

# Avoiding Costs From Oversizing Datacenter Infrastructure

**White Paper # 37**



## Executive Summary

The physical and power infrastructure of data centers is typically oversized by more than 100%. Statistics related to oversizing are presented. The costs associated with oversizing are quantified. The fundamental reasons why oversizing occurs are discussed. An architecture and method for avoiding oversizing is described.

# Introduction

This paper will show that the single largest avoidable cost associated with typical datacenter infrastructure is oversizing. The utilization of the physical and power infrastructure in a datacenter is typically much less than 50%. The unused capacity of a datacenter is an avoidable capital cost, and it also represents avoidable operating and maintenance costs.

This paper is constructed in three parts. First, the facts and statistics related to oversizing are described. Next, the reasons why this occurs are discussed. Finally, an architecture and method for avoiding these costs is described.

## Facts and Statistics Related to Oversizing

Anyone in the Information Technology or Facilities business has seen unused datacenter space and observed unused power capacity or other underutilized infrastructure in data centers. In order to quantify this phenomenon, it is important to define the terms used for discussion.

### Definitions related to Oversizing

For purposes of this paper, the following terms are defined as follows:

Term	Definition
Design Lifetime	The overall planned life of the datacenter. Typically 6-15 years. 10 years is the assumed typical value
Design Power Capacity	The maximum load for which the power system is designed. All or part of the equipment needed to support this load may be installed at start-up.
Ultimate Installed Power Capacity	The load capability of the power equipment ultimately installed. Equal to or less than the Design Power Capacity
Nameplate Power Capacity	The nameplate rating of the UPS system installed. May be greater than the Design Power Capacity due to planned de-rating or due to equipment redundancy
Ultimate Actual Power Requirement	The maximum load, which actually occurs during the design life of the datacenter. Less than or equal to the Design Power Capacity.
Start-up Design Power Capacity	The load capability of the power equipment installed at Start-up. Equal to or less than the Design Power Capacity
Estimated Start-Up Power Requirement	The load capability of the power equipment installed at the commissioning of the system. Less than or equal to the Ultimate Installed Power Capacity
Actual Start-Up Power Requirement	The actual Start-Up load power requirement at the commissioning of the system. Typically significantly less than the Estimated Start-Up Power Requirement.

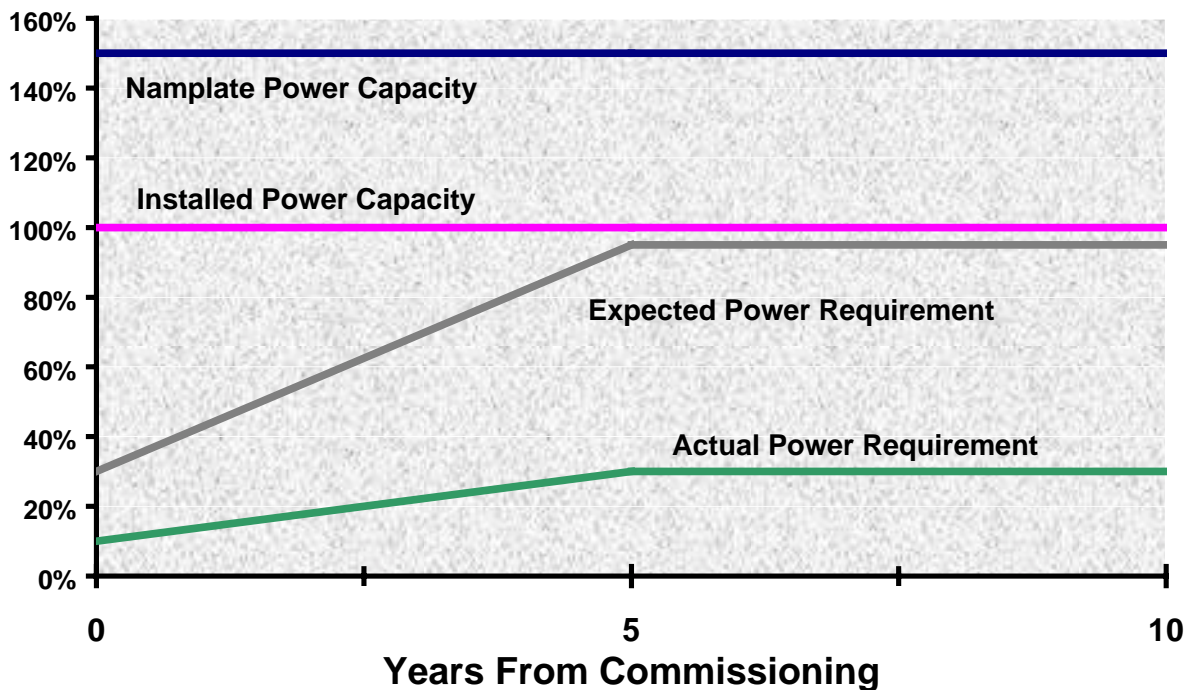
### Modeling assumptions

In order to collect and analyze data related to oversizing, APC surveyed users and developed a simplified model to describe infrastructure capacity plans for data centers. The model assumes the following:

- That the design life of a datacenter is 10 years
- That a datacenter plan has an ultimate design power capacity and an estimated Start-up power requirement
- That in the typical lifecycle of a datacenter the power requirement is planned to increase linearly from the Actual Startup Requirement and achieve the Design Power Capacity halfway through its expected lifecycle.

