Virtualization: Optimized Power and Cooling to Maximize Benefits

By Suzanne Niles

White Paper #118

Includes Interactive Tool

APC TRADEOFF TOOLS™

APC by Schneider Electric

Executive Summary

Data centers are routinely and unknowingly missing a great portion of their entitlement from virtualization. Beyond virtualization's undisputed IT benefits – from reduced rack footprint to disaster recovery – is the parallel story of a substantial benefit from optimizing the physical infrastructure that supports it. In particular, row-based cooling, correctly sized power and cooling, and real-time capacity management are essential elements in realizing virtualization's full potential in cost reduction, efficiency, and reliability.

Introduction

There are three primary things to understand about virtualization as it relates to the data center's power and cooling infrastructure:

- **Power and cooling technology is available today** to safeguard availability and meet the challenges of density and dynamics brought by virtualization.
- **Power consumption will be less** after virtualizing, as a result of the consolidation and physical reduction of the amount of IT equipment. With optimized power and cooling to minimize unused capacity, power consumption will typically be ***much*** less.
- **Data center efficiency (DCiE) will go down** after virtualizing, as the result of an increase in unused power and cooling capacity. With optimized power and cooling to minimize unused capacity, efficiency can be brought back to nearly pre-virtualization levels – sometimes higher, depending upon the nature of improvements to the cooling architecture.

Virtualization is bringing new extremes of power density and pace of change to the data center, increasing the demands on power and cooling infrastructure (see sidebar) and naturally raising some concern about possible effects on availability. Fortunately, the key characteristic of virtualization – high density – is not new, and effective strategies for supporting it have had time to develop. While virtualization carries consolidated and dynamic computing to extraordinary new levels, the basic power and cooling requirements of virtualized computing are similar to those already introduced by high-density blade servers during the past decade. As a result, technologies are available today to meet the power, cooling, and management needs of a virtualized environment.

Beyond high density

Virtualization is closely linked to high density, but it also introduces issues that go beyond high density alone and which must be considered in power and cooling systems supporting a virtualized environment.

What's different now

- **Increased server criticality** – Virtualization is bringing higher and higher processor utilization to the data center, increasing the business importance of each physical server, which makes effective power and cooling even more critical in safeguarding availability.
- **Hot spots that vary in time AND place** – With virtualization, applications can be dynamically started and stopped, resulting in loads that change both over time AND in physical location. This adds a new challenge to the architecture and management of power and cooling.
- **Reduction in IT load** – The abrupt, sometimes extreme, reduction in IT load that accompanies virtualization presents an opportunity for cost reduction in power and cooling systems that is often missed.

Physical consolidation from virtualizing will always reduce power consumption – directly, from the reduced server population and indirectly, from eliminating a portion of the power consumed by the power and cooling systems (although the latter reduction may be less than expected, as explained later). A parallel upgrade to bring power and cooling into line with the same lean philosophy as the virtualized IT layer will **reduce power consumption even further – often increasing by double (or more) the electrical savings achieved by virtualization alone.**

If power and cooling capacity is not optimized to match the new lower load, data center efficiency will go *down* after virtualizing (even while power consumption is reduced), which reflects the additional overhead of idle power and cooling capacity at the new lower IT load. This paper describes how the latest power and cooling systems not only provide *effective* support for a virtualized environment, but also *efficient* support –

significantly leveraging the IT-layer efficiency gains that are the signature benefit of virtualization. **Properly designed physical infrastructure will not only provide solutions for the specific power and cooling demands of virtualization, but – especially if replacing room-based cooling systems – can raise both power density capacity and overall data center efficiency significantly above what they were before virtualization.**

Challenges to Power and Cooling Infrastructure

Virtualization creates changes in the data center that present new challenges to power and cooling infrastructure, with implications to both *effectiveness* (how well it performs the job of safeguarding the IT load) and *efficiency* (how well it conserves power while performing that job). While an upgrade of power and cooling systems is not necessarily required to make virtualization "work," the greatest benefits from virtualization will be realized with power and cooling that responds to these challenges, which can be characterized as follows:

	Virtualization challenge **to power/cooling infrastructure**	**Solution**
1	Dynamic and migrating high-density loads	***Row-based cooling***
2	Underloading of power/cooling systems	***Scalable power and cooling***
3	The need to ensure that capacity meets demand at the row, rack, and server level	***Capacity management tools***

While these challenges are not new, and not unique to virtualized data centers, the combined simultaneous effects of virtualization are focusing attention on them with a new urgency, especially in light of the expanding interest in energy efficiency.

This paper approaches these challenges and solutions in the context of a virtualized environment. The list of white papers at the end provides additional general and detailed information about these topics in the overall data center context (virtualized or not).

Whole-system approach

Power and cooling issues can be articulated separately for the purpose of explanation and analysis, but effective deployment of a total virtualization solution requires a system-level view. The shift toward virtualization, with its new challenges for physical infrastructure, re-emphasizes the need for integrated solutions using a holistic approach: consider everything together, and make it work as a ***system***. Each part of the system must communicate and interoperate with the others. Demands and capacities must be monitored, coordinated, and managed by a central system, in real time, at the rack level, to ensure efficient use of resources and to warn of scarce or unusable ones – this centralized management issue is addressed below under "Challenge #3."

Row-Based Cooling

Variable cooling near the load

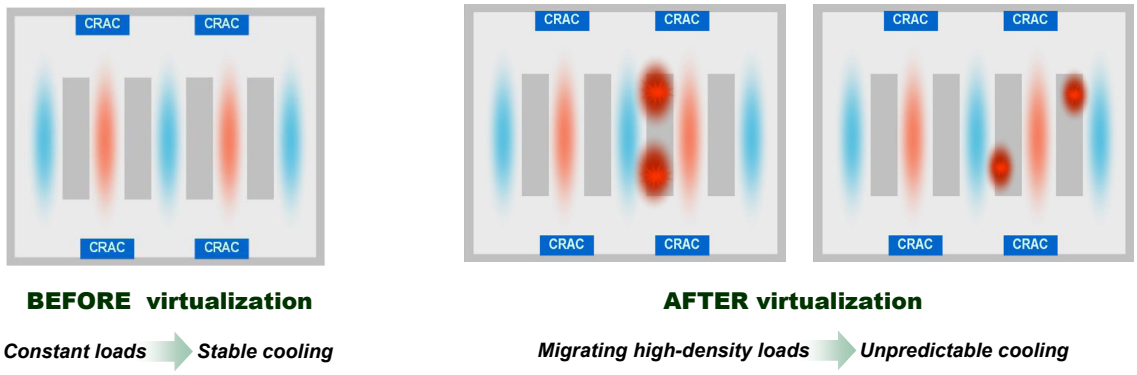
While virtualization may reduce *overall* power consumption in the room, virtualized servers tend to be installed and grouped in ways that create localized high-density areas, or "hot spots." These new densities can be an order of magnitude greater than the densities in a typical unvirtualized data center. Not only are densities increasing, but virtualization also allows applications to be dynamically moved, started, and stopped – the result can be loads that change both over time AND in their physical location in the room.

Meeting challenge #1

Dynamic and migrating high-density loads

Before virtualization brought dynamic allocation of server loads, localized high-density hot spots stayed put. If perimeter, room-based cooling and air distribution could be configured to adequately cool the hot spots, it would stay that way until physical servers were added, removed, or relocated. The thermal state of the room was typically determined by a walk-through with a thermometer, and cooling was adjusted by moving vented floor tiles. With dynamic loading of servers, the thermal profile of the room can shift, unseen, with no visible physical changes in equipment (**Figure 1**).

Figure 1 *– High-density "hot spots" that vary both in power density AND location can result from dynamic virtualized IT loads*

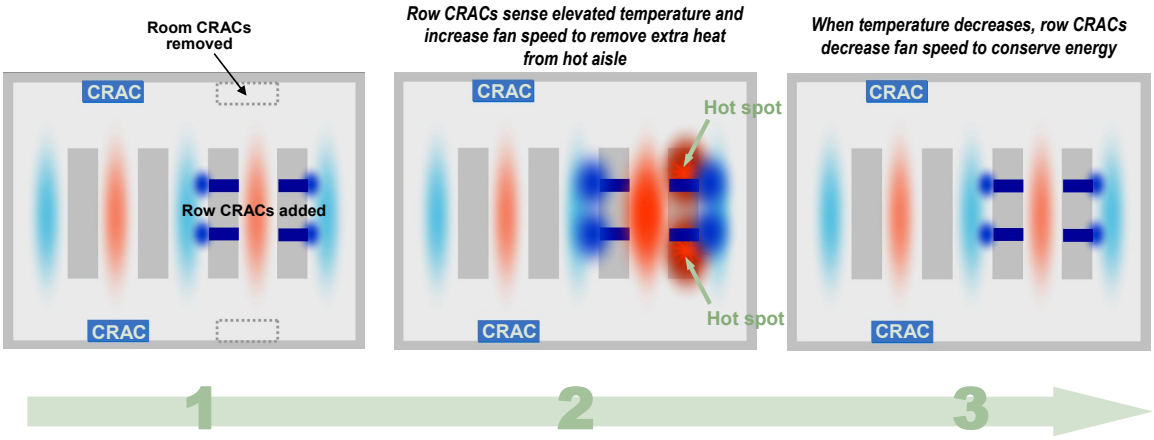


Predictable and efficient cooling calls for a system that comprehends these variations and automatically matches cooling – both in location and in amount – to changing power densities. The key characteristics of such a cooling system are

- Short air path between cooling and load
- Dynamic response to load changes

Cooling units located within the rows, and instrumented to sense and respond to temperature changes, meet the above two essential criteria. Row-based cooling substantially increases efficiency by providing cooling only *where* needed, only *when* needed, and only *in the amount* needed (**Figure 2**).

