# Avoiding Costs From Oversizing Data Center and Network Room Infrastructure

White Paper # 37

Legendary Reliability™

Revision 4

# Executive Summary

**The physical and power infrastructure of data centers and network rooms 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 data center and network room infrastructure is oversizing. The utilization of the physical and power infrastructure in a data center or network room is typically much less than 50%. The unused capacity of data centers and network rooms 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 data center 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 data center. Typically 6-15 years. 10 years is the assumed typical value. |
| **Room Capacity** | The maximum load the room is capable of. All or part of the power and cooling equipment needed to provide this capacity may be installed at start-up. |
| **Installed Capacity** | The load capability of the power and cooling equipment installed. Equal to or less then the Room Capacity. |
| **Expected Load** | The estimated power required at the commissioning of the system and over its lifetime. Typically changes with time and increases from time of commissioning. |
| **Actual Load** | The actual power required at the commissioning of the system and over its lifetime. Typically changes with time and increases from time of commissioning. |

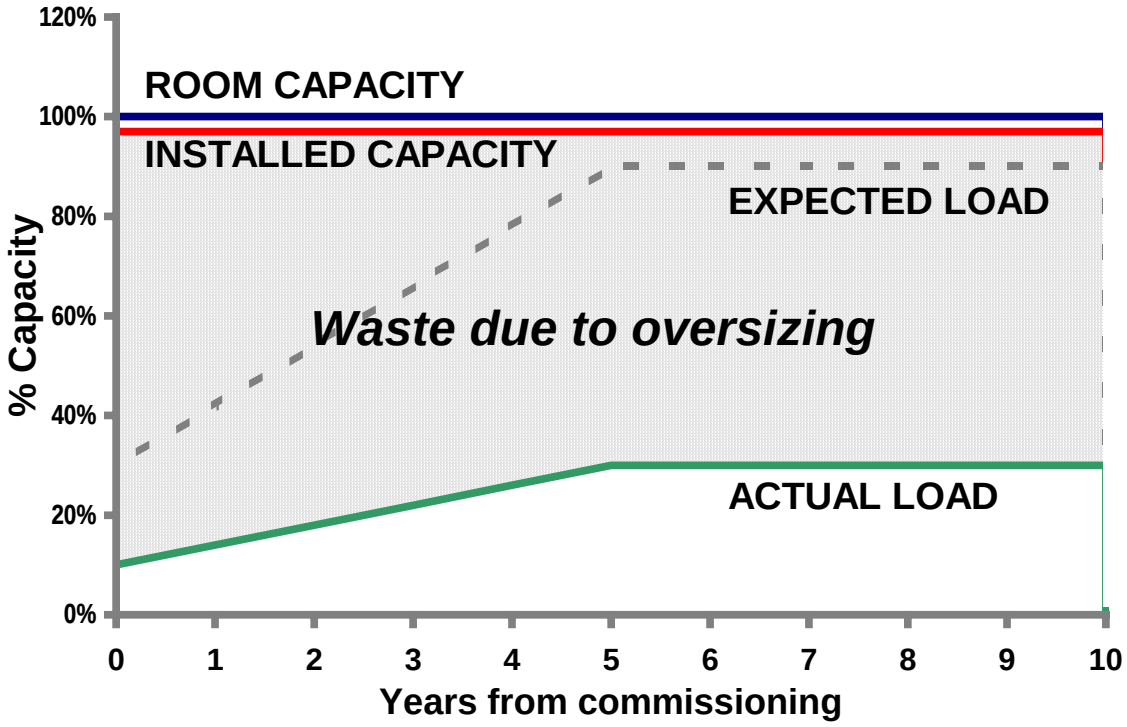
## Modeling assumptions

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 data center is 10 years
- That a data center plan has an ultimate Design Power Capacity and an Estimated Start-Up Power Requirement
- That in the typical lifecycle of a data center the Expected Load is estimated to increase linearly from an expected start-up load and achieve a final ultimate value 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 data center*




The Figure shows a typical planning cycle. The Installed Capacity of the power and cooling equipment installed is equal to the Room Capacity. In other words the system is completely built-out from the beginning. The plan is that the Expected Load of the data center or network room load will start at 30% and ramp up to a final Expected Load value. However, the Actual Start-Up Load is typically lower than the

Expected start-up Load, and it ramps up to an ultimate Actual Load, which is considerably less than the Installed Capacity (note that the Nameplate Power Capacity of the actual equipment installed may be larger than the Installed 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 start-up Expected Load is typically 30% of the ultimate Expected Load. It was further found that the start-up Actual Load is typically 30% of the start-up Expected Load, and that the ultimate Actual Load is typically about 30% of the Installed Capacity. This data is summarized in Figure 1. The average data center is ultimately oversized by 3 times in design 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 by the shaded area of Figure 1. The shaded area in the figure represents the fraction of the system capacity that is not utilized in an average installation. The excess capacity translates directly to excess capital costs. The excess capital costs include the costs of the excess power and cooling equipment, as well as capitalized design and installation costs including wiring and ductwork.

