DATA CENTER KNOWLEDGE

THE INFORMATION SOURCE FOR THE DATA CENTER INDUSTRY

EXECUTIVE GUIDE SERIES – PART 3

Data Center Energy Efficiency

by Julius Neudorfer

This is the third of a six part series of our Executive Guide whitepapers:

1. Data Center: Build vs. Buy
2. Total Cost of Ownership
3. Data Center Energy Efficiency
4. Creating Data Center Strategies with Global Scale
5. Custom Data Centers
6. Data Center Designs

Brought to you by

DIGITAL REALTY
Data Centre Solutions

Introduction

Data Center Energy Efficiency, third in the Data Center Executive Guide series

In today's world, it is almost impossible to discuss any business operation without considering efficiency. In the data center realm, maximizing the energy efficiency without impacting the reliability should be the goal of virtually every data center owner or operator. The path to that goal involves some basic design and practices issues, as well as more sophisticated methodologies.

This third part of the Executive Guide series will examine the more detailed aspects, such as the tradeoff of energy efficiency vs. redundancy, the potential energy saving by the expanded use of "Free Cooling" and the potential use of sustainable energy sources, as well as other data center specific energy related issues.

Executive Overview

Energy represents one of the most significant operating costs in the data center, as was discussed in part two of the Executive Series, Total Cost of Ownership (TCO). Moreover, based on current and foreseeable trends, the basic price of energy will continue to rise over time, and may become constrained as global demands rise, making energy use and efficiency a long term business priority.

Nonetheless, it is important to differentiate the price of power from overall energy use and efficiency, since direct prices may vary widely by market conditions, geographic location and the underling energy source (coal, natural gas, nuclear, hydro, as well as wind and solar, etc.). Moreover, the overall carbon footprint, which has become a high visibility, globally recognized issue, is based on the total quantity of energy consumed, as well as its source, not the price of the energy.

Carbon footprint and greenhouse gases are also becoming subject to governmental regulations and taxes. While not yet fully implemented in the US as yet, there are now 28 types of energy policies affecting data centers in 12 EMEA countries, according to a 2012 report released by The Green Grid. The report states;

> *"Because of the continued proliferation of data centres with energy intensive requirements, our industry is particularly affected by legislation driving efficiency of design, build and operations — above and beyond the business impacts of increasing costs of energy and of carbon."*

Data Center Energy Efficiency

Defining Data Center Efficiency

Energy Efficiency is impacted by many factors in a Data Center. However, before we delve into the details, it is important to define how energy efficiency is defined and calculated.

Common general definitions of "efficiency":

- ***The ratio of the effective or useful output to the total input in any system.***
- ***The ratio of the useful work delivered by a system to the energy supplied to it.***

When we think in terms of energy efficiency in the data center, we can easily measure the energy that the entire site uses and compare it to its "output" (processing data), which unfortunately is not so easily defined or measured.

From the overall level, data center energy usage can generally be divided into two primary categories, IT systems (hardware and software) and the required supporting infrastructure (power and cooling systems) of the data center facility itself.

This whitepaper will focus on understanding the factors that impact total energy use and the opportunities for improving energy efficiency in the data center.

IT System Efficiency

In the first category IT systems, it should be noted that Moore's law, (which is based on Gordon Moore's 1965 original whitepaper) predicted that chip densities would double every two years. In 1975 it was restated by Intel that processing power would double every 18 months due the improvement in transistor speed, as well as density. This prediction has essentially proven true over the last five decades. Performance has continued to increase exponentially with each new generation of IT hardware, and power usage has also increased as well.

Nonetheless, clearly defining a universally accepted metric of "useful work" for "computing" has been a highly debated topic for many years. Although, many people and organization have put a lot of effort into trying to quantify it, this has proven to be very complex and difficult to get a consensus of a methodology or commonly accepted metric of the overall IT system's "useful work" and therefore its energy efficiency.

Even within each of the major sub-system categories, such as hardware (Servers, Storage and Network), as well as the software (Operating Systems and Virtualization, as well as the myriad of standardized and custom Applications), there are widely differing opinions by vendors and users. Even related standards organizations have yet to create and fully agree to a commonly accepted set of metrics. In particular, hardware and software performance are interdependent on each other and therefore each unique combination of layers of software (operating system and applications) running on the various underlying hardware platforms, will be different in its actual overall throughput and energy use.