Figure 2 *– Row-based CRACs[1] work together to remove extra heat from hot aisle*



The placement of cooling units close to the servers provides the essential foundation that is the key to efficient cooling: ***short air paths***. A short air path between cooling and the load enables a number efficiency and availability benefits:

- Reduced mixing of cold supply air with hot return air
- Increased return temperature (increases rate of heat transfer to coil)
- Targeted cooling that can respond to localized demand
- Conservation of fan power
- Reduced – often eliminated – need for make-up humidification (to replace condensation formed on a too-cold coil resulting from a too-low set point)

Dynamic power variation in virtualized IT loads is a major reason why the industry is moving away from room cooling and toward row or rack cooling. For more about row-based and rack-based cooling, see APC White Paper #130, *The Advantages of Row and Rack-Oriented Cooling Architectures for Data Centers.*

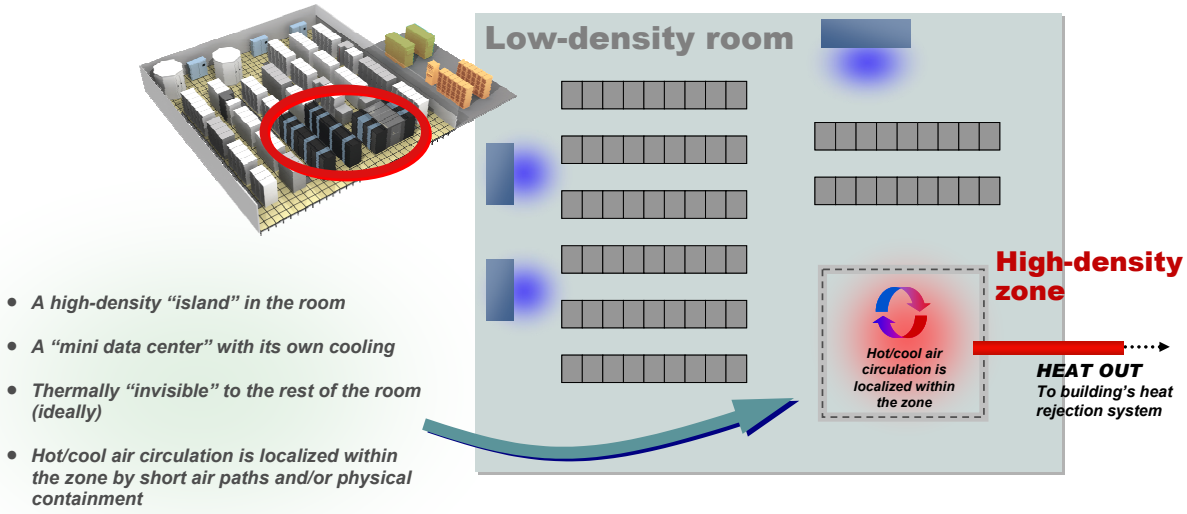
High-density zones

Virtualization projects often result in the deployment of a cluster of servers, such as blade servers, into an existing low-density data center. Row-based cooling – the localization of targeted cooling close to the loads – provides a technique for retroactive deployment of high density into an existing low-density data center: the ***high-density zone***.

[1] The term "CRAC" *(computer room air conditioner)* used in the figures might not be considered technically correct if the actual exchange of heat takes place outside the room, as with chilled water systems – in which case the unit is a CRAH *(computer room air handler)*. The term is used here because it is commonly used to refer to either a true air conditioner (a CRAC) or an air handler (a CRAH).

A high-density zone is a physical area of the data center allocated to high-density operation, with self-contained cooling so that the zone appears thermally "neutral" to the rest of the room – requiring no cooling other than its own, and causing little or no disturbance to the existing airflow in the room. **Figure 3** illustrates the concept of a high-density zone. The zone can be cooled and managed independently of the rest of the room, to simplify deployment and minimize disruption.

Figure 3 *– The high-density zone is an option for deploying virtualization in an existing data center*



For more about high-density zones, see APC White Paper #134, *Deploying High-Density Zones in a Low-Density Data Center*.

The benefits of variable, short-air-path cooling are only part of the advantage of row-based over room-based cooling. Other significant benefits arise from the fact that row-based cooling is **modular** and **scalable**, which addresses the second challenge when virtualizing: increasing efficiency by deploying *correctly sized power and cooling capacity*, discussed in the next section.

Scalable Power and Cooling

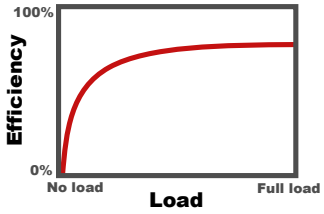
Matching power and cooling to the load

The reduction in IT load as a result of server consolidation offers a new opportunity to take advantage of modular, scalable architecture for power and cooling. Until now, the usual argument in favor of scalable architecture has been the ability to start small and grow as needed, to avoid over-investment and wasted operational cost from infrastructure that may never be used. With virtualization, scalable architecture now allows scaling *down* to remove unneeded capacity at the time of initial virtualization, with the later option to re-grow as the new virtualized environment re-populates. Whether scaling up or scaling down, the idea is the same – power

Meeting challenge #2

Underloading of power/cooling systems

and cooling devices are less efficient at lower loading, so it is wasteful to be running more power or cooling than you need. "Right-sized" infrastructure keeps capacity at a level that is appropriate for the actual demand (taking into account whatever redundancy and safety margins are desired).



For more about efficiency as a function of load, see APC White Paper #113, Efficiency Modeling for Data Centers

Why is oversizing wasteful?

Running more power and cooling equipment than needed is like leaving your car running at idle when you're not using it – energy is consumed, but no useful work is accomplished. All power and cooling devices have electrical losses (inefficiency) dispersed as heat. A portion of this loss is ***fixed loss*** – power consumed regardless of load. At no load (idle), fixed loss is the only power consumed by the device, so 100% of the power consumed is electrical loss (heat) and the device is 0% efficient, doing no useful work. As load increases, the device's fixed loss stays the same and other losses that are tied to the amount of load on the device, collectively called ***proportional loss***, increase in proportion to the amount of load. As load *increases*, fixed loss becomes a smaller and smaller portion of the total energy used, and as the load *decreases*, fixed loss becomes a larger portion of total energy used. The crucial role of fixed loss in underloading (oversizing) is discussed in the later section **Virtualization's Effect on Power Consumption and Efficiency**.

Impact of virtualization on oversizing

Since virtualizing can significantly reduce load, oversizing is an important efficiency issue in a virtualized data center. Even without virtualization, oversizing has long been a primary contributor to data center inefficiency. Server consolidation and server power management, by reducing load even more, will shift efficiency further toward the low end of the efficiency curve, if power and cooling systems stay the same. While the electric bill will indeed go down because of the lower IT load and less air conditioning needed to cool it, the *proportion* of utility power that reaches the IT loads – in other words, efficiency – will drop, which signifies wasted power that could be conserved to further reduce energy consumption. [2]

Virtualization is a new chance to take advantage of scalable infrastructure. Power and cooling devices that can scale in capacity will reduce fixed losses and increase

Effects of extreme underloading

In addition to its efficiency benefits, optimally sized power and cooling will protect against a number of detrimental effects that can result from ***very*** low loading. In a data center that is already at low loading because of redundancy or other reasons, virtualization can reduce loading to exceptionally low levels. Unless power and cooling are downsized to bring loading back within normal operating limits, these effects could result in expenses that negate some of the energy savings or, in some cases, pose a risk to availability.

Cooling *(too-low thermal load)*

- Safety shutdown because of high head pressure on compressors
- Short-cycling of compressors from frequent shutdown, which shortens compressor life
- Possible voiding of warranty from consistent operation below lower load limits
- Cost of hot-gas bypass on compressors to simulate "normal" load to prevent short-cycling

Generator *(too-low electrical load or too many generators)*

- Unburned fuel in the system ("wet stacking"), which may result in pollution fines or risk of fire
- Cost of unneeded jacket water heaters to keep engines warm
- Cost of storage, testing, and maintenance of excess fuel

[2] EPA and The Green Grid are now helping to educate the user community about the significant efficiency value of "right-sizing" physical infrastructure to more closely follow the IT load.

efficiency. Scalable architecture will facilitate not only downsizing to follow IT consolidation, but also subsequent regrowth to follow expansion of the now-virtualized IT load (**Figure 4**).

