TechTarget

E-Guide

Energy Essentials: Rethinking Power and Cooling for the Modern Data Center

Tactics for creating an energy-efficient IT strategy

Contents

Traditional infrastructures weren't designed for *today's new workloads. As a result, many IT professionals have been left frustrated by issues related to troubleshooting, capacity planning, power and cooling, provisioning and configuring their complex, heterogeneous environments. In this e-guide, hear from the experts at SearchDataCenter.com as they offer practical advice for improving data center designs to optimize energy efficiency across the IT landscape. Learn more about current energy concerns and latest techniques for reducing consumption in your data center.*

New data center cooling strategies to improve efficiency, lower costs

By: Clive Longbottom

What are your data center cooling strategies? Do you run a few computer-room air conditioning units continuously to push cooled air through perforated plates in your raised floor to maintain a suitable temperature for all the racks in the room? If so, you are likely wasting large amounts of energy -- and money. There are several new cooling approaches and technologies that could give a short capital payback coupled with ongoing financial benefits.

First, make sure you're not running the data center temperature too low. The American Society of Heating, Refrigeration and Air conditioning Engineers (ASHRAE) has been tracking recommended and allowable temperatures for defined types of data centers since 2004.

The maximum high-end temperature in 2004 was set at 77 degrees Fahrenheit (25 degrees Celsius). By 2008, it had risen to 81 F (27 C). By 2011, ASHRAE had created a range of data center types, and although the recommended temperature stayed at 81 F (27 C), it raised the maximum allowable temperatures – the temperatures at which increased equipment failure rates may be seen -- to 113 F (45 C). By matching need against risk, a

Contents

higher-temperature data center requires far less cooling; this leads to fewer computer room air conditioning (CRAC) units and less energy used.

The easiest way to save money is to reduce the number of running CRAC units. If half the amount of cooling is required, turning off half the CRAC units will give a direct saving in energy costs -- and in maintenance costs. Using variable-speed instead of fixed-speed CRAC units is another way to accomplish this, where the units run only at the speed required to maintain the desired temperature. The units run at their most effective levels only when they run at 100%, and some variable speed systems don't run at a fully optimized rate when operating at partial load.

Running standard, fixed-rate CRAC units in such a way as to build up "thermal inertia" can be cost- effective. Here, the data center is cooled considerably below the target temperature by running the units, and then they are turned off. The data center then is allowed to warm up until it reaches a defined point, and the CRAC units are turned back on. Through this process, the units are run at full load and are operating at their highest operational efficiency. When they are turned off, they are running very energy-efficiently -- there is no draw of energy at all.

Nevertheless, straightforward volumetric approaches to cooling a data center remain wasteful, no matter how the initial cooling of the air is carried out; the majority of the air being cooled won't come into close enough contact with any IT equipment to cause any effective cooling.

Keeping cool with air

The use of hot aisles and cold aisles requires less volume of cooling air if they're set up properly (see Figure 1).

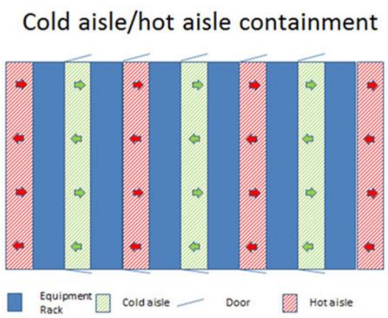

Figure 1. Hot/Cold Aisle

The spaces between facing racks are enclosed, with a "roof" placed over the racks and doors at either end. Cold air is

Contents

blown into the enclosed space. Blanking plates are used to prevent cold air leaking from within the racks. Ducts direct cold air to the hottest parts of the equipment.

The hot air either vents into the data center, then to the external air, or it is collected and used for heating other spaces. The hot air can also be used for heating water via a heat pump. These systems can be highly engineered, or they can be implemented quite effectively through a home-grown approach using polycarbonate to cover the racks as a roof and polypropylene sheeting as the doors at each end.

