A Comparison of HP BladeSystem with Thermal Logic Technologies to Competitive Systems

Prepared for:
Hewlett-Packard Corporation

By:
Sine Nomine Associates
43596 Blacksmith Square
Ashburn, VA 20147

HP-2360-02
February 15, 2007
SNA Proprietary

Table of Contents

List of Tables

List of Figures

1 Introduction

Hewlett-Packard Corporation has introduced a line of blade server products with a radically redesigned power and cooling system. According to HP advertising, the "HP Thermal Logic" power and cooling system represents a new approach that significantly reduces the power consumption and required airflow for data centers requiring fully-featured, high-density servers.

This paper presents a laboratory study comparing "HP Thermal Logic" blade server power and cooling design versus a competitive blade platform as well as two traditional 1U rack servers, and evaluates the power consumption and cooling performance of each system under a workload that simulates a light and heavily-utilized data center.

The paper is divided into several sections. Chapter 1 is this introduction. Chapter 2 documents the rationale and methodology of this test suite, explaining what is being tested and how. Chapter 3 presents the results in detail, and Chapter 4 lays out the test team's conclusions.

At the end of this document are several Appendices that include helpful or interesting background information, including an overview of the physics of heat generation and removal in modern servers.

1.1 Why Old Thinking is No Longer Enough

As the density of servers in a modern data center increases, factors that could previously be overlooked or solved with simplistic designs become more significant. Nowhere is this more apparent than in power and cooling design, which is concerned with minimizing the rate of energy consumption of a system and simultaneously ensuring that the energy it does consume – and then releases as heat – is efficiently removed from the ambient air so that the system does not overheat and fail.

1.1.1 Limitations in the Physical Facility

In the past, data center designers could rely on having more power and more cooling capacity available as needed, but a number of factors have rendered this baseline assumption no longer valid.

In a densely-populated urban area, it is no longer safe to assume that one can simply add more power and more air conditioning to the facility. Many data centers have no more room for electrical switchgear, backup generators, and so on, nor for additional air conditioning equipment. In the largest cities, there are plenty of locales in which the local electrical utility is already supplying all the power it can allocate to a given street address due to limitations in the distribution grid or, ultimately, the generating capacity.

Even in less-urbanized settings, reducing power consumption is increasingly important as energy costs rise over time. With the cost of hardware and software remaining flat or even decreasing, and wage cost increases moderating in recent years, the cost of energy is becoming a more significant portion of the total cost of ownership of a large data center.

Although server studies -- including this one -- typically focus on large data centers, similar factors comes into play when discussing smaller installations, but for different reasons. For example, a small business or a branch of a larger business may need a "data center" that is no bigger than a closet, holding perhaps one blade system with servers and a few network components. Often these server closets are rooms that were never designed for this purpose -- sometimes having literally been closets before -- and thus power and HVAC capacity are not unlimited. In an older building, or in a leased office suite, there may be no opportunity to increase either available power or available airflow, and so energy efficiency becomes much more important than one might at first assume.

In the manufacturing environment, it is not uncommon for a small data center to be created in a plant floor "mezzanine" enclosure or in an engineering/supervisory office next to the factory. Again, these small facilities were often constructed without expectation of housing a data center, and therefore the power and HVAC constraints may be non-negotiable (or very costly to change).

For the reasons cited above, this study is applicable (with appropriate scaling) to a small installation even though it's primary focus is on a larger data center.

1.1.2 Changes in Server Architecture

Server density has increased dramatically, with the introduction of 1U (one rack unit) standalone servers and high-density "blade" enclosures that combine a dozen or more servers into an enclosure of about 10U height, thus achieving a density of greater than one server per rack unit for the first time.

Due to increasing gate count and clock speed of the processor chips, and increasing number of processors within the typical server, the power needed per server has also increased. Efficiency per unit function increases, but the quantity of functional units increases even faster, with the net effect being more power needed per server.

Virtual servers, using technology such as VMware, Xen, or similar systems, allow consolidating lightly-loaded physical servers by virtualizing them and then running the images on a smaller number of physical servers. Due to the load consolidation, these servers will naturally be more heavily utilized. Although the overall power consumption may be lower for a given workload, that consumption will be in a more-concentrated physical space, which again presents challenges to data center design. Add to that the natural tendency of businesses to fill the space freed up by server consolidation by adding more applications (and thus more servers), and the net effect often will be a much greater output of useful business tasks but also a much greater power and cooling requirement for the data center.

1.1.3 New Approaches to System Design

Thus, a convergence of factors has made power efficiency and cooling requirements increasingly critical parameters in data center design. The emergence of blade servers presents new challenges, but this legacy-free enclosure also offers opportunities to rethink the power and cooling design. This study, then, evaluates and compares the effectiveness of two such designs.

2 Scope and Methodology of Study

This study examines the power and airflow requirements of the systems under test in a typical rack-mounted data center environment.

The two blade systems compared in this study were the IBM BladeCenter-H with HS21 server blades ("BCH") and the HP BladeSystem c7000 enclosure with ProLiant BL460c server blades. Both of the systems are high-quality machines from first-tier hardware manufacturers, and both are similar in terms of processor specifications and memory type. Both IBM and HP provide a dedicated enclosure-level supervisory processor that manages and monitors all of the server blades in the system.

One differentiator of the two units, according to HP, is the Thermal Logic technology in the BladeSystem c7000 enclosure which allows it to adjust its fan speed over a wide range, responding dynamically to the thermal conditions sensed by an array of temperature sensors all around the enclosure and the circuit boards. HP also emphasizes the efficiency of airflow control in their design, and claims that their system's ducted fans and tighter air seals improve overall cooling efficiency. The purpose of this study, then, is to put those claims to the test in a direct comparison of overall power consumption and external airflow requirements in a simulation of a typical heavily-utilized data center workload.

A pair of standard 1U rack-mount servers, one from IBM and the other from Dell, were included in the comparison in order to ascertain the relative power and airflow requirements per unit server in traditional versus blade mechanical configurations.

2.1 Philosophy of the Design of Experiment

Data centers have to be provisioned to handle the maximum expected thermal and electrical load from the number of servers planned for installation at full capacity. For this reason, power testing of servers needs to reflect each unit's power consumption (which equates to heat emission) at its worst-case or near-worst-case level. Any decision to rely on load levelling – that is, an assumption that not all servers would be running at full capacity at any moment in time – can be made only by the data center designer with a full understanding of the enterprise's application mix, work scheduling, and strategic plans for new systems.

Modern servers consume much more power under compute-intensive workloads than under disk-intensive workloads. This is because drive manufacturers have improved the performance of mechanical parts and thereby reduced power consumption of disk spindles, while at the same time CPU power consumption has risen because of increased density and clock speed of the processor. Thus, the testing for this analysis was done under a CPU load that approaches maximum.

The phrase "approaches maximum" is used here, because in order to truly maximize CPU power consumption it is necessary to use a "pyrovirus" type program, that is, essentially a tight interrupt-disabled assembly language loop program. These are available for testing, but there are no real-world applications that work that way, without reference to data in RAM. A compute-intensive load simulator that taxes both the CPU and main memory was chosen as a reasonable approximation of the worst load that would ever be encountered in a useful business application.

2.2 Common Configuration and Methodology

In this study, there were two separate types of experimental testing: power and airflow. This section documents the hardware that was included in the test runs, as well as configuration details that were common to all tests. Very light utilization efficiency was similarly briefly studied, and since growth potential is also important, future power and cooling expandability was investigated

2.2.1 Hardware Under Test

The following hardware devices and enclosure configurations were tested and compared in this study:

Figure 1: HP ProLiant BL460c blade server

- HP BladeSystem c7000 enclosure with ProLiant BL460c server blades (a 16-slot enclosure) with 16 blades installed, in the following configurations:
  - 8 2GB DIMMs (16GB total) per blade, with interleaving of memory banks
  - 4 2GB DIMMs (8GB total) per blade, with interleaving of memory banks
  - 4 2GB DIMMs (8GB total) per blade, with no interleaving of memory banks
  - The HP server density has 16 servers in 10U for a greater than 1U density of 1.6 servers per U.
- IBM BladeCenter-H enclosure with HS21 server blades (a 14-slot enclosure), in the following configurations:
  - 4 2GB DIMMs (8GB total) per blade, without interleaving, 14 blades in enclosure. This is referred to as the "IBM single-wide" configuration.
  - 8 2GB DIMMs (16 GB total) per blade, with interleaving, 7 blades in enclosure plus one expansion module per blade (14 total slots filled by the 7 blades). This is referred to as the "IBM double-wide" configuration.

