of power in a datacenter can correlate to the IT load that it can support (as we will show), making it important for computing software and hardware to be aware of these issues and adapt accordingly. Until now, there has been very little effort [2, 20, 22] at studying datacenter power infrastructure availability from the IT perspective (e.g. how do we design a power infrastructure to ensure we can support 95% of the IT load in the most cost-effective fashion, and can we place workloads accordingly?). This paper takes an important step in this direction.

Before getting to the IT equipment, power flows through several infrastructural components each serving a different purpose. These components range from sources (e.g. diesel generators), to storage reservoirs (UPS batteries/units), modulators/transformers, and distributors. Many of these are expensive components, especially for the high wattage requirements mandated by datacenters, and replicating these components to provide redundancy (for high availability) amplifies the cost substantially. To some extent, specific datacenter power hierarchies [29] have gained acceptance in different deployments. For instance, Tier-2 datacenters which are most common, have a backup diesel generator (DG), and redundant UPS units before the incoming voltage is stepped down and the power is routed through PDUs (Power Distribution Units) to the different racks. The traditional approach to enhancing its power availability is by adding redundancy in one or more of these stages, which multiplies the cost. 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 which is the most cost-effective for meeting the availability targets for a given load in the 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/feasibility and resulting availability for a given datacenter load. If one visualizes the infrastructure as a hierarchy, the first consideration is the issue of where in the hierarchy to place each of these infrastructure components. For instance, the Tier-2 configuration described above, uses centralized UPS units (of higher capacities), while there are datacenters (such as those at Google [14]) 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 will also place a crucial role on availability (e.g. [22]). Further, each of these knobs is not independent, and can interact in complex ways with each other, 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 for a specific IT load?

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, 20, 22]. This is the gap that this

\(^1\) Availability of power in a datacenter may depend on several factors including IT equipment failures [18, 27], power infrastructure failures, power spikes causing circuit-breakers to trip [12], and unavailability of cooling power. In this paper, we restrict our focus on power infrastructure failures.
paper looks to fill in the important area of datacenter power infrastructure optimization by making the following key contributions:

- We provide a systematic approach to quantify the trade-off between availability and cost involved in constructing a datacenter power infrastructure as a function of component redundancy, power-feed connection topologies and placement of these components in the hierarchy.
- We present a markov-chain based modeling technique for datacenter availability, which takes into consideration the UPS battery capacities, in addition to failure and recovery rates of different components. As the number of batteries and charge capacity increases, it can lead to state explosion. We present a hierarchical aggregation technique to reduce this state space to get a computationally tractable solution.
- Concurrently, we show how the power hierarchy can be modeled using reliability block diagrams (RBDs) based on conditional probabilities of component failures that occur for a specific topology. We demonstrate that this technique can closely approximate the results obtained from the markov-chain modeling technique, especially when battery runtimes exceed 5 minutes (which is usually the minimum capacity in most commercial UPS units). More importantly, our approach can quantify availability as a function of the IT load that needs to be sustained.
- With these tools, we show how we can model a wide spectrum of datacenter power infrastructure configurations and quantify their availability together with the associated costs. We find that the availability spans a wide spectrum, ranging from two 9’s of availability, to six 9’s, with the costs rising almost 50% in the conventional infrastructures used today. On the other hand, a Google-like approach of distributing batteries is a much more cost-effective option, but is not able to provide meaningful availability for a sustenance of 100% IT load. However, it does very well (provides six 9’s availability when the load to be sustained drops marginally to 99%). We also explore options between these extremes, using UPS units at rack and PDU granularities.
- Based on these results, we show that a hybrid strategy which combines the UPS per server extreme, with a rack-level UPS unit, provides availability commensurate to the best centralized UPS scheme at just two-thirds of its cost. These arguments hold across a wide range of datacenter sizes, component failure/recovery rates, and cost models.

2 Background

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 select/switch between these two sources. Upon an utility failure, when the ATS attempts to switch to DG power, it takes about 10-20 seconds (called as 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 power failure and DG activation. UPS units store energy (charged during power availability) using batteries which typically have a runtime (reserve charge level) of about 10 minutes to power the datacenter during this transition. Datacenters typically employ double-conversion UPSes, which have zero transfer-time (unlike standby UPS) to batteries upon a utility failure. In a double-conversion UPS, the incoming power is always converted from AC to DC (to store in batteries) and is again converted from DC to AC (to power IT equipment). Since the UPS units are always involved in this double-conversion process (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) whose sizes are typically between 100-200 KW. The PDUs have transformers that step down the voltage and route power to several racks, each of which houses IT
equipment. We refer to the set of racks associated with a given PDU as a rack cluster in Figure 1 (typically 5-10 racks are connected to each PDU). The power infrastructure consisting of 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 lower 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.

**Power Infrastructure Availability** 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, in Figure 1(a), whereas an ATS failure would render the entire IT equipment unavailable, the failure of a single PDU would only cause the associated rack cluster to be unavailable with the remaining being operational. Therefore, we find it useful to define the availability of power infrastructure as a function of the fraction of IT equipment whose power needs it can support (akin to performability [30]). While we note that “reliability” and “availability” are both metrics for high assurance system qualification, we mainly focus on the latter since it is not only important to optimize mean time between failures, but also reduce the recovery time when these failures do happen in the datacenter context. Power infrastructure availability depends on several parameters related to its design and construction that we discuss next.

**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. For example, the configuration of the UPS units in Figure 1(a) corresponds to a 1+1 redundancy, where each UPS is sized to power the entire datacenter and therefore can tolerate failure of one UPS unit.

**UPS Placement:** Centralized UPS placement denotes the configuration where all the UPS units are connected using a parallel bus and are placed above the PDUs as shown in Figure 1(a). Due to the scalability limitations associated with connecting multiple UPSes to the parallel bus (parallel bus becomes unreliable when connected to more than 8 UPS units [24, 2, 19]), large datacenters increasingly adopt distributed UPS placement (primarily above the PDUs) where the UPS units are independently connected to the underlying level. In this paper, we consider a variety of distributed UPS placement settings, at the PDU, rack and server-levels. Figures 1(b) and 1(c) both show a distributed UPS placement at the PDU level, where each PDU is directly connected to one UPS unit through which the power is routed to the rack cluster. Similarly rack-level and server-level UPS placements employ UPS units at each rack and each server, respectively. It is important to understand that depending on the UPS placement, failure of UPS units may result in unavailability to different portions of the power infras-
structure. For example, whereas failure of centralized UPS units affect the entire power infrastructure, the failure of server-level UPS units will only affect the particular server the UPS is connected to. We would like to point out that server-level UPS units have additional benefits of avoiding the inefficiency due to double-conversion (as used in Google servers [14]). In this paper, our focus is more on the availability and capital cost impact of these different UPS placement strategies.