The goal is to make each rack its own contained system, so that the volume of air that requires cooling is minimized even further. Cooling can be engineered and targeted even more. This is where things like Chatsworth Towers come in. These self-contained systems have a 19-inch rack inside, but take cooling air from bottom to top without the air touching the rest of the air in the data center.

In certain climates, running at higher temperatures could allow the option of using free air cooling without the need for CRAC units. For example, if the choice is to run a data center at 86 F (30 C), an external air temperature of lower than 77 F (25 C) might be cool enough to require no additional cooling -- as long as moisture levels remain within required limits.

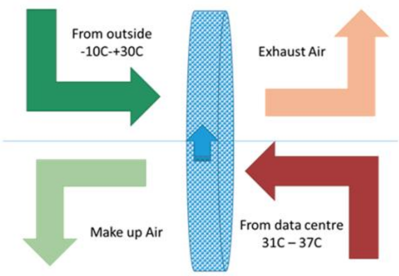

Figure 2. The Kyoto

However, the basic approach of simply ducting external air can lead to inefficiencies, such as thermal hot spots or dust and contaminants getting into the data center. This is where new designs, such as the Kyoto Wheel, come into play (see Figure 2).

In this scenario, a wound, corrugated metal wheel approximately 10 feet (3 meters) in diameter rotates slowly through a two-compartment space. The hot air from the data center flows through one space, transferring its heat to

Contents

the metal in the wheel. Cold external air flows through the other and takes the heat out of the metal and exhausts it to the outside air. The cooled air from the data center is fed back to the data center for use in cooling the equipment.

The data center loop is enclosed, and the small volume of air that gets transferred as the wheel rotates ensures that only very small amounts of particulates or moisture are mixed from one compartment to the other, with the wheel itself acting partially as a filter.

The benefit here is that the low-speed fans and motors used by the Kyoto cooling method require little maintenance, and the overall system runs with very small amounts of energy, often from solar power and a battery backup. Such a system can last for many years -- it is expected that 25 years will be a low-end lifetime -- and maintenance can involve just a quick clean of the wheel every few months, along with general motor maintenance.

When there's no chill in the air, try water

Aside from these air-based methods, another approach is adiabatic cooling, using the cooling effect of water as it evaporates (see Figure 3).

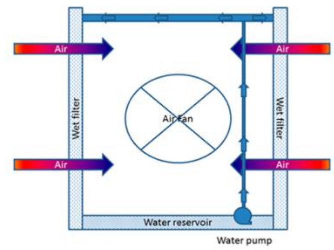

Figure 3. Water Cooling

Water cooling is effective in warmer climates, where direct environmental heat can be used against wet filters to evaporate the water and cool the air being pulled through the filters. This is a two-chamber system, with the filters providing the break between the outside air and the internal data center. However, the filters need to be changed on a regular basis to remove particulate contaminants. In addition, the air's moisture content may need to be adjusted to prevent condensation on IT equipment.

For companies that want to run at extreme equipment densities with high thermal profiles in warm climates, direct water cooling could be one way of solving the problem. IBM has used water cooling in the past, but has

Contents

managed to advance the method to extraordinary levels in its Aquasar and Liebniz SuperMUC supercomputer systems. The system gets around the problems that come with mixing water and electricity in a data center: Negative pressure is used to suck the water around the system, instead of pumps being used to push it. Therefore, if a leak occurs, air is pulled into the system instead of water coming out into the data center. Advanced sensors are used to identify rapidly where a leak has occurred, and modular construction allows for repairs while the rest of the system continues to run.

The system uses a hot water inlet for the cooling liquid, which might seem strange. But in a highly targeted system, water at a temperature more than 86 F (30 C) can cool CPUs to within operating parameters; the outlet's water temperature can be around 113 F (45 C). High-temperature water coupled with heat exchangers makes the hot water available in other parts of the building. Besides lowering energy usage by around 40%, this system can lead to further savings in the energy used to heat water for the rest of the building.