In the past, IT equipment manufacturers and their customers were focused more on maximum performance, not energy efficiency. While the newest generations of computing hardware have continued to increase in performance, they have also become highly focused on energy efficiency.

In fact it has been shown that in many cases the cost of energy for an older commodity server is higher over a three year period than the cost of the server itself.

This is especially true in older data centers when the cost energy for the supporting power and cooling infrastructure is added to the total energy cost *(see Understanding PUE in the facilities section, page 5)*.

To help simplify and comparatively quantify the hardware aspect of energy usage and efficiency and promote this concept, the US Environmental Protection Agency (EPA) instituted an "Energy Star" program for data center IT equipment. The Energy Star program initially introduced the first version of the Server specification in 2009 and continues to update and expand its list of included equipment. Currently, they are also working on Energy Star for Storage and Networking equipment specifications and it is expected to be released later this year or early 2013.

According to the US EPA ***"Computer servers that earn the ENERGY STAR will, on average, be 30 percent more energy efficient than standard servers".*** However, it should be noted that in some cases the energy requirements for Energy Star rated servers such a the widely used "1U" volume server, (with a single CPU, one hard drive and one power supply) can be substantially better than the average figure of 30%. In many cases it could use as much as 80% less energy, when compared to a comparably equipped 2–3 year old typical "commodity" server. As such, an IT hardware refresh can significantly reduce the total energy required in the data center, and in fact may offer a very short ROI timeframe.

Furthermore, according to the US EPA "If all servers sold in the United States meet the Energy Star specification, energy cost savings would approach $800 million per year and prevent greenhouse gas emissions equivalent to those from over one million vehicles." A list of Energy Star related resources and listed equipment can be found at **www.energystar.gov**.

The European Union (EU) also has an Energy Star program which is not quite as focused data center equipment, but which generally follows the precepts of the US EPA program. This is expanded relationship is related to the creation of the Global Data Center Energy Efficiency Task Force and the Global Harmonization Agreement of 2010, of which the US EPA, US DOE, The Green Grid and the EU are key members.

Ultimately, the goal is to continuously reduce the energy required (and related overall carbon footprint) to deliver a higher level of computing, storage and networking performance. This will help meet the ever rising demand for more information (including entertainment), delivered to more consumers and businesses over more devices, including televisions and mobile devices, while minimizing the energy required.

For example, a modern smartphone has far more raw computing capacity than an early mainframe of the last century, and it in turn accesses and requires far more data and network resources from the data centers almost everywhere. Whether it is for the financial industry, commerce, healthcare, social media or simply to watch movies over the internet, there is an ever rising requirement for more computing capacity and performance. Of course, this in turn drives up the need for more power in data centers everywhere.

As mentioned, hardware alone is only one part of the IT energy efficiency factor, the software (Operating Systems and Virtualization, as well as the myriad of standardized and custom Applications), are also critical factors in determining and improving the overall computing performance and therefore efficiency, as well as total energy use. In the past, there were high numbers of individual applications running on individual servers, which were not constantly or highly utilized. However, these servers typically used 60–70% of the full load power while sitting idle.

The introduction of virtualization software and server consolidation has helped to improve the ratio of idle servers drawing significant power with little not productivity. Moreover, one of the important features of Energy Star servers is active power management, which significantly reduces the power the server draws while idle.

Furthermore, the advent of the so called "Bladeserver" which is a central chassis with redundant shared power supplies (which improves energy efficiency), that can hold many individual server "blades", which allows packing more computing power into a smaller space. The bladeserver has also helped drive server consolidation and virtualization projects for many organizations.

The entire IT industry is now driven to improve energy efficiency. The IT hardware manufacturers have recognized that not only do they need to inherently improve the energy efficiency in their products, they are well aware that the largest use of energy in the data center facility itself is used by the cooling systems. They have worked to make their hardware more robust to allow it to operate at higher temperatures, which in turn reduces the cooling systems energy requirements. *(see The Impact of ASHRAE 2011 in the facilities section, page 8)*

This is not just done out of a purely altruistic motivation. In the past, as IT Hardware heat densities increased, it became a significant problem for cooling systems, and in many cases limited the installation of more IT equipment. The hardware manufacturers also realized that by improving their inherent energy efficiency and also decreasing the energy required to cool the equipment by being able to operate in a warmer environment, that more power will be available to support more IT equipment. This helped spur sales of more hardware. Nonetheless, the net resulting increase in energy efficiency benefits everyone.