The power and cooling systems in a typical 100kW data center have a capital cost on the order of $500,000 or $5 per Watt. 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 data center! 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 data center or network room, 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 data center example, this wasted cost is on the order of $250,000 over the system lifetime.

Excess electricity costs are significant when data centers or network rooms are oversized. The idling loss of a data center or network room power system is on the order of 5% of the power rating. When cooling costs are factored in, this becomes 10%. For a 100kW data center, oversized to typical values, the wasted electricity over the 10 year system lifetime is on the order of 600,000 kWHr, equating to on the order of $55,000.

The total excess costs over the lifetime of the data center or network room will on average be around 70% of the cost of the power and cooling infrastructure. This represents an entitlement that could theoretically be recovered if the 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 data center and network room 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 estimated power requirements of the load. The Room Capacity and Installed Capacity are made slightly larger than the ultimate Expected Load. 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 making the Installed Capacity larger than the ultimate Expected Load for a data center is reflected in Figure 1. 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 data center and network room planning process. These include:

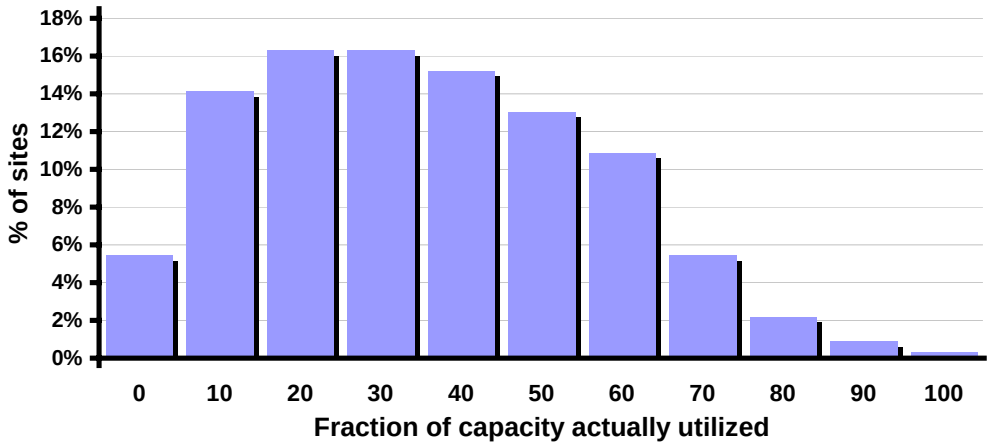
- The cost of not providing sufficient capacity in the data center or network room is very high and must be eliminated.
- It is very costly to increase capacity partway through the data center or network room lifecycle.
- The work associated with increasing data center or network room capacity during the lifecycle creates a large and unacceptable risk of creating downtime.
- All of the engineering and planning for the ultimate data center or network room capacity must be done up-front
- The load requirement of the data center or network room will increase but this increase is cannot be reliably predicted.

The result of these assumptions is that data centers or network room are planned, engineered, and built out up-front to meet an unknown need, and the capacity of the data center or network room 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 that, 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 2 shows the distribution of ultimate utilization fraction for actual installations, that is, the ultimate Actual Load divided by the ultimate Installed Capacity.

***Figure 2*** *–Ultimate utilization fraction of typical data centers*




A study of this data provides the following insights:

- The expected value for the actual utilization fraction is approximately 30 %
- The expected value of surplus or unnecessary power capacity is 70%
- The actual utilization fraction varies considerably, suggesting on average a very poor ability to predict the future during the design process.
- If the Installed Capacity were routinely set to the expected value of 30%, 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 Load by reducing the likelihood that the system will fail to meet the load requirement during its lifetime.