The model as defined above gives rise to the planning model shown in figure 1. This is assumed to be a representative model for how systems are planned.

**Figure 1: Design Power Capacity and requirement over the lifetime of a datacenter**



The figure shows a typical planning cycle. The initial power capacity installed is equal to the Ultimate Installed Power Capacity and is 100% (the system is completely built-out from the beginning). The plan is that the actual load will start at the Estimated Start-Up Power Requirement of 30% and ramp up to the Ultimate Actual Power Requirement, which is typically equal to the Design Power Capacity. However, the Actual Start-Up Power Requirement is typically lower than the Estimated Start-Up Power Requirement, and it ramps up to an Ultimate Actual Power Requirement, which is considerably less than the Design Power

Capacity. Note that the Nameplate Power Capacity may be larger than the Design Power Capacity due to redundancy or user-desired de-rating margins.

### Data from actual installations

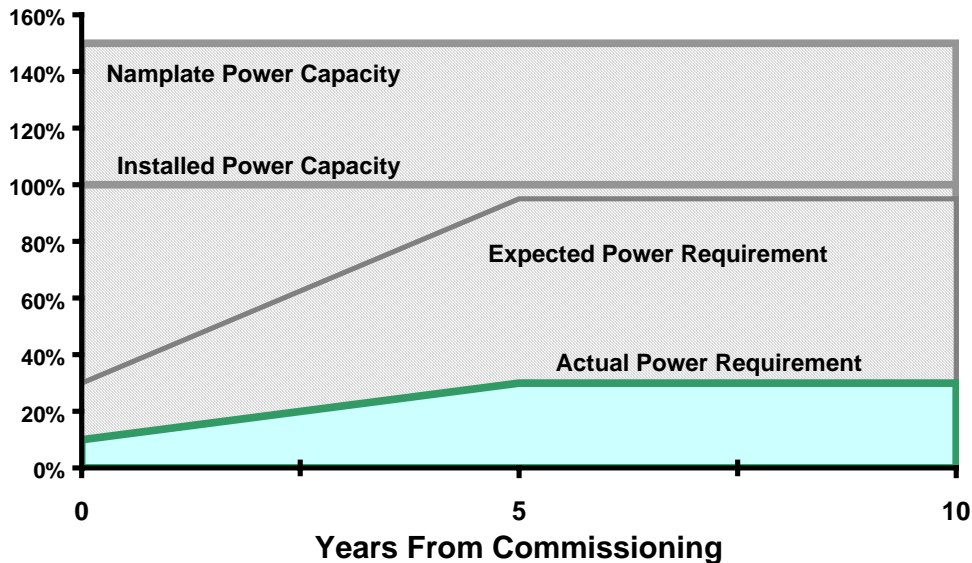
To understand the actual degree of oversizing in real installations, APC collected data from many customers. This data was obtained by a survey of actual installations and through customer interviews. It was found that the Expected initial Start-up load is typically 30% of the Design Capacity. It was further found that the Actual Start-up load is typically 30% of the Expected Start-up load, and that the Actual Ultimate Power Requirement is typically about 30% of the Design Capacity. This data is summarized in Figure 1. The average datacenter is ultimately oversized by 3 times in design value, and over 4 times in nameplate value. At commissioning, the oversizing is even more dramatic, being typically on the order of 10 times.

### Excess cost associated with oversizing

The lifecycle costs associated with oversizing can be separated into two parts: The capital costs and the operating costs.

The excess cost associated with capital is indicated in Figure 2. The shaded area in the figure represents the fraction of the system capacity, which is unutilized in an average installation.

**Figure 2: Excess cost associated with oversizing  
(Shaded area represents excess cost)**



The excess capacity translates directly to excess capital costs. In addition to the costs associated with the power system, excess capital costs include infrastructure such as raised floors, as well as cooling system infrastructure.

The typical 100kW datacenter costs on the order of \$500,000 or \$5 per kilowatt. This analysis indicates that on the order of 70% or \$350,000 of this investment is wasted. In the early years, this waste is even greater. When the time-cost of money is figured in, the typical loss due to oversizing nearly equals 100% of the entire capital cost of the datacenter! That is, the interest alone on the original capital is almost capable of paying for the actual capital requirement.

