Five Reasons your Data Center's **Availability**, **Capacity** and **Efficiency** are being **Compromised**

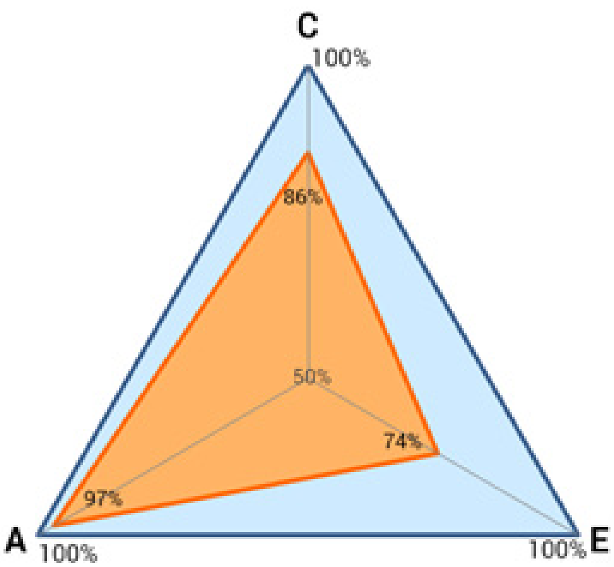
Future Facilities' White Paper

FFL-001 Revision 1.2

AKHIL DOCCA engineering & product manager, North America

DAVE KING senior data center engineer, EMEA

STEVE DAVIES product marketing manager



Executive Summary

Data center owner-operators are increasingly looking for solutions to minimize total cost of ownership, cost per kW of IT load, and downtime. This paper explains the five main contributors to runaway data center costs, then introduces the **ACE performance score** and the **continuous modeling** process. Using both, this paper briefly explains how they are helping owner-operators save millions of dollars annually per data hall.

future facilities

Introduction

Could '**minimize**' be the verb that best sums up a data center owner-operator's ultimate objective?

"The **ACE performance score** *considers the three interconnected variables that ultimately dictate how costly a DC is to run"*

Think about it, whatever business you're in, and whichever type of data center(s) you own, you almost certainly want to minimize one or more of the following:

- **Cost overruns**
- **TCO (total cost of ownership)**
- **Cost per kilowatt ($/kW) of IT load**
- **Downtime**

In an industry where the average TCO overspend is around $27m per MW, where $/kW can spiral out of control within just a few short years of entering operation, and where the average cost of downtime is $627k per incident, owner-operators want solutions.

Poor planning and inefficient use of power, cooling and space represents a significant threat to your efforts to minimize costs. Yet it is precisely this that so often forces you into a corner – build a new facility to take the strain, or invest in a major overhaul. Neither 'solution' is attractive, so why are owner-operators so frequently in a position where their aspirations are never realized?

In this paper, we set out to not only answer that question, but to also offer a solution going forward.

First, we identify the five major contributors to increased costs and downtime. Then we propose that the greatest opportunity to minimize these can be achieved by adopting a simple, inexpensive solution: the **ACE performance score**.

The ACE performance score is a unique way of assessing and visualizing the three critical indicators of data center performance, as described below. It works by mapping data from DCIM toolsets into a powerful 3D virtual facility model. With that automated process accomplished, it simulates the resulting distribution of airflow and temperature in the space. This confluence of predictive modeling and DCIM data is called **Predictive Modeling for DCIM**.

The ACE performance score can be used from inception through operation, and it considers the dynamic interrelationship of the three variables - ACE - that ultimately dictate how well a data center performs and, by extension, how costly it is to run:

- **Availability** (A) of IT, including during power and cooling failures

- How much **capacity** (C) is available to install, power and cool additional IT
- How **efficient** (E) the cooling delivery is to the IT

With the ACE performance score introduced and explained, we conclude by introducing a simple business process through which ACE can be easily applied: **continuous modeling**.

A follow-up paper, *From Compromised to Optimized: An ACE Performance Assessment Case Study*, provides clear real world examples of how the ACE performance score and continuous modeling are being used today by global owner-operators to annually save millions of dollars per data hall.