Give hardware a cooling bath

For even more thorough liquid cooling, there are fully immersive systems. Companies such as Iceotope and Green Revolution Cooling provide systems that cover the whole server -- or other pieces of IT equipment -- with a non-electricity-conducting but high-heat-conducting liquid to take heat away from any component in the equipment. These systems are ideal for GPU [graphics processing unit] servers with hundreds of cores running in high-density configurations, or for massive computing densities with a high density of hot-running CPUs, and they can deal with more than 100 kW per bath -- which is essentially an enclosed rack on its side. Some immersive systems are being run at a liquid temperature of 140 F (60 C).

Fans are unnecessary in an immersive system, which saves additional energy. Because the liquids used are far better at removing heat than air or water, hardware can run at higher temperatures and allow heat recovery from the liquid to provide heat for the rest of the building.

The data center is cooled, but now what?

Contents

These systems provide the main ways of providing cooling based around different needs and differing environmental conditions. Monitoring must be in place to supervise all thermal aspects of the data center.

This is where data center infrastructure management (DCIM) comes in. The use of thermal sensors and infrared detectors can build a map of hot spots. Computational fluid dynamics enables what-if? scenarios to see how new cooling flows and how different inlet temperatures will work with different systems.

Once the DCIM is in place, continued monitoring ensures hot spots are rapidly identified, which allows systems to be slowed down, turned off and replaced as necessary. In many cases, the sudden appearance of a hot spot indicates an imminent failure in the equipment. Being able to pick this up before it happens means that systems uptime and availability stay high.

The world of the data center continues to change, and taking an old-world view of cooling is one guaranteed way to waste money. Combining newer guidelines with newer approaches results in more effective cooling for much less in capital, running and maintenance costs.

Five questions on data center environmental Monitoring

By: Bill Kleyman

The data center should neither a sauna nor the Sahara be. Environmental sensors can help monitor optimal temperature and humidity levels to keep the data center equipment working reliably.

Safety controls such as fire suppression equipment and water detection sensors also need to be located properly and routinely tested to protect the company's investments. Our expert Bill Kleyman shares his expertise on the proper data center environmental monitoring techniques in this Q&A.

Q. What parameters should be part of a data center environmental monitoring scheme?

Contents

Bill Kleyman: A data center environment monitoring scheme can take many different forms. It all depends on the type of organization. Many companies demand high levels of security and visibility into the server room's physical environment. An administrator should try to monitor common environmental factors including:

- **Temperature:** No server or server rack should operate above a maximum inlet temperature. Many administrators look for rack air flow exhaust metrics, internal system temperatures and even CPU temperatures. The more information that is provided by the temperature control mechanisms, the faster an engineer can nip issues before they become serious problems.
- **Humidity and water control:** Water leaking from chiller lines or other sources can damage electronic systems. Similarly, excessively humid air can condense into liquid water that poses an equal danger to servers and systems. Deploy water sensors and humidity sensors at strategic locations both inside and outside of the rack.
- **Aisle conditions:** Containment has emerged as an important part of cooling management, so administrators should distribute environmental sensors within the hot and/or cold aisles. This can offer a broader picture of what's happening within a cooling zone.
- **Fire suppression monitoring:** Many large data centers have a comprehensive fire suppression system, but these systems can sit idle for years. Will they work if a fire starts? Short of testing the mechanisms, administrators should have an active monitoring system to ensure the equipment is working.
- **Static electricity sensors:** Excessively dry air, as well as improperly grounded equipment or personnel, can accumulate potentially harmful levels of static charge. Static electricity monitoring equipment located at intervals around the data center can report on the presence of potentially large charges.
- **Server room and rack entry:** This is all about physical security: doors opening and closing. Room and rack entry sensors send alerts

if the data center or rack has been improperly entered. Advanced environments can activate a camera that points to the rack where the cage has been opened.

Contents

Q. Are there any established guidelines that specify the server room environment? Certainly ASHRAE makes recommendations for temperature and particulates, but are there any industry standards for overall environmental conditions?

