# Rightsizing Servers to Achieve Cost and Power Savings in the Datacenter

Today more than ever, IT departments need to make sure that the servers they select and deploy are as efficient as possible in terms of acquisition cost and energy consumption. This paper describes how the Microsoft Global Foundation Services team that manages and operates the company's vast datacenter rightsizes its servers to achieve maximum efficiency. The process focuses on collecting detailed performance data using representative workloads, and then analyzing that dataset to select balanced servers that are optimally sized for production scenarios. IT departments that follow this methodology can stretch their purchasing budgets significantly to help achieve organizational goals even in times of constraint.

Published: December 2009



### Introduction

How do you make sure that the servers you purchase and deploy are most efficient in terms of cost and energy? In the Microsoft Global Foundation Services organization (GFS)—which builds and manages the company's datacenters that house tens of thousands of servers—we do this by first performing detailed analysis of our internal workloads. Then by implementing a formal analysis process to rightsize the servers we deploy an immediate and long term cost savings can be realized. GFS finds that testing on actual internal workloads leads to much more useful comparison data versus published benchmark data. In rightsizing our servers we balance systems to achieve substantial savings. Our analysis and experience shows that it usually makes more sense to use fewer and less expensive processors because the bottleneck in performance is almost invariably the disk I/O portion of the platform, not the CPU.

#### About the server environment in Microsoft datacenters

Is our experience applicable to your environment? A few words at the outset may help you decide. Global Foundation Services supports Microsoft's entire Online, Live, and Cloud services environment, which is quite varied and not monolithic by any means. Most people know about our e-mail and search offerings, but we actually support more than 200 online services, with applications ranging from Bing, Hotmail, Microsoft Office SharePoint Online, and Xbox Live. For efficiency sake we consolidate on a small line-up of servers that work across all our services. These servers are running Windows Server 2003, Windows Server 2008 or Windows Server 2008 R2. This is a familiar scenario for many IT professionals, who must support a wide variety of very different internal applications, such as those for Engineering, Finance, and Human Resources. At Microsoft the applications fall into several primary buckets. We need to consider what's good for Bing (formerly Live Search) and for Windows Azure, and just about everything else falls under Web (IIS), File, and Database, where in our case we use Microsoft SQL Server for transaction processing.

#### What kind of performance is important to you?

Typically in the industry, when you say "performance" most people think about speed. But when you look up the term in the dictionary, performance is defined as how well something performs the functions for which it's intended. In the case of servers, we're concerned about far more than speed. The traditional metrics for speed are throughput, response time, and latency. We also take cost-effective measures into account.

As an example, consider an automobile. People who are concerned about economy don't look just at how fast the car will go. They also want to know about miles per gallon, maintainability, and other economically based criteria. When we look at servers, we think broadly and look at performance per dollar, performance per watt, and performance per dollar per watt. We explore actual energy consumption and how much work gets done for that energy consumption. In addition, we also take reliability and maintainability into account.

# How much credence should you give published benchmark data?

When you look at the traditional metrics for performance—throughput and response time—a number of industry-standard benchmarks exist that are very useful. A majority of these are developed by two major industry organizations: the Standard Performance Evaluation Council (SPEC), and the Transaction Processing Performance Council (TPC).

One of the most commonly used benchmarks is SPEC CPU2006. It provides valuable insight into performance characteristics for different microprocessors central processing units (CPUs) running a standardized set of single-threaded integer and floating-point benchmarks. A multi-threaded version of the benchmark is CPU2006\_rate, which provides insight into throughput characteristics using multiple running instances of the CPU2006 benchmark.

But important caveats need to be considered when interpreting the data provided by the CPU2006 benchmark suite. Published benchmark results are almost always obtained using very highly tuned compilers that are rarely if ever used in code development for production systems. They often include settings for code optimization switches uncommon in most production systems. Also, while the individual benchmarks that make up the CPU2006 suite represent a very useful and diverse set of applications, these are not necessarily representative of the applications running in customer production environments. Additionally, it is very important to consider the specifics of the system setup used for obtaining the benchmarking data (e.g., CPU frequency and cache size, memory capacity, etc.) while interpreting the benchmark results since the setup has an impact on results and needs to be understood before making comparisons for product selection.