Toward that end they worked to together with ASHRAE *(see ASHRAE sidebar for more details, page 8)* to help standardize the equipment categories based on the environmental operating ranges into four classes; defined as A1- A4, essentially each higher category can safely operate at higher temperatures and wider humidity conditions. Effectively, class A1 and A2 equipment commonly exist today and have been available for the last 5 years. New A3 equipment is beginning to be offered by some manufacturers now and A4 equipment is not readily available yet, but is on product roadmap.

Moreover, newer technologies such as Solid State Drives (SSD) has driven the trend to replace the traditional hard drive. The expanded use of SSD also helps reduce energy since inherently they use less energy that traditional hard drives. In addition, they can also operate at much wider environmental ranges than the more temperature sensitive mechanical spinning disk hard drives, thus requiring less cooling.

So what does this mean to the data center facility and it cooling system design and operation? Data centers have historically kept very tight environment conditions to help ensure the reliability of the IT equipment. This was originally driven by older equipments susceptibility to temperature and humidity changes as well as a very narrow range of "recommended" environmental conditions mandated by the equipment manufacturers themselves. While there IT equipment has clearly become more robust, the data center industry as a whole is more conservative and has not fully taken advantage of the cooling energy saving opportunities the new equipment offers.

A final note on IT energy usage; based on all these improvements, one would tend to assume that since each generation of IT equipment offers greater efficiency that the overall energy usage for IT systems would begin to decline. However, an insatiable demand for more features, applications and more data that is be generated, read and viewed by both consumers and businesses keeps IT system expanding, ultimately, driving up the need for more power for IT equipment.

Facility Energy Efficiency

In contrast to the difficulty in trying to quantify the energy efficiency of IT systems, the efficiency measurement of the data center facility infrastructure is now is a well defined metric. It is known as Power Usage Effectiveness (PUE), which was introduced by The Green Grid in 2008, The Green Grid is a global consortium of companies, government agencies, and educational institutions dedicated to advancing energy efficiency in data centers and business computing ecosystems. PUE became an internationally agreed upon metric in 2011, when the US Dept of Energy, EPA, European Union and Japan, agreed to it as a mutually satisfactory metric.

Understanding the PUE Metric – Power Usage Effectiveness

The basis of the PUE metric is relatively straightforward; it is the ratio of the Total Energy being used by the facility (including the IT Energy) divided by the Energy used by the IT Equipment. This is measured on an annualized basis. The range for PUE measurement is 1.0 (theoretically perfect – 100% efficient) with no upper limit (very inefficient). As can be seen by this example at a PUE of 2, the facility power and cooling systems have used as much energy as the IT equipment.

Example of PUE calculation: PUE = 2

Total Data Center Energy: 1,800,000 KWH

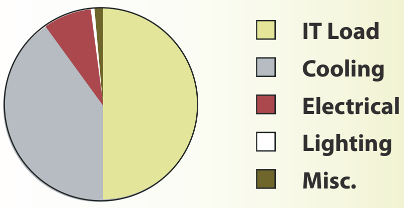
(Facility: 900,000 KWH + IT equipment: 900,000 KWH)

Total IT Equipment: 900,000 KWH

Typical Energy Use at a PUE of 2

Insofar as the physical data center facility and the related power and cooling infrastructure, the ever rising power requirements have strained many older data centers so that they are either incapable of fully supporting the higher power and higher density cooling loads of newer IT hardware, or they are just marginally supporting the loads, but are doing so very inefficiently.

Energy Use at a PUE of 2

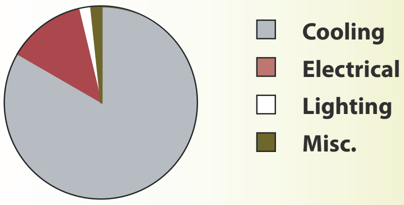


Cooling typically represents the majority of facility energy use and offers the greatest opportunity for improvement.

Previously data centers were primarily focused on reliability, and not on energy efficiency. In reality, *the average older data center used twice as much energy as was delivered to the computing equipment* (this represents a PUE of 2.0 or 50 percent operating efficiency).

In some cases the PUE of some older sites are even worse with a PUE of 2.5 – 3.0. This meant that they were using far more energy for power and cooling systems used to support the computing equipment than the energy actually used by the IT equipment.

Cooling Uses Majority of Facility Energy



New mainstream data centers are being designed and built currently that typically have much better operating efficiencies, with a PUE range of 1.3 – 1.6.