Bill Kleyman: It's hard to get specific with environmental monitoring best practices because each server room has different sizes and equipment and, therefore, different requirements. However, there are core environmental conditions that should be observed. The factors you measure will depend on the size and complexity of the infrastructure.

- **Temperature:** Gauging temperature will always be a key component. The optimum range for equipment stability had been recommended between 70 to 74 degrees Fahrenheit (21 to 23 degrees Celsius). In 2008, ASHRAE increased the range to 59 to 89.6 degrees Fahrenheit. In 2011, ASHRAE raised the limits again — for the very latest equipment — to 41 to 113 degrees Fahrenheit.
- **Humidity:** A humidity sensor is fairly standard in any sized environment. Relative Humidity (RH) is the ratio of moisture in a given sample of air at a given temperature when compared to the maximum amount of moisture the sample could contain at that temperature. The recommended RH is between 45% and 60%. More humidity can condense into liquid water, while less humidity can cause electrostatic discharge (ESD) — either scenario can damage equipment.
- **Water:** There should be no liquid water standing in a data center, so sensors at low points within racks and aisles can report on leaks and condensation problems.
- **Airflow:** Maintaining good airflow is vital for both temperature and humidity control. Good airflow recommendations vary depending on the size of the environment, and are expressed in volume as cubic feet per minute (cfm). The actual volume of air required is related to

Contents

the moisture content of the air and the temperature difference between the supply air and return air. Avoid turbulent airflow as this will be felt as a draft. This is where size of the environment becomes very important: In a dense server room, the number of air changes per hour may be several times greater than that of a smaller environment.

- **Computer room air conditioner/handler:** CRAC units must be monitored constantly. This includes supply and return temperatures, internal humidity statistics and air-loss percentages. Any faults in this unit need to be dealt with immediately.
- **PDU and electrical system status:** Electrical circuits within the environment should be watched continuously for unexpected fluctuations or disruptions. Any disruptions would be a serious emergency (not to mention a potential hazard to the facility). In addition, the data collected from power distribution locations is often used to calculate power usage effectiveness (PUE).

Q. What about environmental sensors? How durable and reliable are today's environmental sensors? Should sensors be maintained, tested or replaced, and if so, how often?

Bill Kleyman: Environmental sensors really report on the "health" of the overall facility, but no sensor is guaranteed to work forever. That's why it's important to have a redundant sensor environment. Intelligent data center monitoring tools will observe all of the sensors and can look at multiple sensors at the same time in case one has failed. Sensor redundancy can eliminate false-positives when a sensor fails. Environmental monitoring also requires careful alerting: if a sensor fails, the right administrator and technician must be notified promptly. This is the same if the device starts to post incorrect information or triggers false alarms.

Q. How should environmental sensors be placed? Are there any tools to help optimize sensor placement in and around servers, or is this still a manual hit-and-miss process?

Contents

Bill Kleyman: Since each environment is unique, there are no tools to determine the optimum sensor placement. A certain amount of empirical trial-and-error is a normal part of environmental sensor placement. However, working with a data center and HVAC professional can help an organization plan the best deployment. There are a few general placement guidelines:

- Temperatures will be greater at the rear and top of a rack, so measure exhaust temperatures and air flows there.
- Temperatures will be lower at the front and bottom of a rack, so measure intake temperatures and air flows there.
- Humidity will be highest in warmer air which can hold more moisture, so place humidity sensors higher in open areas away from any distinct heating or cooling sources.
- Place water sensors in low locations such as along rack bottoms or near drain openings, usually below any raised floor.

Q. Let's talk about integration. How do environmental monitoring tools interoperate with administrator-focused systems (server) management tools and facilities-oriented building management tools? How do these all fit together to give companies a complete picture of what's happening in the data center?

