Realtime publishers

The Essentials Series: Building a More Energy-Efficient Data Center

The Business Drivers for Energy-Efficient Data Centers

sponsored by

viridity

by Don Jones

Introduction to Realtime Publishers

by Don Jones, Series Editor

For several years now, Realtime has produced dozens and dozens of high-quality books that just happen to be delivered in electronic format—at no cost to you, the reader. We've made this unique publishing model work through the generous support and cooperation of our sponsors, who agree to bear each book's production expenses for the benefit of our readers.

Although we've always offered our publications to you for free, don't think for a moment that quality is anything less than our top priority. My job is to make sure that our books are as good as—and in most cases better than—any printed book that would cost you $40 or more. Our electronic publishing model offers several advantages over printed books: You receive chapters literally as fast as our authors produce them (hence the "realtime" aspect of our model), and we can update chapters to reflect the latest changes in technology.

I want to point out that our books are by no means paid advertisements or white papers. We're an independent publishing company, and an important aspect of my job is to make sure that our authors are free to voice their expertise and opinions without reservation or restriction. We maintain complete editorial control of our publications, and I'm proud that we've produced so many quality books over the past years.

I want to extend an invitation to visit us at http://nexus.realtimepublishers.com, especially if you've received this publication from a friend or colleague. We have a wide variety of additional books on a range of topics, and you're sure to find something that's of interest to you—and it won't cost you a thing. We hope you'll continue to come to Realtime for your educational needs far into the future.

Until then, enjoy.

Don Jones

Introduction to Realtime Publishers i
The Business Drivers for Energy-Efficient Data Centers 1
Why Data Center Energy Efficiency Has Always Been Difficult to Measure 1
Prime Business Motivators for Efficient Data Centers 3
Better Utilization of Existing Capacity 3
Intelligently Consolidating Servers 3
Retiring or Refreshing Underutilized Servers 4
A Two-Pronged Approach to Efficiency 4
Finding Unused Cooling and Power Capacity 5
Finding and Using (or Eliminating) Underutilized Server Capacity 5

Copyright Statement

The Business Drivers for Energy-Efficient Data Centers

Most businesses have *always* been at least slightly interested in improving the energy efficiency of their data centers. After all, *more efficient* also means *less expensive,* and what business doesn't want to save a little money here and there? Until fairly recently, however, very few businesses actually *pursued* the idea of improved energy efficiency in the data center. Instead, they bought—and powered and cooled—whatever servers they needed, and simply tried to ensure that they were supplying sufficient cooling and power to make everything run smoothly and reliably.

Today, that attitude is starting to change. As businesses tighten their belts and look for more savings, they're perhaps wondering why their data centers need *so much* cooling capacity and whether they need to power *that many* servers every day. Many businesses also have "green" initiatives, where they seek to reduce their energy consumption for environmental reasons. Many of those companies quickly discover, however, that actually reducing data center energy consumption is more difficult than it initially appears to be.

Why Data Center Energy Efficiency Has Always Been Difficult to Measure

One reason data center energy efficiency has always been difficult to measure is that data center energy efficiency is, frankly, rather difficult to measure. It's tough to find out how much you're actually *using*, meaning it's difficult to determine how much—if any—you'll save by taking any particular actions.

Consider cooling capacity, one of many data centers' larger energy uses. In theory, you need enough cooling capacity to cover the maximum heat output of all your data center equipment. You might even want a bit extra, in case one cooling unit fails. Further, you have to actually size your cooling capacity somewhat larger than that because most cooling units can't operate continually at their peak capability—you have to let them operate in a somewhat lower range for the best efficiency. In theory, then, you add up all the heat produced by your servers and other equipment, add a pad of some amount, and there's how much cooling you need.