Another commonly used set of benchmarks is published by the Transaction Processing Performance Council (TPC) to address system-level performance for transactional workloads (e.g., databases). TPC-C, TPC-E, and TPC-H are well known in this category. These benchmarks are very useful in gauging relative performance between different hardware platforms for a given software stack. However, based on the published benchmarking data available on the TPC Web site, it is challenging to make true comparisons between competing products, especially because the hardware components are not always similar, are typically configured differently, and may use different database software and operating systems. Additionally, the system configuration is often highly tuned to ensure there are no performance bottlenecks. This typically means using an extremely high performing storage subsystem to keep up with the CPU subsystem. In fact, it is not uncommon to observe system configurations with 1,000 or more disk drives in the storage subsystem for breakthrough TPC-C or TPC-E results. To illustrate this point, a recent real-world example involves a TPC-C

result for a dual-processor server platform that has an entry level price a little over \$3,000 (Source: http://www.tpc.org). The result from the published benchmark is impressive: more than 600,000 transactions per minute. But the total system cost is over \$675,000. That's not a very realistic configuration for most companies. Most of the expense comes from employing 144 GB of memory and over a thousand disk drives.

As with CPU2006, several caveats should be considered when interpreting TPC benchmark results. The primary caveat relates to how closely the benchmarked configuration reflects deployments in customer production environments. This is critical to understand before giving credence to any published benchmark data or comparisons between competing products. Next to consider are elements of the software stack – the operating system and the database software – and how they were configured, and comparing that to actual software configurations deployed in the customer environment.

Industry benchmarks are certainly valuable in establishing a standard methodology for comparing performance between competing products. But it must be noted that published benchmark configurations are designed to elicit the best performance that can be potentially delivered by the system. When using benchmark data for purchasing decisions, we advise caution before interpreting the results verbatim, and strongly recommend close analysis of the benchmarked systems and understanding how they relate to the customer scenario. Using a published benchmark result as-is for buying decisions is somewhat analogous to observing how a Formula 1 car performs on a race track and then using that data to purchase a commuter car that you intend to drive on the way to work.

#### CPU is typically not your bottleneck: Balance your systems accordingly

So how should you look at performance in the real world? First you need to consider what the typical user configuration is in your organization. Normally this will be dictated either by the capability or by cost constraints. Typically your memory sizes are smaller than what you see in published benchmarks, and you have a limited amount of disk I/O. This is why CPU utilization throughout the industry is very low: server systems are not well balanced. What can you do about it? One option is to use more memory so there are fewer disk accesses. This adds a bit of cost, but can help you improve performance. The other option—the one GFS likes to use—is to deploy balanced servers so that major platform resources (CPU, memory, disk, and network) are sized correctly.



Figure 1: Trend in CPU performance vs. disk performance

The chart above (Source: Intel white paper) shows that CPU performance has improved tremendously over the past 10 years while disk performance has remained relatively flat. To keep the CPU pipelines busy, the system architecture needs to be designed so that the other platform components (memory and disk) are capable of providing adequate data bandwidth to the CPU. More memory bandwidth means additional CPU pins to attach to memory channels, which drive up system cost and power. More drive bandwidth requires additional drives in high performance configurations (RAID10, RAID5), which drives up system power, cost, and rack footprint. If memory or disk bandwidth is under-provisioned for a given application, the CPU will remain idle for a significant amount of time, wasting system power. The problem gets worse with multicore CPUs on the technology roadmap, offering further increases in CPU pipeline processing capabilities. A common technique to mitigate this mismatch is to increase the amount of system memory to reduce the frequency of disk accesses.

While this relentless increase in CPU performance (courtesy <u>Moore's Law</u>) and the disparity it creates with the rest of the platform (notably memory and disk) means significant challenges for performance tuning, GFS views this as a great opportunity for server rightsizing. Multicore CPUs with increasing performance drive the commodity market towards a smaller number of CPUs in the server. What used to be a four-socket system can now be achieved with two sockets, and what used to be a two-socket server can now be a one-socket server. The challenge is to find the right balance of CPU/memory/disk for a given workload category for a given platform generation. That essentially is the end goal of the server rightsizing exercise.