There are two main categories and related major sub-systems that comprise the energy usage of the data center facility; Power and Cooling

Power System Efficiency

The power system is considered the most critical element of the data center. Effectively it is a chain of key electrical sub-systems and components that are the essential elements for delivering uninterrupted, conditioned power to the IT equipment. The typical power chain consists of; Utility Transformer, Automatic Transfer Switch, Back-Up Generator, Distribution Switch Gear, Uninterruptable Power Supply (UPS) and the downstream Power Distribution System going to the IT equipment cabinets.

The above is a simplified list. Like a break in a link in a real chain, ***the failure of any component in the power chain could cause the loss of power to the IT equipment.*** In order to allow for continuous availability there must be redundant power system components, as well as power paths, to prevent downtime in the event of any equipment failure. Moreover, there are also system bypass systems to allow for proper maintenance of equipment (known as concurrent maintainability). This redundancy effectively means that there are typically twice as many power components in a Tier 3 or 4 data center.

While each of the components in the power chain incur a relatively small loss, the overall efficiency in of the power system is primarily affected by the efficiency of the Uninterruptible Power System "UPS", as well the overall smaller downstream power distribution losses.

In the case of redundant power, its greatest impact is on the energy efficiency of the UPS systems. Virtually all type of UPS, and in particular the "double conversion" online UPS system (which is the most commonly used type of UPS used in a data center), will be less efficient when operated at very low load levels (i.e. under 30% of rated load capacity). Nonetheless, this is the normal and common operational range for most 2N or 2(N+1) power systems which are the heart of upper tier level (tier 3 & 4) data centers.

It is not uncommon to see older UPS units in a redundant configuration operating at only 65 – 75% energy efficiency.

This severely impacts older data centers especially those with older technology UPS systems. It is not uncommon to see older UPS units in a highly redundant configuration operating at only 15-25% of their rated capacity. Their efficiency at these lower loads may be as low as 65-75%. Moreover, since the 25-35% of the lost energy is directly converted to heat, this requires that an additional amount of cooling system energy must be used to cool the UPS waste heat.

The US EPA now also has an Energy Star Program for UPS systems. This was just finalized and goes into effect in August 2012. New UPS units are more efficient overall (typically 92-95% at full load), but more importantly, are far more efficient at the lower load ranges. In comparison, a new UPS can still operate at 85-90% efficiency even when operating at only15-25% of their rated capacity.

In addition, ***a feature of the newest generation of UPS systems is the so-called "hybrid" or "eco" mode of operation.*** This permits the industry standard "double-conversion" UPS to operate in an internal by-pass mode, which allows it to operate at 98-99% efficiencies, even at very low loads. The UPS will still automatically revert nearly instantaneously (typically less than 4 ms) to full "double-conversion" mode if there are any power anomalies, to ensure that the computing equipment will not be impacted. Not everyone in the data center community is completely comfortable with this mode of operation. However, it should be noted that while this feature will be seen on most of the newest units, it does not have to be operated in that mode, the UPS can be configured to operate in the more conservative full "double-conversion" mode at all times.

New UPS units are also now being offered without transformers (which were required for older technology UPS systems), to improve efficiency. Besides the eliminating transformers from the UPS, the use of new higher efficiency transformers in the downstream distribution can also improve the overall power chain efficiency.

It should be noted that the greatest risk of an outage in a data center is primarily from power systems failure. Unlike broader environmental conditions, which may or may not impact long term reliability of IT equipment, any issue that can potentially disrupt power can cause an immediate outage in a data center. Toward that end, the entire power system must always be designed with a mandate for maximum availability and high levels of fault tolerance. ***While improved energy efficiency is a worthy goal, it should never be done at the expense of overall reliability.***

Cooling System Efficiency

Of all the factors that can impact the energy efficiency cooling represents the majority of facility related energy usage in the data center, outside of the actual IT load itself.

While there are several variations on cooling systems, they generally fall into two categories, the Computer Room Air Conditioner "CRAC" wherein each unit has its own internal compressor, and Computer Room Air Handler "CRAH" which is primarily a coil and a fan which requires externally supplied chilled water. From an energy efficiency viewpoint, the CRAH which is usually supplied by a water cooled central chilled water plant, is more efficient than an air-cooled CRAC units. However, the air-cooled CRAC unit has one advantage over a centralized chiller system; they are all autonomous and therefore offer inherent redundancy and fault tolerance, in that there is no single point of failure (other than power failure).