Future Facilities White Paper Link

From Compromised to Optimized: An ACE Performance Assessment Case Study

The Data Center Challenge

Of the many challenges an owner-operator faces in managing a data center, constraining TCO, driving down $/kW and minimizing downtime are key.

The challenge is illustrated in our example of a facility with a 1MW (megawatt) IT load. Let's single out the estimated TCO to make the point. For such a facility, TCO should be $32m over 15 years... the reality, however, is completely different: a skyrocketed $59m (**Figure 1**).

"...most data centers will realize a capacity utilization of only 70%"

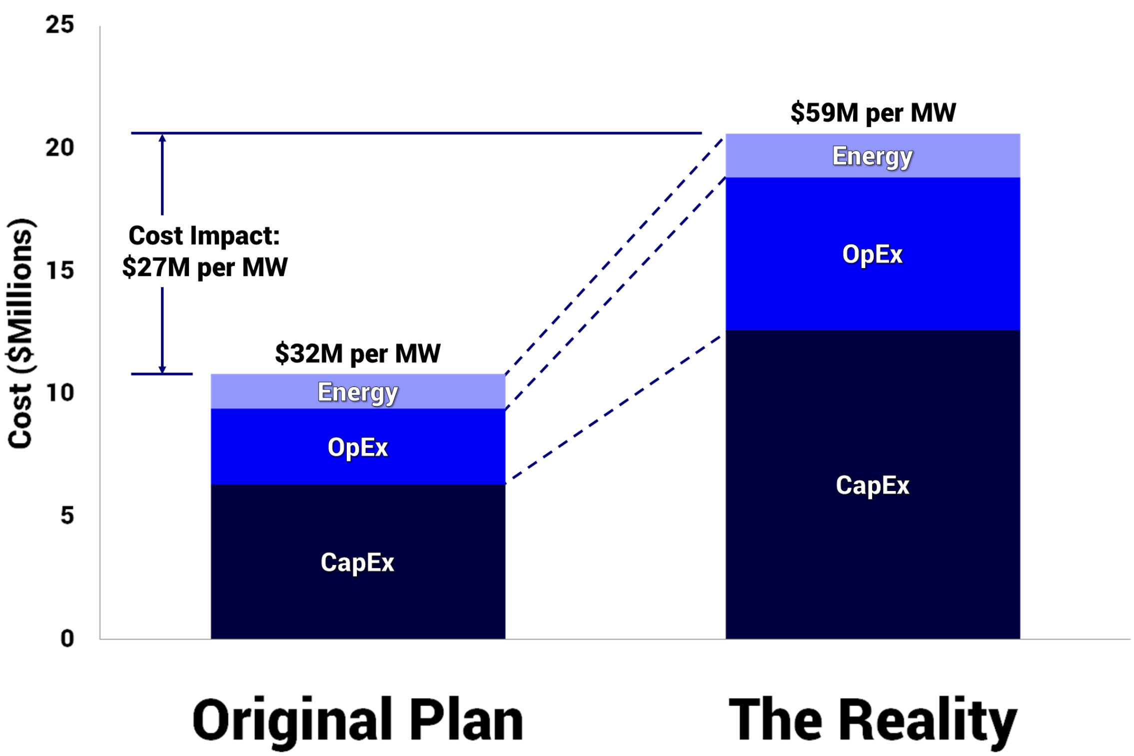


Figure 1

The high-level breakdown of a typical Tier III data center's costs. The expected TCO never matches the actual TCO. The reality is that energy, operational and capital costs all rise during the life of the data center. In addition, $/kW will also rise over time.

So, what happened to make the costs almost double in **Figure 1**? In short, there was a discrepancy between the physical capacity that you thought you were getting, and what you actually got.

It's clear then that data centers have the potential to be financial black holes. But why is this? The answer is simple: because of poor availability, capacity and/or efficiency. There are five main causes of this...

1. Designers and the Design Chain

When you, the owner-operator, put your DC designs out to tender (RFI/RFP), you are unwittingly creating an environment where a single product (the facility) is being supplied by multiple independent vendors.

