RC24996 (W1005-088) May 17, 2010
Computer Science

IBM Research Report

The Case for Full Throttle Computing: An Alternative Datacenter Design Strategy

José E. Moreira, John Karidis
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

Research Division
Almaden - Austin - Beijing - Cambridge - Haifa - India - T. J. Watson - Tokyo - Zurich

The Case for Full Throttle Computing:
An Alternative Datacenter Design Strategy

José E. Moreira   John Karidis
jmoreira@us.ibm.com   karidis@us.ibm.com
IBM Thomas J. Watson Research Center, Yorktown Heights NY 10598

Introduction Our message can be summarized as follows: minimize the cost of computing by maximizing server utilization; maximize server utilization by consolidating latency-sensitive workloads onto relatively large servers; fill overhead capacity with latency-insensitive (batch-like) work sold at a significantly reduced price. When the total of both types of work falls well below the total capacity of all servers, migrate workloads to a smaller number of machines running at as high a capacity as possible and turn off remaining machines.

Cost of computing in a datacenter For a datacenter to operate and deliver useful computing, one needs a building, power and cooling infrastructure, IT equipment, labor to manage that IT equipment, and electricity to power the IT equipment and corresponding cooling. Electricity and infrastructure costs are somewhat proportional to the average amount of computing, while the other costs are independent of the computing load. Idle servers typically consume ½ of the electricity of fully utilized servers, but there is a push today to develop truly "energy-proportional" equipment in which the electricity consumption more closely matches the utilization of the servers [2],[3]. Such truly proportional equipment drives down both variable and fixed costs since it reduces power requirements of servers operating below peak utilization [2].

In general, the fixed costs above dominate the variable costs [3]. The bar graph on the left side of Figure 1 plots the cost per hour for each kW of computing equipment installed, as a function of the average server utilization, for a typical United States datacenter [4],[7]. That cost varies from $0.26/h/kW for idle servers to $0.33/h/kW for fully utilized ones. The difference is only $0.07/h/kW or about 20% of the total cost. Since the amount of computing delivered increases with server utilization, the cost of computing decreases with utilization. This is shown in the line plot on the right of Figure 1. The plots are very similar whether the servers' power consumption is truly proportional to utilization or not. For the truly energy-proportional case we make the electricity and infrastructure cost scale perfectly with average utilization. We adopt the same server cost, whether the servers are truly energy-proportional or not.

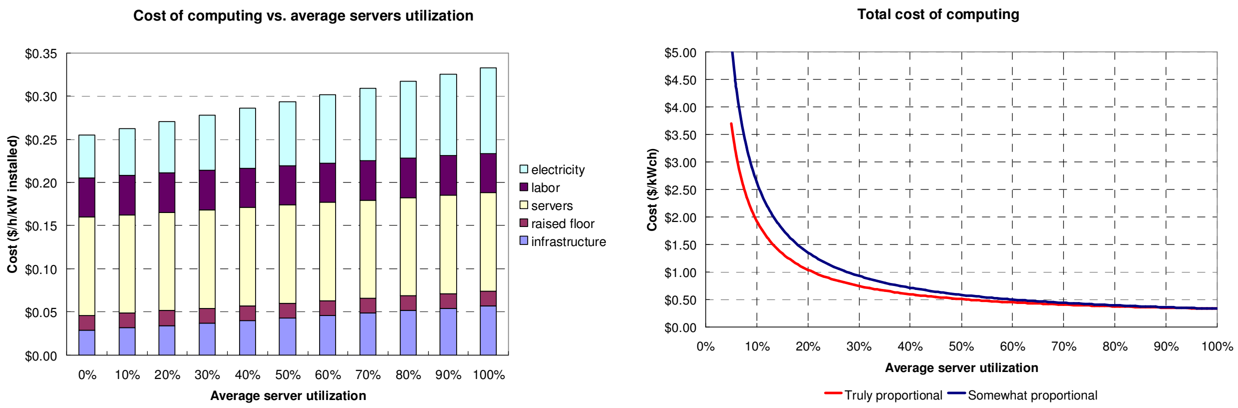

Figure 1: Cost of computing for a data center (left). The cost is shown in units of dollar per hour for each kW of IT equipment installed. Most of the cost is fixed and does not change as more or less work is performed in the data center. Only electricity use (and cost) increases as the utilization grows. As server utilization increases, the cost per unit of computing delivered goes down (right).