Regardless of the type of cooling system, the amount of cooling required and therefore the energy required is reduced if the data center temperatures can be increased. Moreover, tightly controlled humidity is another area where a lot of energy is used, in many cases quite needlessly.

So what does this mean to the data center facility and its cooling system design and operation? Data centers have historically kept very tight environmental conditions to help ensure the reliability of the IT equipment. This was originally driven by older equipments susceptibility to temperature and humidity changes as well as a very narrow range of "recommended" environmental conditions mandated by the equipment manufacturers themselves. *(see ASHRAE sidebar for more details, page 8)*

In 2011 ASHRAE in conjunction with the consensus of major IT equipment manufacturers radically hoped to change the direction of the data center industry's view toward cooling requirements by openly stating that:

> *"A roadmap has been outlined to facilitate a significant increase in the operational hours during which economizer systems are able to be used,* ***and to increase the opportunity for data centers to become "chillerless," eliminating mechanical cooling systems entirely,*** *in order to realize improved Power Usage Effectiveness (PUE)."*

The 2011 version ASHRAE's guidelines, openly endorsed "free cooling" *(see Free Cooling sidebar, page 9)*. This would have been considered heresy by many only a few years ago, and some are still in shock and have difficulty accepting this new outlook toward less tightly controlled environmental conditions in the data center.

The opportunity to save significant amounts of cooling energy by moderating the cooling requirements and the expanded use of "free cooling" is enormous. However, due to the highly conservative and risk adverse nature of the industry this will take a while to become a widespread and common practice. Clearly some have begun to slowly explore raising the temperatures a few degrees to gather some experience and to see if they experience any operational issues with the IT equipment. Ultimately, it is a question of whether the energy (and cost) saved, is worth the risk (perceived or real) of potential equipment failures due to higher temperatures (and perhaps wider humidity).

There are clearly some legitimate reasons to keep lower temperatures; the first is a concern of loss of thermal ride-through time in the event of a brief loss of cooling, this is especially true for higher density cabinets, where an event of only a few minutes would cause an unacceptable high intake IT temperature. This can occur during the loss of utility power, and the subsequent transfer to back-up generator, which while it typically takes 30 second or less, will cause most compressors in chillers or CRAC units to recycle and remain off for 5–10 minutes or more. While there are some ways to minimize or mitigate this risk, is a valid concern.

The other concern is also another common issue; the wide variations in IT equipment intake temperatures that occur in most data centers due to airflow mixing and bypass air from less than ideal airflow management. Most sites resort to overcooling the supply air so that the worst areas (typically end-of-aisles and top of racks) of higher density areas do not overheat from re-circulated warm air from the hot aisles.

However, if better airflow management is implemented to minimize hotspots, it would allow intake temperatures to be slowly raised beyond the conservative 68–70°F. This can be accomplished by a variety of means such as; to the spreading out and balancing rack level heat load, and adjusting the airflow to match the heat load, as well as better segregation of hot and cold air via blanking panels in the racks and the use of containment systems. If done properly, it is more likely that within one to two years, 75–77°F in the cold aisle would no longer be a cause for alarm to IT users. The key to this is to improve communications and educate both the IT and facilities management about the importance of air management and the opportunity for energy savings, without reducing equipment reliability.

The Potential Energy Efficiency Impact of New 2011 ASHRAE Expanded Thermal Guidelines

The American Society of Heating, Refrigeration and Air Conditioning Engineers (ASHRAE) is considered the reference standard for all types of buildings in the US. In particular, there is committee within ASHRAE known as the Technical Committee 9.9, (TC9.9) which is focused solely on the data center. TC 9.9 created the first edition of its now widely accepted Thermal Guidelines for Data Centers in 2004. It defined the recommended and acceptable temperature and humidity ranges for computer equipment operating in the data center.

For many years prior to the first TC 9.9 Thermal Guideline, the major IT equipment manufacturers specified the environmental conditions, and in most cases, they were very narrow and rigid, (typically 68–70°F and 45–55% Relative Humidity "RH"), since some computer equipment were very sensitive to environmental conditions. Some of these tight requirements could be traced back to the days of the original mainframe computers and the fact that there was paper everywhere in the data center, including punch-cards and paper tape that were used to enter the data into the computer, as well as huge reams of paper fed into high speed mechanical printers. Even small changes in temperature, and especially humidity, could cause paper jams — a minor catastrophe at the time.