In the vast majority of cases, these supply chain vendors do not talk to each other, and that often leads to problems when the data center is built and handed over. This will cost you physical capacity, eventual downtime and long-term cooling inefficiency.

2. Aspirational vs. Actual: Design never matches IT Operations

There is a yawning gap between aspirational design requirements and the reality of the built facility. You are paying for one thing, but you are actually getting a lot less.

The overall data center budget (CapEx and OpEx) is financed based on the ability to fill the facility with IT to 100 percent of the design load. According to industry experts Gartner and 451 Group, this figure is simply not realistic - the reasons for this are made clear in this paper. Indeed, these two independent bodies agree that most data centers will realize a 'capacity utilization' of only 70%. Put another way, most, if not all, modern mission critical data centers suffer from low capacity utilization.

Techspeak

capacity utilization

the use of as much physical capacity in a data center as possible - this is capacity that an owner-operator has already paid for. Failing to use this capacity will incur future capital expenditure

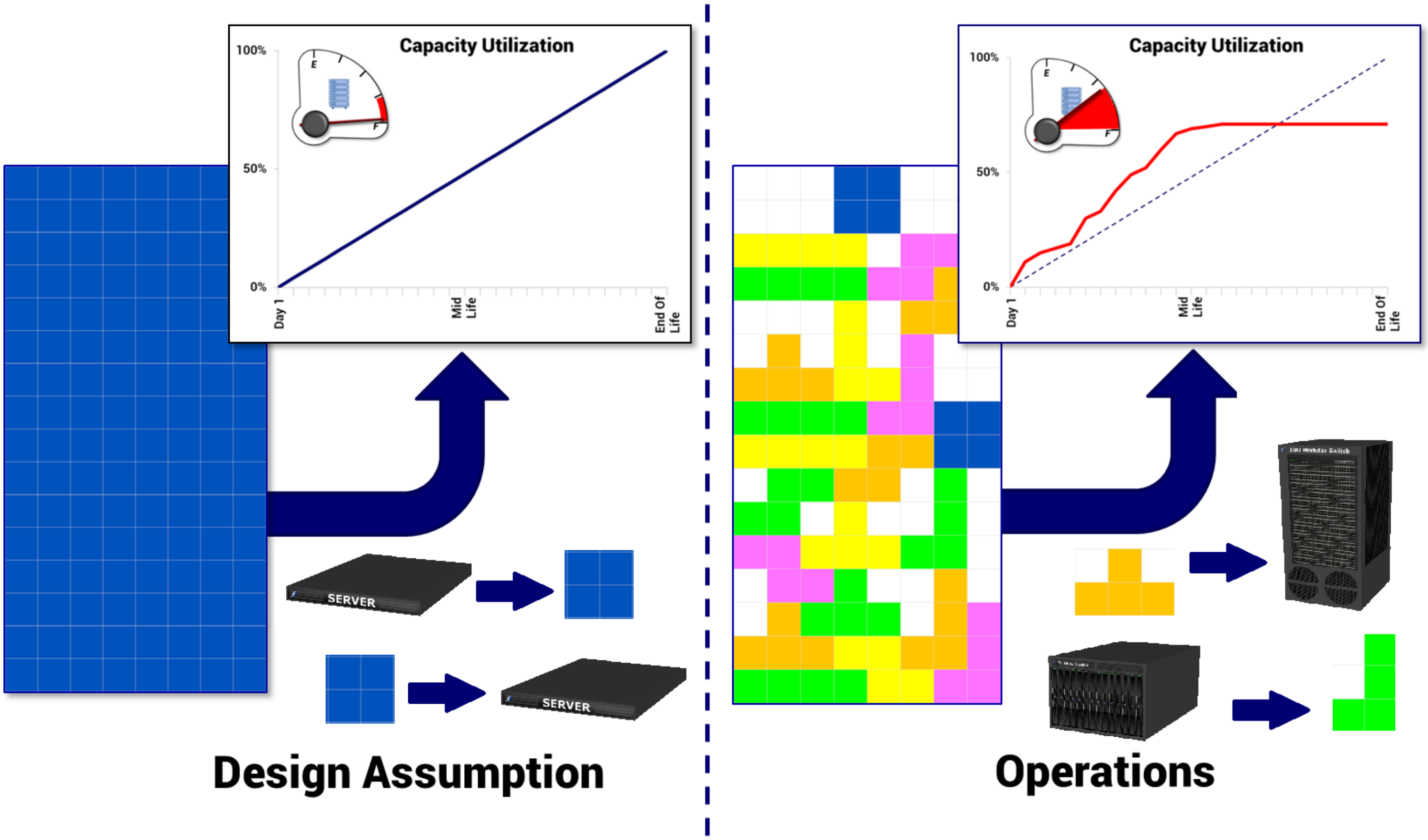
Low capacity utilization increases the costs associated with utilization of floor space, infrastructure, operation and energy for a given IT load. Combined, this dramatically drives up the TCO and cost/kW, as shown in **Figure 1**.