Power-feed topology and component capacity: Power-feed topology refers to the connectivity configuration between components placed in two consecutive levels of the power hierarchy. Let \( X \) and \( Y \) denote two consecutive levels in the power hierarchy with \( Y \) directly below \( X \). The connectivity between levels \( X \) and \( Y \) dictates how failures of components in level \( X \) affect component availability in level \( Y \). For example, consider the connectivity between (PDU, rack cluster) in Figure 1(a) which is called a one-one power-feed topology since exactly one PDU is connected to each rack cluster. Here one PDU failure will result in failure of the associated rack cluster. It is possible for level \( Y \) to tolerate one or more failures in level \( X \) using a combination of power-feed connection topology between levels \( X \) and \( Y \) and the capacity of the individual components in level \( X \) (over-provisioning). For example, consider the connectivity between (UPS, rack cluster) in Figure 1(b) which is called a wrapped power-feed topology, where each rack cluster is connected to two neighboring UPS units, with the last rack cluster wrapping around to connect to the first UPS. Also shown in the figure is the capacity of the UPS units relative to that of the rack cluster capacity, \( C_{R} \). As we can see, the UPS units are over-provisioned by 50%. Such a connectivity configuration can tolerate any one UPS failure without causing unavailability to any of the rack clusters. This is because, each rack cluster is now connected to two UPS units. Hence upon failure of one of the connected UPS units, the rack cluster draws power from the other UPS unit which is possible due to the 50% capacity over-provisioning for each of the UPS units [22]. In general, the number of components in level \( X \) to which a single component in level \( Y \) is connected and the component capacity in level \( X \) determines the number of failures that can be tolerated and hence the overall availability. In this paper, apart from one-one and wrapped, we consider two other power-feed topologies that have been recently proposed [22], namely serpentine and fully-connected. Serpentine topology connects every component in level \( Y \) to 4 neighboring components in level \( X \) and fully-connected connects every component in level \( Y \) to every component in level \( X \). Figure 1(c) depicts a fully-connected topology between UPS and rack clusters. For clarity, only one rack cluster is shown in the figure. As shown in the figure, with a component over-provisioning of just 33% (compare to 50% with wrapped) this topology can tolerate any one UPS failure without compromising on availability. As we increase the number of connections between a component in level \( Y \) to components in level \( X \), the over-provisioning in level \( X \) can be reduced. It is important to note that highly connected topologies like serpentine and fully-connected have feasibility limitations related to wiring which have to be considered for realistic deployments. In this paper, we evaluate the efficacy of these power-feed topologies at different levels of the power hierarchy on overall datacenter availability and cost.

We would like to point out that the power infrastructure configuration parameters that we consider extend beyond the scope of what is defined by the datacenter Tier standards [29]. In our evaluation, we compare our infrastructure configurations with those of the Tier standards whenever applicable.

3 Modeling the Availability of Power Infrastructure

Existing work on modeling availability of the power infrastructure has employed two main techniques - Markov chains (MCs) or reliability block diagrams (RBDs). These techniques work with the mean-time-to-failure (MTTF\(_i\)) and mean-time-to-repair (MTTR\(_i\)) for each component \( i \) of the system as well as the following well-known derived quantities: failure rate \( f_i = \frac{1}{MTTF_i} \), recovery rate \( r_i = \frac{1}{MTTR_i} \), and availability \( a_i = \frac{MTTF_i}{MTTF_i + MTTR_i} \). We obtain failure and recovery rates of the power infrastructure components described in Section 2 from the IEEE Gold-book [10] and other resources [2] and present those in Table 2.
### Component Reliability Parameters

<table>
<thead>
<tr>
<th>Component</th>
<th>Reliability parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Utility</td>
<td>( f_u = 3.89 \times 10^{-3}, \quad r_u = 30.48 )</td>
</tr>
<tr>
<td>DG</td>
<td>( f_d = 1.03 \times 10^{-4}, \quad r_d = 0.25 )</td>
</tr>
<tr>
<td>UPS</td>
<td>( f_d = 3.64 \times 10^{-5}, \quad r_d = 0.12 )</td>
</tr>
<tr>
<td>ATS</td>
<td>( f_{ats} = 9.79 \times 10^{-6}, \quad r_{ats} = 0.17 )</td>
</tr>
<tr>
<td>PDU</td>
<td>( f_{pdu} = 1.80 \times 10^{-6}, \quad r_{pdu} = 0.016 )</td>
</tr>
</tbody>
</table>

Table 2: Failure rates \((f_x)\) and recovery rates \((r_x)\) of power infrastructure components (source: [10, 2]). The rates presented indicate failure/recovery events per hour.

### 3.1 Choosing A Modeling Technique

The key non-trivial aspect of our 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 when the UPS is powering IT in the following two scenarios: (i) both utility and DG have failed or (ii) utility has failed and DG is in the process of starting up. If the UPS gets completely discharged in this process before the utility or DG recover, the effect on the availability of IT is the same as if the UPS had actually failed. 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 fit this requirement and have been used in some existing research [2].

We assume that the UPS will not be used to power the datacenter upon a failure of either the ATS or the PDU. That is, the failure of these two components are assumed to cause unavailability of associated IT equipment. The reason for this assumption is that ATS and PDU recovery times are orders of magnitude higher than that for the utility (see Table 2), which implies that incorporating scenarios when UPS powers IT upon ATS/PDU failure makes negligible contribution to the estimate of overall availability. Therefore, we make the simplifying assumption that the failure processes for ATS and PDUs are independent of those for the combination utility+DG+UPS. This allows the MC to only focus on the latter. Its output can then be combined with the failure characteristics of ATS/PDUs using combinatorial techniques that apply to components with independent failures. Note that our methodology of modeling utility+DG+UPS using MC and then combining it with remaining components applies for the centralized as well as the various distributed UPS placement schemes.