The original 2004 guidelines were written to still account for the older legacy systems, and while they had not used punch-cards for over 20 years, many computer systems were still somewhat sensitive to temperature and humidity. And so even in 2004, data centers were still designed and operated at or near 68°F and 50% RH.

In 2008, ASHRAE TC 9.9 released the second edition, which broadened and raised the recommended environmental ranges up to 80°F and 60% RH. While this was a step forward to improve cooling systems energy efficiency, it was still relatively conservative in relation to the newer computing equipment's tolerance to broader and higher temperatures and humidity. In reality, despite the new guidelines, most data center designers and operators continued to maintain tight temperature and humidity conditions as an industry norm.

In 2011 ASHRAE basically upended the data center industry with release of the 3rd edition entitled "Thermal Guidelines – Expanded Data Center Classes and Usage Guidance", which openly declared that whenever possible data centers should consider using "Free Cooling" to save energy. So called "Free Cooling" refers to avoiding or minimizing the use of mechanical cooling systems to cool the data center by taking advantage of ambient conditions whenever possible. The 2011 edition was the result of the collective work of virtually all the major computer equipment manufacturers. Data Center Energy Efficiency and Free Cooling became the underlying theme and it reflected the environmental ruggedness of the newer IT equipment. It allowed the operation of data centers at much warmer temperatures than previously suggested, with the support of IT equipment manufacturers. By recommending "Free Cooling" via the expanded environmental ranges, it offered data center designers and operators a significant opportunity to save cooling energy.

What is "Free Cooling"

Over the last few years, we have used and heard the term "Free Cooling" with increasing frequency. However, the correct term is an economizer system. The purpose of the economizer is to reduce the amount of run time and energy used for "mechanical" cooling (typically a compressor based system). These economizer systems can be part of a water cooled evaporative system or an air cooled system or even a combination of both.

Most recently, there has been an increased interest in so called "fresh air" or "air-side" economizers systems. Previously, if a data center had an economizer system it was most likely a "water-side" economizer system used in conjunction with a chilled water system. The concept of a direct "air-side" economizer is simple; just bring in outside fresh air into the data center when the outside temperatures are within the temperatures required by the IT equipment, to "cool" the data center and then extract the hot air from the IT equipment out of the building. Only when the outside air is too warm is mechanical cooling required.

However, before the release of the 2011 ASHRAE Expanded Thermal Guidelines, this concept of bringing in outside air was considered heresy in the data center world. Typically the mainstream data center is designed to be a closed system with very little outside air permitted, to keep environmental conditions stable (temperature and humidity) and to avoid contaminants and dust from affecting the IT equipment.

The issue of contaminants and dust can be mitigated by filtering the air before entering, and the modern IT equipment now has a broader environmental range. This makes direct "fresh air" a very interesting opportunity to radically reduce the energy required to "cool" a new data center.

In particular, the direct "fresh air" economizer is making those ultra-low PUE headline generating numbers possible. Although primarily associated with Internet Search and Social Media organizations, whose SLA's are not really defined, nonetheless these energy saving low PUE numbers are real. Of course, your own organization may have very well defined and tight SLA requirements, especially if you are a large enterprise or financial firm, or a co-lo whose customers expect a very stable environment.

So is it really possible to use outside air for "free cooling", with wider temperature and humidity swings, and yet still have the "safe" and reliable operating conditions that a "traditional" closed airflow loop cooling system offers?

The simple answer is yes, however it requires that the data center be designed specifically with that capability. However, it cannot be easily retrofitted into an existing building. Nonetheless, the energy efficiency gains are quite attractive. For reference, The Green Grid published a 2012 updated set of maps based on the ASHRAE 2011 Expanded Thermal Guidelines, which showed that even when staying within the A2 "Allowable" environmental parameters (50–95°F, 20–80% RH, virtually all modern IT equipment except tape, is rated A2), that 75% of US sites could operate on air-side economizers for 97% of the year. Moreover, 99% of European locations would be able to operate on "free cooling" all year long.

Note that it is not necessary to go to the extreme edges of the "allowable" range. It is possible to stay within the "recommended" range and still get substantial energy savings. In many moderate climates it is possible to save substantial cooling system annualized energy costs, while maintaining a conservative environmental range. As stated previously, while the IT equipment has clearly become more robust, the data center industry as a whole is more conservative and has not fully taken advantage of the cooling energy saving opportunities the new equipment offers.

Factors Impacting Energy Efficiency

Site Location

Of course picking a site location that is physically secure and has reliable access to power, water and communications is an important first step. Moreover with the rising awareness of "free cooling" The average and highest temperatures (as well as humidity) will directly impact the energy efficiency of the cooling systems.

