Realtime publishers

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

Finding Inefficient Servers and Reducing Data Center Costs

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

Copyright Statement

Finding Inefficient Servers and Reducing Data Center Costs

Servers that aren't being fully-utilized are, to be blunt, a waste. You wasted money buying the server, are wasting money powering and cooling the server, and are wasting money maintaining the server. So how do you find those inefficient servers? What do you do with them when you find them? How much money can you save?

The Trick to Measuring Server Consumption?

Actually measuring server consumption can be incredibly difficult. How do you measure power consumption for what the server is *actually doing*?

The Old Way: Meters, Amps, Watts, and Nameplates

The oldest and easiest way to determine a server's power consumption is to simply read its wattage rating from its nameplate. For example, Figure 2.1 was taken from the back of a computer: What's its energy consumption? 500 watts. Easy enough. Add up all the nameplates, and you know your energy consumption.

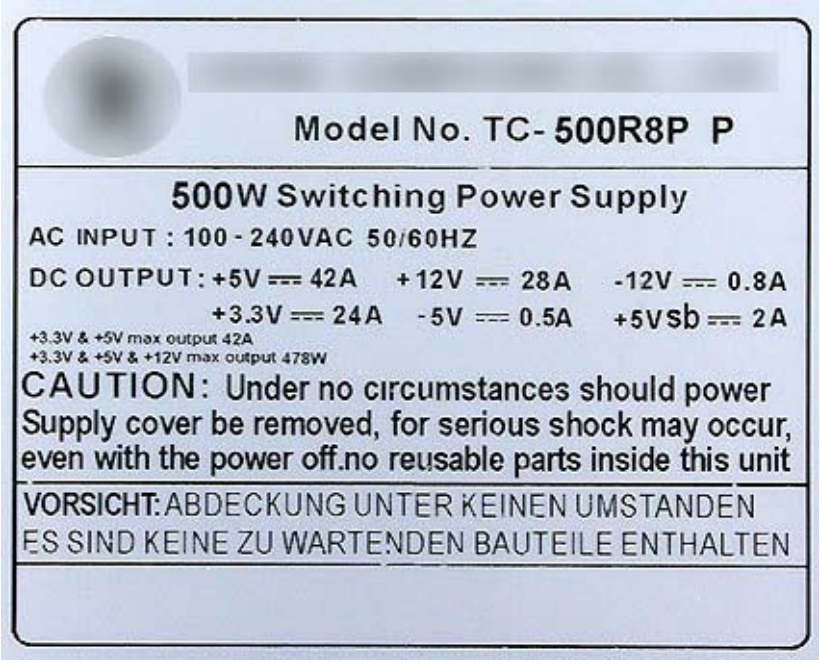

Figure 2.1: Nameplates state power requirements.

Unfortunately, this is exactly how we started over-building our data centers in the first place. I have servers, for example, that consume about 70% of their stated requirement when they're doing nothing; as their workload increases, they move to about 90% of the requirement—but few of them, when I've measured, ever consume the full amount stated on the nameplate. So measuring *actual* energy consumption is much more accurate, and is often accomplished by using a meter of some kind (such as the one shown in Figure 2.2), or by using an intelligent power distribution center (PDC) that has built-in monitoring and measuring capabilities.

Figure 2.2: Using a meter to measure power consumption.

The problem with the metered approach is that it doesn't correlate to server *utilization.* In other words, it's great to know that all your servers consume a certain wattage of power—but how does that relate to how busy the servers are? Are you measuring when they're fully loaded or practically idle?

The New Way: Intelligent, Utilization-Based Measurements

A more useful approach starts with a couple of basic assumptions. First, a given model of server in a given configuration (including power supplies, processors, memory, and so on) will use a certain amount of power when idle. As its workload increases, its power consumption will increase in a certain fashion. Therefore, if you know the server's configuration information, and you know its utilization, you can pretty accurately state its power consumption—without having to spend a lot of time connecting and reading meters.

The trick with making those assumptions usable is in gathering a database of server configuration and power consumption information. If you had that, you'd be set: You could install a small software agent on each server to measure utilization and report back to a central console, or potentially even use an agentless approach and collect server utilization information across the network. The result might be a console display like the one in Figure 2.3, which shows an inventory of the data center along with a display—at the bottom—of how much power is actually being consumed in the data center.