Figure 2: IBM HS21 blade server, plus memory - I/O and storage expansion blades

  - 8 2GB DIMMs (16 GB total) per blade, with interleaving, 4 blades in enclosure plus one expansion module and one I/O module per blade (12 total slots filled by the 4 blades). This is referred to as the "IBM triple-wide" configuration.

  - The IBM single-wide configuration has 14 servers in 9U for a greater than 1U density of 1.56 servers per U. The double-wide configuration has 7 servers in 9U for 0.78 servers per U, less dense than rack mount 1U servers. The triple-wide has 4 servers in 9U for 0.44 servers per U, less dense than rack mount 2U servers at .5 servers per U.

- IBM x3550 Rack-Mount Server, a 1U standalone server
  - 8 2GB DIMMs (16 GB total) , with interleaving
  - The IBM x3550 has 1 server in 1U for a density of exactly 1.00 servers per U.
- Dell PowerEdge 1950 Rack-Mount Server, a 1U standalone server
  - 8 2GB DIMMs (16 GB total) , with interleaving
  - The Dell PowerEdge 1950 has 1 server in 1U for a density of exactly 1.00 servers per U.

The nominal speed of the processors in each blade and in the standalone servers was 2.33 GHz, and there were two dual-core Intel Xeon 5140 processors in each server (that is, each 1U unit or each blade in the enclosure).

The memory modules (all 2GB DIMMs) were randomly selected from a bin of identical DIMMs.

Note that the IBM BladeCenter-H HS21 server blade has only four DIMM sockets on the main blades, and thus requires an expansion blade to support the eight-DIMM configurations. Furthermore, the IBM blade includes disk control on its main board, but this is not hot-pluggable. Thus, on the one hand the IBM system should need only one main board plus one expansion board per server to support eight DIMMs, but since the HP blades include hot-pluggable drive support on the main board, this would have meant the capabilities of the units were not identical.

Some applications will need only four DIMMs and will not need hot-pluggable disk storage, and therefore can use the full complement of 14 blades in the IBM enclosure and no expansion boards. Other applications will require more DIMMs and/or hot-pluggable disk control, in which case it is necessary to add expansion boards to the IBM enclosure, and thus reduce the number of servers the BladeCenter-H can support.

In the interest of fairness, all three IBM configurations were tested separately, and compared against the closest matching HP configuration. The reader is invited to select the test results most relevant to his or her own requirements when evaluating these two systems.

2.2.2 Load Simulation Software

Both the power tests and the airflow tests were run with the servers heavily loaded, so that power consumption and airflow requirements would be near their maximum values.

A program called PRIME95, which uses Fourier transforms as a filtering mechanism to identify very large prime numbers, was selected as the load simulator for these tests. PRIME95 attempts to locate Mersenne Primes, that is, prime numbers with specific mathematical properties. The

software creates large tables in main memory for the Fourier transforms and then performs extensive computations against this data, thereby exercising the CPU and its registers heavily.

As indicated in the chart, one interesting characteristic of the PRIME95 algorithm is that its load is predictably cyclical, varying in a smooth sawtooth-like pattern that repeats at intervals of approximately 3.25 hours on a 2.33 GHz Intel Xeon 5140 processor. The algorithm repeats a core set of computations but uses different-sized memory tables with each iteration. The more the processor accesses memory, which is slower than its own internal registers, the lower the power consumption. The difference in power consumption between small and large memory tables is both measurable and repeatable, rapidly rising to maximum at the beginning of the cycle and then gradually tapering off until the end of the cycle. In the systems under test, this overall cycle was set to repeat indefinitely until interrupted by an operator at the conclusion of the test period.

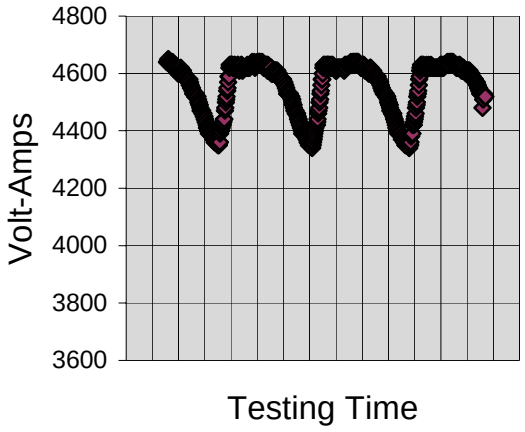

Figure 3: Typical PRIME95 Load Pattern

Because of the cyclical load, each power consumption test was scheduled to run for a minimum of 4 hours in order to capture at least one entire cycle of the PRIME95 algorithm. PRIME95 is publicly available at www.mersenne.org.

2.2.3 Test Execution and Instrumentation

Each test consisted of several specific sequence steps, the timing of which was controlled and carefully logged:

1. The operating system was booted (if not already running) on all blades and allowed to run at an idle state for at least 20 minutes, which was the amount of time empirically determined necessary to stabilize the system's power consumption in the idle state after start-up. During this phase, the PRIME95 software was prepared for execution in its multiple instances (four per blade) but not actually started.

2. The PRIME95 application was started on each blade, in backplane-position order, as rapidly as possible. The beginning and ending of the start-up period was logged and treated as a start-up transient when analyzing the result data.

3. The actual test period continued for a minimum of four hours, during which time the systems ran the PRIME95 application exclusively. No other operator-initiated tasks were performed on the servers except for minimally-intrusive checking to make sure the system had not crashed or issued errors.

4. The test period was officially stopped, and the time carefully noted in the log. Only after the end of the test period was the execution of PRIME95 interrupted by the operator.

If at any time during a test there was a hardware or software failure resulting in interruption of PRIME95, the entire test was repeated after correcting the problem.

2.3 Test Methodology and Instrumentation -- Power

2.3.1 Environment

The power tests were run in an environmentally-controlled chamber that represents a typical rack area in a hypothetical large data center. The chamber's available HVAC resources were such that there was no possibility of the servers under test requiring more airflow than could be delivered by the HVAC, and the temperature of the air at the inlet of the enclosure under test was monitored independently using a digital thermometer.

Figure 4: Test environment

The units under test were installed in separate racks, and the airflow conditions were carefully evaluated to ensure that there was no significant reflow of hot air from the back of the racks (the "hot aisle") to the front of the racks (the "cool aisle"). The two blade enclosures were installed in their respective racks at approximately the same height.

2.3.2 Instrumentation

The power consumption of the servers was monitored by their own internal supervisory modules, whose capabilities varied by brand and model, and by external Voltech model PM300 power monitors. The Voltech PM300 units were identical in configuration and were assigned at random to the servers under test to eliminate any measurement bias. All calibration was current, and the PM300 units were cross-compared at the beginning of this study and found to agree with each other within expected tolerances.

Figure 5: Voltech PM300 power monitor

Data logging of events such as start and end of test phases, and data collected from the real-time self-monitoring of the systems under test, was done manually by Sine Nomine Associates (SNA) personnel, and data from the Voltech power monitors was collected by an automated IEEE-488 (GPIB) based data acquisition system under close SNA supervision.

At the conclusion of each test, the Voltech PM300 logs were examined to ensure that they contained the expected time ranges. Any extraneous data after the official end of the test was discarded, because it was not desired to measure transients during the shutdown of the systems under test.

The watt and volt-ampere readings from the Voltech power monitors were analyzed over the test period using several statistical approaches, so that they could be compared in various ways. The absolute minimum and maximum values were obtained, with the minimum value occurring (without exception) during the near-idle state of the system just as the test began. An overall average was computed for all samples over the entire test run, and also a five-sample running

average was computed over the test run and its peak identified and reported. The five-sample peak algorithm was designed to eliminate artificially high "maximum" values that might be reported if there was a measurement transient or other artifact.

The power factor of each system was also monitored over the entire test run, and its average, minimum, and maximum values computed.

Ambient air temperature and humidity were continuously monitored throughout the testing period, using multiple instruments and probe locations including the internal capabilities (if applicable) of each system under test. The maximum variation in air temperature at the inlet of the enclosure under test was approximately 2 degrees Celsius over the entire study period, and humidity variation was negligible.