#### Overview of A MC-based Model

Such a Markov chain consists of states representing different component available/failed combinations, each corresponding to a certain overall IT availability. The transition rates between these states are given by the failure and recovery rates of the different components. With such MCs, one can create states to also capture the charge within UPS units, which then allows us to remember the failure of other power sources (utility and DG) when the UPS discharges completely. Obviously, these MCs require the model developer to assume that the transition processes between these states are governed by exponential distribution. Component \( i \) failure rates typically follow a “bath-tub curve” behavior where (i) it experiences a rather high failure rate \( f_i \) that decreases with time during an initial “infant” phase, (ii) followed by a long-lasting phase with lower and time-invariant \( f_i \), and (iii) concluding with a “wear-out” phase where \( f_i \) increases sharply [26]. MC models make the reasonable assumption that all components are past their infant phase and will be replaced well before they reach the wear-out phase, i.e., they spend the majority of their operation in phase (ii) with time invariant \( f_i \), which allows us to assume exponentially distributed inter-failure durations. Thus, we find that MCs suit our availability analysis requirements. We develop a MC-based model for modeling datacenter availability and describe the techniques we adopt to deal with scalability issues that arise for our setting.
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.

### 3.2 Markov Chain-based Availability Model

We present a continuous-time MC-based model for a power infrastructure with one utility, one DG unit, and $N$ identical UPS units $b_1 \cdots b_N$ ($b$ for battery). It is easy to extend the proposed technique to incorporate multiple utility and DG units as well and we omit this due to space constraints. 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. Finally, we assume that all UPS units are operated (charging or discharging) identically, implying they possess identical charge at all times.

**States and Transitions** 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 or charged, respectively. For clarity, we present our discussion 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 events pertaining to 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 first column in Figure 2 contains those five states where $(i + 1)$ UPS units are fully charged. We label these $C_1 \cdots C_5$. $C_1 : \{1, 2, (i + 1, 1)\}$ denotes the state with the utility powering the datacenter, the DG available but not being used, and $(i + 1)$ UPS units fully charged. A utility failure at the rate $f_u$ from this state causes the datacenter to be powered from the UPS while the DG starts up, indicated by the transition from $C_1$ to $C_2 : \{0, 2, (i + 1, 1)\}$. Conversely, the transition from $C_2$ to $C_1$ represents the utility being repaired before the DG starts up which occurs at the recovery rate of utility $r_u$. $C_3 : \{0, 1, (i + 1, 1)\}$ represents the state which arises if the DG starts up and powers the datacenter before the utility recovers. The transition to this state occurs from $C_2$ at the DG startup rate $s_d$ as shown. Once in this state, a utility recovery causes a transition back to state...
the utility and DG has failed and the UPS batteries are used to power the datacenter. From $C_4$, both the utility and DG have failed and the UPS batteries are used to power the datacenter. From $C_4$, either the utility can recover causing a transition to $C_5 : \{1, 0, (i+1, 1)\}$ at the rate $f_d$; or the DG can recover causing a transition back to $C_2$ at the rate $r_d$. Finally, from $C_5$, the system can transition back to $C_1$ (when the DG recovers) at rate $r_d$ or back to $C_4$ (when the utility fails) at rate $f_u$.

Transitions along states within a row fall into four types that are related to UPS (i) charge, (ii) discharge, (iii) failure, and (iv) recovery. Charging occurs when the UPS is empty (columns two and four) and either the utility is powering the datacenter (e.g., the transitions from $\{1, 2, (i+1, 0)\}$ to $\{1, 2, (i+1, 1)\}$ and $\{1, 0, (i+1, 0)\}$ to $\{1, 0, (i+1, 1)\}$) or DG is powering the datacenter (e.g., the transition from $\{0, 1, (i+1, 0)\}$ to $\{0, 1, (i+1, 1)\}$), both transitioning to a state where the UPS becomes fully charged. Discharging occurs when the UPS is fully charged (columns one and three). This can happen in two cases, first when the utility has failed and DG is starting up and UPS in powering the datacenter (from $\{0, 2, (i+1, 1)\}$ to $\{0, 2, (i+1, 0)\}$) and second, when the utility and DG has failed and UPS is powering the datacenter (from $\{0, 0, (i+1, 1)\}$ to $\{0, 0, (i+1, 0)\}$). Failure and recovery can happen at all the states and is shown in the figure only for the states in the top row for clarity. Failure and recovery rates from a particular state with $i$ active UPS units are $i \cdot f_b$ and $(N - i) \cdot r_b$, respectively, where $N$ is the total number of UPS in the datacenter.

### Addressing State Explosion

The number of states in our model is proportional to the number of UPS units (specifically, $10 \times N$). Consequently, this model can become computationally intractable when the power infrastructure employs a large number of UPS units. For example, for distributed server-level UPS placement, where the number of UPS units scale with the number of servers, there could be hundreds to tens of thousands of UPS units. The matrix geometric method is a popular technique for reducing such state explosion in Markov chains [21]. However, it does not work for our modeling as the the repeating column structure shown in Figure 2 does not have the same transition rates for different number of active UPS units. Specifically, the transition rate from the column corresponding to $i$ UPS units to the one corresponding to $i$ UPS units is independent of $i$. E.g., the rate from $\{1, 2, (i+1, 1)\}$ to $\{1, 2, (i, 1)\}$ is $(i+1) \cdot f_b$ whereas the rate from $\{1, 2, (i, 1)\}$ to $\{1, 2, (i-1, 1)\}$ is $i \cdot f_b$ (transition not shown in the figure). We employ a hierarchical aggregation-based approach to contain the state space which consists of the following steps. First, we reduce our Markov chain into a simpler birth-death process, for which a closed-form solution can be obtained by solving balancing equations [5]. We do this by aggregating all the states corresponding to the same number $i$ of active UPS units (which comprise two neighboring columns in Figure 2) into a single “super-state” $S_i$. We are able to do such aggregation because the outgoing and incoming transition rates to all these states are identical. For examples, all the states within the first two columns in Figure 2 have $i$ active UPS units. Therefore, all of these have the same incoming transition rate of $(i+2) \cdot f_b + (N-i) \cdot r_b$ and the same outgoing rate of $(N - i - 1) \cdot r_b + (i+1) \cdot f_b$. We would aggregate all into the super-state $S_{i+1}$. Similarly, states within the last two columns in Figure 2 would be aggregated into the super-state $S_1$. Such aggregation reduces our system to the birth-death Markov chain shown in Figure 3 with the steady-state probability $\pi_i$ of being in $S_i$ given as follows:

$$
\pi_i = \left( \frac{N}{N-i} \right) \left( \frac{f_b}{r_b} \right)^{N-i} \frac{1}{\left( 1 + \frac{f_b}{r_b} \right)^N}
$$