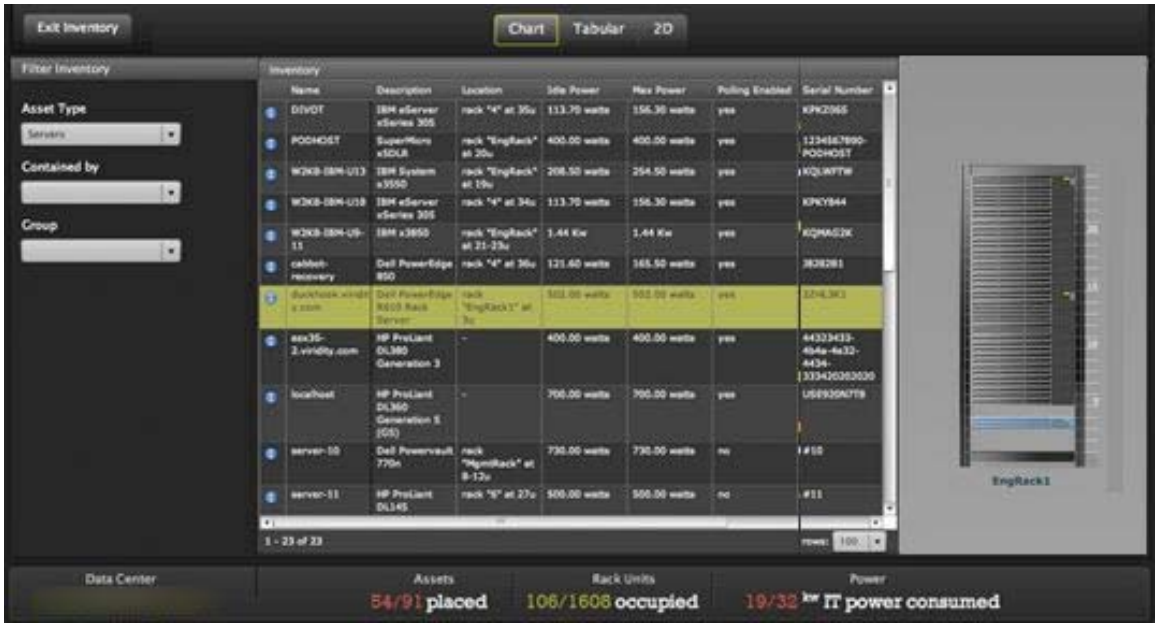

Figure 2.3: Accurately measuring consumption-based on utilization.

By breaking down this information into a per-server utilization graph, you can see which servers have free capacity, along with how much energy they're using.

Do You Know Your Servers' Efficiency Scores?

You might even make up an *efficiency score* for your servers. Servers that are performing the most work and using the least energy get a higher score; servers that are working less but still using more energy receive a lower score. Using that score, you can start to intelligently target less-efficient servers for retiring, refreshing, or for consolidation. You might even learn a bit about which brands and models of servers are more energy-efficient, helping guide future purchasing decisions.

A chart like the one in Figure 2.4 can help, too. This shows server utilization in green, and power consumption in orange; a big orange bar next to a smaller green bar is a server that's taking more than it gives. Any small green bar represents a server that isn't working to full capacity—and that's a server you should be looking to do something with.

Figure 2.4: Charting server efficiency.

What to Do with an Underutilized Server

Once you've identified servers that have unused capacity, what can you do with it?

- Leave it. It's okay to *make a conscious decision* in favor of inefficiency; you may know something that your measurement tools don't, like the fact that once a quarter a particular server works a lot harder. However, that example does illustrate the importance of getting a good representative measurement of servers' long-term utilization before making any permanent decisions.
- Consolidate it. Move the server's workload somewhere else, and either retire the server or put different services onto it. Typically, you'll consolidate through virtualization, but that might not always be the case.
- Add services to it. In other words, let the under-utilized server participate in consolidation by taking on some *other* server's workload.
- Refresh it. Depending on how much a server is costing you to operate, it may be less expensive to retire it early and replace it with a newer, more-efficient piece of hardware—one that can perhaps take on additional workloads and be better-utilized, even.

Modeling Efficiency for Server-Level Changes

A key task in all of this is being able to model proposed changes to the data center. In other words, if you're going to be retiring servers, or consolidating servers, or whatever else, what impact will that have on your energy and cooling use?

This is where a good tool can help. If you have a power-utilization tool that works primarily from a database of known power consumption models, then all you have to do is tell it the type of hardware you're considering adding to the data center, and how much you expect that hardware to be utilized. The tool can produce estimates on that machine's power requirements for the workload you've specified, and you can see how that fits into your energy efficiency plan.

The idea here is to make every asset in the data center work as hard as possible for the least amount of power and cooling possible. The industry term for this is *maximal data center density,* and it's the topic of the last article in this Essential Series.