But even doing that isn't easy. How much heat does a server produce? A server with two 1000-watt power supplies won't produce 2000 watts' worth of heat; those power supplies will rarely run at that full capacity, and not every watt of energy turns into waste heat. The server's other components—especially the processors and hard drives—will also produce waste heat. But how much? An industry rule of thumb is to add the wattage from the servers' data plates and multiple by 3.5. So, a server with 2000 watts in power supplies will need 7000 watts of cooling, or—following the general rule—about 24,500 British Thermal Units (BTUs) of cooling. Worse, if your data center has outside-facing windows (fortunately, few do), you'll have to calculate the heat load from solar input through those windows. Will there usually be people in the data center? Add about 400 BTUs per person. 1000 watts of lighting in there? Another 18,000 BTUs.

But these types of numbers are just *estimates*, and they're on the very high side. They don't consider *actual consumption.* For example, many companies don't leave their data centers completely lit all day—after all, if nobody's in there, why waste the light? Even more importantly, *most servers don't run at full capacity all the time.* That means they're not generating their maximum heat output. In fact, *most* servers don't run at maximum capacity *ever*, meaning most data centers have vastly more cooling capacity than they ever need. Sure, that means the cooling units won't need to run at full-tilt all the time, but larger cooling units tend to consume more power than smaller ones, even when those larger ones aren't running at maximum capacity.

And then there's the energy requirement. Electricity is obviously an on-demand thing, meaning you pay for the exact amount you use. It might be okay to have excess electrical capacity, because other than the initial investment in building it, you don't pay much for it if you don't use it (however, data center power supplies and distribution centers have their own overhead, so having a great deal of extra capacity *will* cost you). But *servers* always use a certain amount of electricity, and in today's world of high-powered, multi-processor, multi-core computers, we tend to under-load our servers. For example, most companies have several "infrastructure" servers that provide core services to the network: authentication, dynamic network address assignment, computer name resolution, and so on. These servers *rarely* run at full capacity. If your data center has four of these, each running at 25 to 30% capacity, you're wasting 280% of your computing capacity. That means you could cut back to one or two of those servers—which means you'd reduce your energy consumption for infrastructure services from about a quarter to even half.

But how do you discover that? How can you put exact numbers around how much electricity your servers already use? That's one of the prime questions that this Essentials Series seeks to answer.

Prime Business Motivators for Efficient Data Centers

Before you start trying to track energy and heat usage, however, you need to be very clear about why you care. If you're just out to reduce consumption for publicity purposes, for example, then exact measurements of your savings might not be critical. However, if you're under a corporate directive to measurably reduce consumption, more precise measurements are probably very desirable. So what are some of the most common business motivators behind greater data center efficiency?

Better Utilization of Existing Capacity

One major driver is *computing efficiency.* In addition to wasting a lot of electricity and cooling capacity, most of today's data centers have a lot of wasted computing capacity. Some cautiously-built data centers—those where servers were built to handle theoretical workloads that have never actually materialized—can have 40%, 50%, or even more wasted capacity. Imagine: If you're wasting 50% of your computing capacity, you can reduce the number of servers you have by 30 to 40% and still have room for growth and workload surges. *One third fewer servers.*

That doesn't mean you'll need to throw away those unneeded machines. However, as businesses grow and demand more services from IT, it would be nice to be able to provide those services from *your existing capacity* rather than having to add more servers, more racks, more power supplies, more cooling, and so on. In other words, finding unused computing capacity can provide a very inexpensive way to support the company's growing needs.

Of course, we have to recognize that one reason we tend to have that unused capacity is that IT professionals prefer single-use servers. That is, a server does only one thing: it's a database server, or a directory server, or a messaging server. Designing servers that way helps keep support and maintenance easier, and helps improve stability and reliability. The cost, of course, can be vast amounts of unused capacity. The solution, of course, is *consolidation.*

Intelligently Consolidating Servers

Ten years ago, asking a server administrator to consolidate servers would probably have resulted in him or her having a fit. Loading multiple services—databases, messaging, infrastructure, and so forth—is generally considered a poor IT management practice because it makes maintenance and patching more difficult and decreases server reliability.