Figure 4 *– Using scalable power and cooling to minimize the inefficiency of unused capacity during consolidation and regrowth*

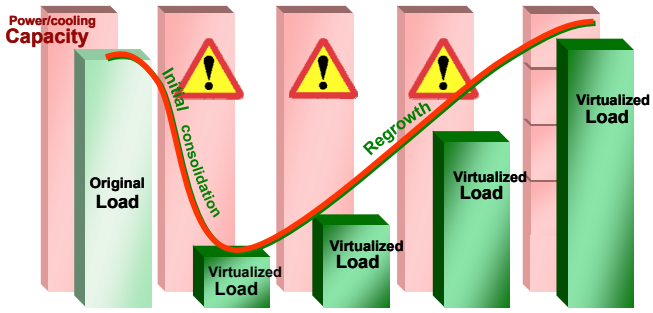


Fig 4a – Power and cooling not downsized after virtualizing
Unused capacity is a significant source of inefficiency (low DCiE)

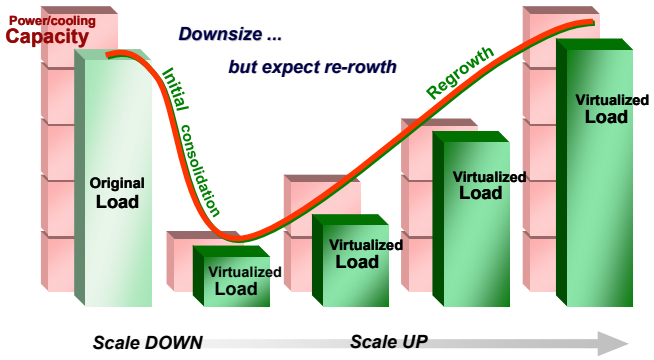


Fig 4b – Right-sized power/cooling
Scaled capacity maximizes efficiency

Capacity Management

Knowing what's going on, in real time

The dynamic nature of virtualized computing demands accurate, timely, and actionable information about power and cooling capacities to ensure that power and cooling are keeping up with a changing load profile that can shift from one day to the next.

Capacity management provides instrumentation and software for real-time monitoring and analysis of information about the three essential capacities of the data center:

- Power
- Cooling
- Physical space

Meeting challenge #3

Ensure that capacity meets demand at the at the row, rack, and server level

For any location where new or reconfigured IT deployment is under consideration, all three of these resources must be available *at that location* and in sufficient capacity in order to support the desired deployment. If there is insufficient capacity of any one of the three, the deployment cannot move forward.

Capacity management enables these resources to be utilized effectively and efficiently throughout the data center, through continuous real-time visibility to capacities *at the rack and server level.* With this data, management software can identify locations where there is available capacity of one or more resources, where a capacity is dangerously low, or where there is unusable (stranded) capacity of one or more resources (see box). Stranded capacity is an important efficiency issue in a highly dynamic data center, not

only because it directly contributes to inefficiency – as resources paid for but unused – but also because unmanaged change can *create* stranded capacity.

An effective capacity management system uses automated intelligence and modeling to monitor power, cooling, and physical space capacities at the room, row, rack, and server level, to suggest the best place for adding equipment, to predict the effect of proposed changes, and to recognize conditions or trends in time for corrective action to be taken. Capacity management that comprehends server locations and loads, power and cooling capacity available to servers, temperature fluctuations, and power consumption not only protects against downtime from localized shortages of power or cooling, but also ***increases data center efficiency by optimizing the use of available resources***. A holistic system like this can

- Model the system-wide effects of proposed server changes
- Compare alternative layouts using detailed design analysis
- Confirm, before deployment, that a proposed change will not cause a power or cooling overload
- Verify that a change was made as planned
- Reserve power, cooling, and rack space so new equipment can be installed quickly

Stranded capacity = inefficiency

When one or two of the three essential resources – power, cooling, or space – is insufficient in a particular location, that location cannot be used even though there may be available capacity of the other resource(s). The resources that *are* available in this location, but cannot now be used because of lack of the other one(s) – are called **stranded capacity**. For example, in a location where existing rack space and power capacity are unusable because there is not enough cooling, there is stranded capacity of space and power. In a location where there is extra cooling capacity but no floor space for racks or no power available, there is stranded cooling capacity.

Stranded capacity amounts to wasted resources – islands of power capacity, cooling capacity, or rack space that you can't use. These wasted resources can be the result of either original design error or subsequent unmanaged change. Finding and reallocating stranded capacity directly increases data center efficiency by increasing the amount of IT equipment that can be supported by the same resources.

The need for capacity management is greatest when there is *change,* the hallmark of today's virtualized data center – a changing server population, varying power density, load migration, the steady advance of new technologies, and increasing pressure to conserve energy. Unmanaged change – in *any* data center – can compromise availability, thwart planning, and waste resources. Effective capacity management comprehends the mechanics and wide-ranging effects of change, allowing the data center to utilize its power, cooling, and physical space to maximum advantage. With this intelligence, virtualization can fulfill its potential for efficiency and business value.

Virtualization introduces or escalates a number of factors which, taken together, provide a checklist for what the capacity management system must be able to handle:

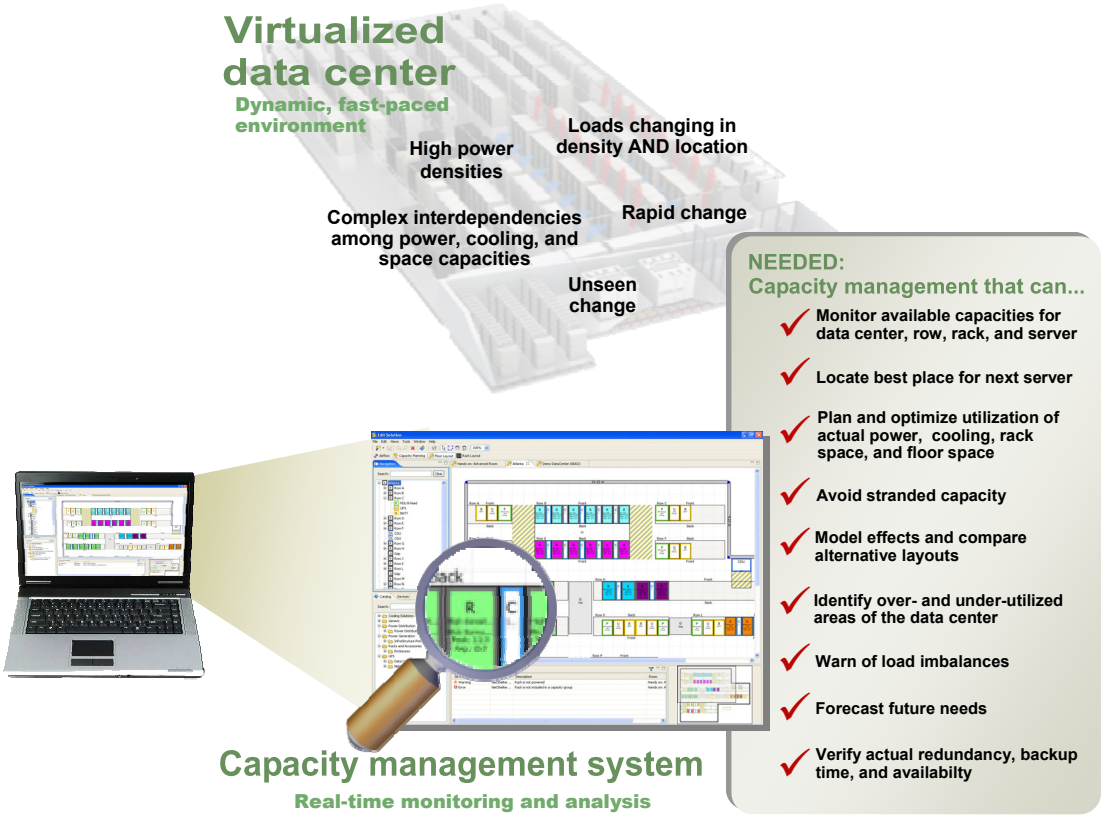
- **Loads changing in density *and* location –** Dynamic utilization can create hot spots even without adding new physical servers. Hot spots can occur in new places as the virtual machines migrate across servers, and as server power management shuts down some physical servers and powers up others.
- **Increased pace of change** – In the fast-moving climate of virtualization and constantly changing technology, the latest change is barely settled before the next one is happening. Maintaining system

stability becomes paramount, especially if multiple parties are making changes without centralized coordination.

- **Complex interdependencies –** Virtualization brings a new level of sophistication to the shared dependencies and secondary effects in the relationships among power, cooling, and space capacities. It can be difficult to predict the system-wide effects of adding, removing, or relocating servers.
- **Unseen change –** Even without adding new physical servers, virtualization creates unseen changes in location and demand for power and cooling. Without visibility to the warning signs of potential trouble, an unnoticed irregularity can escalate to overload, overheating, or loss of cooling redundancy.
- **Lean provisioning of power and cooling** – If power and cooling infrastructure is optimized as it should be for maximum efficiency, supply and demand will be closely aligned, reducing tolerance for unexpected changes on either side.

Effective management of this environment requires a system that knows the physical layout of the room (to track physical space capacity), knows device-level characteristics of the power/cooling supply and demand, and uses an integrated model for interpretation of present conditions, recognition of trends, and forecasting of future requirements (**Figure 5**).

Figure 5 *– Meeting the management challenge in a virtualized environment*



A fully informed, analytical, and interactive management system is the lifeblood of an integrated infrastructure, where subsystems communicate with a central coordinator that can correlate, analyze, advise, warn, and predict. Such a system knows the current state and limits of the infrastructure *right now* and can predict, through modeling, its future state as changes occur.

For more about capacity management see APC White Paper #150, *Power and Cooling Capacity Management for Data Centers*.

Virtualization's Effect on Power Consumption and Efficiency

This paper stresses that virtualization in an existing data center, with no changes to power and cooling infrastructure, always *reduces* data center efficiency (DCiE). This section explains why this efficiency reduction occurs, how to quantify it, and how to prevent it.

