

In the data center, power and cooling costs more than the IT equipment it supports

Christian L. Belady, P.E.

Hewlett-Packard

Historically, the cost of energy and the cost of the data center power and cooling infrastructure have not been on the radar for most Chief Financial Officers (CFO) and Chief Information Officers (CIO) and have not been considered in TCO (Total Cost of Ownership) models. As a result, almost all of the focus has been on driving down the cost of IT equipment in the data center. This was a reasonable assumption during the 90's when server power and energy costs were substantially lower. However, as shown by the ASHRAE power trends curve in Figure 1, power density has been increasing at an alarming rate. During this same period of rapid power growth, server costs have stayed virtually flat and raw performance has increased substantially as shown in Figure 2. During the eight-year period shown in Figure 2, performance has increased 75 times or, in other words, the servers are providing 75 times more performance now for the same hardware cost of eight years ago. In addition, the performance per Watt has increased 16 times during the same period. Thus, for every unit of energy, the customer is getting 16 times the throughput as they did eight years ago. Roughly speaking, the performance/Watt of a server doubles every two years. From a CFO's perspective, the fact that both the cost per performance and performance per unit of energy are going down should be very pleasing, but this is not the whole picture. To complete the picture one must consider the following points:

- The cost of data centers is going up as a result of the increased power capacity required. ASHRAE data shows that the server power density will continue to increase and data centers will have to scale their infrastructure to support this increase.
- Despite the massive improvements in the performance of servers, business application needs are outstripping the performance improvements in servers. The result is that the number of server units per year is growing.
- Data centers are becoming more mission critical for business operations, resulting in the need for more expensive fault tolerant designs.

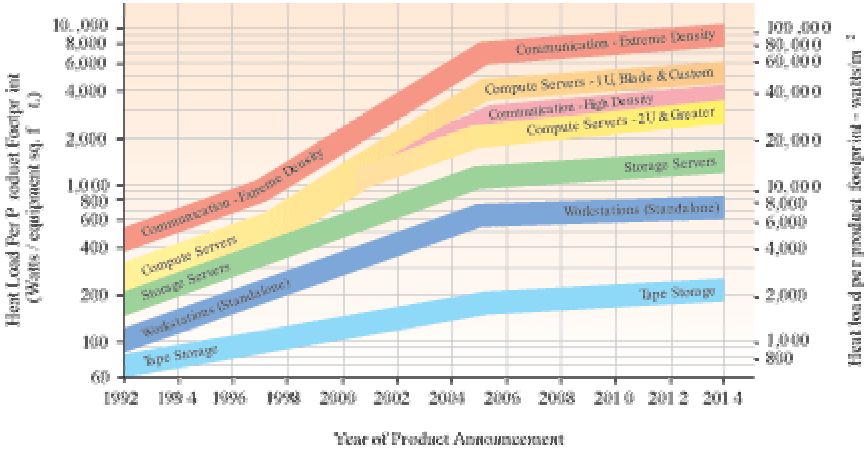


Figure 1. ASHRAE TC 9.9 equipment power projection [1].

CFOs are struggling with the combined effects of these points: the rapid need for more data centers to house the growing server population due to the insatiable need for more business applications. As a result, the cost of these data centers and the energy usage is showing up on the radar for businesses and they do not understand why. This article is meant to shed some light on why the focus is shifting.

In a recent article, the 3-year Energy Cost to Acquisition Cost ratio (EAC) [3] was introduced as a metric to understand the cost of energy relative to the cost of the server. Today, for 1U servers this is approaching unity and comes as a surprise to most data center managers. To convince the reader, it is a simple calculation to determine the energy cost of the server:

3-yr Energy Use Cost = 3 yrs x (8760 hrs/yr) x (\$0.10/kWhr) x (Server Power in kW) x PUE (1) where PUE is the Power Usage Effectiveness [3] or the Data Center Electrical Load over the IT Electrical Load. For a well managed data center this value is usually about 2.0 (or less), which implies that for every Watt of server power, an additional Watt is consumed by the chillers, UPSs, air handlers, pumps, etc. Indications are that for some data centers this value can be as high as 3.0 [4] and in some cases higher. Usually, this variation is completely due to how well the cooling environment is designed in the data center and has a direct relationship to the energy cost [5].

Using equation (1) for a 1U server (which, when fully configured, costs about \$4,000 and consumes about 500 W) and a PUE of 2.0 results in a cost of energy of \$2,628. This is almost as much as the server itself, but the reality is that in many cases the cost is much higher. In Japan, energy costs are twice as much, so this number would be double. To make matters worse, in data centers where the cooling design is poor (PUE = 3.0, for example), the cost of energy would be 50% higher.

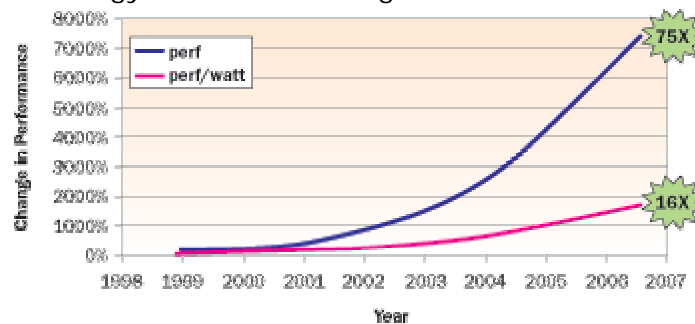


Figure 2. Raw Performance and Performance/Watt increase in a typical server [2].