2.4 Test Methodology and Instrumentation -- Airflow

Numerous factors affect the total energy cost of a fan-cooled server in a data center. As the CPU and devices draw less power, there is less total heat to dissipate. More efficient fans (those with better motors, or a better blade design) require less power to move the same volume of air. A well designed thermal architecture, including more efficient heat sink design and ducted airflow paths, allows the same amount of heat to be removed with a smaller airflow volume. A dynamic fan system that operates at the level of effort required to remove its heat load, and no higher, will consume less power than one which runs at full capacity regardless.

In addition to power demand, tests were run to measure volumetric airflow, measured in cubic feet per minute (CFM). Airflow is determined by the physical characteristics of the chassis thermal design and the speed of all operating fans.

The tested server hardware included a number of differences in fan design, technology, and thermal architecture. Fan sizes varied; some used squirrel cage designs, and others used ducted fan designs. Fan speeds between different server hardware are therefore not directly comparable, but CFM values are.

For hardware supporting fan speed modulation, tests were run over a variety of fan speeds. The other hardware tested showed a consistent fan speed in the power tests; CFM measurements were made in comparable conditions, including the CPU load simulation.

2.4.1 Equipment and Configurations Tested

Airflow volume tests the rate of airflow through the system at the same amount of atmospheric pressure due to the fan speed. An airflow volume test is used to determine the performance of a fan or blower.

For hardware that provides manual control of fan speeds, the CFM test measured airflow generated by the fans at a range of speeds that were observed in the power tests. For hardware that did not allow manual control of the fan speeds, the CFM test simulates airflow generated by the fans under peak processor load. As in the power tests, Prime95 was used to simulate the saturated processor load.

The airflow test chambers are also used to measure the pressure required to force a given volume of air through the system. The server hardware is mounted on the front of the chamber and the

volume flow is derived from the static pressure differential of a calibrated nozzle inside the chamber when the chamber pressure is maintained to equal atmospheric pressure.

2.4.2 Environment of Airflow Test

Ambient air temperature varied between 21 and 25C. Temperature readings were taken during each test and used in the calculations to ensure accurate measurements of airflow.

2.4.3 Mechanical Installation

Volumetric airflow tests were performed in an isolated airflow testing apparatus. All tests were run using a 4 inch diameter calibrated nozzle. A temperature reading was taken inside the airflow testing chamber each time a test was performed.

2.4.4 Metrology and Lab Instruments

The amount of air being moved through the system by the fans was measured using a 1200 CFM Chamber manufactured by Airflow Measurement Systems. The pressure measurements were taken using a Setra Pressure Transducer, which was last calibrated on September 15, 2006.

For each tested configuration, the pressure was allowed to stabilize, and pressure values were recorded manually from the Setra.

For each tested configuration, the server was allowed to stabilize, which took between 5 and 10 minutes in each case. Then the ingress air pressure was manually adjusted to match ambient pressure, which took an additional five minutes. Once ingress pressure was stable at ambient pressure level, the measured pressure was read.

A pressure measurement is taken in a pre-ingress air chamber in the testing apparatus. CFM values are derived from the pressure differential between measured and ambient pressures.

This formula is given by the Airflow Management System for running this test to calculate the volume of airflow.

$$V = Y\ C\ A\ \sqrt{(2g\ (\Delta P)/\rho)}$$

where:

V : volume of airflow in CFM

Y : net expansion factor for compressible flow through a nozzle

C : flow coefficient for nozzles

A : Area of the nozzles

g : gravitational constant

ΔP : Differential Pressure

ρ : Density of Air

This is a standard formula for calculating volume airflow, and is documented in detail by the manufacturer of the airflow test chamber.

2.4.5 Airflow Test Procedure

The team tested three different IBM configurations, one Dell configuration, and one HP configuration.

Fan speeds cannot be controlled manually in the IBM or Dell configurations. For this reason, all of the IBM and Dell systems were tested with PRIME95 running, since there had been negligible fluctuation of fan speed and fan power during the earlier power tests.

HP's fans varied significantly during the power tests as the control software adjusted fan speeds to meet the heat load. The HP system was tested over a range of manually configured fan speeds to determine the relationship between fan speed and CFM.

All tests measured total airflow, including power supply fans and blade fans, where appropriate. The airflow is then multiplied by a factor to give the equivalent to 224 machines. 224 machines are used to equalize the number of servers, given that the HP enclosure has 16 servers while the IBM has 14 servers.

3 Test Results

3.1 Power Tests

Since there were varying numbers of active servers in the different test configurations, useful comparisons of power utilization are best made on a per-server basis rather than a per-enclosure basis.

The power data logs were analyzed to obtain the peak values at maximum load as well as a "peak-average" value, that is, the maximum value obtained by a five-sample running average. Since samples were taken once per minute, this represents a five-minute sustained load figure that eliminates measurement transients due to starting and stopping of fans and similar causes.

As shown in the chart below, which lists the absolute peak power consumption of each server, the HP BladeSystem BL460c consumes the least power per server in the 4-DIMM non-interleaved configuration that is directly comparable against the IBM BladeCenter-H in the "single-wide" configuration. The difference in power consumption between the HP and IBM blade systems in this configuration is not statistically significant.

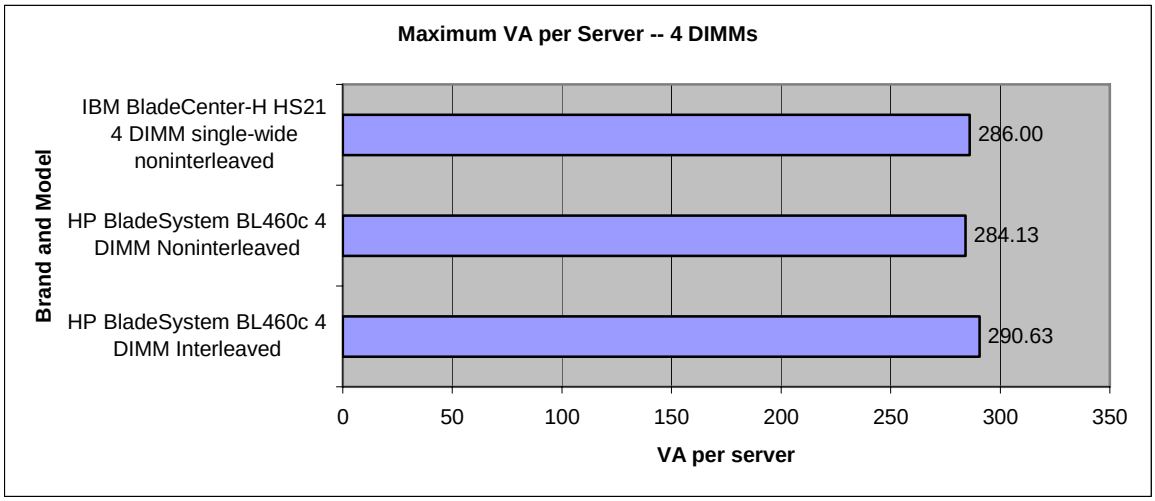

Figure 6: Maximum VA per server (4 DIMMs)

The next chart provides the same measurement for the 8-DIMM configurations tested, including the 1U rack servers.

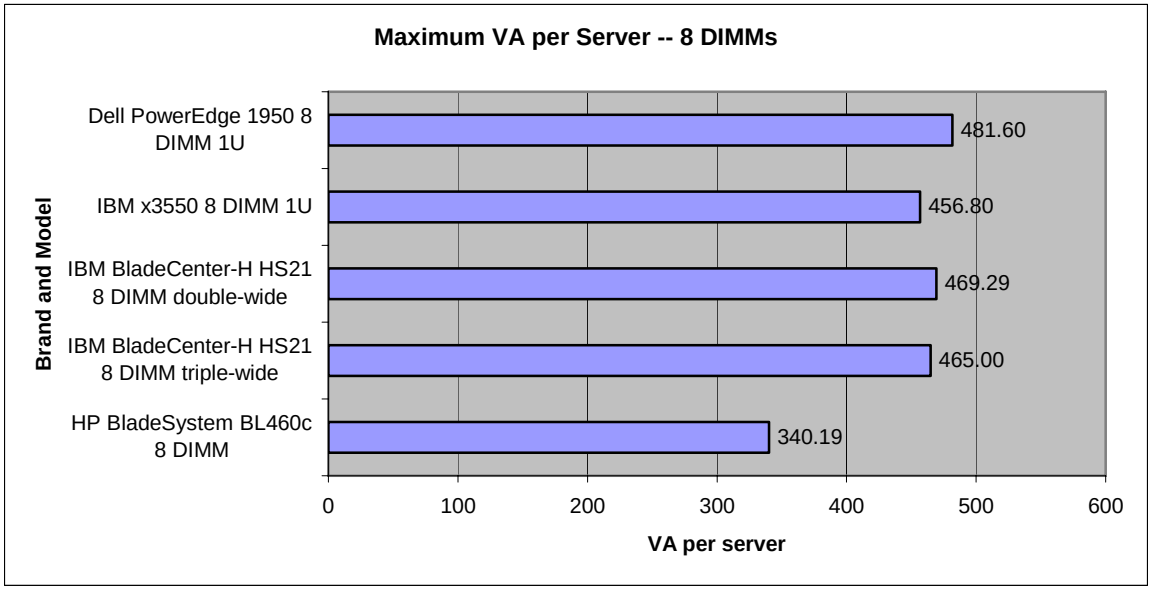

Figure 7: Maximum VA per server (8-DIMMs, all with interleaving)