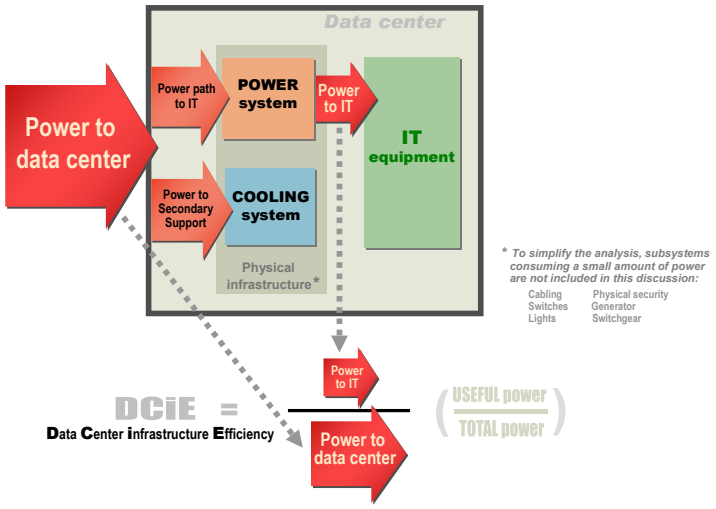
A primary motivation for virtualization is increased *computing* efficiency – more computing per watt of power consumed by the data center. The reduced power consumed by the consolidated IT load itself, however, is only the beginning of the savings entitlement that can be claimed when virtualizing. In the majority of existing data centers, there is a significant additional savings potential in the power and cooling systems that support the new, now smaller, virtualized IT load. These additional savings go beyond the immediate savings from the reduced load placed on the power and cooling systems – which tends to be disappointing because of ***fixed losses***, described below – but rather, they come from the efficiency of a reconfigured, streamlined power and cooling architecture that aligns more tightly with virtualization's reduced, and varying, demand.

Virtualization affects power consumption and efficiency in a number of ways, some of which can seem counter-intuitive. The concepts are not difficult, but they depend upon a clear understanding of basic definitions and the fundamental relationships among power, loss, and load. To illustrate the effects of virtualization on power consumption and efficiency, it will be helpful to review data center efficiency and identify the principal consumers of power in the data center.

What is "data center efficiency"?

"Data center efficiency," as the term is now being used in discussions of data center energy consumption, refers to the efficiency of the data center's physical infrastructure, whose major components are the power and cooling systems. The metric for this is DCiE (*data center infrastructure efficiency*), which quantifies the "useful work" performed by physical infrastructure and is defined as the proportion of total data center input power that is delivered to the IT load (**Figure 6**).

Figure 6 *– Definition of data center physical infrastructure efficiency (DCiE)*

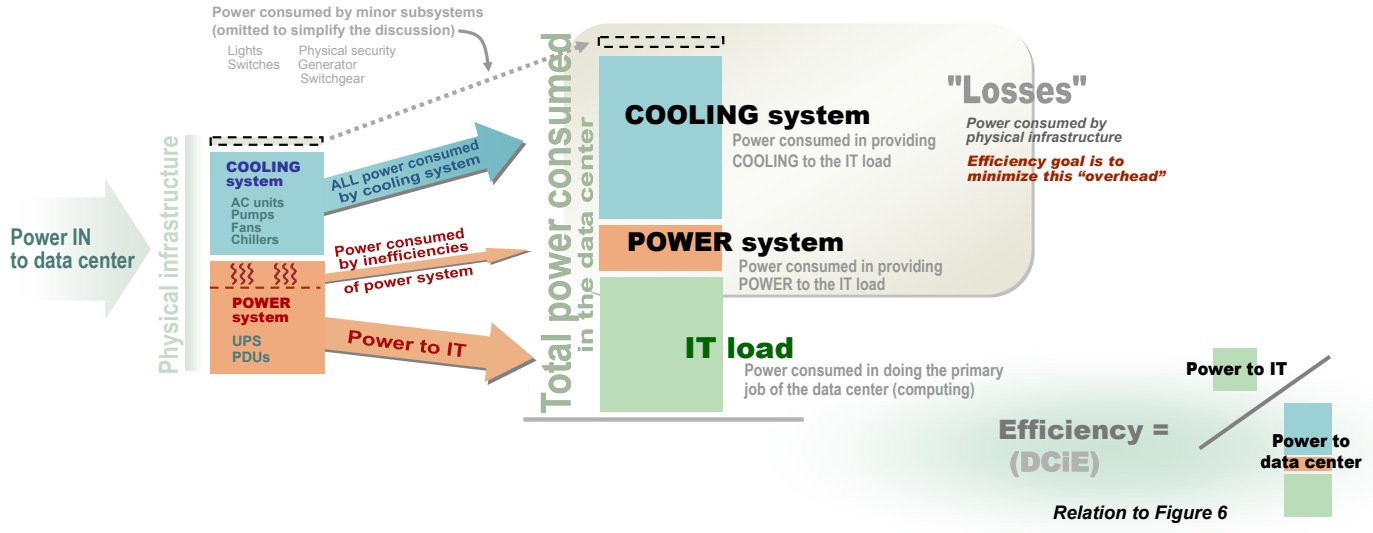


In this context, *all other power* consumed in the data center – in other words, all power *not* consumed by the IT load – is considered "loss." These non-IT power consumptions, or losses, include

- The internal inefficiencies of the power system (power path devices such as UPS, PDUs, wiring, etc.), dispersed as heat
- All power consumed by the cooling system
- All power consumed by other data center physical infrastructure subsystems (small by comparison, and not shown in **Figure 6**)

Figure 7 illustrates these losses in the context of total power consumed in the data center. For more about the concept of data center efficiency and the distinction between "loss" and "useful work" see APC White Paper #113, *Electrical Efficiency Modeling for Data Centers*.

Figure 7 *– Definition of "losses" in data center power consumption*



Fixed vs. proportional loss

Of the power consumed by the power and cooling systems – the "losses" in **Figure 6** – some stays the same no matter how large or small the IT load is, and some varies in proportion to the size of the IT load. These two components of consumed power (loss) are called **fixed loss** and **proportional loss**.[3]

- **Fixed loss** – Always the same amount no matter what the load. Fixed loss is power that is consumed whenever the device or system is running, regardless of how much load is present. Reducing the load does not change this fixed component of loss. Examples of devices with a large component of fixed loss are transformers and fixed-speed fans. The presence of fixed loss is the reason efficiency is greater at high loads (where fixed loss is a *small* proportion of total power) and lower at low loads (where fixed loss is a *large* proportion of total power) – see **Figure 9** below. **Reducing fixed loss, by improved device efficiencies and/or better system configuration, is the most effective way to increase efficiency.**
- **Proportional loss** – Directly proportional to the load on the device. Doubling the load will double the proportional loss. Reducing the load 75% will reduce the proportional loss 75%. Examples of devices with a large component of proportional loss are variable-speed fans and pumps.

As the following sections will show, fixed loss is responsible for limiting both the power savings and the efficiency that can be achieved by virtualization, because fixed loss does not change *no matter how much the IT load is reduced*.

Fixed losses limit power savings

Consolidation makes power consumption go down. The key question when contemplating virtualization is "How ***much*** will my power consumption go down?" The answer is closely linked to the fixed losses that are present in the power and cooling infrastructure. **Figure 8** illustrates the relationship between consolidation and power savings, and how fixed losses in the power and cooling infrastructure limit the power savings that are possible.

The key to increasing the savings from consolidation is *reducing fixed losses* in power and cooling infrastructure. Reducing fixed losses can be accomplished either by *eliminating* some of them (such as eliminating rehumidification cost by improving cooling system design), *reducing* some of them (such as switching to a more efficient UPS), or *converting* some of them to proportional losses (such as switching from fixed-speed to variable-speed fans and pumps). **Figure 9** shows the effect of reducing fixed losses.

[3] There is also a third, usually much smaller, type of loss – square-law loss – which varies with the *square* of the load. For more about the three types of loss, see APC White Paper #113, *Electrical Efficiency Modeling for Data Centers.*