Bill Kleyman: Large data centers must have clear visibility into the entire environment. This isn't just environmental information — this means server metrics as well. There are tools to watch power consumption, CPU, RAM and other vital components in conjunction with environmental monitoring systems. For example, AVTECH Software Inc. provides a variety of appliance-based monitoring tools as well as sensors for comprehensive monitoring. Other tools such as up.time Software will help an administrator monitor distributed data centers and gauge resource utilization.

Ultimately, how well a large infrastructure runs is due to the communication between the data center teams, not because of the tools. Alerts must go to the correct engineer and manager from server, data center and virtualization teams, in a coordinated effort to create an optimally functioning environment. Data center consolidation has been an ongoing initiative with many

Contents

organizations. This means that larger servers are supporting more workloads. Teams with monitoring capabilities must collaborate with one another to create an environmental diagnostic plan should an event occur with any system. Integration of major systems should be done with a provider capable of handling the environment and organizational needs of the customer.

An introduction to the ARM server

By: Brien Posey

A new generation of servers has appeared in data centers. These new servers, fitted with ARM processors, promise extremely high processing scalability while keeping power and cooling demands in check. In spite of the promise, however, ARM processors are not quite ready for wide adoption in data centers. This tip explains the current benefits and limitations of an ARM server and shows you what to look for as this technology develops.

Emergence of the ARM processor

Traditional processors are general-purpose devices capable of handling hundreds of different instructions. The problem is that general processors need hundreds of millions of transistors to handle the countless logical conditions needed to support so many instructions. This translates directly into expensive manufacturing costs: The latest Intel Xeon processors routinely cost more than $1,000 each. Even more substantially, each processor has significant power and cooling demands, which drives up the ongoing operating costs of data center servers.

By comparison, ARM processors, based on reduced instruction set computing (RISC) architectures, overcome many of these obstacles. By reducing the number of instructions, the processor is simpler, cheaper, uses far less power and runs with little (if any) substantial heat. The reduction in transistor count also improves the processor’s performance because there are fewer logical stages needed to process instructions. ARM processors are not new – they have been around for decades and can be found in smartphones, printers, digital cameras and other commercial electronic devices.

Contents

ARM processors in servers

Today, ARM processors have slowly begun to appear in some cutting-edge servers. The key factor driving the adoption of ARM processors is scalability. It's important to understand that data centers are changing quickly. Internet-based companies, such as Google and Facebook, have discovered that it is generally more cost effective to fill data centers with vast amounts of inexpensive commodity hardware than to rely on high-performance, overpriced servers. The idea of many general-purpose servers is sometimes called "hyperscaling," and many of the cloud service providers have also adopted the concept of hyperscaling their data centers.

While hyperscaling may offer a low-cost alternative to large, high-end servers, it also poses some problems. Perhaps the biggest issue that organizations may face as a result of hyperscaling their data centers is that of power consumption. The cost of the electricity required to keep a server online is usually trivial. However, when you multiply that cost by the hundreds of thousands of servers in a large data center, the cumulative server power cost can become staggering. And all of those servers produce heat, so cooling the data center can easily increase the cost of electricity by another 50 percent.

Organizations are discovering that ARM servers can be ideal for handling the power and cooling challenges of hyperscaling. To put the power savings into perspective, Hewlett-Packard Co. estimates that servers equipped with ARM processors could potentially consume up to 90% less power than their Intel counterparts.

While each ARM processor core typically provides less raw processing power than a traditional Intel or AMD chip, the low power and cooling requirements allow a level of server scalability that would be impossible with traditional processors. For example, the Tilara TILE-Gx family of processors can provide from 16 to 100 cores on a single chip.

In addition, a large array of processors can be assembled in a single server chassis. For example, the SeaMicro SM10000-64-HD includes dozens of Intel Atom chips, providing 768 cores in a single chassis. Compare this to a common 1U rack server with two 8-core (or even 12-core) processors. Ideally, systems like the SeaMicro SM10000-64-HD can replace 60 traditional servers with a quarter of the power and weight and one-sixth of the space.