The root cause of the gap is this: when a data center is specified, very high level parameters are used (total IT kW or IT kW per rack, for example). Based on these parameters, design consultancies produce a sensible design fulfilling requirements for availability, capacity and efficiency. However, the assumptions under which the designers have operated are detached from the actual operation of a data center.

The reality is this: the IT build out during operations never matches the original design assumptions; it will change over time as the needs of the business change. In short, there is a disparity between the aspirational design of the facility and its actual operational performance.

To understand exactly why capacity utilization is so low, the Tetris® game provides us with an excellent analogy. **Figure 2 (p.6)** represents the design stage of the data center life cycle: all the blocks (representing IT equipment) are known well in advance, which makes the game predictable! However, **Figure 3 (p.6)** reflects the reality of an operational data center: the blocks are not only different from those used in design, but also arrive with very little time to spare. The time pressure to place them on the board causes fragmentation of space and capacity.

It is not possible for conceptual design to guarantee performance in normal operation due to changing IT and the tactical build out of the facility over time.



3. Siloed Operations

A data center is a complex, multi-layered system that serves the needs of multiple stakeholders with mutually exclusive vested interests. IT operations, corporate real estate, facilities engineering etc. all plan and execute actions in their respective silos that have a profound effect on performance of the facility.

Such silo-based operations lead to fragmented operational processes, which in turn leads to fragmentation of physical capacity, as the Tetris® analogy demonstrates. Here are three examples:

1. The procurement teams that dictate to the IT teams what equipment they can buy are not concerned with details such as cooling airflow direction. As a result, *both* IT and facilities teams are left with IT that does not really 'fit' into the DC

2. Energy saving measures are initiated by the facility teams. This often results in cooling problems that can have a profound effect on availability and capacity

3. With the advent of virtualization, IT teams can run more applications on a single piece of hardware. Additionally, they are also able to move applications where needed, based on utilization rates. This creates varying heat loads in the facility due to server utilization. The facilities team must effectively respond to these equipment load changes.

Figure 2 (above left)

A Tetris® board as you'd like to see it: the shapes (representing IT) are all known in advance, making them simple to place!

Figure 3 (above right)

A Tetris® board as you'll likely see it. The blocks are differently shaped, signifying different IT configurations. Most critical, though, is that there is very little time to place them. Time pressures in the real data center force placement errors, and these directly contribute to the loss of capacity.

4. Lack of Capacity Tracking

Your physical capacity is dictated by the resource that is least available – space, power, cooling or networking – and is tied to the connected IT in the data center. However, there exists a major misconception that tracking power is the same as, or equates to, knowing how much physical capacity you have left – it is not.

"...there exists a major misconception that tracking power equates to knowing how much physical capacity you have left – it does not"

Data Center Infrastructure Management (DCIM) tools, which perform multiple functions that include asset tracking and workflow management, provide a powerful means to monitor and track space and power. As a result, owner-operators like you are investing in them heavily. However, in doing so, you are at risk of being lulled into a false sense of security.

Resources fragment due to IT deployments that deviate drastically from the design intent, resulting in an asynchronous use of resources, as seen in the Tetris® analogy. For example, when cooling is utilized faster than space and power, the data center reaches the end of its life far quicker than anticipated - the least available resource (cooling) is no longer available.