The excess lifecycle costs associated with oversizing also include the expenses of operating the facility. These costs include maintenance contracts, consumables, and electricity. Maintenance costs are typically slightly less than the capital cost over the lifetime of a datacenter, when the equipment is maintained per the manufacturers instructions. Since oversizing gives rise to underutilized equipment that must be maintained, a large fraction of the maintenance costs are wasted. In the case of the 100kW datacenter example, this wasted cost is on the order of \$250,000 over the system lifetime.

Excess electricity costs are significant when a datacenter is oversized. The idling loss of a datacenter power system is on the order of 4% of the power rating. When cooling costs are factored in, this becomes 8%. For a 100kW datacenter, oversized to typical values, with a nameplate rating in excess of the design rating as it is in a typical datacenter, the wasted electricity over the 10 year system lifetime is on the order of 600,000 kWhr, equating to on the order of \$30,000.

The total excess costs over the lifetime of the datacenter will on average be around 70% of the system cost. This represents an entitlement that could theoretically be recovered if the datacenter infrastructure could adapt and change to meet the actual requirement.

For many companies the waste of capital and expense dollars becomes a lost opportunity cost, which can be many times larger than the out-of-pocket cost. For example, Internet hosting companies have failed when the unutilized capital tied up in one installation prevented its deployment in another opportunity.

## Why does oversizing Occur?

The data indicates a very large and quite variable amount of oversizing of datacenter infrastructure occurs in real installations. Naturally, the question arises as to whether this oversizing is planned and expected, whether it is due to faulty planning, or whether there are fundamental reasons why oversizing must occur.

### Planned Oversizing

Interviews with the managers of typical installations indicate that data centers are planned to meet the maximum future expected power requirements of the load. The planned Nameplate Power Capacity of the system will typically be larger than the expected power requirement due to two factors. First, most power

systems of 100kVA or larger are designed with N+1 fault tolerance; that is, a 100kW system is comprised of 3 separate 50kW UPS modules such that if one fails the load requirement will still be met. Second, many customers have a standard practice of de-rating the power system and utilizing only a fraction, such as 80%, of the rated capacity; this is done with the idea that operating the system at less than full power will maximize overall reliability.

The practice of planning the nameplate power to be larger than the design power for a datacenter is reflected in the upper curve in figure one. This represents a planned an intentional form of oversizing. This type of oversizing is a form of underutilization although it is not the largest contributor to overall excess cost.

## Planning Process and Defects

A number of assumptions regarding future requirements are incorporated into the typical datacenter planning process. These include:

The cost of not providing sufficient capacity in the datacenter is very high and must be eliminated.

It is very costly to increase capacity partway through the datacenter lifecycle.

The work associated with increasing datacenter capacity during the lifecycle creates a large and unacceptable risk of creating downtime.

All of the engineering and planning for the ultimate datacenter capacity must be done up-front

The load requirement of the datacenter will increase but this increase is cannot be reliably predicted.

The result of these assumptions is that data centers are planned, engineered, and built out up-front to meet an unknown need, and the capacity of the datacenter is planned to be conservatively to the high side of any reasonable growth scenario.

### Fundamental reasons for Oversizing

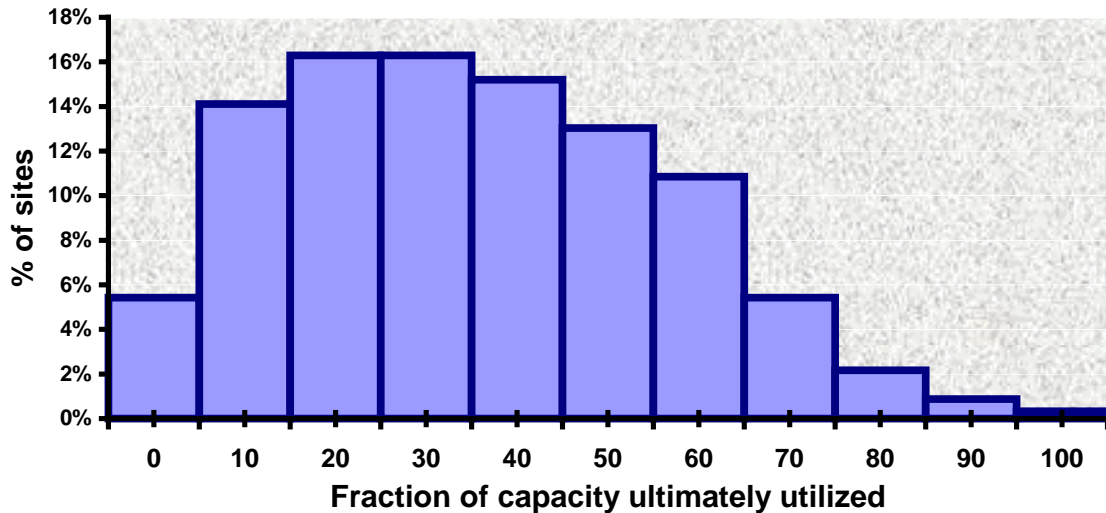
The planning process gives rise to plans, which, on average, yield a very poor utilization as demonstrated by actual results and must be judged a failure on economic terms. Yet the above examination of the planning process does not yield any fundamental defect. This apparent contradiction can be reconciled by a closer study of the data and the process constraints. Figure 3 shows the distribution of ultimate utilization fraction for actual installations, that is, the Ultimate Actual Power Requirement divided by the Ultimate Installed Power Capacity (Note: the utilization fraction would be lower if the Nameplate capacity had been used).