ARM processors not quite ready

But hold onto those purchase orders. ARM technology is improving quickly, but the technology isn't quite ready yet. Before you add an ARM server to your next technology refresh plan, it's important to understand some of the current limitations.

First, there are currently competing ARM standards under development. ARM processors are developed differently than mainstream processors. If you buy a server with an Intel or AMD processor, then you can rest assured that the server's processor was manufactured by Intel or AMD. If you were to purchase an ARM server or other device with an ARM processor, that processor was not manufactured by ARM Holdings. Instead, the ARM architecture is licensed to manufacturers who develop their own ARM-based CPUs.

There are several manufacturers that are developing ARM processors, and these processors will vary in capability because each manufacturer puts its own spin on the ARM standard that it has licensed.

Performance is another stumbling block that ARM processors must overcome. Current ARM processors are 32-bit, which is a poor choice for busy data center servers. However, ARM Holdings announced in October that it is developing a 64-bit core. While ARM processor licensees will invariably adopt the 64-bit license, it's impossible to know exactly how the licensees will develop the new ARM architecture.

As one example, Applied Micro is developing an ARM-based chip known as the X-Gene that will function as a "system on a chip." This single chip will

Contents

Contents

feature multiple ARM cores (although the number of cores has not yet been released), twin 10-gigabit Ethernet ports, SATA storage control and even virtualization support. Although the X-Gene chip isn't ready for purchase just yet, Applied Micro expects to be producing X-Gene chips by late 2012.

And finally, it's important to consider the impact of operating systems (OSes) and workloads. ARM processors and ARM-based servers will need a suitable OS. Consumer devices may use an OS like Android or Apple's iOS, but these are not suited for servers. Linux will support ARM, but Windows Server versions, such as 2008 R2, will not support ARM processors until (at least) the release of Windows 8. Also, current ARM processors tend to favor less-sophisticated workloads, such as Web servers, so deploying high-performance database applications on an ARM server may not be possible for some time.

The future of the ARM server is closer than you think
In spite of the challenges, highly scaled ARM servers are appearing in projects like HP's Project Moonshot, using 2,800 Calxeda Energycore processors in a single rack. Future efforts, such as HP's Redstone project, may look to use other ARM-type processors like the Intel Atom. When development of the hardware and software finally utilize the sheer number of ARM processors available in a single chassis, data center operators may see an extremely competitive computing platform for at least some workloads.

The realities of PUE and PUE v2

By: Robert McFarlane

The measurement of power usage effectiveness (PUE) has provided data centers with a common standard for comparing and improving power usage. But the standard provides lots of room for interpretation, allowing many organizations to slant results to their own advantage. In this tip, we'll examine the potential flaws in PUE and explain how the more recent PUE 2 standard can close some of these loopholes.

Contents

The flaws in PUE

PUE was meant to be a means of evaluating and tracking how effectively our data centers utilize the massive amount of power they consume and help us measure the effectiveness of our facility improvements. But we quickly turned PUE into an "extreme sport" contest, where enterprises published and advertised PUE numbers that almost no one believed. No one provided details as to how the numbers were obtained or where measurements were taken, so we didn't know how they were derived or what they really meant. And PUE is a very easy number to manipulate.

On October 29, 2009 The Green Grid released Version 2.1 of their PUE white paper, which people now refer to simply as "PUE v2." It was meant to provide a more accurate means of reporting PUE, but confusion still remained. On May 17, 2011, the Data Center Efficiency Task Force (DCETF) published a paper that clarified the recommended means for measuring and reporting overall PUE.

What has PUE 2 clarified, and what PUE reporting errors is it supposed to correct?

As most everyone knows by now, the PUE definition is deceptively simple. It is the ratio of total power going into the data center divided by the power actually used by the IT computing equipment.

$$\text{PUE} = \frac{\text{Total data center power in}}{\text{IT load power}}$$

A PUE of 1.0 would be perfect, and would mean that all incoming power is going to the IT equipment. It would mean no cooling; no lights; not even wiring, in which some power loss always occurs. That's technically impossible. Most legacy data centers are probably have PUE levels between 2.5 and 3.5, but claims have been made of new data centers and even containerized structures with PUE's less than 1.1. That's pretty amazing!