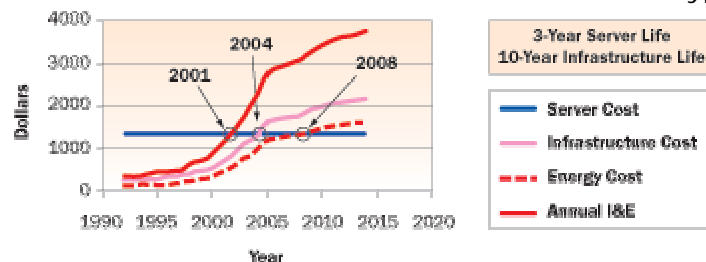


Figure 3. Annual Amortized Costs in the Data Center for a 1U server.

This means that the energy cost would be \$3,942 in the U.S. and \$7,884 in Japan. Clearly, there can be huge savings in this energy cost by focusing on optimizing the cooling in the data center as shown in the articles identified earlier [3,5].

Unfortunately, the energy usage is not the only cost driver that scales with power. In 2005, fundamental research [6] was published showing that the infrastructure cost is a big portion of the TCO and quantified the real cost drivers in the data center, which included the amortized cost of the power and cooling infrastructure. This research shows

that a fundamental problem is the over-provisioning of cooling due to poor cooling and the lack of understanding of the environment. In addition, The Uptime Institute has also introduced a simplified way for estimating the cost of data center infrastructure [7] based on Tier ratings. For brevity, only Tier IV data centers (with dual redundant power throughout) will be examined since this is the recommended approach for mission critical operations. The Uptime Institute's Infrastructure Cost (IC) equation for a Tier IV data center is as follows:

$$IC = \text{Total Power} \times (\$22,000/\text{kW of UPS output for IT}) + (\$2370/\text{m}^2 \text{ of allocated raised floor for IT}) \quad (2)$$

While, admittedly, the authors state that there is a large error band around this equation, it is very useful in capturing the magnitude of infrastructure cost. In fact, it is this author's contention that this equation could be fine tuned for more accuracy using PUE because poor cooling will mean that more infrastructure will be needed.

However, that discussion is beyond the scope of this paper. Again, looking at the 500 watts of power consumed by the 1U server and using equation (2) and ignoring the IT space the server occupies, the cost of infrastructure to support that server would be enormous at \$11,000. The reality is that this cost would be amortized over 10 to 15 years so real annual cost of the infrastructure is \$1,100 per year. For the 3-year life of the server, this equates to \$3,300 or again close to the cost of the server. Note that there is also an adjustment in the cost as a result of the space occupied by the server, but its calculation is beyond the scope of this discussion.

Using the ASHRAE data [1] and the conservative data center projection outlined at IThERM 2006 [8], we can use the concepts outlined in this paper to project where costs will lie in the data center over the next few years in Figure 3.

The graph shown in Figure 3 is subject to a number of assumptions and qualifications, such as:

- Server cost has stayed constant, though, if anything, they are going down.
- ASHRAE data defines the actual server power growth rate.
- PUE = 2. More data is needed to define where data centers truly lie.
- Tier IV data center practices. Many data centers are not quite at the Tier IV level yet.
- U.S. energy costs are constant at \$0.10/kWhr. In all likelihood this will increase over time.

The refinement of these assumptions will need to be fleshed out by the industry over the next few years but Figure 3 does invalidate one assumption from the 90's: IT equipment costs are all that matter in the data center. Instead the following realities are setting in:

- Energy costs alone will exceed the cost of the servers in 2008.
- Infrastructure costs alone have already exceeded the cost of the server in 2004.
- The combined cost of the Infrastructure and Energy (I&E) exceeded the cost of the server back in 2001.
- Infrastructure and Energy Cost (I&E) will be 75% of the cost in 2014 and IT will be only 25%. That is a significant shift of 20% I&E and 80% IT in the early 90's.

As a result, more effort will be required to keep energy costs and infrastructure costs down. This will mean that better TCO modeling tools that include these costs, but more

importantly, optimize the environment will be clearly warranted. For example, data centers need to be engineered with computational fluid dynamics to eliminate over-provisioning and waste to lower costs and improve efficiency. In addition, the use of technologies such as liquid cooling can enable more efficient designs and thus lower power and cooling costs overall. The importance of these tools and technologies will play an important role in optimizing costs in the future.

Conclusion

Hopefully, this article has shown that a paradigm shift has occurred in the data center. The cost of IT equipment is no longer the bulk of the cost, but rather the cost of the power and cooling infrastructure has crept in to be the primary cost driver. As with the CIOs and CFOs, we need to internalize this fact and recognize that there are huge cost management opportunities in the data center. Judicious design practices that have been applied to servers for over a decade will now have to be applied to the data center environment to curb costs. The demand for new technologies that reduce overall TCO and the demand for technically savvy engineers in this space will grow rapidly; the thermal analyst is no exception. The future is indeed bright for the thermal engineer!

Christian Belady

Distinguished Technologist

Hewlett-Packard

3000 Waterview Pkwy

Richardson, TX 75080

Tel: 972-497-4049

Fax: 972-497-4500

E-mail: christian.belady@hp.com