Research Brief

IBM BladeCenter Reliability/Availability Evaluation

Executive Summary

In July, 2007, *Clabby Analytics* received an invitation from IBM to visit an IBM BladeCenter benchmarking and testing laboratory in Raleigh, North Carolina. The purpose of this visit would be to audit a suite of tests comparing the IBM BladeCenter H architecture with the Hewlett-Packard (HP) BladeSystem c-Class architecture — and then provide a written evaluation of the test results.

Now, for those of you who know *Clabby Analytics* (that's me), you know that I am generally hesitant to do sponsored research. I accepted this invitation because IBM told me that, under certain stress workloads, they could show that Hewlett-Packard's blade memory modules run 10°-15° Fahrenheit hotter than the uppermost range typically recommended by memory manufacturers. If this proved true, this situation could have implications on the reliability of HP blades. So I made the trip to Raleigh…

In Raleigh, I personally audited HP and IBM blades being stress tested using Agilent and HP test equipment. I can verify that IBM's observations about HP blades are true — HP blades run hotter than IBM BladeCenter under the same workload. (I'll describe the specific test environment later).

> ***What this means is that, in heavy workload environments, HP may be "cooking" its memory modules (running memory out of spec for extended periods of time). And for those of us who have ever fried memory, processor, or disk components, we know that this kind of situation can lead to some serious reliability/availability problems...***

While in Raleigh, *Clabby Analytics* also requested an in-depth BladeCenter availability/-reliability design review (a BladeCenter "tear-down"). What I discovered was that IBM has other reliability/availability blade advantages in power design (with a redundant power backplane); in availability (with redundancies to reduce the possibility of a single point of failure); in disk (as it replaces mechanical disks with solid-state disk drive); and in storage integration (IBM's Direct Attach Storage subsystem places mechanical storage under external control). Further, IBM's Open Fabric and Open Fabric Manager serves to make management of blades easier by allowing switches and LAN settings to be preconfigured, as well as automating the failover of blades — helping improve overall blade availability.

> ***Based-upon actual lab tests and a thorough tear-down of IBM's BladeCenter chassis, Clabby Analytics concludes that the IBM BladeCenter design provides a superior reliability/availability design (especially in the areas of memory cooling and power redundancy) when compared to HP's BladeSystem design. For enterprises looking for the more reliable/available design, IBM's BladeCenter has the clear, undisputed edge.***

Finally, IBM engineers and strategists described the company's overall blade design philosophy (fewer moving parts for fewer mechanical failures; open network fabrics for easier management; and more). A more in-depth discussion of this design philosophy is also contained in this report.

The Test Environment and Results

The equipment that was tested included:

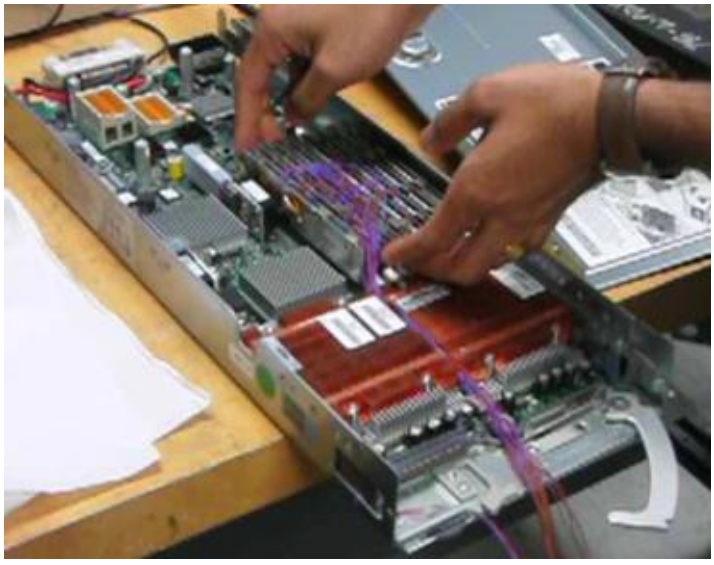
- *An IBM Blade Environment:* 14 HS21xm Blades: 2x2.33 GHz Intel Xeon 5345 CPUs, 8x2GB DIMMS, 1x73GB 10k rpm HDD, 0xI/O cards, IBM BladeCenter-H; 4 Power Supplies, 2 Blowers, 0 Switch Modules, 1 Management Module.