Contents

So what makes a PUE number really low? Simple: reducing the numerator or increasing the denominator makes the PUE quotient smaller. Like a golf score, a lower number is supposed to say the data center is better. But golfers can cheat, and PUE numbers can be manipulated. How? For starters, you might take instantaneous power draw measurements at your best operating time, when it's cool outside, all the lights are off, but your computers are cranking out research or most of your customers are online. You might even shut down redundant cooling for a few minutes while you take readings. The result could be a really good number, but would it really mean anything?

There are more ways still to distort the PUE figure. Start with what is included in "Total data center power in." How about your network operations center (NOC), or all the IT offices? Leave those out and the total power becomes smaller and PUE becomes better. Where do you measure the "IT load power?" If it's from the read-out on the uninterruptable power supply (UPS) output, then you've added power distribution unit (PDU) transformer losses, wiring losses, and things like cabinet fans into your IT load calculation, even though they don't contribute a thing to the actual processing work. So the denominator gets bigger and your PUE looks even better. And what if you have a shared facility with a common power feed for the building and all its offices, conference rooms, food service and lobbies? Or a shared chiller plant or cooling tower? If they're not all separately metered, some "creative estimating" of the portion of incoming power supposedly used by the data center can dramatically affect your PUE.

The improvements in PUE 2

In short, PUE was not originally intended to compare dissimilar data centers. It was meant as a means of tracking how an individual data center is performing over a period of time. In this case, there would be no benefit at all to manipulating the numbers, since the purpose was strictly for internal improvement. The Green Grid maintains that PUE is not meant to compare different facilities – but since people are going to make comparisons anyway, it should at least be done on a uniform basis, with all the right measurements included. Enter "PUE v2."

PUE Version 2 makes four major and very important changes in the PUE metric:

Contents

1. PUE v2 establishes different types of PUE measurement. The Green Grid calls them Levels 1, 2 and 3; and labels them Basic, Intermediate and Advanced. The DCETF adds a fourth even more basic level, and identifies the categories as PUE0 to PUE3. Each higher level or category is more rigid and exacting than the one before it, and it is now required that any stated PUE value identify the measurement level or category that was used to derive it. For the rest of this article, we will discuss the DCETF categories of PUE0, PUE1, PUE2, and PUE3.
2. With the exception of PUE0, all measurements must now be converted to energy, which means an included time component like kilowatts per hour instead of just kilowatts. And it must include all incoming energy sources, corrected to a common equivalent (e.g., the gas burned for heating must be equated to kilowatts per hour.). Further, all measurements must be made at specified points for each category, and taken over at least a one-year time frame.
3. PUE v2 now stipulates what amenities (such as offices and NOCs) must be included in the PUE calculation for different types of facilities (e.g., data centers, containerized assemblies, and so on) – if it's IT-related, it's in.
4. PUE v2 also makes clear that renewable energy sources, such as wind and solar – while certainly valuable ways of reducing carbon footprint – have no bearing or influence on PUE. The PUE metric is meant strictly to quantify the efficiency with which the compute facility utilizes the energy it consumes, regardless of where or how that energy originated.

Let's briefly summarize the differences among the four new PUE categories to better understand what is necessary to meet each level, and why PUE measurements in different categories shouldn't be compared.

PUE Category 0 ("PUE0"): This is essentially the original PUE, in that it still measures "Total data center power in" and "IT load power" at discrete points

Contents

in time. The major improvement is that it now specifies that the readings be taken "during peak IT equipment utilization." Therefore, PUE0 does not show the effects of varying loads, so may still be significantly skewed depending on when measurements are made. Further, "IT load power" is taken at the output of the UPS, so will add any non-IT overhead such as PDU transformer losses and cabinet fans to the real IT load. And PUE0 is still based on power rather than energy, so it can only be used for data centers that are 100% electrically powered. There is no means of correcting for different fuels. In short, it can be useful for tracking the effects of changes in an individual data center if measurements are taken consistently and compared realistically. But PUE0 should never be used to compare different data centers, even if they are supposedly of the same designs.