A study of this data provides the following insights:

- The expected value for the ultimate utilization fraction is approximately 30 %
- The expected value of surplus or unnecessary power capacity is 70%
- The ultimate utilization fraction varies considerably, suggesting on average a very poor ability to predict the future during the design process.

- If the Design Power Capacity were routinely set to the expected value, instead of the typical values chosen, then 50% of systems would not be able to meet the load requirement during their lifetime.
- The current technique for sizing is a logical tradeoff where an oversized system protects against the high degree of variability in the Ultimate Actual Power Requirement by reducing the likelihood that the system will fail to meet the load requirement during its lifetime.

**Figure 3: Ultimate utilization fraction of typical data centers**



The surprising conclusion is that given the constraints of design and the unpredictability of the future power requirements, the current method of planning data centers is logical. If the cost to the business of creating a datacenter, which fails to meet the load requirement, is high, then, given the conventional way of creating data centers, the best way to minimize the overall expected cost of the system is to oversize it substantially.

## Architecture and Method to avoid oversizing

The fundamental uncertainty of future requirements during the planning process for datacenter infrastructure is an insurmountable challenge that cannot be solved without predicting the future. Given this situation, the clear solution is to provide datacenter infrastructure responsive to the unpredictable demand.

### Barriers to adaptability

The question that naturally arises after a review of the magnitude of the oversizing problem is: Why is datacenter infrastructure built out in advance rather than built out to track the actual load requirement?

In fact, many data centers do have some phased growth designed in. For example, the deployment of equipment racks is frequently phased. The deployment of the final leg of power distribution to the

datacenter space is frequently phased. In some cases the deployment of a redundant UPS module may be phased. These approaches give rise to some savings in overall lifetime datacenter costs. However, in many cases the extra costs associated with installing this equipment later is much greater than if the equipment had been installed up-front, so that many planners choose to do a complete up-front installation. Therefore in practice only a small amount of the cost savings entitlement is obtained.

## Method and approach to creating adaptable infrastructure

The ideal situation is to provide a method and architecture that can continuously adapt to changing requirements. Such a method and architecture would have the following attributes:

- The one time engineering associated with the datacenter design would be greatly reduced or eliminated
- The datacenter power infrastructure would be provided in pre-engineered modular building blocks
- The components could be wheeled in through common doorways and passenger elevators and plugged in without the need for performing wiring operations on live circuits
- Special site preparation such as raised floors would be eliminated
- The system would be capable of operating in N, N+1, or 2N configurations without modification
- Installation work such as wiring, drilling, cutting would be eliminated
- Special permitting or regulatory procedures would not be required in order to increase capacity.
- The equipment cost of the modular power system would be the same or less than the cost of the traditional centralized system
- The maintenance cost of the modular power system would be the same or less than the cost of the traditional centralized system.

## Practical and achievable levels of adaptability

An example of an adaptable power system meeting the requirements above is the APC PowerStruXure architecture. A complete description of this system is not presented here. In the PowerStruXure architecture, over 70% of the power system can be deployed in a manner, which tracks the growth of the datacenter requirement. In practice, the only part of the power system that is completely deployed up-front is the main input switchgear and main power distribution panels, which are sized to meet the Ultimate Design Power Requirement. The UPS, Battery system, Power Distribution Units, Bypass switchgear, and Rack power distribution wiring are all deployed in a modular fashion in response to the changing load.

Note that this discussion has focused on the attributes associated with the power system, which is a primary contributor to overall datacenter infrastructure costs. The same analysis can and must be extended to comprehend the need for physical space, cooling requirements, fire protection requirements, and security requirements in order to be complete.

## Conclusions

Data centers are routinely oversized to three times their required capacity. Oversizing drives excessive capital and maintenance expenses, which are a substantial fraction of the overall lifecycle cost. Most of this excess cost can be recovered by implementing a method and architecture, which can adapt to changing requirements in a cost-effective manner while at the same time providing high availability.