As expected, the configurations with 8-DIMMs consume more power than the 4-DIMM configurations, regardless of brand or model. Note, however, that the 8-DIMM IBM blades consume approximately the same power per server as the 1U standalone units, whereas the HP blade required only about 75% as much power as the other 8-DIMM configurations, be they standalone or blade.

In the above chart, a measurement anomaly caused the IBM "double-wide" blade to have higher peak utilization than the "triple-wide" configuration next to it. This is counter-intuitive, and close examination of the test data reveals a sharp but brief power transient in the "double-wide" data stream that affected only one sample, but thereby artificially raised the absolute peak value for that test.

If the "peak-average" values, which nullify such measurement artifacts, are used, the data more closely follows the intuitive notion that a server with three boards should (and does) consume more power than the same server without one of those boards. The following chart shows the same data as above, but this time using the peak-average analysis method, which the test team considers to be more accurate.

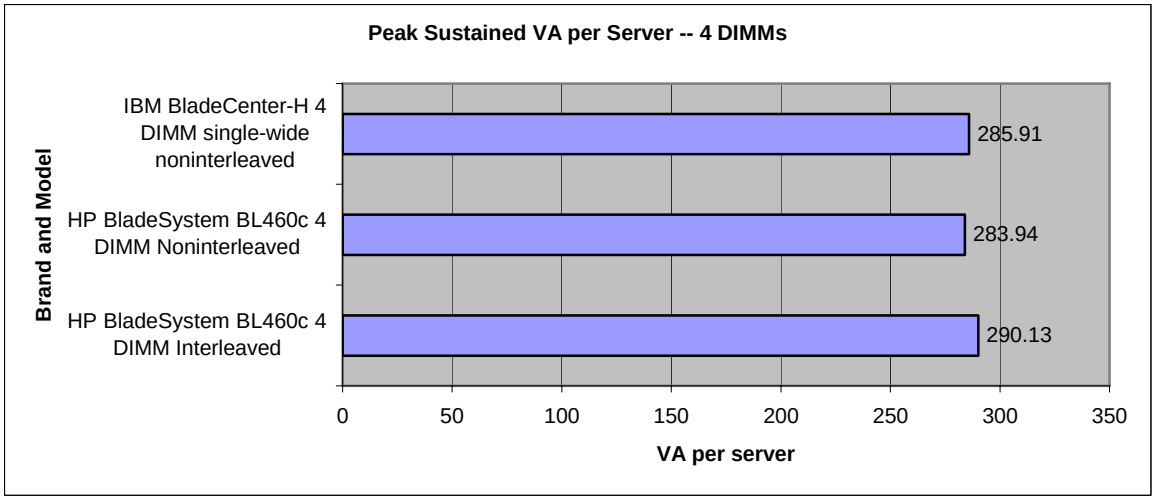

Figure 8: Peak Sustained Volt-Amps per Server

When only 4-DIMMs are needed and there is no requirement of hot-pluggable disk drives, the IBM and HP blade servers are closely competitive with one another for power consumption per server, differing by less than 1% in these tests (non-interleaved memory).

However, as soon as more memory or hot-pluggable drives are needed, IBM's use of expansion boards costs dearly, and the HP BladeSystem with ProLiant BL460c significantly outperforms the IBM BladeCenter-H with HS21. In these tests, the BL460c used only about 340 VA/server versus 407 VA/server for the BladeCenter-H in its minimal 8-DIMM configuration, meaning the HP blade consumed about 16.5% less power than the IBM blade, per server. The following chart shows the various 8-DIMM configurations, again tracking the peak sustained VA.

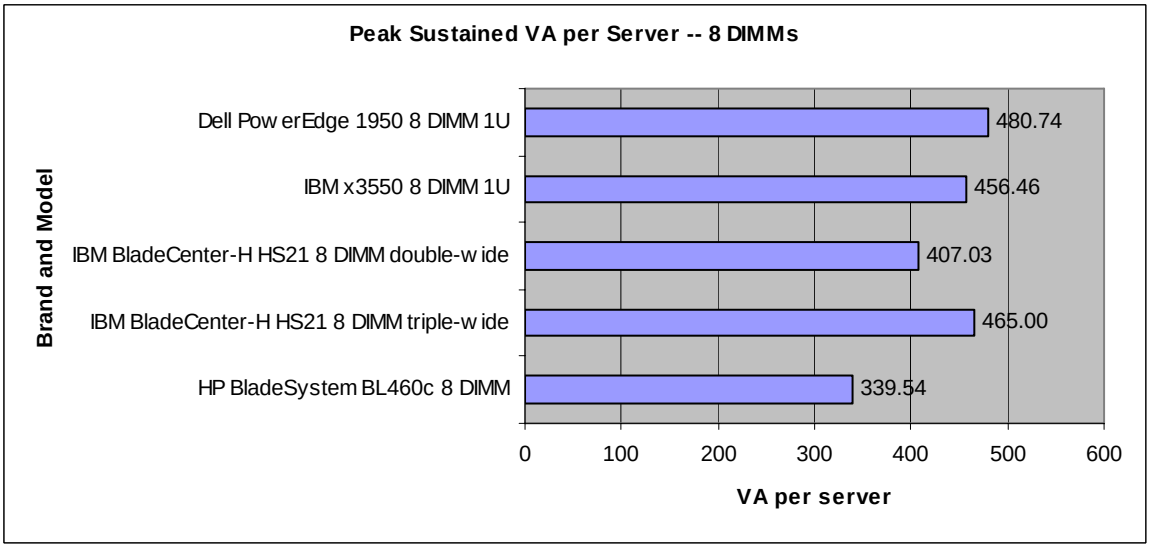

Figure 9: Peak Sustained VA per Server (8 DIMMs, all interleaved)

If hot-pluggable drives and larger memory are both needed, then the triple-wide IBM configuration is required and the power usage rises to 465 VA/server, which means the HP BladeSystem uses 26.8% less in this instance.

As in the previous chart, the IBM and HP blade servers equalled or beat the 1U traditional servers with regard to power usage per server, thus disproving the claims sometimes heard that blades are more power-hungry than traditional 1U servers.

3.1.1 Effect of Fan Power on Overall System

The HP BladeSystem's supervisory console (HP Onboard Administrator) provides a readout of the speed and power consumption of each of the enclosure's ten fans, and this data was logged during the test runs of the HP BladeSystem. The total power consumed by all ten fans ranged from a low of 133 VA to a high of 208. This difference of 75 VA represents the power that is saved by the Thermal Logic supervisor throttling down the fans when less cooling is needed. The lower values were observed at idle time, not during the test run, and represent an additional power savings opportunity (if application conditions permit) that is above and beyond the overall enclosure power figures shown previously.

3.1.2 Power Consumption at Idle and Moderate Loads

The test team also measured the power consumption of the systems at idle state (operating system booted but no applications running) and at the lowest "trough" of the power levels during the PRIME95 test, which represents a moderate-to-high load that falls well short of resource saturation. The following charts present these results for the 4 DIMM and 8 DIMM configurations (this data was not collected for the HP BladeSystem c7000 with ProLiant BL460c with 4 DIMMs interleaved).