Today, however, virtualization offers a way to combine multiple services on a single machine *without* having them overlap. Each virtual machine is its own independent entity, but it can be sized—even dynamically—for the exact workload *that it actually performs.* In other words, you can load several virtual servers onto a single physical machine, and more intelligently allocate the physical resources—processor, memory, and disk—across those virtual machines. Less wasted computing capacity!

But consolidation has to be *intelligent*. For example, suppose you currently have six older computers that consume about 1000 watts each. Each of them is about 60% utilized, which means they have about 40% wasted capacity. You propose consolidating them onto a new virtualization host. That machine can handle all six servers' actual workload, with about 20% capacity to spare—a good pad for future growth and to handle unforeseen spikes in workload. But that new machine and its supporting equipment (such as a storage area network—SAN) consumes 10,000 watts of energy—meaning that you're using more energy than you used to, although you have less wasted computing capacity. Is that a smart move? That's why you need to be able to precisely measure your existing usage and accurately estimate usage from proposed changes so that you can decide whether a particular consolidation move is the right one. Of course, if you do find situations where consolidation makes sense, you'll find yourself with some extra servers on hand.

Retiring or Refreshing Underutilized Servers

When you find underutilized servers, you have some complex questions to ask. For example, does retiring them make sense? The answer to that question might depend on who you ask and how old the server is.

Many companies, for example, depreciate servers' costs over at least 5 years (the minimum under US rules), and many may choose longer depreciation periods. If a server hasn't been fully depreciated, that means the company hasn't yet gotten the maximum financial use from it, and your CFO might not be willing to let it go, yet. However, the equation here isn't as simple as depreciation period alone. You need to look at the amount of money you're paying to *operate* this server. If a server is costing you $120 a month in power and cooling—not an uncommon amount—and your company only has $1000 of the server's cost left to depreciate, the server might already be costing you more than you're saving. It might make sense to move its workload to a more modern server that can accomplish the same thing with less energy and heat. That's a *refresh*: Bringing in newer hardware to replace an existing server. Or, it might make sense to consolidate the old server's workload onto a virtualization host, retiring the old server completely.

A Two-Pronged Approach to Efficiency

Creating a more efficient data center requires you to look for savings in two areas: unused utility capacity and unused computing capacity.

Finding Unused Cooling and Power Capacity

Utility capacity refers primarily to power and cooling. The fact is that few companies are going to *reduce* their cooling and power capacities: Once you've put the equipment in place, it isn't cost-efficient to remove it. However, by identifying excess capacity, you can do two things:

- Shut it off. If you have a great deal of unused capacity, you can turn off cooling units, reserving them as backups for when an active unit fails or needs maintenance. Unused power capacity may lead you to turn off uninterruptible power supplies, power distribution centers, and so on.
- Use it. Finding unused capacity offers room for growth *without additional capital investment,* so this is how many companies utilize their excess utility capacity. Rather than expanding the data center, get better use out of the one you have.

It's probably relatively easy to measure your existing *capacity*. After all, *someone* knows how much cooling capability you have, and you can add up nameplate information to see how much power distribution and battery backup capacity you have in place. Finding out how much of that you're actually using—well, that's usually trickier.

Finding and Using (or Eliminating) Underutilized Server Capacity

Identifying unused computing capacity enables you to take pretty much the same options:

- Get rid of it. If you can move a server's workload elsewhere, you can eliminate that server, eliminating its heat load and electrical use at the same time.
- Use it. If a server has enough excess capacity, it can take on additional workloads—perhaps through virtualization techniques—and help consolidate workloads so that older servers can be retired.

What's tricky here is that it's tough to measure your computing capacity at all, let alone measure how much of it you're using. Keep reading the remaining articles in this Essentials Series to learn about the techniques you'll need to use.