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

Maximizing Data Center Density by Finding Power and Cooling Capacity You Didn't Know You Had

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

Realtime
publishers

Copyright Statement

Realtime
publishers

Maximizing Data Center Density by Finding Power and Cooling Capacity You Didn't Know You Had

The ultimate data center contains servers that run near-100% utilization (you always want a *little* pad for spikes in demand) and sip—rather than gulp—power. Such a data center has exactly enough cooling to handle the servers' heat output and whatever other overhead heat load exists—such as from lighting, people, and other equipment. That's called *maximal data center density:* Every square foot performing maximum useful work with no unnecessary excess capacity.

> **Note**
>
> I used the phrase *unnecessary excess capacity* very deliberately: *Necessary*, or planned, excess capacity is fine—if you've made a decision to absorb the cost for a valid business reason.

You may not even realize the capital you've spent on your data center that isn't being used. Maximizing data center density is, ultimately, about getting the maximum return from that investment. And, as your business needs to grow and expand, you can start putting that excess investment to use, supporting growth and expansion with much less incremental investment.

The Trick to Measuring Data Center Capacity

A difficulty in doing so, however, is in figuring out how much capacity your data center *really has.* Traditionally, we've overbuilt our data centers—pretty significantly in some cases—without really knowing how *much* we were overbuilding. If you don't know your data center's total capacity, it's tough to really maximize it.

One limiting factor that you'll always have to contend with is that you can't really manage at the data center level. You have to manage at the rack level: Server racks have physical capacities and electrical capacities that have to be considered. If every rack is full, it doesn't really matter if—for example—you have excess cooling capacity because you've got no place to put any new servers.

Realtime
publishers

The Old Way: Nameplates, Guesstimates, and Over-Engineering

The old guesswork and estimation methods that I outlined in the previous article in this Essential Series apply here, too. Traditionally, we've simply added up the power requirements of our servers by reading their nameplates. Then we multiply by some arbitrary number to account for overhead and such, then multiply again for a safety pad, and that tells us how much power and cooling we'll need. The result? Massively overbuilt data centers with no real way of telling how much of our capacity is being used and how much is available.

Worse still, most of us have to make a lot of up-front, permanent decisions *before we put a single server in the room.* A facilities person measures the available space. They figure they can have a specific number of server racks in that space, which is usually a straightforward calculation. Each server has 42U of space, and some facilities-planning manual tells the person that each 4U of space will contain a server with a couple of 1000-watt power supplies. That's 21,000 watts per rack. That's about 72,000 BTUs, or about six tons of cooling. That's a *lot.* Following those numbers, a small 10-rack data center would need 60 tons of cooling. Fortunately, most facilities planners apply some common sense to that number and "round down," but the bottom line is that a lot of resources are being dedicated based on guesswork—not on what the data center will actually contain.

The New Way: Actual Utilization, Hard Facts, and Precision

You can't change what's been done: Your data center's cooling capacity, for example, is fixed for all practical purposes. Your electrical capacity isn't fixed, but it can be expensive to add higher-amperage service. However, once you're using that data center, you can certainly start utilizing that capacity more efficiently, likely increasing your density beyond the original estimates.

The way to do so is to measure your *actual consumption* based on your *actual utilization.* Again, directly measuring server consumption is difficult, requiring the use of meters and so forth. A far better approach is to measure the data center's total capacities—which are known figures like tons of cooling and amps of electrical service—and to measure *server utilization.* Once you know server utilization, you can use a knowledge base to draw conclusions about server consumption at those levels of utilization.

Finding and Creating Data Center Capacity

A good toolset can actually find unused capacity for you. For example, Figure 3.1 shows a "placement assistance" tool. Based on what you've told it about which servers are in which racks, it helps you find physical space for a proposed new server. It bases that information in part on physical capacity and in part on the availability of power to that rack.
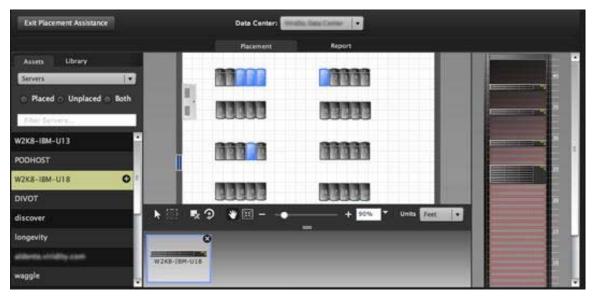
Realtime
publishers

Figure 3.1: A server placement assistance tool.

This type of tool can help you find and utilize excess capacity, making it possible to safely put as many servers as possible into the data center.

Creating Capacity through Consolidation and Refreshes

Once you know where your capacity is being used, you can choose to *create more capacity* by consolidating multiple servers—often onto virtualization hosts—or by refreshing older servers with newer, smaller, or more-efficient models. As you remove older, underutilized, or less-efficient servers from the data center, your planning tools can help you re-use that space to accommodate new servers. Or, in some instances, you may choose to actually downsize the data center somewhat (although in practice it's rare to do so).

How Extra Data Center Capacity Helps and Hurts Your Business

But wait—isn't having some excess capacity good? Don't we overbuild for a reason? Is it really wise to use up *all* our excess data center capacity?

Extra Capacity = Wasted Money

In most people's minds, extra capacity equals room for growth. There can be two problems with that mindset: First, we tend to plan for *too much* extra capacity. Second, we're sometimes reluctant to *use* the extra capacity we have, for whatever reasons.

Extra capacity is, in the final evaluation, wasted money. You paid to build it, you're paying to maintain it, and it does nothing for you. If you had the "extra capacity" discussion about your company's payroll, you'd find most managers far more willing to either reduce capacity—through layoffs—or put that extra capacity to work, perhaps by retaining and reassigning personnel. When it comes to a data center, however, companies don't tend to have that same viewpoint—and they *should.* Either let your excess capacity go or put it to work.

For example, I've worked with one company that was able to eliminate about a quarter of the racks in their data center. By rearranging the remaining racks, they created a contiguous space in the room where all the empty racks sat. They were then able to construct an inexpensive wall, temporarily downsizing the data center and enabling them to cut back on the amount of cold air being pumped into the remaining space. That excess capacity still existed, but now it cost significantly less and could be brought into utilization when needed. They had simply found that, through virtualization and newer hardware, they could get by with markedly fewer physical machines.

Using Extra Capacity = New Business Services

It's more common, however, to see that excess capacity put to work. I've had a couple of experiences where we went *looking* for extra data center capacity because we thought we were full-up and needed room for new machines. Through some consolidation, rearranging, and technology refreshes, we were able to free up space, increase our data center density, and accommodate the new machines. The data center was doing more work overall, meaning the company was seeing a better return on that very expensive long-term investment.

Conclusion

Whether you're looking for a "greener" IT environment, interested in saving some money on electricity, or looking to accommodate additional business services out of your existing facilities, creating a more efficient data center can help. The trick is to correlate actual server utilization with server energy consumption, helping you identify and act on low-efficiency servers, increase data center density, and enjoy the business benefits of greater efficiency.