Second, we obtain the steady-state probabilities for the 10 states within a super-state. We only consider the transitions among these 10 states, without the transitions across super states. Let $\pi^s = (\pi^s_1, \ldots, \pi^s_9)$ denote the steady state probability vector associated with these 10 states whose sum is equal to $\pi_i$. Therefore, we can solve the specific steady-state probabilities for 10 states by $\pi^s Q = 0$, $\sum_{j=1}^{10} \pi^s_j = \pi_i$, where $Q$ is a $10 \times 10$ generator matrix of MC within a super state.

**What Does MC-based Modeling Buy Us?** Recall that our main reason for choosing an MC-based approach was to be able to capture failure scenarios that arise when the UPS batteries get discharged when they are
called to power IT upon a utility+DG failure or utility failure+DG startup. Barring this complexity, one could use the simpler availability modeling technique of a reliability block diagram (RBD). An RBD works with the simplifying assumption that the failure processes for all components are independent of each other. This allows an RBD to view various components (or groups) as being arranged in “series” (e.g., utility and ATS in Figure 1(a)) or “parallel” (e.g., utility and DG in Figure 1(a)). That is, components $x$ and $y$ in series have a combined availability of $a_x \times a_y$; when in parallel, these have a combined availability of $1 - (1 - a_x) \times (1 - a_y)$. Clearly, an RBD cannot capture failure scenarios arising from the dependencies between battery discharge and utility/DG state. Discounting these failure scenarios implies that utility and DG must both fail for IT unavailability. This suggests an RBD model that considers utility and DG to be in parallel and UPS as in series with them. This RBD amounts to the following two assumptions: (i) UPS batteries always have enough charge to bridge the gap between utility failure and DG startup time. (ii) UPS will not be used to power the datacenter upon failure of both utility and DG. To quantify the impact of these assumptions on overall availability, we use the numbers in Table 2 to parametrize both our MC-based model and our RBD (for utility+DG+UPS component in the configuration of Figure 1(a)) and compare their outputs. We present the results in Table 3. The tabulated results also uses the following MC parameters: DG startup rate $s_d = 20/3600$ corresponding to 20 seconds of the DG startup time, UPS discharge rate $d_b = x/60$, where $x$ is the battery runtime in minutes, and UPS charge rate $c_b = d_b/10$ to indicate that UPS charge time is about a factor of 10 slower than discharge duration which is indicative of real UPS charge processes. We find that availability estimated by our RBD closely matches that of MC for different battery runtimes. In particular, for UPS units with more than 5 minutes of battery runtime (most datacenters have about 8-20 minutes of battery runtime), RBD matches with that of MC up to the sixth 9 in availability. In addition to this, we have also evaluated availability with a very conservative DG startup time of 1 minute and still find the RBD matching our MC-based model with a very high degree of accuracy. Given this, in the rest of this paper, we work with RBDs. Note that this does not render our MC modeling useless, since it is instrumental in allowing us to ascertain that RBD is a adequate modeling technique for our purposes. In fact, there can be power infrastructure configurations (i) with backup sources that are less reliable or slower to start up or (ii) drawing upon energy stored in UPS more aggressively thereby leaving it in lower charge states. In both such settings, an RBD might fail to match the MC-based model and the latter should be relied upon for availability estimates.

<table>
<thead>
<tr>
<th>Battery runtime</th>
<th>Markov Availability</th>
<th>RBD Availability</th>
<th>Absolute Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 minute</td>
<td>0.999999590</td>
<td>0.999999948</td>
<td>0.000000405</td>
</tr>
<tr>
<td>2 minutes</td>
<td>0.999997700</td>
<td>0.999999948</td>
<td>0.000000225</td>
</tr>
<tr>
<td>5 minutes</td>
<td>0.999999000</td>
<td>0.999999948</td>
<td>0.00000095</td>
</tr>
<tr>
<td>10 minutes</td>
<td>0.999999494</td>
<td>0.999999948</td>
<td>0.00000046</td>
</tr>
</tbody>
</table>

Table 3: Availability of the utility+DG+UPS subsystem for configuration shown in Figure 1(a). We see that RBD satisfactorily approximates Markov chains for battery runtimes above 5 minutes. We observe similar results for other power infrastructure topology/redundancy configurations.

Figure 3: Aggregation technique accumulates the 10 states that correspond to $i$ number of active UPS units in Figure 2 in to a single super state, $S_i$. This reduces the original Markov chain in to a simple birth-death process.
3.3 RBD-based Availability Modeling

Recall from Section 2 that we are interested in computing the fractional IT availability \( a_{IT}^f \) - that is, what is the probability that fraction \( f \) of the overall IT load can be powered (\( 0 \leq f \leq 1 \)). Computing \( a_{IT}^f \) involves the following two steps:

1. **Identify failure scenarios:** Enumerate all possible component failure scenarios within the power infrastructure that leaves a fraction \( f \) of IT to be available.

2. **Compute \( a_{IT}^f \):** Compute and combine the probabilities of these failure scenarios corresponding to IT load \( f \) using RBDs to obtain \( a_{IT}^f \).