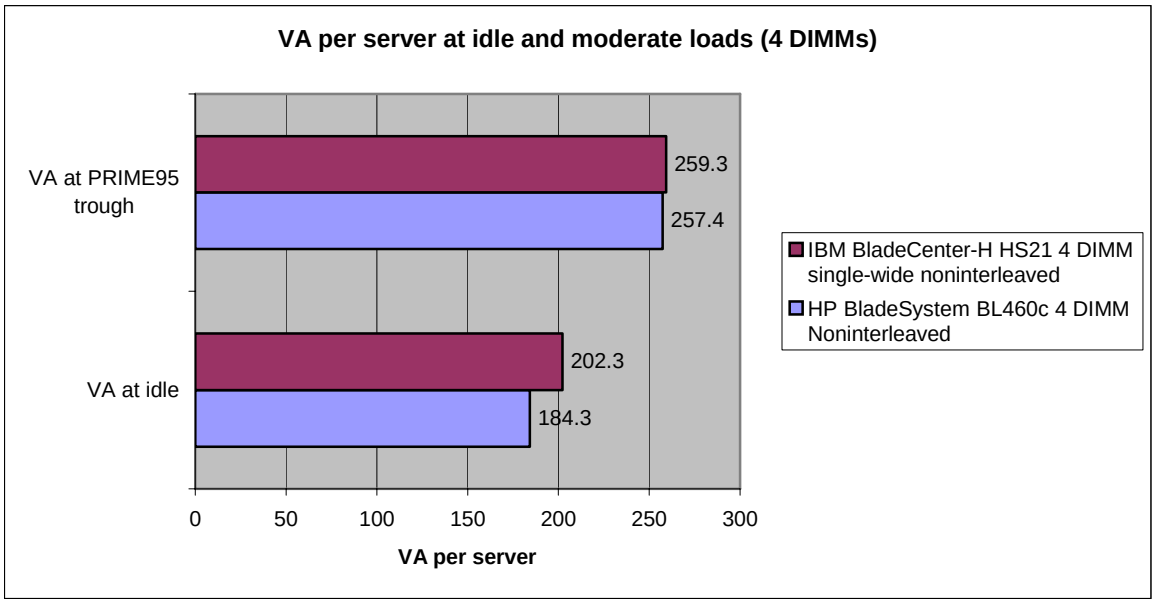

Figure 10: VA per server at idle and moderate loads (4 DIMMs)

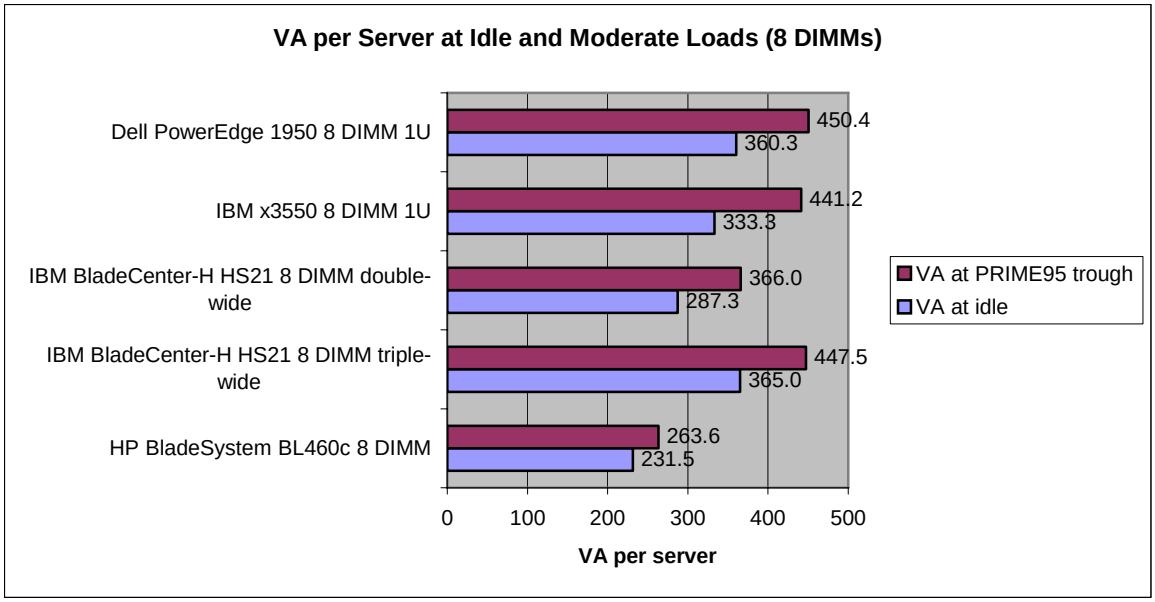

Figure 11: VA per server at idle and moderate loads (8 DIMMs)

The HP BladeSystem consumed slightly over 20% less power per than the IBM BladeCenter-H in its 4 DIMM configuration. In the 8-DIMM configuration, the HP BladeSystem consumed less power per server than any of the other configurations.

As with the full-load measurements (the PRIME95 peaks), the IBM BladeCenter-H performs considerably better if configured in the double-wide configuration than in the triple-wide.

Both the IBM BladeCenter-H double-wide and the HP BladeSystem performed better in these tests than the IBM x3550 or Dell PowerEdge 1950 standalone servers, consuming less power per server than their standalone counterparts.

3.1.3 Accuracy of Internal Power Metering (HP BladeSystem only)

As an additional test, the team evaluated the accuracy of the internal power metering for the HP BladeSystem's supervisory system, comparing the firmware-reported power readings against the readings from the externally-connected Voltech power monitor. The chart in Figure 12 presents the results, with the horizontal axis representing the time (in minutes) since the beginning of the test run.

When interpreting this chart, it should be noted that the readings on the external meter were manually logged, and that there was some inherent test jitter resulting from that logging process because the displayed numbers were constantly changing and it was not practical to record them at the exact same instant as one recorded the firmware-reported reading. Even so, the two sets of readings correlate quite closely, with the worst error being about 4.7% and errors of less than 2% being typical of most of the data points.

This test was run only for the HP BladeSystem enclosure, and not for the other systems, so no direct comparison is possible at this time. This test does, however, illustrate that the HP

BladeSystem's firmware-reported power readings are more than adequate for normal system administration purposes.

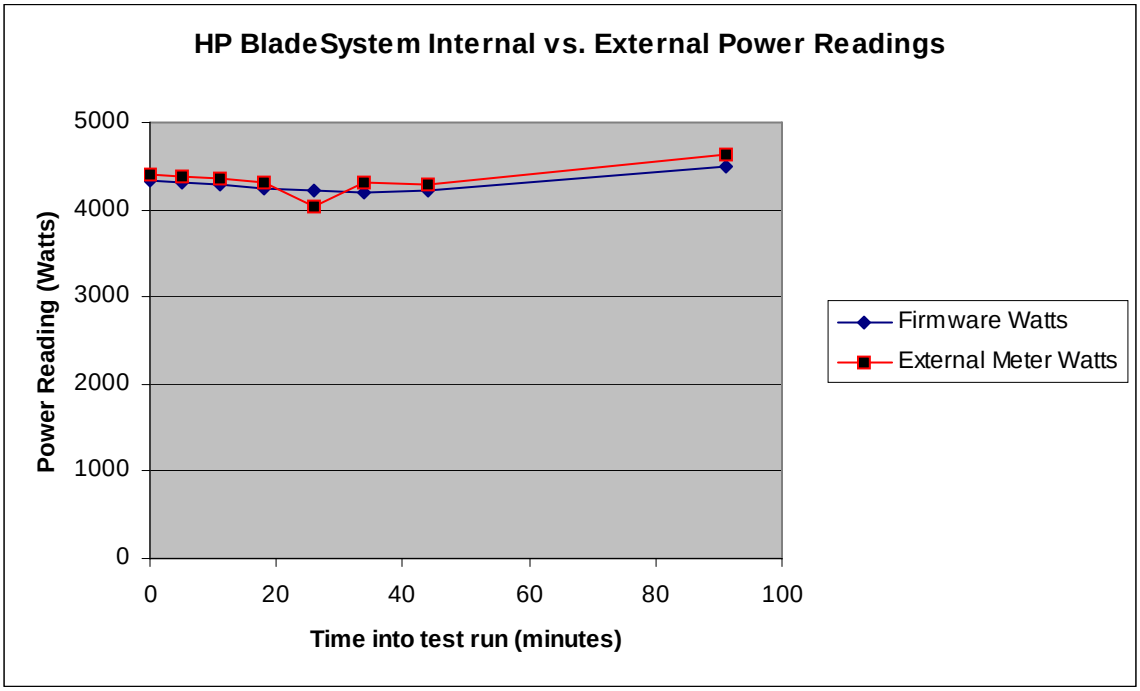

Figure 12: Firmware-reported power vs. externally-measured power

3.2 Airflow Tests

An analysis of the fan speeds on the HP BladeSystem c7000 enclosure shows the fan speeds to be highly linear in respect to the resulting airflow volume. During testing, the fan speeds varied but never approached the maximum speed of which they are capable. This test was not run on the IBM BladeCenter-H because that system does not allow manually setting the fan speed from the supervisory console.