To summarize, DCIM cannot:

- Model and track cooling availability
- Relate the distributions of space, power, cooling and IT to each other to show capacity
- Predict the impact of future IT plans on power and cooling collectively.

5. Variable IT within a Fixed Infrastructure

The IT hardware refresh that comes every few months or years is another variable that compromises your attempt to minimize TCO. It is unrealistic to expect a fixed infrastructure to adapt to ever-changing IT hardware and software requirements.

Such refreshes are the result of growing business needs that demand performance and availability to both internal and external customers, and for as long as possible. Newer IT hardware can have completely different requirements for space, power and cooling resources that could never have been accounted for by the original design. The effects of these equipment changes are only visible much later in operations, at which point hotspots occur.

Future Facilities White Paper Link

The Elephant in the Room is Lost Capacity

An example of such disruption is that most, if not all, designs assume 'front-to-back' IT equipment airflow patterns (see **Figures 4** and **5**, **p.8**). However, it's not uncommon to see IT equipment developed that actually requires more (or less) cooling than forecast. Similar to cooling, IT hardware may often require more than two power cords, which is again a deviation from design intent, but both "break" the guidelines that were in place when the data center was designed, requiring an operational remedy.

In summary, the unpredictability of IT equipment, coupled with the requirement to fix problems immediately, exacerbates the $/kW problem.