We use the configuration in Figure 1(a) to illustrate the above two-step process for computing \( a_{IT}^{10} \) (100% IT load). In the first step, we identify the following six failure scenarios that leave the entire IT available (labeled \( f_{s1}, \ldots, f_{s6} \)); (\( f_{s1} \)) no failed component: \( a_{util} \cdot a_{dg} \cdot a_{ats} \cdot (a_{ups})^2 \cdot (a_{pdu})^4 \); (\( f_{s2} \)) only utility failed: \( (1 - a_{util}) \cdot a_{dg} \cdot a_{ats} \cdot (a_{ups})^2 \cdot (a_{pdu})^4 \); (\( f_{s3} \)) only DG failed: \( a_{util} \cdot (1 - a_{dg}) \cdot a_{ats} \cdot (a_{ups})^2 \cdot (a_{pdu})^4 \); (\( f_{s4} \)) only one UPS failed: \( a_{util} \cdot a_{dg} \cdot a_{ats} \cdot (2) a_{ups} \cdot (1 - a_{ups}) \cdot (a_{pdu})^4 \); (\( f_{s5} \)) utility and one UPS failed: \( (1 - a_{util}) \cdot a_{dg} \cdot a_{ats} \cdot (2) a_{ups} \cdot (1 - a_{ups}) \cdot (a_{pdu})^4 \); (\( f_{s6} \)) DG and one UPS failed: \( a_{util} \cdot (1 - a_{dg}) \cdot a_{ats} \cdot (2) a_{ups} \cdot (1 - a_{ups}) \cdot (a_{pdu})^4 \). In the second step, we compute \( a_{IT}^{10} \) by adding the probabilities associated with these six failure scenarios. Computing \( a_{IT}^{0.75} \) is similar except that it only requires 3 PDUs among the total 4 to be available. Similarly, availability for lower IT load levels can be computed.

Additionally, performing the above two steps has to incorporate the following configuration parameters: (i) placement of UPS units, (ii) number of components at each level, (iii) the power-feed topology between neighboring levels and the associated over-provisioning in component capacity. Impact of parameters (i) and (ii) on overall availability are relatively straightforward. For example, consider the configurations (a) and (b) in Figure 1 which differ both in the placement and redundancy of UPS units. In (a), failure of both the centralized UPS units would cause the entire IT (100% of IT) to become unavailable (a single UPS failure do not cause any unavailability). Whereas in (b), failure of a single UPS unit would only affect the associated racks (25% of IT). Incorporating the impact of parameter (iii), however, is more challenging and we focus on this in more detail below.

**Incorporating Topology and Component Capacity** Let us denote two neighboring levels in the power infrastructure hierarchy as \( X \) and \( Y \), with \( Y \) directly below \( X \). As discussed in Section 2, the number of level \( Y \) components affected due to failure of level \( X \) components is determined by the power-feed topology and associated component capacity over-provisioning at \( X \). Let us denote by \( C_{XY} \) the ratio of the capacity of a component at layer \( X \) to the capacity of a component at layer \( Y \). The degree of over-provisioning at layer \( X \) is then given by \( 100(C_{XY} - 1)\% \). The power feed topology connecting \( X, Y \) and \( C_{XY} \) are chosen based on certain availability guarantees for layer \( Y \) in the event of failure(s) at layer \( X \). In order to evaluate the availability, \( a_{IT}^f \) of the power infrastructure, we should be able to estimate the impact of power-feed connectivity and \( C_{XY} \) on the availability of components at level \( Y \). This turns out to be a non-trivial exercise for the power-feed topologies that we are considering.

We illustrate the difficulty by performing the following exercise for neighboring levels \( X \) and \( Y \) in the power infrastructure. We denote by \( C_{XY}^{\min} \), the minimum over-provisioning required to tolerate failure of at least one component in level \( X \). We estimate the number of failures that can be tolerated at level \( X \) for \( a_{IT}^{10} \) availability for the different power-feed connection topologies at \( C_{XY} = C_{XY}^{\min} \). (i) One-One: No failures can be tolerated irrespective of \( C_{XY} \). (ii) Wrapped: With \( C_{XY}^{\min} = 1.5 \), this configuration can tolerate any number of failures in \( X \) as long as the failures are spaced by at least two active components. Then, \( \frac{n}{n-2s} \binom{n-2s}{j} \) indicates the number of ways in which \( j \) components can fail in \( X \) from a total of \( n \) components such that no component in \( Y \) is affected. (iii) Serpentine: With \( C_{XY}^{\min} = 1.17 \), this configuration can tolerate any number of failures as long as the failures are spaced by at least six active components. Then \( \frac{n}{n-6s} \binom{n-6s}{j} \) indicates the total number...
of ways in which \( j \) components can fail from a total of \( n \) components in level \( X \) such that no component in \( Y \) is affected. (iv) Fully-connected: With \( C_{XY}^{\text{min}} = 1 + \frac{1}{n-1} \), it can exactly tolerate 1 failure in \( X \). The above expressions are derived from existing work on combinatorics that estimate possibilities for placing items in a circular arrangement such that no two items are adjacent \([16, 6]\). We use similar calculations to extend the availability estimates for the other values for \( C_{XY} \) and fractional IT load, \( f \).

4 Evaluation

4.1 Design Space

The configuration parameters that we consider for our evaluation is presented in Table 4.

<table>
<thead>
<tr>
<th>Datacenter size</th>
<th>512 KW (1024 servers) and 4 MW (8192 servers)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Power-feed topology</td>
<td>one-one/wrapped/serpentine/fully-connected</td>
</tr>
<tr>
<td>UPS placement</td>
<td>centralized, distributed (PDU, Rack and Server-level)</td>
</tr>
<tr>
<td>Redundancy</td>
<td>( N, N+1, N+2, N+3, \ldots, 2N )</td>
</tr>
<tr>
<td>Cost sensitivity</td>
<td>Current cost trend, linear model</td>
</tr>
<tr>
<td>Availability sensitivity</td>
<td>UPSes with high failure rates</td>
</tr>
</tbody>
</table>

Table 4: Our Design Space

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. For our datacenter sizes, we find that computing the availability \( a_{IT}^f \) for lower fractional IT loads (say smaller than 75%) requires enumeration of an unmanageably large number of failure scenarios. We address this concern by noting that even a slight drop in required IT load (say 90%) sharply boost availability, typically going beyond seven 9’s. Given this, we are able to restrict our analysis only to regions offering up to seven 9’s of availability (which is comparable to the availability offered by highly available datacenters \([2]\)). We assume one utility, one DG and 1+1 ATS redundancy across all our configurations. Given that the ATS is a relatively cheap component (less than $20000 for large datacenters) that has poorer availability properties compared to others, the cost/availability trade-off is rather straightforward - it is a reasonable choice to use a redundant ATS unit to eliminate the possibility of it becoming an availability bottleneck.