Of course, while not directly related the facility energy efficiency per se, attention to the cost of power and it fuel source should not be overlooked. Energy costs are highly location dependent and are based on local or purchased power generation costs (related to fuel types or sustainable sources such as, hydro, wind or solar), as well as any state and local taxes (or tax incentives) incentives which can provide lower costs, energy efficiency rebates or tax benefits.

Facility Occupancy and Power Loading Rate – Design vs. Actual

The energy efficiency of any data center will be directly affected by the actual percentage of the design load being used. The lower the load utilization compared to its design maximum, the lower the efficiency. This is directly related to its occupancy rate. If the site will not be heavily occupied for the first several years, a modular design should be considered to mitigate the impact of underutilization. In addition, most data centers never operate a 100% of design load capacity primarily for ensuring equipment reliability and maintaining uptime. Depending on the organization culture, typically systems are operated at no more than 80-85% of design ratings (some may push to 90%) before it is considered "full". This is a necessary, but prudent compromise of reliability vs. energy efficiency.

Oversize Design Capacity Impacts Efficiency

When deciding on the design capacity of the data center there are many competing factor that influence the decisions. The fear of making it too small and running out of space or power in only a few years is a very realistic scenario and fear. In recent years the growth in computing power demands and power density has made many data centers that were built less than 10 years ago functionally obsolete, a real risk of then looking for additional space or power that is not readily resolvable with a dedicated single site. Conversely, oversizing will mitigate that risk, but decrease the energy efficiency.

Modular Design

One method to mitigate the potential of over or undersized data centers is modular design. Capacity Planning and modular capacity designs can help mitigate the risk of capacity or functional obsolescence. In some designs, the total space and utility capacity is designed and built upfront, but only individual sections are fully outfitted with the UPS, generators and cooling equipment. This saves both upfront capital cost and recurring maintenance expenses. Moreover, it also help improve energy efficiency at each stage, since the smaller sections are more fully occupied and operate at a higher efficiency. This modular design still allows for planned expansion, without the energy efficiency penalty of underutilization .

Proper and Continuous Maintenance

Of course, regardless of how well designed and built, the equipment must be maintained for proper operation. In the past, this was mostly driven by the need to ensure reliability to avoid system failure. Today, while this is still the key requirement, ensure optimum energy efficiency is also part of the maintenance goals. This is particularly true for cooling systems, whose efficiency and effectiveness falls off rapidly, if filters become clogged and cooling towers are not rigorously cleaned, as well as other required maintenance and system optimization for changing internal heat loads and external ambient weather conditions.

Power Density

Power density which is reflection of how much computing equipment can be placed in each rack. A data center with a lower power density would mean that you may need to use more racks (and whitespace) to house the same amount of computing equipment than at a higher density site. Power density is typically expressed in two ways; watts per square foot or Kilowatts (kW) per rack, or sometimes both. This is primarily based on the design and type of the data center cooling system. Many older data centers cannot effectively or efficiently cool more than 5 kW per rack (some even less), and in some cases their energy efficiency goes down beyond 3 kW per rack. Even today, not all newer data centers can accommodate medium (5–10kW per rack) or "high-density" racks which require 10 kW or more per rack.

Other Efficiency Related Metrics

Energy efficiency and usage its related sustainability issues has seen rising awareness in the consumer, business and government sectors and continues to become more important as the increased public focus on data center and their growing energy demands spreads globally.

Beside PUE, The Green Grid also released several other metrics, three of which have particular long term significance to the data center industry:

Carbon Usage Effectiveness "CUE"

As its name implies, Carbon Usage Effectiveness is based on the annual CO_2 generated by the energy used by the data center. It is based on the total of CO_2 emitted by the local powerplant (and the type fuel used) to generate the power, as well as any grid supplied energy and the mixture of fuels or sustainable resources (Wind, Solar, Hydro, etc.) used to generate the energy reused by the data center.

Water Usage Effectiveness "WUE"

Water is used as part of the cooling process in many data centers. Large water cooled data centers can consume millions of gallons per month. Until recently this has never been considered as part of it "efficiency". However, it does take energy to process and pump water to the data center. Fresh water is rapidly becoming a constrained resource in many areas, which can affect data centers that depend on water in the future.