Figure 8 *– Fixed losses limit power savings from consolidation*

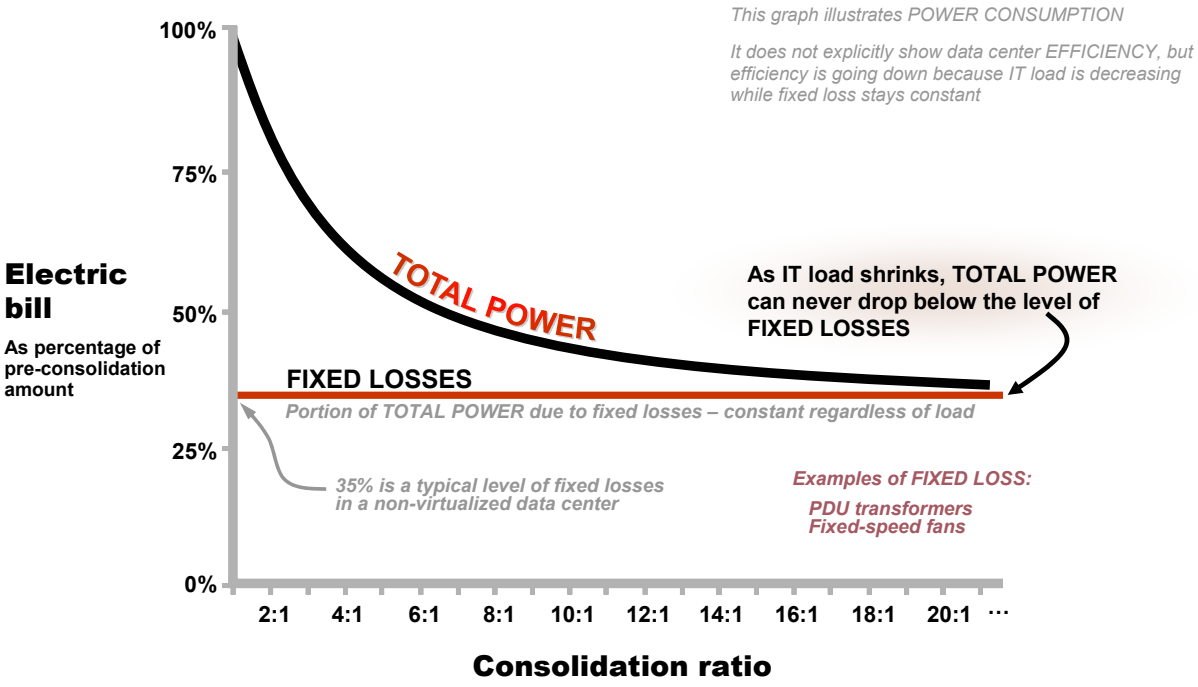
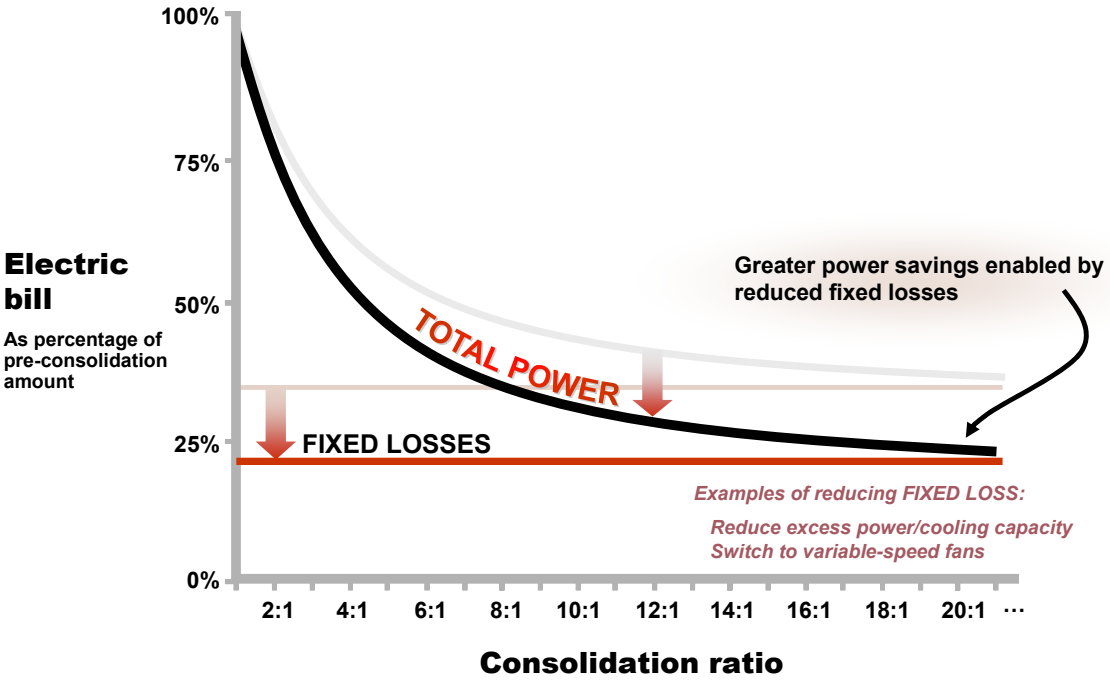


Figure 9 *– Reducing fixed losses enables greater savings from consolidation*

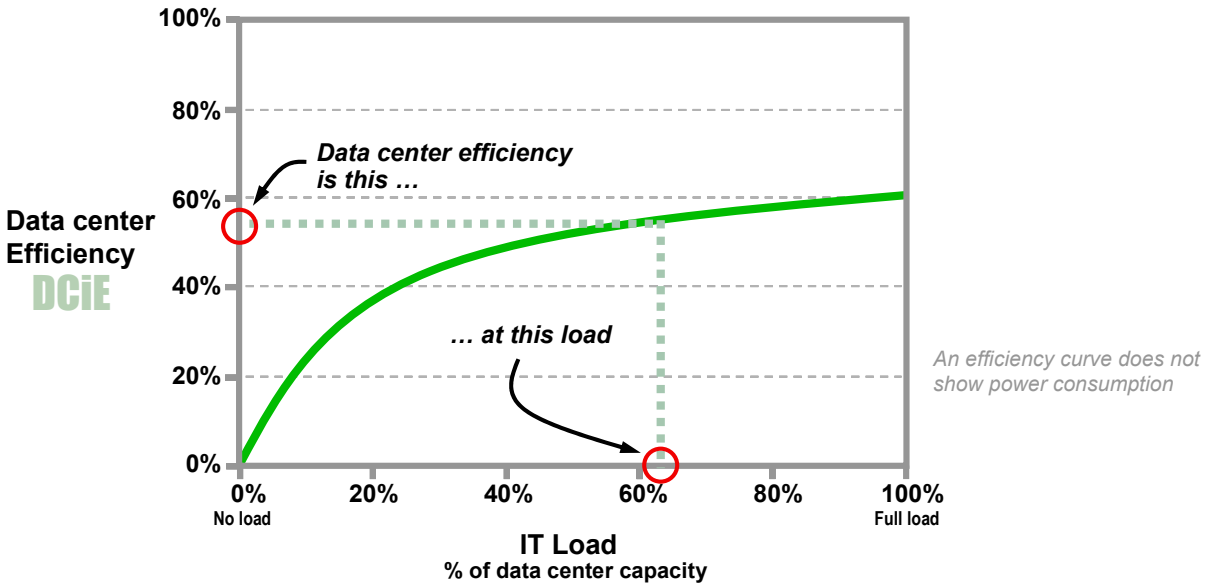


Data center efficiency is a function of IT load

In the context of data center efficiency, the "losses" shown in **Figure 7** consist of the aggregate of all power consumed by physical infrastructure systems in the data center – that is, the power consumed in *supporting* the IT load, but not the IT load itself. As with the power consumed by an individual device such as a UPS, some of these losses are ***fixed loss*** that stays the same regardless of load – power that is consumed all the time, whenever the system is turned on. The rest of the losses are ***proportional loss***, which varies in proportion to the IT load, and consists of devices such as variable speed fans and pumps.

If all of the losses were proportional losses (going up or down with the IT load) data center efficiency would simply be a single number – the same for any IT load. However, for data centers this is never the case because fixed loss is *always present* in data centers, resulting in efficiency that is always higher at high loads and lower at low loads. Data center efficiency, therefore, is always a curve – *efficiency is a function of load*. **Figure 10** shows a typical data center efficiency curve.

Figure 10 *– Typical data center efficiency curve*

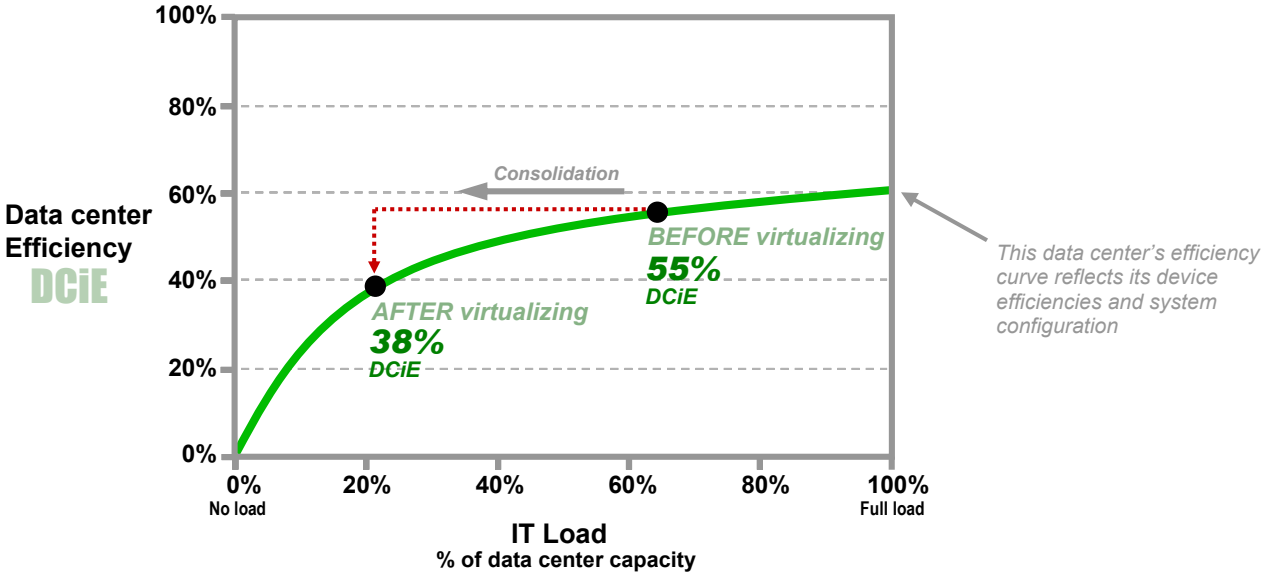


Each data center will have a higher or lower efficiency curve – depending upon the efficiency of its individual devices and the efficiency of its system configuration – but the curve always starts at zero and has this same general shape.