Cost Model

The cost of power infrastructure includes utility substations, DG, ATS, UPS, and PDUs. We only present cost numbers for the UPS and PDU subsystem, since we do not vary the configuration parameters associated with utility, DG and ATS. Also, the UPS and PDU subsystem constitutes to a significant portion of
Figure 4: Cost (Source: APC [1]) of UPS and PDU units for different capacities that we explore (y-axis is in log scale). Also shown is a linear cost model for the UPS units assuming $500 per KW of UPS capacity, which is used in Section 4.5 for sensitivity analysis due to cost.

<table>
<thead>
<tr>
<th>Configuration</th>
<th>Cost (Million $)</th>
<th># of 9’s for 100% IT</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Centr N, 1-1 PDU”</td>
<td>3.42</td>
<td>2</td>
</tr>
<tr>
<td>“Centr N+1, 1-1 PDU”</td>
<td>3.72</td>
<td>2</td>
</tr>
<tr>
<td>“Centr N+1, wrapped PDU”</td>
<td>4.04</td>
<td>5</td>
</tr>
<tr>
<td>“Centr N+2, wrapped PDU”</td>
<td>4.34</td>
<td>6</td>
</tr>
<tr>
<td>“Centr 2N, wrapped PDU”</td>
<td>5.82</td>
<td>6</td>
</tr>
</tbody>
</table>

Table 5: Cost and availability (at 100% IT load) 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.

overall power infrastructure costs [2, 19]. Since wiring costs are difficult to obtain and are known to contribute to only a small fraction of the overall power infrastructure cost [22], we ignore differences in wiring cost across different power-feed topologies. However, we do include the additional cost due to capacity over-provisioning resulting from different power-feed topology configurations. We obtain cost numbers for UPSes and PDUs of different capacities from APC [1] and present these in Figure 4. The shown price range for UPS and PDU are similar across other major vendors. An interesting observation concerns the higher $/KW for larger UPS capacities - potentially due to lower market demand for these. For example, we see that a commodity server-level UPS (less than 750W) has a much lower $/KW value compared with UPS units with higher capacity. This suggests that distributed UPS units (especially server-level) are preferable from a cost perspective. For most of our evaluation, we assume these cost numbers from APC. In Section 4.5, we evaluate the impact of a linear UPS cost model, assuming $500 per KW of UPS capacity (also shown in Figure 4) on the cost/availability trade-off.

Availability of Individual Components Apart from using component availability numbers from Table 2, we also evaluate the impact of varying UPS availability across different UPS capacities (eg., large centralized UPS with high availability vs small server-level UPS with lower availability) in Section 4.5.

4.2 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). Figure 5 shows the availability for our large datacenter for different configurations that vary these knobs. The configuration shown as “Centr. N, 1-1 PDU” employs no redundancy in its UPS/PDU levels. This corresponds to the configuration used in Tier-1³ datacenters [29]. 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 for supporting 100% IT load, since it requires all 32 PDUs and 8 UPS units to be available. The availability of this configuration increases to

³Note 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.
five 9’s for 85% load and to seven 9’s for 75% load since it can accommodate failures both at the UPS and at the PDU level while still supporting these lower load levels. Obviously, this configuration with no redundancy has the lowest cost (see Table 5) among other centralized configurations (discussed below) that employ some form of redundancy.

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. This increases the availability for 95% IT load to five 9’s compared to just two 9’s for Tier-1 (‘Centr. N, 1-1 PDU’). The availability at 100% load 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 Figure 5), the availability at 100% load 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 at 100% can be improved beyond five 9’s. For this, we investigate 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 at 100% IT load and offer seven 7’s at lower loads beyond 95%. The figure also suggests any redundancy beyond N+2 becomes unnecessary.

Next, we consider if a denser topology than wrapped help further improve availability. We evaluate the serpentine and fully-connected topologies at PDU level. Both these result in adding just one more 9 at the 100% load, and offer the same number of 9’s as wrapped for lower IT loads. This reiterates the intuition offered by the expressions in Section 3.3 that wrapped topology tolerates a high number of failures and performs as well as denser topologies. We consider this as good news since serpentine and fully-connected topologies present deployment difficulties [22]. In the remainder of our evaluation, we will only consider wrapped topology at the PDU level. Finally, we compare the UPS+PDU costs and the achievable availability at 100% IT load for these configurations in Table 5. We see 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).

Key insights: (i) Wrapped PDU suffices even for supporting higher IT load-levels (achieving five to seven 9’s for loads); (ii) N+1 is good enough for centralized UPS (five 9’s at a small additional cost); and (iii) for supporting less than 75% load, even the Tier-1 configuration (lowest cost) suffices.

4.3 UPS Placement in the Hierarchy

In this section, we study distributed UPS placements at three levels of the hierarchy: PDU-level, rack-level and server-level. We start with a one-one topology between UPS and the level below it and present results in Figure 6. We compare the availability of these distributed UPS placements with that of centralized N+1 (Tier-2) configuration discussed above, denoted as ‘Centr. Tier-2’. We see that as we move from centralized to PDU-level
Table 6: Cost and availability (at 100% IT load) 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.