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 and network rooms is logical. If the cost to the business of creating a data center or network room that fails to meet the load requirement is high, then, given the conventional way of creating data centers and network rooms, 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 data center and network room infrastructure is an insurmountable challenge that cannot be solved without predicting the future. Given this situation, the clear solution is to provide data center and network room 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 data center and network room 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 data center 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 data center 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:

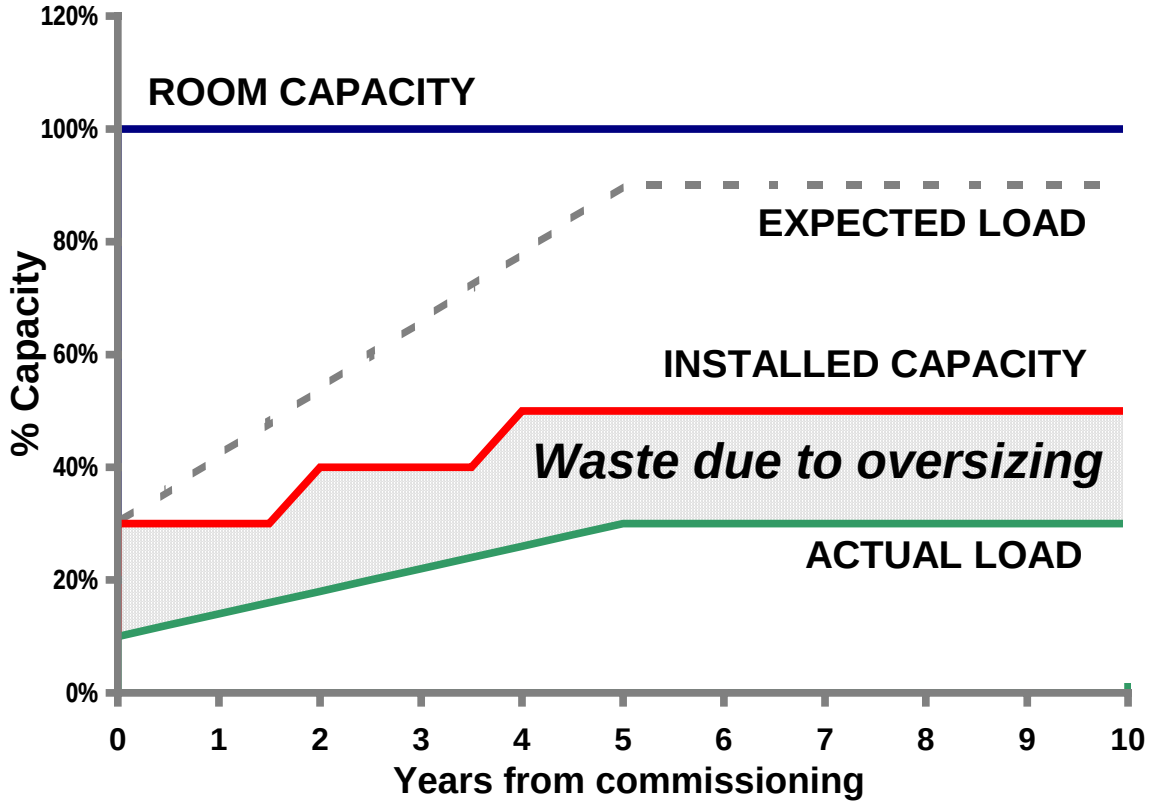
- The one time engineering associated with the data center and network room design would be greatly reduced or eliminated
- The data center or network room power infrastructure would be provided in pre-engineered modular building blocks
- The components could be wheeled in though common doorways and 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

When an adaptable system for physical infrastructure is deployed, the waste due to oversizing shown as the shaded area of prior Figure 1 can be reduced substantially. This savings is shown in Figure 3 below.

***Figure 3*** *– Design Power Capacity and requirement over the lifetime of a data center*




Note that the Installed Capacity is not built out to the Room Capacity at startup and that the Installed Capacity changes to track the Actual Load. This figure should be contrasted with the scenario described by prior Figure 1.

An example of an adaptable system meeting the requirements above is the APC InfraStruXure architecture. A complete description of this system is not presented here. In the InfraStruXure architecture, over 70% of the power system can be deployed in a manner that tracks the growth of the data center or network room

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 Room Capacity. 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 and cooling systems, which are primary contributor to overall data center and network room infrastructure costs. The same analysis can and must be extended to comprehend the need for physical space, fire protection requirements, and security requirements in order to be complete.

# Conclusions

Data centers and network rooms 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 that can adapt to changing requirements in a cost-effective manner while at the same time providing high availability.

# References

Mitchell-Jackson, J.D., Koomey, J.G., Nordman, B., Blazek, M., "Data Center Power Requirements: Measurements From Silcon Valley", May 16, 2001. Master's Thesis, Energy and Resources Group, University of California. Berkeley, California. Available at http://enduse.lbl.gov/Projects/InfoTech.htm