Virtualization's track on the efficiency curve

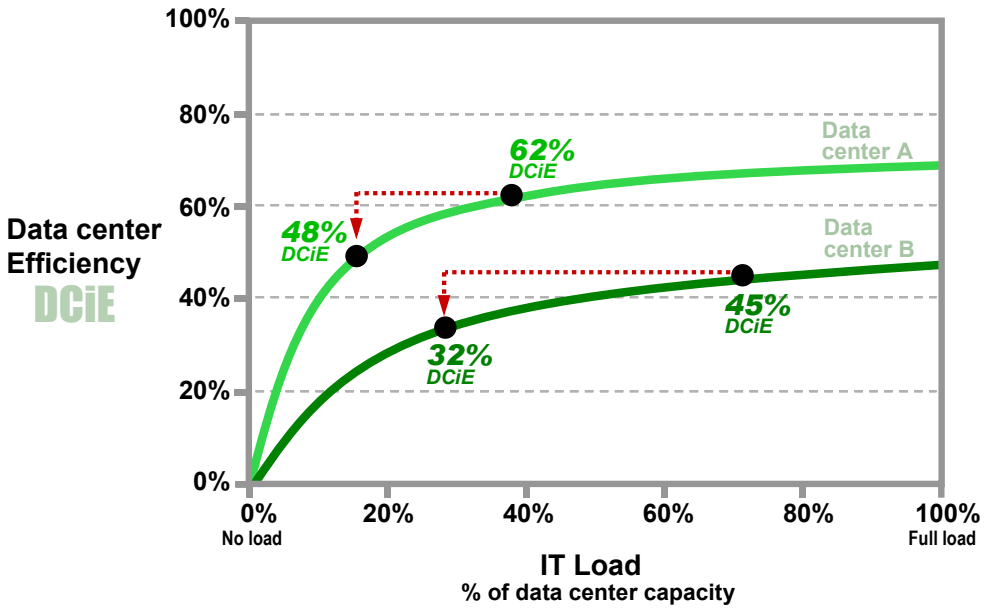
Virtualization will always reduce power consumption due to the optimization and consolidation of computing onto a smaller number of physical devices. However, if no concurrent downsizing is done to power and cooling infrastructure, the data center's efficiency curve will remain the same, and efficiency (DCiE) will move down on the curve because of the new, lower load (**Figure 11**).

Figure 11 *– Consolidation reduces load and moves efficiency down the curve*



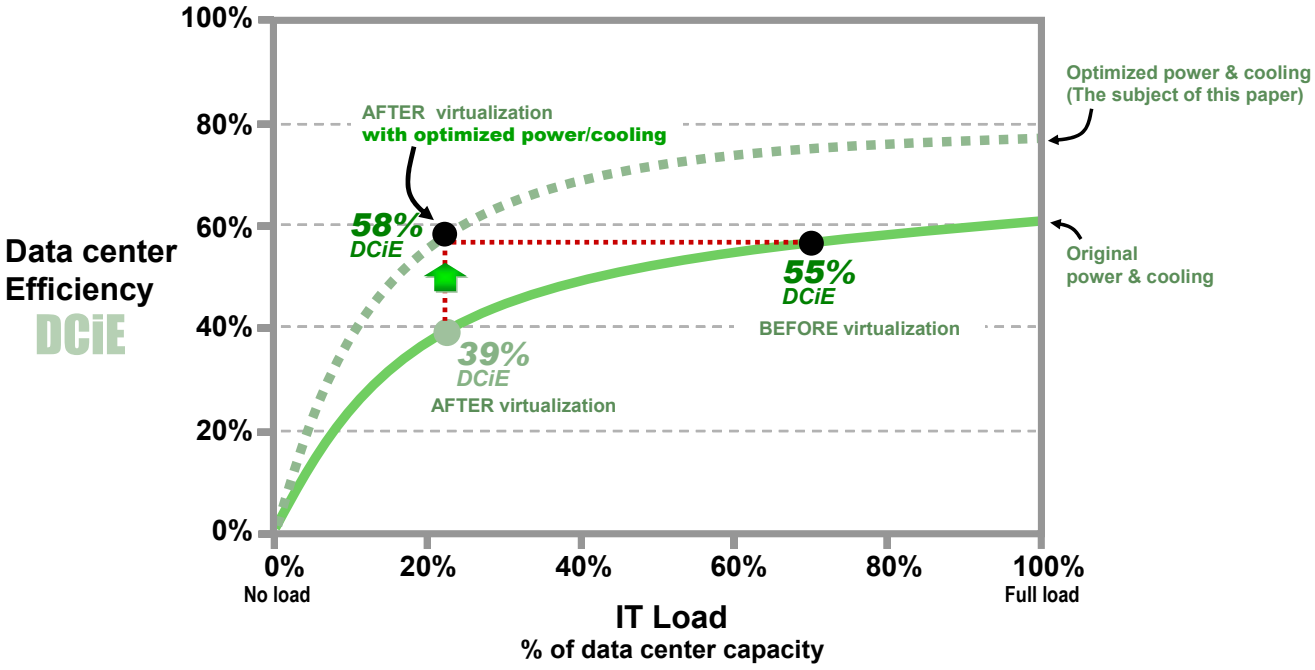
Note that efficiency has dropped, even though power consumption has been reduced. Efficiency doesn't measure how much power is being *used* – rather, it indicates the magnitude of power being *wasted* compared to what is being used. It is, essentially, a measure of "room for improvement." This efficiency drop will happen regardless of the data center's particular efficiency curve (which will always have a *shape* similar to this) and regardless of DCiE before virtualizing. If no change is made to power and cooling systems – to raise the efficiency curve – DCiE will be lower after consolidation *for any data center* (**Figure 12**).

Figure 12 *– For any data center, consolidation reduces efficiency if the data center's efficiency curve remains unchanged*



To improve post-virtualization DCiE, *the data center's efficiency curve must be raised* by optimizing power and cooling systems to reduce the waste of oversizing and align capacity with the new, lower load (**Figure 13**) – this optimization is the subject of this paper. The greatest impact on the efficiency curve can be made by going from room-based to row-based cooling and by "right-sizing" the power and cooling systems. **In addition to improving efficiency, optimized power and cooling will directly impact the electric bill by reducing the power consumed by unused power and cooling capacity.**

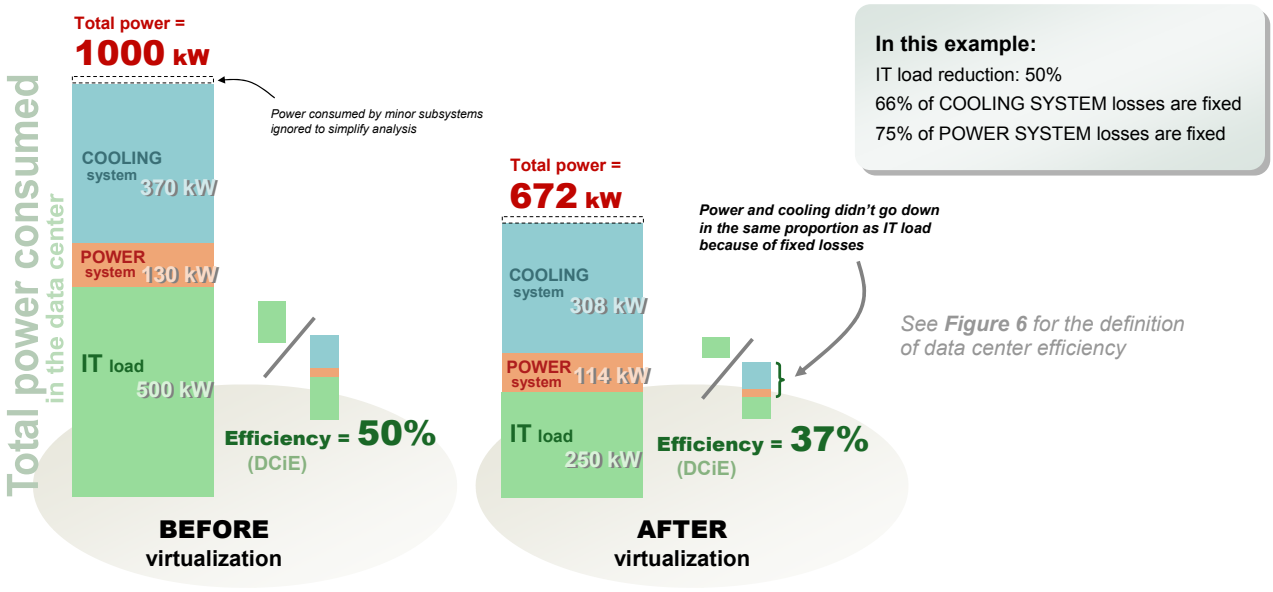
Figure 13 *– Optimized power and cooling raises the efficiency curve*



Reduced power consumption but lower efficiency

Power consumption will always be less (better) after virtualizing, due to reduced server population and *some* concurrent reduction in power consumed by the power and cooling systems (the *proportional losses*, described earlier). Data center efficiency, however, will be lower (worse) if power and cooling systems are not downsized and optimized to align with the new, smaller IT load. In other words, physical infrastructure that doesn't "slim down" to match the lower IT load will still be consuming power (*fixed losses*) to maintain excess or misdirected capacity that is not useful in supporting the reduced IT load. **Figure 14** illustrates this typical outcome – reduced power consumption combined with lower efficiency.