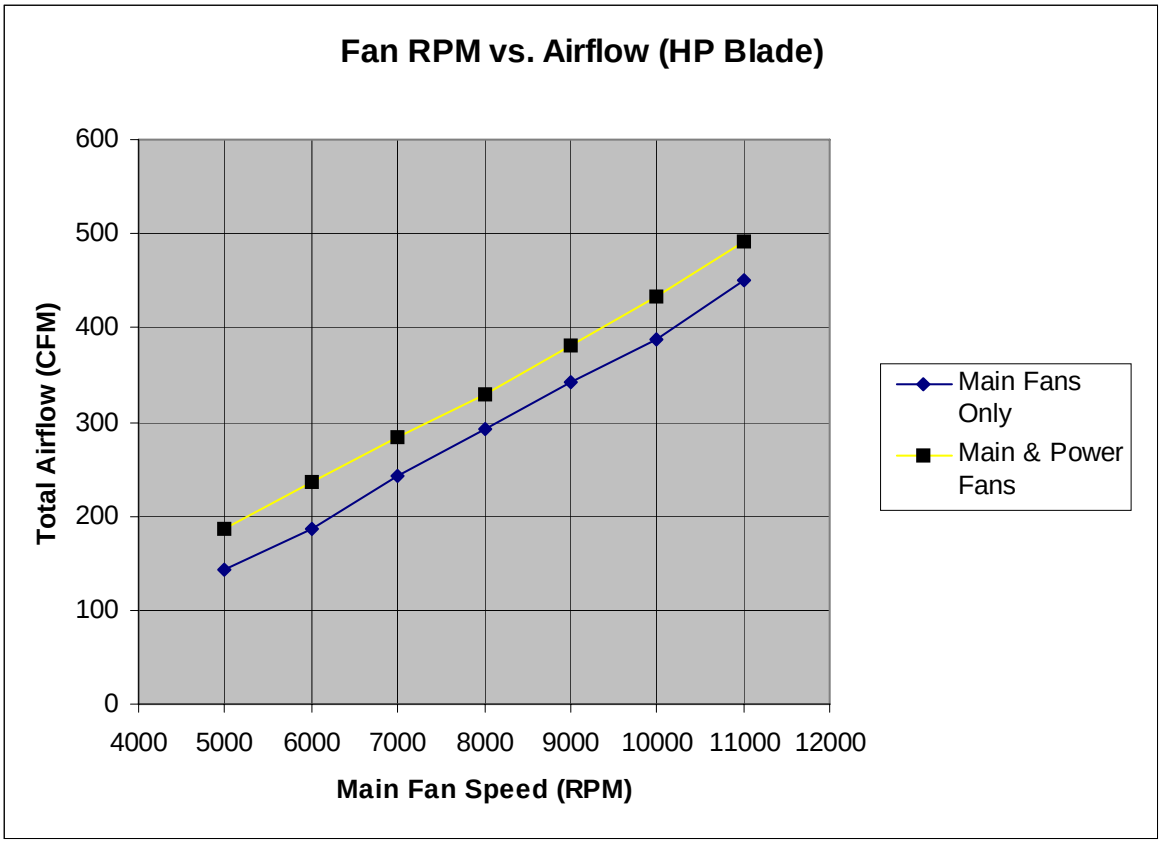

Figure 13: Fan Speed vs. Airflow for the HP BladeSystem

The raw data points, for reference, are:

Fan RPM	CFM (w/o Power Supply)	CFM (with Power Supply)
5000	144	186.5
6000	185.6	236.5
7000	242.7	283.2
8000	291.6	329.2
9000	342.4	381.5
10000	386.7	433
11000	449.6	490.7

Table 1: Fan RPM vs. Airflow Data

A linear regression gives us the relationship CFM = 0.05 fan speed – 66.74 with a regression coefficient of 0.9988. This allows us to accurately interpolate CFM for the HP BladeSystem c7000 from any given fan speed. Growth capacity is important for customers to future-proof their implementations. The PRIME95 HP configuration did not draw over 344 CFM, but the linear

regression in Figure 13 shows HP's future capacity of over 800 CFM. The IBM BladeCenter-H also did not "max out" its fans during the PRIME95 tests, but the linear regression projection is not available because it was not possible to manually set that system's fan speeds.

There is a limit on the accuracy of the airflow measurements on the 1U servers due to the fact that the enclosure for these machines are much less airtight than the blade servers. There was noticeable airflow around these machines during the test, detectable by placing a hand near the machine and feeling the air escaping through the holes in the enclosure. This would cause an under reporting of the actual airflow drawn by the fans in these machines.

Machine Type	Airflow (CFM) for unit	Airflow (CFM) for equivalent of 224 servers
Dell PowerEdge 1950 (1U)	40.4	9049.6
IBM x3550 (1U)	27.4	6137.6
IBM BladeCenter-H HS21, double-wide (7 per enclosure, 8 DIMM)	375.3	12009.6
IBM BladeCenter-H HS21, single-wide (14 per enclosure, 4 DIMM)	404.1	6465.6
For the HP BladeSystem with ProLiant BL460c server blades with 8 DIMMs the fan speeds ranged from 7230 rpm to 8208 rpm during the power tests. Using the linear relationship shown above this gives a CFM range of 294.8 to 343.7 over the course of the power test. The maximum fan speed is used here.		
HP BladeSystem 16 blades 8DIMM	343.7	4811.24
For the HP BladeSystem with ProLiant BL460c server blades with 4 DIMMs non-interleaved the fan speeds ranged from 5951 rpm to 6928 rpm during the power tests. Using the linear relationship shown above this gives a CFM range of 230.8 to 279.7 over the course of the power test. The maximum fan speed is used here.		
HP BladeSystem 16 blades 4DIMM	279.7	3915.2

Table 2: Measured Enclosure-Level Airflow

3.3 Fan Failure Test

The test team simulated failure of cooling fans on each of the two blade enclosures. In the case of the IBM BladeCenter-H, there are two main fans, and the simulation involved failing one of the main fans (a 50% failure). For the HP BladeSystem c7000, there are ten main fans, and the test

involved failing them one by one (at intervals of approximately ten minutes) until only four were running.

During the test interval, the input power (watts) was monitored for each enclosure to observe the impact on overall power consumption when cooling fans fail. The results are reported on a per-server basis (16 servers for the HP BladeSystem, and 14 or 7 servers for the two tested configurations of the IBM BladeCenter-H).

The IBM system continued to operate successfully with only one of its two fans, though the remaining fan increased speed dramatically, causing the system's power consumption per server to rise, as shown in the following chart. The enclosure-level supervisory system issued an operator alert when the fan was disabled, but otherwise the system continued to operate normally.

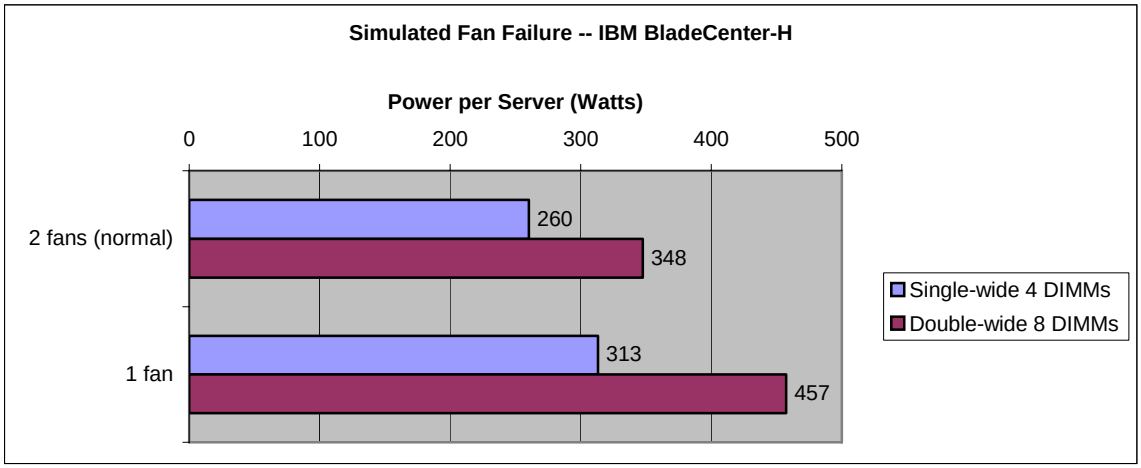

Figure 14: Effect of fan failure on power consumption for IBM

The HP system also continued to operate successfully, even with only four of its ten fans running. The remaining fans increased in speed to compensate for the failure, but the power did not increase as dramatically with the HP BladeSystem as it did with the IBM BladeCenter-H. The power consumed by the remaining fans increasing in speed was approximately equal to the power "saved" by the absence of the disabled fans. The enclosure-level supervisory system issued alerts as each fan was disabled, and the remaining fans increased speed to compensate, but otherwise the system's operation was normal until the sixth fan was removed.