Energy Reuse Effectiveness "ERE"

This is the ultimate efficiency issue. Regardless of how much we improve the facility PUE, virtually all energy used by the computing equipment is turned into waste heat. This of course impacts global warming. There are opportunities to make effective reuse of the waste heat from data centers for other purposes, such as heating adjacent or nearby buildings. This can obviously have very positive environment benefits, as well as potential economic advantages from reduced heating energy costs in other buildings.

Energy Monitoring and Management

As the need to improve energy efficiency became apparent, it also became evident that not all data centers had the basic systems in place to monitor or measure the actual energy efficiency. Early efforts in many cases were simply "snap-shots" based of a one time power reading which was then used to calculate the site's PUE. In many cases the readings were taken under unusually or artificially low energy conditions (such as the coldest days of winter when the cooling systems were only minimally operating). This created many false or highly inaccurate PUE claims that were sometimes used by the marketing departments of some organizations to hype their data center's perceived efficiency. Today, the current version of the PUE metric is based on measuring the annualized energy used by a data center, and therefore represents a true and realistic picture of the facility's energy efficiency.

Virtually every major data center hardware and software vendor now offers a multitude of tools to monitor and manage the data centers energy usage and efficiency. This has become a major category in the industry; known as Data Center Infrastructure Management (DCIM). Rising demand and commitment to these tools by both the vendors and the customer's acceptance, underlines the importance and economic benefit of monitoring and improving energy efficiency. However it begs the question: Who should monitor and manage Energy Efficiency, IT or Facilities? The long term answer is both groups, working together.

The Bottom Line

In order to maximize the opportunity to save energy in the data center, it will take a holistic approach by both the organization's IT and the facilities departments working together. Only by exchanging knowledge and mutual cooperation will the newest and most efficient technologies be successfully deployed. The driving force for this needs to come from senior management and be driven from the top down.

There is no doubt that energy usage is rising in the data center, and is likely to continue to grow for the foreseeable future. Moreover, the price of energy is a significant and rising cost of operating a datacenter, and that it should be given serious consideration when selecting a site and designing its infrastructure systems. Moreover, the growing concern for sustainable source energy is also a factor, which can impact the mainstream data center industry.

It is important to note that we have been discussing the use of industry standard IT equipment typically used by most enterprise organizations, operated in higher tier level data centers. This should not be compared to some Internet search and social media organizations who may use custom built servers for hyper-scale web server farms and are housed in data centers that operate with different environmental parameters. Moreover, they tend to have lower power and cooling systems redundancy at the facility level (in come cases comparable to Tier 1 or 2). Unlike typical enterprise organizations, which can be severely impacted by a critical IT system failure, their IT architecture is radically different in that they are more tolerant of equipment and systems failures and can redirect web user requests to other sites in the event of a failure. This allows them the liberty to have ultra low PUEs, which while enviable (and can make headlines), should not be directly compared to a highly redundant enterprise data center.

However, for enterprise class data centers and co-location operators, the highest levels of availability and system uptime are still the first and highest priority. In order to achieve that goal, a high level of redundancy will extract a small but necessary efficiency price. There are now tremendous gains in efficiency possible, when compared with previous generation data centers. This is a relatively minor compromise necessary to support the system and component level redundancy to ensure fault tolerance, which also allows for concurrent maintenance capabilities. However, as was discussed in this whitepaper, there are various ways to significantly reduce or mitigate the rise in energy consumption, without impacting the reliability and availability of the computing systems.

So when evaluating the energy usage vs. risk, for data center facilities design, equipment and operations, investigate carefully and weigh the projected energy usage and cost saving vs. your own organization's tolerance to exposure of potential downtime.

This is the third of a six part series of Executive Guide whitepapers

Next in the series is:

4. Creating Data Center Strategies with Global Scale
5. Custom Data Centers
6. Data Center Designs

Julius Neudorfer Bio

Julius Neudorfer is the CTO and founder of North American Access Technologies, Inc. (NAAT). Based in Westchester NY, NAAT's clients include Fortune 500 firms and government agencies. NAAT has been designing and implementing Data Center Infrastructure and related technology projects for over 20 years.

Julius is a member of AFCOM, ASHRAE, BICSI, IEEE and The Green Grid, as well as a Certified Data Center Design Professional "CDCDP" designer and instructor. Most recently, he is also an instructor for the US Department of Energy "Data Center Energy Practitioner" "DCEP" program.

Julius has written numerous articles and whitepapers for various IT and Data Center publications and has delivered seminars and webinars on data center power, cooling and efficiency.