This can be achieved by system load testing to understand system scalability for the workload scenarios. In a later section, we show how this can be done for Microsoft application environments. Again using a car analogy, why opt for the V-8 engine if you intend to drive mostly on city streets? Going with the four-cylinder engine is the far more efficient option in this scenario.

Another aspect to consider is shown in Figure 2 below. If you look at performance as measured by frequency for any given processor, typically there is a non-linear effect. At the higher frequency range, the price goes up faster than the frequency. To make matters worse, performance does not typically scale linearly with frequency. If you're aiming for the highest possible performance, you're going to end up paying a premium that's out of proportion with the performance you're going to get. Do you really need that performance, and is the rest of your system really going to be able to use it? It's very important from a cost perspective to find the sweet spot you're after.



Figure 2: Non-linear price performance of commodity dual-socket processors

#### The many costs of power

Power is an important consideration for a number of reasons, first of all because it is an ongoing, accumulating expense. If you look at the average power rate for commercial customers across the United States, it's about 10 cents per kilowatt/hour. When you factor that in over the lifecycle of the server, the numbers add up. Another factor is that power consumed in the server leads to power consumed elsewhere in the datacenter. One way to measure this is through a metric known as Power Usage Effectiveness, or PUE.

It looks at the fact that for every watt consumed by the server, there is a certain overhead of power consumed to power and cool the server, as shown in Figure 3 below. That overhead can be anywhere from 50 percent to over 100 percent. If you're in an older datacenter that's not very efficient and your PUE is 2.0, for every additional watt of power that you consume in the server, the datacenter infrastructure is adding another watt.

The rough rule of thumb is that every watt you save translates to roughly \$4-\$5 over the typical five-year life of the server. That's a ballpark number that obviously changes depending on the rates you pay for power. But it adds up when you have lots of servers.



Figure 3: Power Usage Effectiveness (PUE), measuring power and cooling overhead of servers in a datacenter

Normally power costs are thought of as operating expenses, but power consumption also generates substantial capital expenditures. In looking at the cost of building a typical datacenter, the general industry estimate is that it costs about \$10-\$20 million per megawatt consumed by the servers. We recommend that companies amortize that capital investment across the servers consuming the power. The chart shown in Figure 4 on the next page shows the three-year total cost of ownership (TCO) of a basic 1U server. Slightly less than half is the cost of purchasing the equipment. The blue section represents what you pay to the energy company after factoring in the PUE, the orange section is the depreciation of the datacenter incurred for the amount of energy consumed for that particular server, and the green section represents other operating costs. Over time it doesn't take long for the purchase price of the server to become less than half its total cost of ownership. In a five-year TCO the purchase slice gets even smaller.



*Figure 4: Three-year total cost of ownership of a basic 1U server* 

The rightsizing principles described in this document can help you achieve higher utilization, which means more efficient use of the datacenter. That in turn can help your current datacenter space meet your needs for a longer period of time so you can push out building or leasing a new datacenter.

#### Our approach to rightsizing servers

So how does GFS determine what type of servers to deploy? First and perhaps most importantly, we focus on understanding our datacenter workloads. Taking this approach involves measuring and analyzing our key applications and then modeling those characteristics for server selection.

Think back to the benchmark example of the system that delivered 600,000 transactions per minute with 1,000 disk drives. When you look at what happens to performance when you have 8, 24, or 57 drives, the number of transactions per second comes down by a factor of 10. See Figure 5 on the next page shows CPU utilization increasing with disk count as the result of the system being disk limited. As you increase the number of disk drives, the number of transactions per second goes up because you're getting more I/O and consequently more throughput. With only eight drives CPU utilization is just 5 percent. At 24 drives CPU utilization goes up to 20 percent. If you double the drives even more, utilization goes up to about 25 percent. What that says is that you're disk I/O limited, so you don't need to buy the most expensive, fastest processor. This kind of data allows us to rightsize the configuration, reducing both power and cost.