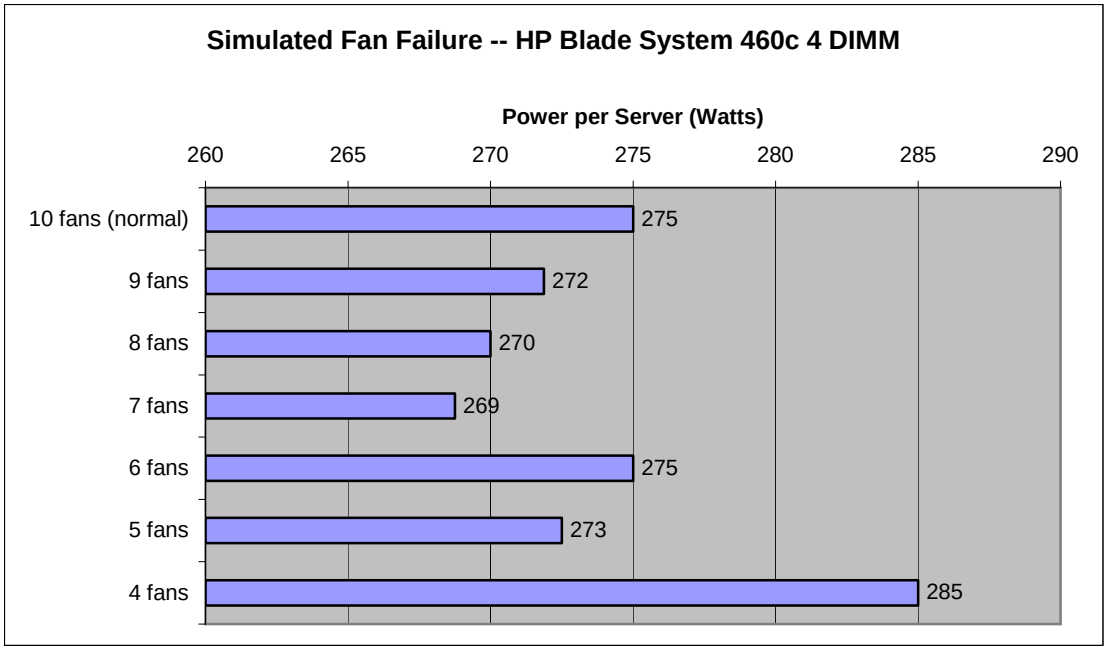

Figure 15: Effect of progressive fan failure on HP BladeSystem c7000

As the sixth fan was disabled in the HP, the remaining four fans quickly sped up to compensate for the lost airflow, which finally caused the system's power consumption per server to increase sharply.

WARNING: This series of tests was conducted under controlled laboratory conditions in order to measure the effect of major component failures. These results do not imply that customers should run either of these systems for extended periods without a full complement of fans according to the manufacturers' respective specifications.

3.4 Subjective Comments

Although the tests themselves have been made as objective as possible, the test team noted a few subjective observations during the laboratory sessions that are worth repeating here.

There was a significant difference in the level of detail available from the HP blade supervisory module versus the IBM blade supervisory module. The IBM console provided much less data, for example offering RPM values only for the power supply fans, which didn't really vary in speed, rather than the adjustable main fans. The HP supervisory console provided the opposite, with data available on the ten main fans but not the two power supply fans.

Any time the IBM supervisory unit was restarted, the main fans in the BladeCenter-H went into maximum speed "failsafe" mode for a minute or so. The amount of air moved by those fans, and the noise levels they emitted, were both extraordinarily high.

Examination of the interior of each blade enclosure left the test team somewhat disappointed with the IBM BladeCenter-H, which seems to fall short of IBM's usually outstanding mechanical design. The interior air gates, which are meant to close off airflow to unused enclosure slots, did not fit well, nor did they close tightly, and blades sometimes would catch on the gate mechanism

when being inserted into the enclosure. By contrast, the HP c7000 enclosure had very snug air gates that opened using a sturdy pushrod mechanism that never fouled during the tests.

The IBM BladeCenter-H has an acceptable level of redundancy with its two main fans, demonstrating that it can run successfully with only one of them. The HP BladeSystem BL460c exceeds this by providing a more gradual failure process in the unlikely event of multiple simultaneous fan failures.

Both the Dell and the IBM standalone servers impressed the test team with their excellent quality of mechanical design and easy access to removable components such as disk drives and memory modules.

4 Conclusions

4.1 Power Tests

Both the IBM and HP blades consumed less power than the standalone servers on a per-unit basis, although of course the overall power numbers for a blade enclosure are of course higher than for a single server. This clearly refutes the sometimes-heard opinion that blades are inherently less power efficient than standalone servers.

For a small server configuration, with non-hotplug disk drives (i.e., no hot-swappable RAID mirror) and with modest RAM requirements, there is very little difference in power consumption between the IBM BladeCenter-H and the HP BladeSystem c7000. As the server configuration increases, however, the HP blade outshines the IBM for power utilization by either 16% or 27%, depending on the need or lack of need for hot-swap disk drives. Clearly, IBM's decision to rely on expansion boards for systems needing more memory or I/O capability hurts the BladeCenter-H's power consumption figures.

In terms of server density, the multi-board IBM solution suffers an even larger disadvantage versus the HP single-board design. This is exacerbated by IBM's odd-sized 14-slot enclosure, which wastes two slots if the blades are configured for hot-plug drives and large memory, that is, in the triple-wide mode. Only 12 out of the 14 slots are used in this configuration, though one could presumably put one double-wide or two single-wide blades into those slots. Even so, this would only bring the total to 4 triple-wide blades plus 1 double-wide or 2 single-wide.

As predicted, the HP BladeSystem suffers a slight power penalty in the 4-DIMM configuration if memory banks are interleaved, versus non-interleaved. This is a result of the CPU being more fully utilized when the memory is interleaved, and the system administrator can choose to trade off performance or energy usage depending on the needs of the application.

The ability of HP's Thermal Logic to throttle back the enclosure cooling fans when less airflow is needed results in a measurable power savings, which amounted to 75 VA in our tests between the idle state versus the fully-loaded state of the enclosure when all blades were running the test load.

4.2 Airflow Tests

Looking at the CFM data it is clear that the in the situation where 8-DIMMs are needed, the IBM BladeCenter-H generates 2.5 times as much volumetric airflow as the HP BL460c. In the situation where only 4-DIMMs are used, the IBM BladeCenter-H generates 1.7 times as much volumetric airflow as the HP BL460c.

This represents wasted effort by the fans in the IBM blade server, which is a component of the higher power draw of that hardware.

Appendix A: Overview of Cooling Physics

From the view of a mechanical engineer, a computer is a device that converts electrical energy into thermal energy – in other words, a space heater. Every Joule of energy that enters a computer's power supply will eventually need to be removed as heat, and this of course requires further energy expenditures in the heating, ventilation, and air conditioning (HVAC) systems of the building.

In order to maintain constant temperature, the quantity of heat produced must equal the quantity of heat removed from the facility. This chapter will therefore examine both terms of the heat equation.

4.3 Heat Generation in the CPU and Memory

Most modern CPUs and memory chips employ complementary metal oxide semiconductor (CMOS) technology, which is capable of very high speeds and extremely high power efficiency. The output stage of a typical CMOS logic gate comprises two series-connected transistors of complementary polarity (hence the "C" in "CMOS"). One transistor pulls the output positive (usually logic "1") when activated, and the other pulls the output to ground (usually logic "0") when activated. The figure below illustrates a vastly-simplified example of a logic gate and its output-side electrical load model.