Table: Cost (Million $) and # of 9’s at 100% IT

<table>
<thead>
<tr>
<th>Configuration</th>
<th>Cost (Million $)</th>
<th># of 9’s at 100% IT</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Centr. Tier-2”</td>
<td>4.04</td>
<td>5</td>
</tr>
<tr>
<td>“Dist. Server-level”</td>
<td>2.57</td>
<td>0</td>
</tr>
<tr>
<td>“Dist. 2N Server-level”</td>
<td>3.38</td>
<td>3</td>
</tr>
<tr>
<td>“Dist. Rack-level”</td>
<td>4.40</td>
<td>1</td>
</tr>
<tr>
<td>“Dist. PDU-level”</td>
<td>4.50</td>
<td>2</td>
</tr>
<tr>
<td>“Hybrid”</td>
<td>2.68</td>
<td>6</td>
</tr>
</tbody>
</table>

Figure 6: Availability of distributed UPS placements as a function of supported IT load. Dist. N Server-level, Dist. Rack-level and Dist. PDU-level has only one UPS unit at server, rack and PDU level respectively and are connected to the level below using one-one topology. Dist. 2N server-level has 2 UPS units per server. The hybrid configuration apart from having one UPS per server, has 3 extra UPS units per rack connected via a parallel bus.

to rack-level to server-level, the availability at 100% load 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. N server-level”) which has only zero 9’s at 100% load, due to a relatively high probability (about 0.9) of at least one UPS unit being down at a given time. It is important to note that “Dist. N Server-level” UPS placement offers higher availability than centralized Tier-2 for slightly lower IT load (99%). We consider improving availability offered by distributed placements at 100% load by assuming redundant UPSes per server (UPS redundancy for distributed PDU-level and rack-level are extremely expensive and we ignore those configurations). The configuration with 2 UPSes per server (labeled “Dist. 2N server-level”) increases the availability to three 9’s at 100% load. We present cost numbers for these configurations in Table 6, which shows that distributed server-level placement (even with 2N redundancy) is much cheaper (36% lower cost) than Tier-2 centralized but has poor availability at 100%. Although PDU-level and rack-level UPS placements offer availability comparable to server-level placement (in fact more at 100% load), current cost trends make these choices more expensive. We also consider varying UPS topologies for the distributed placements (as opposed to one-one discussed above) for improving availability at 100% load. We find that incorporating power-feed connectivity (wrapped and others) at the UPS level for the different placement strategies, though increase availability at 100% IT load, does so with significant cost additions, making them less likely to be adopted in practice.

Key insights: (i) For supporting 99% of the IT load, distributed at server-level offers high availability at significantly lower cost compared to centralized; (ii) Even though distributed UPS placements at PDU-level and rack-level have higher availability than server-level UPS for supporting 100% of IT load, current cost trends suggest that they would incur significantly higher deployment cost.
4.4 Hybrid UPS Placement - High Availability at Lower Cost

We now investigate ways of addressing the low availability offered by distributed UPS placement schemes at 100% load. We propose “hybrid” placement schemes that combine the high availability at 100% load offered by centralized UPS placement with the cheaper cost offered by server-level UPS placement. These hybrid schemes add one or more additional UPS units per-rack in addition to the server-level UPS units (without completely duplicating all units). We find that placing three such additional UPS units per-rack exceeds the availability of ‘Centr. Tier-2’ configuration as shown in Figure 6 for all IT load levels. These three additional UPS units each has a capacity of 500W and feeds into a parallel bus as shown in Figure 7. Each server has one of its dual power feeds connected to its local UPS unit, and the other connected to this 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 100% availability may then be obtained by varying the number of these redundant rack-level UPS units.

Based on the insights gained from above analysis, we present in Figure 8(a)-(b), the minimum cost required to achieve a particular availability target for datacenters that use centralized, distributed server-level and hybrid UPS placements (using APC cost model). We present our results for supporting both 100% IT load in Figure 8(a) and 99% IT load in Figure 8(b). In the centralized configuration, the higher availability targets are achieved by adding redundant UPSes to the parallel bus and also by varying PDU power-feed topology as discussed in Section 4.2. In distributed server-level, we vary the number of UPSes per server from one to two and also incorporate wrapped topology between the UPSes for realizing higher 9’s of availability. For hybrid configurations, we vary the number of UPS connected to the parallel bus at the rack-level from one to three for realizing improved availability and just keep one UPS per server. For distributed server-level placement, we see that for 100% load, beyond three 9’s of availability (achieved using 2 UPSes per server), it incurs significant investments (due to UPS over-provisioning involved in wrapped topology) and hence is even more expensive than the centralized scheme. Hybrid is much cheaper than the other two schemes at both load-levels since adding up to three 500W UPSes at the rack level is much more cost-effective than the redundancy techniques adopted by the centralized and distributed server-level schemes. At 99% load, just using one server per UPS achieves availability up to six 9’s and therefore both distributed and hybrid schemes have the same configuration with just one UPS per server (no rack-level UPS for the hybrid scheme) and proves to be much cheaper than the centralized schemes. We see that, irrespective of the required availability target, current cost trends suggest using our hybrid techniques offer significantly cost savings.

**Key insights:** Hybrid schemes that combine the high availability of centralized (at 100% load) and lower cost of distributed server-level can be constructed that allow significantly higher availability/cost trade-offs than
Figure 8: Cost for achieving different availability targets when supporting (a) 100% IT load, where hybrid scheme reduces the cost by more than 33% compared to centralized and distributed server-level UPS placement schemes, and (b) 99% IT load, where both hybrid and distributed server-level UPS placements realize about 36% savings in cost compared to centralized UPS placement.

4.5 Sensitivity Analysis

**Datacenter Size** We repeat our evaluation for our medium datacenter (1024 servers). Due to a smaller number of deployed UPS units, in Figure 9, we see that the centralized Tier-2 configuration for our medium datacenter achieves six 9’s at 100% load compared to only five 9’s for the large datacenter. For lower load-levels, the availability offered by distributed placements is comparable to that offered by centralized schemes and are much cheaper to deploy. Though not reported, we find that our hybrid scheme with three UPSes at the rack-level proves to be the most cost-effective solution compared to other placement schemes for achieving different availability targets (similar to our large datacenter).

![Figure 9](image)

Figure 9: Centralized Tier-2 achieves six 9’s at 100% IT load for our medium-size datacenter compared to only five for our large datacenter. Hybrid UPS placement continues to be the most cost-effective choice for medium datacenters.

**Component Availability Properties** We consider the effect of reducing the raw availability (see Table 2) by changing server-level UPS availability from three 9s used in all our analysis above to just two 9s (changing their failure/recovery rate) and retain the availability of the centralized UPS to three 9’s and present our comparison results in Figure 10 for our large datacenter. We see that even with a degraded UPS availability, the distributed server-level UPS placement still offers higher overall availability at 99% load compared to centralized placement. One interesting observation is that, it now takes 5 UPS units at the rack-level for the hybrid scheme to reach availability offered by centralized (compared to 3 units earlier).