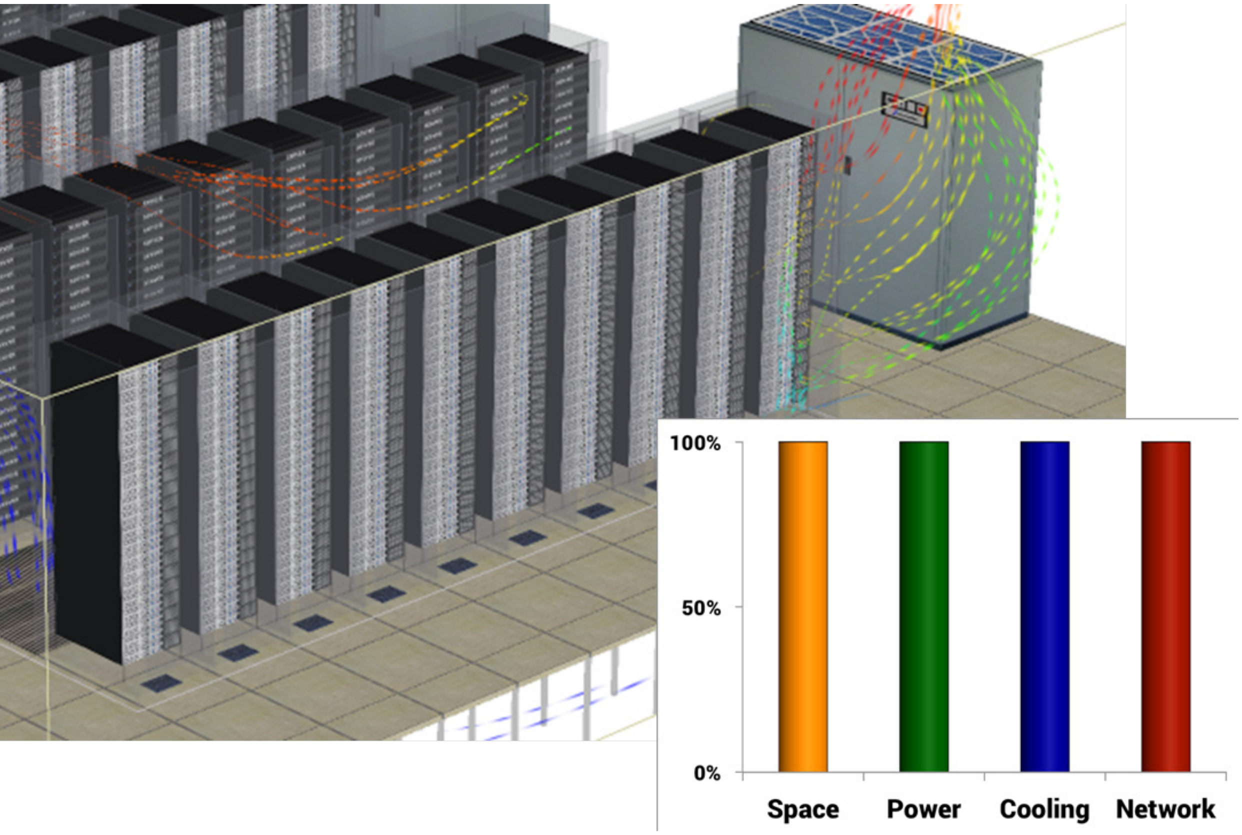


Figure 4

The data center was designed with a specific cooling requirement in mind. Once in operation, if the cooling demand matches the cooling supply, then space, power and cooling would remain unfragmented.

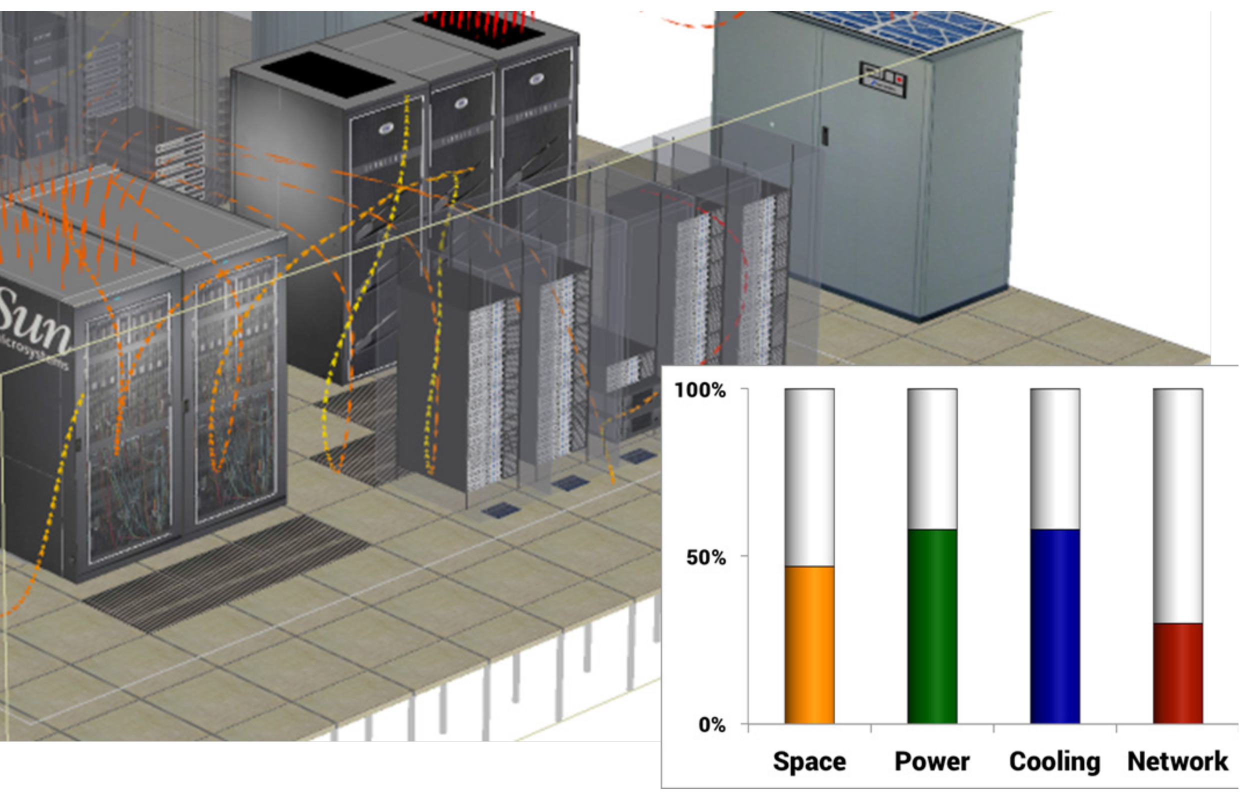


Figure 5

What actually happens once the data center enters operation is that actual equipment proves to be more (or less) cooling hungry. This can cause havoc to the cooling airflow, contiguous space and power, and that in turn threatens availability, capacity and efficiency. It is crucial to recognize that DCIM tools alone simply cannot tell you this - they lack the ability to model airflow!