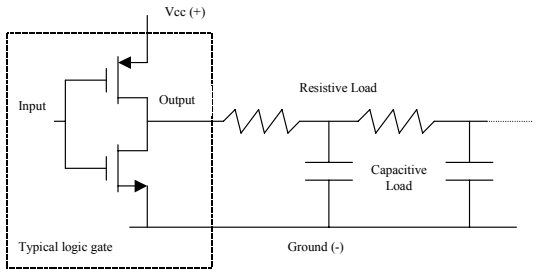

Figure 16: Simplified electrical model of a logic gate and load

As shown in the figure, a circuit board trace acts as a radio-frequency transmission line, commonly modeled by an indefinitely-repeating series of resistors and capacitors. At steady-state conditions, that is, when the logic state is not changing, the capacitors act as open circuits and there is no current flow, and hence no ohmic losses through the resistance (since power equals current squared times resistance, and current is zero). Note also that the input of another logic gate (even one on the same chip) effectively behaves as a tiny capacitor in its own right, with a very slight amount of resistance in series with it. Thus, either within or between integrated circuits in a system, the power dissipation is exceedingly low when no logic state transitions are occurring.

During a state transition, there is a brief instant (on the order of a few picoseconds) during which current flows from Vcc and the upper transistor into the resistive/capacitive load, or from the stored charge in the capacitive part of the load through the lower transistor to ground. In either case, current is flowing through the resistive part of the load and through the tiny-but-finite resistive load of one of the transistors. In addition, the input signal does not transition instantaneously between 1 and 0 or 0 and 1, so there may be an infinitesimal time window during which both transistors are partially "on" and current flows directly from Vcc through both transistors to ground, again dissipating power resistively.

For any single gate, the quantity of energy consumed on a state transition is miniscule, on the order of femtojoules (10^{-15} Joules) in a modern processor chip. Unfortunately, there are hundreds of millions of gates in the processor, and there are billions of clock cycles per second.

To illustrate this, imagine a processor containing 200 million gates at a clock speed of 2 GHz and with 5% of its gates transitioning on every clock cycle. Assume each gate transition consumes only 10 femtojoules of energy. We can then calculate:

$(2 * 10^{+8}) * (2 * 10^{+9}) * 0.05 * (10 * 10^{-15}) = 200$ Joules/sec = 200 Watts

These are of course arbitrary assumptions of energy per gate transition, gate count, and percentage of gates transitioning per clock cycle. The intent is simply to illustrate how seemingly insignificant energy usage is multiplied by the vast speed and complexity of today's integrated circuits into numbers that are no longer insignificant at all. To paraphrase the late Sen. Everett Dirksen, "A femtojoule here, a femtojoule there, and pretty soon you're talking about some real energy."

Notice that the power (that is, the rate of energy consumption) is proportional to the clock speed and the number of gates transitioning in each clock cycle. This implies that the processor will consume the most power when it is running a compute-intensive workload, which is exactly what is observed in empirical studies like this one.

A further factor, worth noting but harder to quantify, is that warmer junction temperatures increase leakage current through some types of gate logic and may also lengthen the transition time of signals, further de-optimizing the circuit's operation from its theoretical ideal. Thus, in almost every case it is desirable to keep the semiconductor as cool as possible. At the same, time, as noted in the next section, allowing a higher junction temperature actually increases the efficiency of the heat transfer to the ambient environment through conduction or radiation. Modern cooling design, then, is a sophisticated balancing act between often-conflicting parameters, not simply a matter of throwing a lot of cold air at the system and hoping for the best.

Memory and CPU heat generation is similar in nature, but much more intensive in the processor. Empirical tests almost invariably show that the overall power consumption is at its worst when the processor is doing only limited memory access, because memory typically introduces delays that cause parts of the CPU circuitry to be momentarily idle, thereby reducing its net power consumption. The aggregate time the processor spends waiting for memory cycles has an effect similar to reducing the processor's clock speed by a few percentage points.

Interleaving of memory banks, that is, installing memory into parallelized channels driven by separate data and address paths with separate controllers, increases the overall performance of memory and allows the CPU to operate with fewer memory-induced delays. This improves the speed of the system, but measurably increases power consumption (other things being equal).

The effects of memory access and memory interleaving are more than anecdotal -- both of these effects were clearly observed and measured during this study.

4.4 Heat transfer

It is important to remember that "heat" and "temperature" are by no means equivalent. The fundamental problem in cooling is to transfer heat energy, not to lower air temperature at the output of the enclosure. In fact, a case can be made that a higher air temperature at the enclosure

exit point may actually improve the efficiency of the building HVAC system, so long as the data center's design does not allow that hot air to recirculate to the inlet points of other equipment. Lowering server airflow may similarly improve the building HVAC efficiency by allowing a higher chiller set point, since servers requiring lower airflow are less likely to recirculate hot air.

In order to keep server equipment running within temperature specifications, heat must be removed from the servers at the same rate that it is created by dissipating electrical energy. All cooling costs are borne to meet this goal. The fundamental goal is to reliably/robustly remove this heat, while spending the minimum on cooling efforts.

Heat is transferred from server racks in three ways: conduction (direct heat transfer by contact to adjacent materials), convection (heat removal by movement of a fluid over the equipment), and radiation (heat removal by spontaneous photon emission). All three means of heat transfer operate most efficiently when the heat difference between the equipment being cooled and the heat sink is the greatest.

In data centers, conduction and radiation are minimal, and forced convection driven by fans is the primary means of removing heat.

The total heat transferred from a rack by all three means is then removed from the data center through HVAC. Cooling costs break down into costs incurred in removing heat from the rack (driving fans), and costs incurred in the HVAC.

4.5 Cooling design implications

Cooling design is a complex balance between multiple parameters and goals, and thus a nuanced analysis is required rather than a brute force approach of simply throwing vast quantities of cold air at the system.

Fan optimization and system thermal design can make a significant difference in the overall power used by the system. If the fans are pushing more air than necessary to cool the system the power input to fans may be considerably higher than necessary. If the fans are not pushing enough air through the system, inadequate heat transfer will result in overheating.

Appendix B: Test Hardware Details

4.6 HP BladeSystem with ProLiant BL460c server blades

2 dual-core Xeon Woodcrest 5140 CPUs

4 2G DIMM interleaved or non-interleaved,
or 8 2G DIMM interleaved, depending on the
test being run

2 hot-plug SAS hard drives

HP 6i SMART array RAID controller

Enclosure s/n USE6331J3J;

Onboard Administrator firmware v1.3

4.6.1 Blade Serial Numbers

USM70103A2	USM70102VK	USM701038S	USM70102VL
USM70103AT	USM65105JM	USM70102WT	USM701039P
USM70102VJ	USM701038T	USM70102VO	USM70103CU
USM70102WY	USM70102WZ	USM70102WR	USM701038D

4.7 IBM BladeCenter-H with HS21 server blades

2 dual-core Xeon Woodcrest 5140 CPUs

4 2G DIMM non-interleaved or 8 2G DIMM interleaved,
depending on the test being run

2 non-hot-plug hard drives; 2 hot-plug drives
depending on the test being run

Management module firmware v1.1

4.7.1 Blade Serial Numbers

YK105069S235	YK115068J2GA	YK115068J2KB	YK115068J2HE
YK115068J2J7	YK115068J2GH	YK115068J2H2	YK115068G1YM
YK115068J2JZ	YK128168L16Y	YK115068J2PL	YK115068J2H9

YK115068J2P8 YK115068J2NN

4.8 Dell PowerEdge 1950

Dell PowerEdge 1950 s/n D417CB1 Model EMU01

2 dual-core Xeon Woodcrest 5140 CPUs

2 PCMCIA slots unpopulated SATA drive

2x Western Digital WD800JD Caviar SE 7.2k 80GB SATA drive

8 2G DIMM

2 power supplies

s/n D417CB1

Mfg 2006-07-17

4.9 IBM x3550

x3550 Machine type 7978 model 61U

2 dual-core Xeon Woodcrest 5140 CPUs

2 PCMCIA slots unpopulated

2x IBM 4DK1043 73.4 GB 15k SAS drive

8 2G DIMM

2 power supplies

s/n KQHA757

Mfg 2006-08-24