The most efficient machine is one that runs at 100% utilization, which is the essence of the title of this paper. We achieve the lowest cost of computing when the servers are running at full throttle, doing useful work. To do that, we need both machines that can operate at high utilization and work to fill them up.

Building machines for higher utilization We use a simple queuing model to illustrate a general trend with server utilization [4]. Consider an m-processor server where all parcels of work go into a single queue from which each parcel is dispatched to one of the processors for execution. (This models an m-way symmetric multiprocessor.) The response time (normalized to service time) as a function of utilization for this queuing model is shown in the left plot of Figure 2 for different values of m. All curves have a characteristic knee shape, with the response time increasing rapidly once a certain threshold of utilization is surpassed. The larger the server, the higher that threshold. Take, for example, the case of response time = 2. A single-socket server (m =1) can be driven to 50% utilization, whereas an eight-socket server (m =8) can be driven to 90% utilization, as per Figure 2. Therefore, a single 8-socket server can deliver 14 times the throughput of a single-socket server ($8 \times 0.9 / 0.5 \cong 14$). For any given response time, larger servers can operate at higher utilization and still achieve that response time. Real-world experience shows that larger servers can indeed be driven to very high utilizations. For example, IBM System z servers (mainframes) achieve better than 80% utilization at customer sites [9].

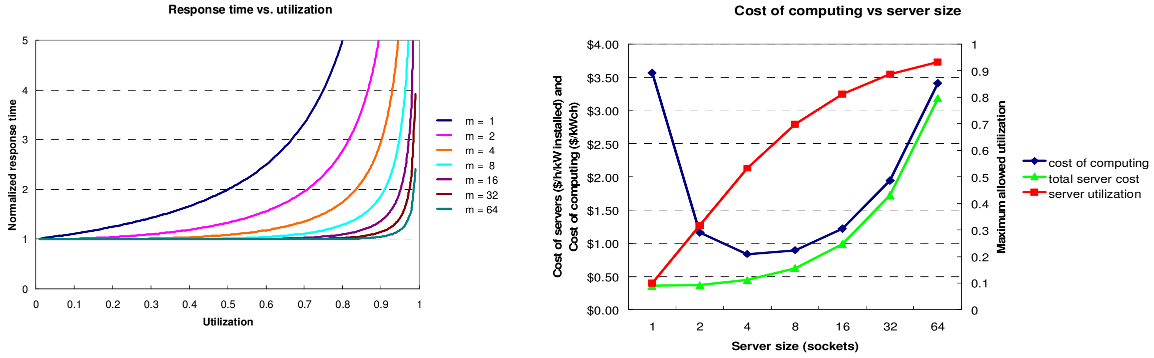

Figure 2: Larger servers can operate at higher utilization for a given response time (left). Server cost increases with the number of processors, resulting in an optimum server size to minimize cost of computing (right).

However, larger servers can be significantly more expensive, on a per processor basis, than smaller servers. The result from these two counteracting forces is that there is an optimal server size for which the cost of computing is at a minimum. The overall shape of the cost of computing as a function of the server size is shown in the right plot of Figure 2. The exact shape of the curve depends on a variety of factors but the important message is that there is an optimal server size that will keep cost at a minimum. For a wide range of parameters that we studied, the optimum server size often falls in the region of 4 to 16 processors.