PUE Category 1 ("PUE1"): This is the first of the new PUEs based on energy or "power consumption" calculations. It requires that readings be made as total kilowatts per hour over a 12-month period, and that all fuel types serving the data center be converted to a common value – usually to the equivalent kilowatts per hour. This is a major improvement over PUE0, but still measures the "IT load energy" at the UPS output, so it retains the same potential for errors in IT load that were noted for PUE0. Depending on actual conditions, these errors could skew the PUE to a somewhat better number than reality. But if there are no PDU transformers, and no cabinet fans, and everything else is measured in full accordance with the PUE1 methodology, it will probably be a pretty accurate number.

PUE Category 2 ("PUE2"): The only real difference between PUE2 and PUE1 is that the "IT load energy" is measured and summed from the outputs of the PDUs. As noted above, if PDUs with transformers are not utilized in the power delivery chain, there should be very little difference between PUE1 and PUE2 numbers.

PUE Category 3 ("PUE3"): This is the ultimate and most accurate PUE measurement method, and the level to which every data center should ultimately aspire to use. It requires having the infrastructure metering necessary to accurately measure cumulative IT load energy over a 12-month period at the power input connection of each IT device. Very few data

Contents

centers as yet have the ability to do this, but it is certainly the most accurate approach. In the opinion of this author, no one should claim a PUE better than 1.3 unless it was developed in full compliance with the requirements of PUE Category 3.

Don't mix PUE categories

So now we should now be seeing PUE numbers that have real meaning – or at least that we know were supposed to have been derived in a way that enables us to assess them with understanding. But the different measurement points specified for each of the PUE 2 categories also show us how meaningless it would be to compare a number derived under one category, such as PUE1, with one derived in accordance with requirements of another category, such as PUE3. They would be "apples to oranges" relationships.

PUE Version 2 removes the ambiguities and suspicions that have been associated with many PUE claims in recent years. It also defines enough different measurement methods to make it possible for most anyone to calculate their PUE, regardless of the level of measurement sophistication they have available. But it also ensures that when PUE numbers are reported, there is an associated understanding of how they were supposed to have been developed, and how close they may be to the kind of truly accurate PUE number that would come from a PUE3 calculation.

But PUE v2 also makes clear that establishing a meaningful PUE number under any of the four categories requires more than just reading the UPS and the building electric meter. A significant amount of accurate sub-metering or shared load allocation is necessary to obtain the required information. For new or renovated data centers, a full range of metering should be incorporated into every design so that data centers can be truly responsible about their energy usage and efficiency.

Contents

Free resources for technology professionals

TechTarget publishes targeted technology media that address your need for information and resources for researching products, developing strategy and making cost-effective purchase decisions. Our network of technology-specific Web sites gives you access to industry experts, independent content and analysis and the Web's largest library of vendor-provided white papers, webcasts, podcasts, videos, virtual trade shows, research reports and more —drawing on the rich R&D resources of technology providers to address market trends, challenges and solutions. Our live events and virtual seminars give you access to vendor neutral, expert commentary and advice on the issues and challenges you face daily. Our social community IT Knowledge Exchange allows you to share real world information in real time with peers and experts.

What makes TechTarget unique?

TechTarget is squarely focused on the enterprise IT space. Our team of editors and network of industry experts provide the richest, most relevant content to IT professionals and management. We leverage the immediacy of the Web, the networking and face-to-face opportunities of events and virtual events, and the ability to interact with peers—all to create compelling and actionable information for enterprise IT professionals across all industries and markets.

Related TechTarget Websites

- SearchDataCenter
- SearchServerVirtualization