Figure 14 *– Example of reduced power consumption but lower efficiency*



To increase efficiency: Reduce fixed losses

The previous sections explain why virtualization causes a decrease in DCiE and that fixed losses in the data center infrastructure are the primary cause of this effect. To compensate for this problem and realize the full energy-saving benefits of virtualization, an optimized power and cooling infrastructure – as described in the first half of this paper – will incorporate design elements such as the following to minimize fixed losses and maximize the electrical efficiency of the virtualization project:

- Power and cooling capacity scaled down to match the load
- VFD fans and pumps that slow down when demand goes down
- Equipment with better device efficiency, to consume less power in doing the job
- Cooling architecture with shorter air paths (e.g. move from room-based to row-based)
- Capacity management system, to balance capacity with demand and identify stranded capacity
- Blanking panels to reduce in-rack air mixing

"Just in time" cooling

The idea of providing a resource at the right time, in the right amount, is not new – but it is new to data centers. The efficiency benefits of the row-based, localized cooling described here have a well-known parallel in manufacturing.

"Just in time" is a manufacturing philosophy developed by Toyota in the 1950s. It is now a cornerstone of management theory that focuses on the elimination of waste by having just enough of the right parts, at the right time, in the right place – just in time for when they are needed. The idea is to eliminate unnecessary storage and movement of inventory, with the goal of a "lean" and steady flow of materials throughout the manufacturing process.

The data center industry has already begun to reap the benefits of lessons learned from other industries in the areas of standardization and modularity. Now with its primary raw material – electricity – becoming a scarce and expensive resource, technologies and strategies for conserving it have become the focus of intense industry and user interest.

Cooling, as a major consumer of electricity in the data center, is a prime candidate for "use only what you need, where you need it, when you need it – and use it *very* efficiently."

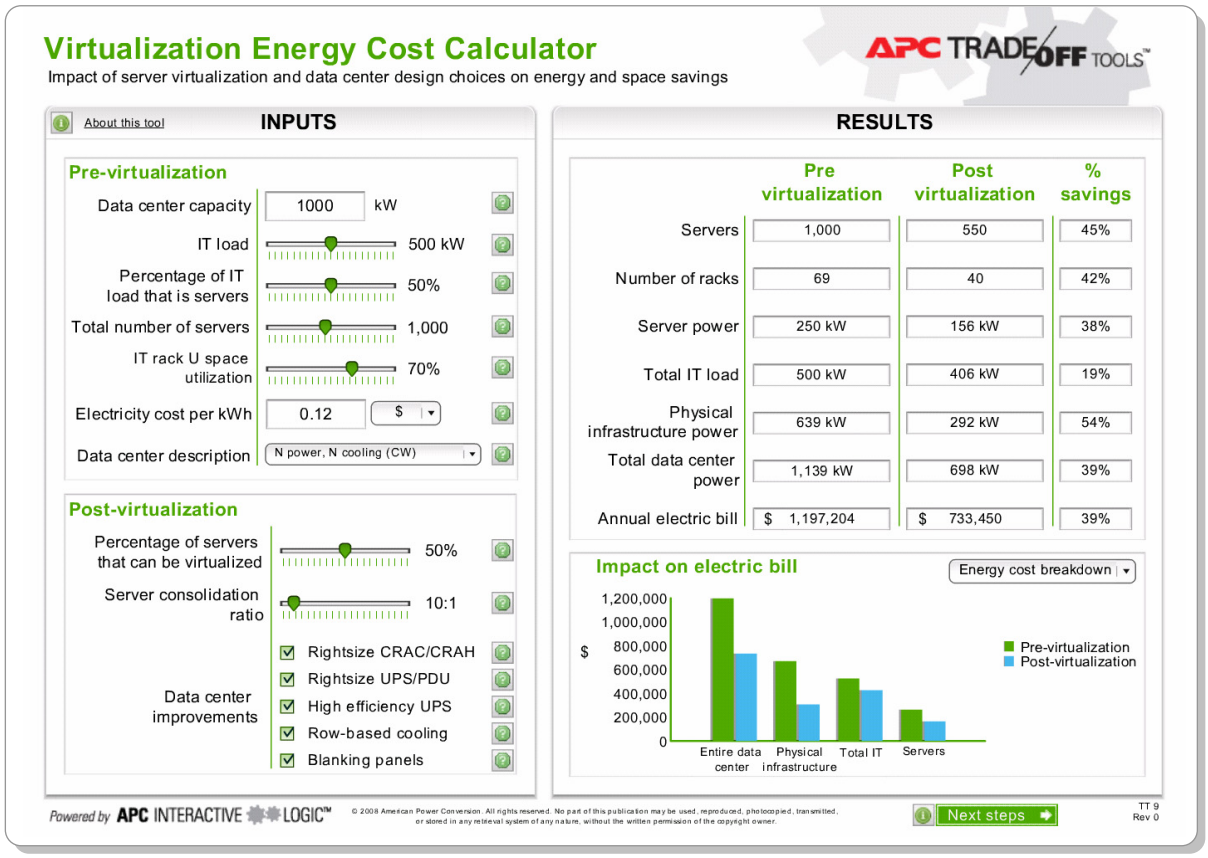
Minimal expenditure of power to do the job – i.e., efficient operation – requires targeted delivery of resources in the right amount, at right time, and in the right place (see box). These are the same principles behind the design of devices and architectures that meet the *functional* challenges of virtualization, presented in the first half of this paper. As a result, increased efficiency comes automatically with the three solutions described earlier – row-based cooling, scalable power and cooling, and capacity management tools.

APC TradeOff Tool™ for calculation of virtualization savings

Figure 15 shows the APC Virtualization Energy Cost Calculator TradeOff Tool™. This interactive tool illustrates IT, physical infrastructure, and energy savings resulting from the virtualization of servers in a data center. The tool allows the user to input data regarding data center capacity, load, number of servers, energy cost, and other data center elements.

Figure 15 *– APC TradeOff Tool for calculating virtualization savings*

Screen image below links to a live version of this interactive tool

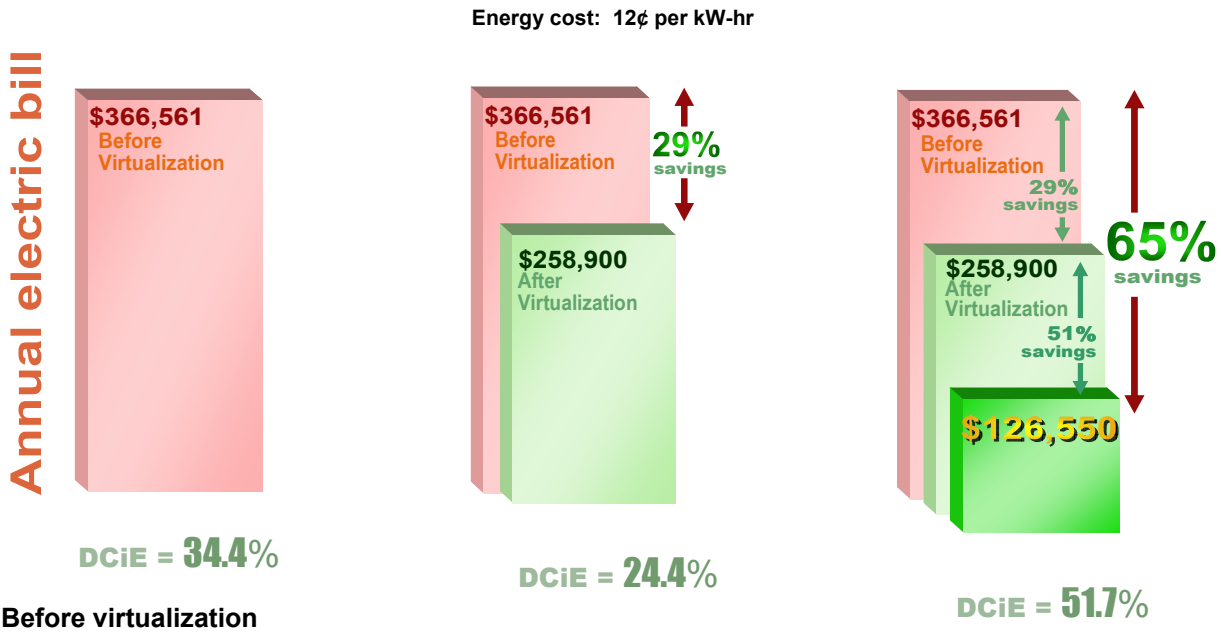


Case study

This case study was performed using the APC TradeOff Tool™ *Virtualization Energy Cost Calculator* (**Figure 15**, above). To simplify the example, the hypothetical data center is assumed to be fully loaded with no redundancy in power and cooling (**Figure 16a**), with an electric bill of $366,560.

The data center is virtualized with a 3-to-1 decrease in server power (67% reduction).[4] Because of the load reduction from virtualizing, we now have a 60 kW data center that is oversized by 100% (120 kW capacity) (**Figure 16b**). As a result of this oversizing, power and cooling infrastructure is now underloaded, and operates at reduced efficiency due to fixed losses, as described in this paper. The total electrical bill only sees a 1.4-to-1 reduction (29%), if no improvements are made to power and cooling infrastructure.

Figure 16 *– Case study showing effect of virtualization with and without power/cooling improvements*



a. Before virtualization

Data center capacity: 120 kW
Data center loading: 100%
Total IT load: 120 kW
Servers: 90kW (75% of load)
Density: 7kW / rack
No redundancy

Room-based cooling
Hot aisle/cold aisle layout
Uncoordinated CRACs
Air-cooled condenser for heat rejection

18" raised floor with 6" cable obstruction
Random placement of perforated tiles

b. After virtualization

Server consolidation only

3-to-1 server reduction
1.4-to-1 electricity reduction

60 kW of servers removed
Data center capacity: 120 kW (no change)
Data center loading: 100% → 50%
Total IT load: 120 kW → 60 kW
Server load: 90 kW (30 kW
UPS: Traditional, 81% efficient at full load

c. After virtualization

With power/cooling improvements

Data center capacity: 120 kW → 60 kW
Power and cooling right-sized to nearest 10 kW
UPS: High efficiency, 96% efficient at full load
Row-based cooling (no containment)
Blanking panels added

Payback period for improvements: < 4 years

[4] 3-to-1 is conservative. Consolidation ratios of 10-to-1 or greater are possible.