What Can You Do About it?

Measure Data Center Performance using the ACE Performance Score, then Continuously Model

The most comprehensive way to measure data center performance is the ACE performance score.

ACE stands for Availability, Capacity and Efficiency. The score is depicted both numerically and graphically, and this allows a visual interpretation of the results. It is today being used by leading data center owner-operators to assess, improve and maintain data centers around the world.

The **ACE performance score** will reduce an owner-operator's total costs in much the same way that virtualization and cloud computing is being used to optimize the value of each physical server in a data center.

It will do this by reducing cost per server (i.e. $/kW of IT) by increasing capacity (thus requiring fewer data centers to support the same number of servers), and/or by allowing those of you whose main interest is in cooling efficiency or downtime reduction (resilience) to prioritize those elements instead.

Employing ACE requires that you start thinking and acting on one simple premise: whatever you do, use **predictive modeling** to simulate changes before you implement such changes in your data center. This is a process that we call **continuous modeling**.

ACE Performance Score

Designed to be utilized at any point in the data center life cycle, from inception through operation, the ACE performance score allows you to see your **ACE performance gap** – the difference between what you paid for and what you can actually expect to get out of your facility.

Once you know your data center's performance gap, you can make more informed decisions about influencing this gap: which variables to protect, which to sacrifice, where to save money and how to reduce the impact of the engineering changes that damage ACE.

The ACE performance score can be used for a variety of *What Ifs*. For example, what if you could...

- Quantify server availability by predictively modeling power and cooling failure?

Future Facilities White Paper Link

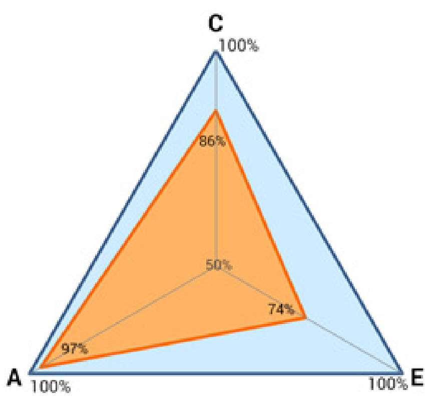
At the End of the Day, It's Lost Capacity

- Quantify how much extra connected IT load your data center can accommodate?
- Visualize airflow and temperature, and quantify your cooling efficiency?
- Perform these three simultaneously, and then assess overall performance in one comprehensive indicator?

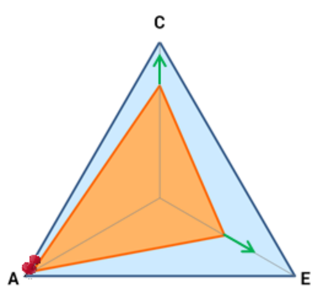
Case Study – In Brief

In 2012, a major investment bank contracted Future Facilities to provide assessment and implementation services for a well-run, 22,000ft^2 Tier IV data center with fully integrated DCIM and live monitoring tools in place.

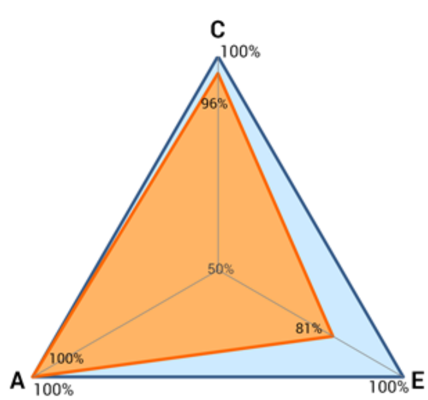
The brief was simple: assess their performance gap, give them choices for how to narrow and balance that gap, then implement the changes they requested.



For this data center, server availability had been well defended, but cooling efficiency and capacity had both suffered as a result of tactical changes from the design load (as explained in points 2 and 5).



The client decided to protect availability (lower left corner). This gave them the choice of prioritizing either capacity or cooling efficiency gains, or placing equal emphasis on each.