![Figure 10](image)

**Linear Cost Model** Current cost trends have indicated the cost-effectiveness of distributed and hybrid UPS placement schemes for realizing high availability. Here, we consider a linear UPS cost model (assuming $500 per KW) and evaluate the availability/cost trade-off between our different configurations. We present the results for our large datacenter supporting 100% IT load and 99% IT load in Figure 11(a) and 11(b), respectively. We see that in spite of assuming a linear cost model, our hybrid schemes continue to be cheaper (but to a smaller
Availability (# of 9’s) vs. Load Supported (%)

- Central Tier-2
- Distributed Server-level
- Distributed Rack-level
- Distributed PDU-level
- Hybrid

Figure 10: Sensitivity to reducing component availability of server-level UPS, while retaining the same availability for centralized UPS. Distributed placements still offer higher availability at 99% IT load at a much lower cost compared to centralized placement. Because of degraded component availability, hybrid now takes 5 UPS units at the rack-level to reach the availability of centralized, compared to just 3 UPS units required earlier. Hybrid is still 30% cheaper.

Cost (Million $) vs. Availability (# of 9’s)

- Centralized
- Distributed Server-level
- Hybrid

(a) # of 9’s for 100% load
(b) # of 9’s for 99% load

Figure 11: Impact of linear UPS cost model. Cost for achieving different availability targets at both (a) 100% and (b) 99% IT loads are shown. Even with linear cost model, hybrid placement continues to be cheaper (but to a smaller extent) compared to centralized placement schemes.

5 Discussion and Future Work

In this section, we discuss possible enhancements, limitations and future directions for our work.

- **Datacenter workload availability**: Interesting workload placement strategies can be developed to leverage fractional IT availability, $a_{IT}^f$, captured by our modeling framework. Simple strategies include, (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 [9]) 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.

- **Modeling simplifications**: We have made two kinds of simplifications. First, our model ignores certain failures due to lack of information about these. Most failures (e.g., failure of the parallel bus connecting UPS units to PDUs) are straightforward to incorporate into our framework. Second, we choose simplicity when it allows us ease of presentation without sacrificing model accuracy. Examples of this include (i) ignoring ATS/PDU in our MC and (ii) synchronous UPS charge/discharge.

- **Validation**: We face two key stumbling blocks in validating our model via a simulation study. First, we have partial information about the failure/recovery processes for different components - only their average failure and recovery rates. Information describing the nature of failure inter-arrivals as well as any correlations between failures would help us conduct a better validation of our models. Second, as
is always the case in such low-probability domains, obtaining statistically meaningful results about our rare failure scenarios is difficult. We plan to explore alternative techniques (e.g., using artificially bloated failure rates) for such validation, with appropriate corresponding changes to our model (e.g., reconsidering aspects of our model that rely on low failure rates).

- **Feasibility**: Even though we discussed some of the simpler feasibility issues when presenting our power infrastructure configurations (those that related to power-feed connections, parallel bus etc...), we would like to study in detail the impact of distributed UPS placement, especially at the server-level and rack-level, on raised-floor real estate, associated cooling inefficiencies and other important feasibility issues that arise from our configurations as part of our future work.

6 Related Work

The main focus of datacenter power management has been on reducing the cost related to energy consumption [7, 23, 8]. Recent research has looked at addressing the impact of peak power on the provisioned capacity of the power infrastructure and hence the datacenter capital costs [12, 15, 22, 25]. Availability issues due to overbooking the power infrastructure (fuse constraints) have been dealt with using workload throttling [13, 12, 15] and power-load balancing techniques [22].

Relatively less attention has been directed towards addressing the cost impact of datacenter availability. A markov chain based availability model for datacenter along with cost analysis is presented in [2]. Our basic markov model is similar to the one proposed in this work which we extend to scale for thousands of UPS units (and avoid state-space explosion) using a two-level hierarchical aggregation technique. A Petri-net based model [20] for capturing datacenter availability along with cost model has been proposed recently. Monte-carlo based simulation studies have also been done to model the impact of UPS units on datacenter availability [4]. All these techniques have only looked at specific datacenter configurations, specifically those that are defined by the Tier standards. To the best of our knowledge, we are the first to comprehensively explore the power infrastructure design parameters including UPS battery placement, power-feed topology and component redundancy systematically and quantify the relative merits of these parameters towards availability/cost of datacenters.

Distributed UPS placements have been considered in datacenters for a variety of reasons: (i) PDU-level UPS placement is used in large datacenters due to scalability limitations of parallel bus required by centralized UPS placement [2, 19], (ii) Rack-level UPS placement is used in datacenter containers which are specially designed for improved scalability and cooling efficiency [28], and (iii) Server-level UPS units (such as those used in Google datacenters [14]) are used for avoiding the inefficiency due to double-conversion. We explore these placement schemes from an availability perspective and intelligently combine these to arrive at hybrid placement schemes which achieves high availability at lower cost. A variety of power-feed topologies between the PDU and the rack cluster are explored in [22], aimed towards reducing the provisioned power capacity. Our work adopts these topologies and explores their impact at different levels (PDU-UPS, UPS-Rack cluster etc.) for varying over-provisioning capacities on availability.

7 Conclusion

We have proposed a systematic approach to evaluate the relative impact of different power infrastructure configuration parameters including UPS placement (centralized and distributed at pdu/rack/server level), connection topology and component redundancy on overall datacenter availability. We have developed detailed availability models using markov chains and reliability block diagrams to quantify the availability of datacenter power infrastructure as a function of the supported IT load. Based on our cost/availability analysis, we find that (Google-like) distributed server-level UPS placement provides comparable, if not higher, availability to that of the conventional centralized UPS placement at a much cheaper cost for supporting 99% of IT load. However, server-level UPS placement offers much lower availability to support 100% IT load. Based on these observations, we propose a
hybrid UPS placement scheme which combines server-level UPS configuration with a rack-level UPS unit and achieves as good availability as existing centralized UPS (even at 100% IT load), at just two-thirds of its cost.

References