Figure 5: Typical system performance: Disk I/O performance is a dominant factor

### Web server testing

Turning to Web servers, the Web Capacity Analysis Tool (WCAT) is a Microsoft tool that we use for testing. It is publicly <u>available for downloading at IIS.net</u>. An alternative is an industry benchmark called <u>SPECweb2009</u>, which GFS has determined doesn't represent our particular usage models. WCAT is a Windows/IIS-related tool that's much more representative of what we do. It is very simple to set up. You take two servers that you are considering and look-up the numbers for those servers. Based on your workload you can choose between a CPU-intensive profile scenario and a disk drive-intensive profile, for instance. Figure 6 below illustrates the setup of the WCAT test environment, including the system under test and a machine that drives it.





Figure 7 below shows the results from a WCAT test. In this case a faster processor did make a substantial difference, assuming that content was served from the memory cache. In this case for Web serving, when we're comparing an older server to a new server, we did see a significant performance improvement with a faster CPU.



Figure 7: Web Capacity Analysis Tool test results

#### Selecting file servers

The next topic is file server performance. First, we look at storage characterization using <u>Event Tracing for</u> <u>Windows (ETW)</u>, a feature included with Windows that collects disk event traces. You can store these traces and then analyze them to draw conclusions about the behavior of file servers in your environment. The tool allows you to generate summary statistics on your transfer sizes, queue depths, input/output operations per second (IOPS), congestions, and so forth. In short the tool allows you to generate extensive statistical information about the ways in which your servers are being accessed. Each and every access that goes to the disk subsystem is captured. GFS captures such traces from our production servers running several different workloads and then use this information to create workload profiles. We also use the profiles to rightsize our servers, making sure we provision the disk subsystem with enough IOPs for the actual workload demand. Some parameters of disk controllers can also be set based on this workload analysis. Figure 8 below shows the analysis we performed on two RAID storage controllers: one with 512 MB of cache and the other with 256 MB of cache. When we looked at performance, beyond a certain queue depth the 512 MB controller began delivering significantly higher performance than the less expensive 256 MB controller. But when we looked at the results from our ETW workload analysis, we found that most of the time our queue depth never goes beyond 8 I/Os. So in our operational area, there is no difference in performance between the two RAID controllers. If we didn't have the workload analysis and just looked at those curves, we might have been impressed by the 10-15 percent performance improvement at the high end of the scale, and paid a premium for performance we would never have used.



Figure 8: Analysis of RAID storage controllers with different cache sizes

Figure 9 below illustrates another application where CPU performance does matter. In this case we're looking at performance scaling across four different processors. The data is normalized to the 2.0 GHz 80 W processor at left. The first bar represents frequency. The second bar represents performance as measured in jobs per second.



Figure 9: Compute workload performance/power/cost tradeoffs

The chart shows that in going from 2.0 GHz to 2.33 GHz, frequency increased 17 percent, but performance only rose 14 percent. So if we measure performance per gigahertz, we see 98 percent efficiency for the 2.33 GHz processor. That's great; but as you go up the frequency curve, the efficiency starts falling. For the highest frequency processor it's below 90 percent. Taking a closer look at that processor, the power (red) bar is 1.5 times that of the 2.0 GHz processor. Furthermore, the CPU list price goes up by a factor of five. So the highest frequency processor is clearly not attractive from a total cost of ownership standpoint. The 50 Watt processor saves power and costs less than the 2.67 GHz processor while delivering a bit lower performance, showing a sweet spot in the middle.

In this illustrative example we've taken a simplistic approach and only shown the list price and the power of the processor. When you're making actual server purchases it is important to take a platform-level view and examine how much the cost and the power of the platform varies with the different processors. But still this example helps show that often the highest overall efficiency is likely to be found somewhere in the middle when you consider price, power, and performance.

## How to ride the technology curve

Advances in processor performance show no signs of abating. How should IT departments deploy the new waves of technology? It helps to look at the history of microprocessor pricing. Basically the manufacturers have offered increased performance at the same price point as previous processors. For instance when dual core processors first came out, the manufacturers in effect said: "We're not going to charge you more just because it's dual core. It's the same price as the old single core that we're replacing. So you get two for the price of one." Then when they went from dual core to quad core they did the same thing; same price, but you got four cores.