With improvements to power and cooling infrastructure, the electric bill reduction becomes nearly 3-to-1 (65%) and efficiency increases to 51.7% (**Figure 16c**).

Effect of reduced power consumption on energy and service contracts

An abrupt reduction in power consumption may have unintended consequences with regard to utility and service contracts. Such contracts will need to be reviewed and renegotiated where necessary, in order not to forfeit data center savings to the utility provider, building owner, or service provider.

- **Utility contract** – Agreements with utility providers may include a clause that penalizes the customer if overall electrical consumption drops below a preset monthly consumption amount.
- **Energy clause in real estate agreement** – Some real estate agreements include the cost of electricity as a flat rate, typically billed on a cost-per-square-foot basis. This agreement may need to be renegotiated, otherwise the savings from virtualization will accrue to the building owner.
- **Equipment service contracts** – Service contracts should be reviewed to remove unused power and cooling equipment, to avoid paying for service on equipment that has been taken out of service through downsizing.

Availability Considerations

Virtualization's higher power densities and variable loads can introduce vulnerabilities characteristic of *any* shift to high density without a reassessment of the power and cooling needed to safeguard availability. This, combined with the increased importance of individual physical servers (with multiple applications now running on each server), raises the stakes in ensuring that power and cooling are effectively aligned with the new demands.

The first half of this paper describes the three major physical infrastructure challenges posed by virtualization – dynamic and migrating high density loads, underloaded power and cooling infrastructure, and the need to manage dynamic interrelationships affecting capacity demand and supply. All of these challenges have implications to availability. A review of the solutions will show that, beyond their efficiency and manageability benefits, they also solve availability issues related to the increased complexity and dynamics of a virtualized environment (**Table 1**).

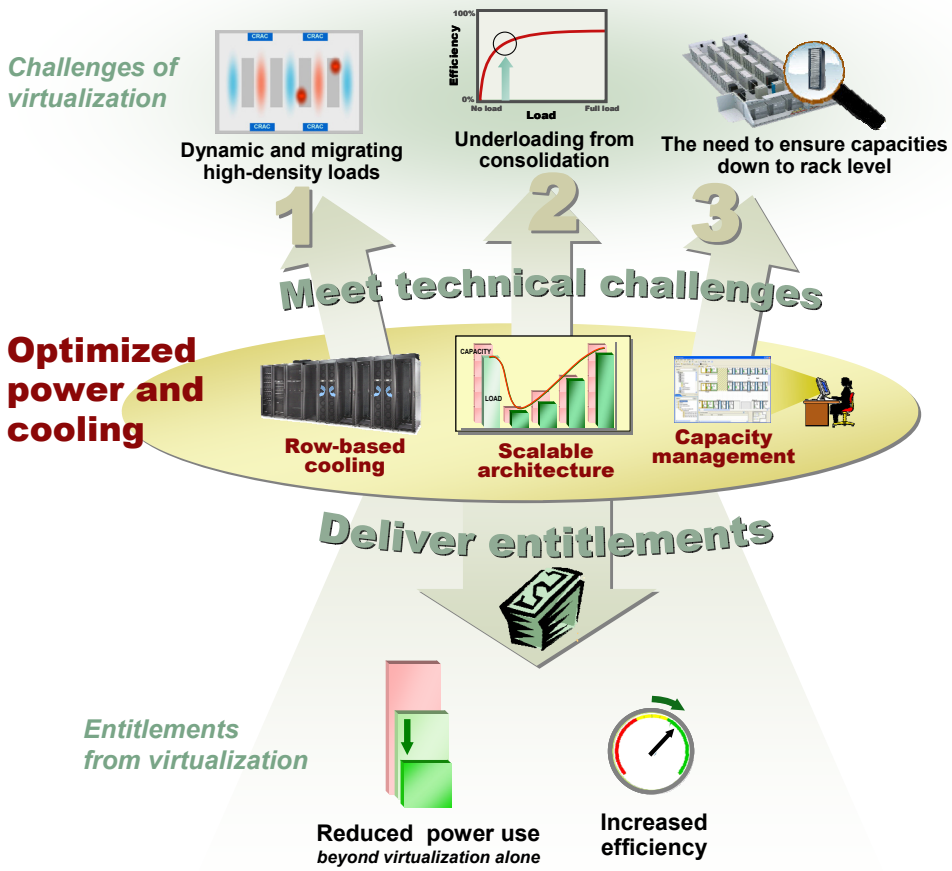
Table 1 *– Availability issues addressed by optimized power and cooling infrastructure*

Availability threat	Why?	How solved
Human error	Human error has historically been a significant cause of downtime in data centers. With more complexity and change – including changes that can't be seen – comes a greater risk of mistakes and oversights.	The systems described in this paper – such as cooling that responds to local demand, rack-level instrumentation, and capacity management – build intelligence into power and cooling, minimizing the need for human interpretation and intervention.
Unpredictable cooling	Traditional room-based cooling is not agile enough to handle the unpredictable dynamic high-density loads of a virtualized environment.	Row-based, managed cooling tightly controls delivery, not only in location but also in amount. Load conditions that threaten to exceed local cooling capacity can be proactively identified by the capacity management system, and corrective action can be taken.
Loss of cooling redundancy	Dynamic loads can locally raise cooling demand to the point where there is not enough capacity to cover scheduled or unscheduled downtime of cooling equipment.	Row-based cooling units can be installed to provide the desired redundancy for that row. The capacity management system can warn of insufficient or lost redundancy due to subsequent changes in loading.
Power overloads	With increased frequency of changes in power demand – due to reconfiguration of physical servers or the migration of virtual servers – comes increased risk of branch circuit loads creeping up close to the breaker-trip limit.	The capacity management system will warn of load imbalances *before* they pose an availability risk.

Conclusion

Virtualization is an undisputed leap forward in data center evolution – it saves energy, it increases computing throughput, it frees up floor space, it facilitates load migration and disaster recovery. Less well known is the extent to which the entitlement can be multiplied if power and cooling infrastructure is optimized to take advantage of the further savings opportunity offered by virtualization. In addition to the financial benefits obtainable, these same power and cooling solutions answer a number of functionality and availability challenges presented by virtualization. **Figure 17** summarizes the effects of the optimized power and cooling infrastructure described in this paper, both as it answers the specific challenges of virtualization and as it provides general performance entitlements.

Figure 17 *– Summary of effects from optimized power and cooling*



The three major challenges that virtualization poses to physical infrastructure are dynamic high density, underloading of power/cooling systems, and the need for rack-level, real-time management of capacities (power, cooling, and physical space). These challenges are met by row-based cooling, scalable power and cooling, and capacity management tools, respectively. All three of these solutions are based on design principles that simultaneously resolve functional challenges, reduce power consumption, and increase efficiency.

The comparison of pre- and post-virtualization power consumption involves two concepts relatively new to data center cost analysis. The first is *fixed loss* – the amount of power consumed by devices and systems regardless of load – which is responsible for the often surprising inefficiency of underloaded systems. The second is the distinction between energy *consumption* and energy *efficiency*, which can confuse a comparison of energy savings. Here is how these two concepts play in the virtualization discussion: Even without a parallel upgrade to power and cooling, virtualization will always lower the electric bill, but (1) not usually as much as might be expected because of the presence of fixed loss in power and cooling systems and (2) in spite of a reduction in power *consumption* by the data center, the data center *efficiency* (DCiE) is typically lower after virtualizing due to the inefficiency of underloaded power and cooling systems. This lowered efficiency indicates room for improvement in power and cooling systems – it is, in effect, a measure of the potential for extracting even more value per energy dollar.

When virtualizing, a parallel upgrade of power and cooling infrastructure will optimize both architecture and operation in a number of ways that safeguard availability, enhance manageability, lower power consumption, and increase efficiency. Properly designed physical infrastructure will not only provide solutions for the specific needs of virtualization, but can also raise both power density capacity and data center efficiency significantly *above* what they were before virtualization.

About the Author

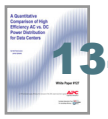
Suzanne Niles is a Senior Research Analyst with the APC Data Center Science Center. She studied mathematics at Wellesley College before receiving her Bachelor's degree in computer science from MIT, with a thesis on handwritten character recognition. She has been educating diverse audiences for over 30 years using a variety of media from software manuals to photography and children's songs.

Related APC White Papers

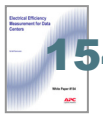
The Advantages of Row and Rack-Oriented Cooling Architectures for Data Centers

Power and Cooling Capacity Management for Data Centers

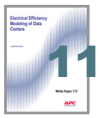
An Improved Architecture for High-Efficiency, High-Density Data Centers



Deploying High-Density Zones in a Low-Density Data Center

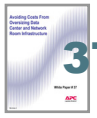


Electrical Efficiency Measurement for Data Centers



Electrical Efficiency Modeling for Data Centers

Implementing Energy Efficient Data Centers



Avoiding Costs From Oversizing Data Center and Network Room Infrastructure

Dynamic Power Variations in Data Centers and Network Rooms