They chose to place equal emphasis on C and E:

Availability increased: to 100% while simultaneously decreasing cooling costs

Capacity reclaimed: 350kW (10%), or $8.75m

Efficiency savings made: $1.15m (4k tons of CO_2) annually - 15% PUE drop

Techspeak:

availability

% of existing connected IT load that will always be available under what if power and cooling failure conditions

capacity

% of design IT load that can be achieved by adding to the present day configuration

efficiency

the effectiveness of airflow and temperature delivery

Techspeak:

ACE Performance Score triangles:

the ACE performance triangle consists of:

1. an outer triangle with a blue outline. This represents the aspirational goal set during design

2. an inner orange triangle with a red outline. This represents actual performance of the data center at any given point in time

3. a light blue gap between the two triangles. This is the ACE performance gap. The more light blue that is visible, the more compromised the facility

Conclusion

Modern mission critical facilities have to stand the test of time. They must also do so in the face of the ever changing technology landscape, the requirement for more and varied compute capacity, with the pressure to minimize downtime.

The deployment of newer, differently-designed IT hardware is the primary cause of resource (space, power and cooling) fragmentation. This forces you to plan and build new data centers much earlier, forcing you to spend your CapEx much sooner than expected.

The **ACE performance score** can be used effectively by both data center designers to deliver the best possible design, and by you to compare current performance of the data center to that of the original design by:

- Minimizing downtime by improving availability
- Controlling lost capacity - and therefore $/kw and TCO - by increasing capacity
- Reducing energy bills by increasing efficiency

Controlling the three interconnected ACE variables in the short-term requires the ACE performance assessment score - a working practice that allows the owner-operator to choose a balance between **availability**, **capacity** and **efficiency.** In the long term, it requires the owner-operator to continuously model all future changes to the data center IT and facility layouts before committing to them.

In the next paper, *From Compromised to Optimized: An ACE Performance Assessment Case Study*, we'll show you how ACE is actually used in operation. And we'll explain just how we saved two well-run, DCIM equipped data halls $10m over two years using the ACE performance score...

"We used predictive modeling to run through a range of what ifs, then presented improvements that would decrease the performance gap, but protect availability".

Future Facilities White Paper Link

From Compromised to Optimized: An ACE Performance Assessment Case Study

References

2013 Cost of Data Center Outages, Ponemon Institute, December 2013.

2013 Study on Data Center Outages, Ponemon Institute, September 2013.

A Simple Model for Determining True Total Cost of Ownership for Data Centers, Jonathan Koomey et al, Uptime Institute, 2006.

At the End of the Day, It's Lost Capacity, Future Facilities' White Paper, 2013.

Billion Dollar Drain, Future Facilities' White Paper, 2013.

Future Facilities' Data Center Lost Capacity Client Survey.

Global Data Center Energy Demand Forecasting, DataCenter Dynamics, September 2011.

Special Report: Data Center Capacity and Energy Efficiency Survey, Kenneth Brill, Uptime Institute, March 2008.

The Elephant in the Room is Lost Capacity, Future Facilities' White Paper, 2013.

United States Department of Energy Statistics.

Uptime Tier Certification: http://www.greenserverroom.org/Tier%20Classifications%20Define%20Site%20Infrastructure.pdf

Use Best Practices to Design Data Center Facilities, Gartner research publication ID#: G00127434.

About Future Facilities Ltd.

For a decade, Future Facilities has provided predictive modeling software and consultancy services to both the world's largest data center owner-operators and to the industry's leading consultancies.

With offices across the globe, we are unique in the market place; the only company offering an holistic solution for the data center lifecycle – from inception through to operation. We call this holistic approach 'the Virtual Facility'.

UK Corporate Headquarters

1 Salamanca Street

London SE1 7HX

Tel: +44 (0) 20 7840 9540

Fax: +44 (0) 20 7091 7171

Email: info@futurefacilities.com

North America

2055 Gateway Place, Suite 110

San Jose, CA 95110

Tel: +1 408 436 7701

Fax: +1 408 436 7705

Email: info@futurefacilities.com

futurefacilities.com