For most customers the natural tendency was to take what had been a single-core platform, simply drop in a dual core, and then a year later drop in a quad core. The price stayed the same and in theory you got more and more performance. It seems like a great deal. But the actual performance doesn't scale linearly as the cores grow. As we noted earlier, if you are memory bandwidth limited or disk I/O bandwidth limited, then simply doubling the CPU core performance doesn't double your performance.

But the increase in number of cores does present an opportunity to rightsize and lower costs significantly. Because the increasing performance drives your CPU utilization down, you can move to two sockets where you previously had four sockets, and you can move to a single socket where you previously used two sockets. Rightsizing thereby provides an opportunity to drive your overall system costs down by taking the opposite approach from what might seem intuitive.

To illustrate how this can save you money, let's say you had a dual-core CPU and four sockets, for a total of eight cores. Let's say eight cores are what you want. Then quad-core processors are introduced and now you only need two sockets, bringing your cost down. The quad-core processors that support only two sockets are priced much more aggressively than the ones that support four or more sockets. The four-socket platforms are more expensive and are typically higher power as well. Once again, in the highest performance ranges cost goes up faster than performance. If your application can be scaled out across multiple two-socket servers, you can save money versus scaling up to a large socket count system. An actual example from our experience involved a dual-socket platform for a file server (see Figure 10 next page). With different configurations the utilization was anywhere from 11 to 35 percent. When the next "new and improved" processor came around, GFS' conclusion was that we could go to a single-socket platform and still get the performance we were after. That's because in the worst case, the 35 percent utilization could rise to 70 percent, but that isn't likely because the new processor is more powerful, so the resulting utilization will probably go to 55 percent. Looking at this kind of data our conclusion was that for file servers, in most cases we don't need two processors.





#### Summary

In conclusion, the first point to emphasize is that there is more to performance than just speed. When your definition of performance includes cost effectiveness, you also need to consider power. The next point is that in many cases processor speed has outpaced our ability to consume it. It's difficult to exploit CPU performance across the board. This platform imbalance presents an opportunity to rightsize your configurations. The results will offer a reduction in both power and costs, with power becoming an increasingly important factor in the focus on total cost of ownership.

It is also important to remember that industry benchmarks may not reflect your environment. We strongly recommend that IT departments do their own workload characterization, understand the behavior of the applications in their own world, and then optimize for that.

**Authors:** Dileep Bhandarkar, distinguished engineer and Kushagra Vaid, principal hardware architect, Global Foundation Services, Microsoft Corporation

#### **Terms & Conditions**

The information contained in this document represents the current view of Microsoft Corporation on the issues discussed as of the date of publication. Because Microsoft must respond to changing market conditions, this document should not be interpreted to be a commitment on the part of Microsoft, and Microsoft cannot guarantee the accuracy of any information presented after the date of publication.

This White Paper is for informational purposes only. MICROSOFT MAKES NO WARRANTIES, EXPRESS, IMPLIED, OR STATUTORY, AS TO THE INFORMATION IN THIS DOCUMENT.

Complying with all applicable copyright laws is the responsibility of the user. Without limiting the rights under copyright, no part of this document may be reproduced, stored in or introduced into a retrieval system, or transmitted in any form or by any means (electronic, mechanical, photocopying, recording, or otherwise), or for any purpose, without the express written permission of Microsoft Corporation.

Microsoft may have patents, patent applications, trademarks, copyrights, or other intellectual property rights covering subject matter in this document. Except, as expressly provided in any written license agreement from Microsoft, the furnishing of this document does not give you any license to these patents, trademarks, copyrights, or other intellectual property.

Microsoft, Bing, Hotmail, Microsoft SharePoint Online, Microsoft SQL Server, Xbox Live, he Microsoft Logo and/or other Microsoft products and services referenced are either registered trademarks or trademarks of Microsoft Corporation, in the United States and/or other countries.

The names of actual companies and products mentioned herein may be the trademarks of their respective owners.

© 2009 Microsoft Corporation. All rights reserved.