Filling up a server with work Once we build a server that can deliver good response time at high utilization, we are still faced with the challenge of filling it up with useful work. It is difficult to fill up a server with a single application, as we have to leave enough headroom in the server capacity to accommodate spikes in that workload. Larger servers can be used in a workload consolidation approach [8], running several distinct applications concurrently. Each application represents a small fraction of total machine capacity, and together they combine to drive the server to high utilization. Because there are more workloads running concurrently, the spikes tend to average out and we can live with smaller headroom.

Even with larger servers, we are still likely to have to operate with some headroom. We can fill that headroom with low priority work that is only performed when there is extra capacity available and can be suspended or shut down otherwise. The only cost of running the low-priority applications is the cost of the additional electricity consumed by the servers, which naturally gives rise to a two-tier model for the price of computing. The high-priority applications are charged for all fixed cost plus their share of the variable cost. The low-priority applications are charged for only their variable cost. We already see this two-tier pricing policy emerging in, for example, the Amazon Elastic Compute Cloud (EC2), where "spot" pricing for low-priority VMs (which run on otherwise unused server capacity) is about one-third of that for higher-priority VMs.

Energy proportionality with full throttle computing There will inevitably be times when all requested computing work (both high- and low-priority) is insufficient to occupy the full computing capacity of all machine is a datacenter. In this case, the choice is to either throttle each machine (which we call "vertical proportionality"), or to turn off some fraction of machines while maintaining the remainder at maximum practical utilization (which we call "horizontal proportionality"). If the workloads are designed to be easily migrated then horizontal proportionality is a viable option. Energy use can scale with overall demand, and each running machine will maintain maximum efficiency by running at high utilization. Vertical proportionality can still be used on the running machines to further reduce the electricity consumption.

Conclusions The most effective approach to reduce the cost of computing in a datacenter is to increase the utilization of the servers in that datacenter. Larger servers are able to operate at higher utilization and "mid-sized" servers are often cost-optimal. Capacity on these servers not consumed by high-priority workloads should be filled with batch-like workload that can be sold at a significantly reduced price. If the total workload is insufficient to fill all machines, energy proportionality can be provided "horizontally" by turning off a fraction of the machines while the remaining machines operate at high utilization.

References

[1] Luiz A. Barroso, Jeffrey Dean, Urs Hölzle. **Web search for a planet: The Google cluster architecture**. *In IEEE Micro Magazine*. April 2003.

[2] Luiz A. Barroso, Urs Hölzle. **The case for energy-proportional computing**. *In IEEE Computer Magazine*. December 2007.

[3] Xiabo Fan, Wolf-Dietrich Weber, Luiz A. Barroso. **Power provisioning for a warehouse-size computer**. *In Proceedings of the ACM International Symposium on Computer Architecture*. San Diego, CA, June 2007.

[4] John Karidis, José E. Moreira, Jaime Moreno. **True value: assessing and optimizing the cost of computing at the data center level**. In *Proceedings of the 6th ACM Conference on Computing Frontiers*. Ischia, Italy, May 18 - 20, 2009.

[5] Kevin Lim, Parthasarathy Ranganathan, Jichuam Chang, Chadrakant Patel, Trevor Mudge, Steven Reinhardt. **Understanding and Designing New Server Architectures for Emerging Warehouse-Computing Environments**. *In Proceeding of the ACM International Symposium on Computer Architecture*. Beijing, China, June 2008.

[6] Chandrakant D. Patel, Amip J. Shah. **Cost Model for Planning, Development and Operation of a Data Center**. Technical Report, HP Laboratories. HPL-2005-107R1. Palo Alto, CA, June 2005.

[7] W. Pitt Turner, Kenneth G. Brill. **Cost model: dollars per kW plus dollars per square foot of compute floor**. *White paper*. Uptime Institute. 2008.

[8] Werner Vogels. **Beyond Server Consolidation**. ACM Queue vol. 6, no. 1 - January/February 2008.

[9] Mark Levin. **Best Practices and Benchmarks in the Data Center.** Mark Levin & Partners, LLC. April 2006.