- *An HP Blade Environment:* 15 BL460x Blades: 2x2.33 GHz Intel Xeon 5345 CPUs, 8x2GB DIMMS, 1x73 GB 10k rpm HDD, 0xI/O cards, HP c7000 enclosure, 6 Power supplies, 10 fans, 0 Switch modules, 1 Management Module.

The test environment consisted of a large room (ambient temperature constant was 77°); a few workbenches; an HP BladeSystem c-Class blade environment and an IBM BladeCenter running side-by-side; and two pieces of test equipment (an Agilent Data Acquisition Switch Unit, and an HP Testmobile Data Acquisition Module). The Agilent unit was used to measure voltage output; the HP Testmobile Data Acquisition Module was used to measure temperature (it converts voltage into temperature measurements).

Thermocouples (wires) were run from the HP and Agilent test equipment to four DIMMS (dual in-line memory modules — the small boards that hold memory chips) located inside HP's BladeSystem (see Figure 1). And once the testing was completed on HP's BladeSystem, these same thermocouples were then attached to 4 DIMMS in an IBM BladeCenter. The wires between the testing devices and the DIMMS were hooked up to two DRAMs (distributed random access memory chips) on each DIMM, as well as each DIMMs advanced memory buffer (AMB). Two thermocouples were also attached to 2.3 GHz processors in each system. Test equipment then automatically displayed the temperature activity on each memory module as well as on the processors. These thermocoupled blades were then placed back into HP's BladeSystem chassis

Figure 1 — Thermocouples to HP DIMMS



Source: Clabby Analytics, July, 2007

Prime 95, an industry standard package used to calculate the highest Mersenne prime number was then used on each blade platform to produce a varied workload environment that uses a lot of memory and central processing unit (CPU) processing computing power. Because Prime 95 generates a variable workload — and because the workload cycles up and down — the HP memory modules reported a low range of temperatures that ranged from just above room temperature — to almost 100° (98.7° was the highest that I measured). Figure 2 shows the HP BladeSystem blade memory test results.

Figure 2 — HP BladeSystem Memory Test

Channel	Description	Location	Actual	Rise	Adjusted	Limit	Delta	Warnings
0 (101) T	Ambient	inlet to the chassi	23.1	0.0	25.0	35	+0.0	
1 (102) T	Tcase1	CPU1	55.3	32.2	57.2	69	+0.3	
2 (103) T	Tcase2	CPU2	55.3	32.2	57.2	69	+0.3	
3 (104) T	AMB	DIMM1	82.2	59.1	84.1	105	+0.2	
4 (105) T	DRAM near AMB	DIMM 1	77.3	54.1	79.1	92	+0.3	
5 (106) T	DRAM	DIMM1	76.5	53.4	78.4	95	+0.2	
6 (107) T	AMB	DIMM2	87.5	64.4	89.4	105	+0.3	
7 (108) T	DRAM near AMB	DIMM2	86.3	63.2	88.2	92	+0.2	HOT
8 (109) T	DRAM	DIMM2	86.1	62.9	87.9	95	+0.2	HOT
9 (110) T	AMB	DIMM3	94.3	71.2	96.2	105	+0.3	HOT
10 (111) T	DRAM near AMB	DIMM3	96.8	73.7	98.7	92	+0.2	EXCEED
11 (112) T	DRAM	DIMM3	94.8	71.7	96.7	95	+0.2	EXCEED
12 (113) T	AMB	DIMM4	93.8	70.6	95.6	105	+0.2	HOT
13 (114) T	DRAM near AMB	DIMM4	93.0	69.9	94.9	92	+0.2	EXCEED
14 (115) T	DRAM	DIMM4	93.5	70.3	95.3	95	+0.2	EXCEED

Source: Clabby Analytics, July, 2007

I audited the exact same test on an IBM blade located in the same room as the HP BladeSystem, using the same software test suite and the same measurement equipment. The IBM memory modules never exceeded the 85° degree mark (the high-end temperature range recommended by most manufacturers) — and in fact, ran closer to 80° throughout the entire test. (I don't show the IBM picture because there's nothing to show — no "HOT" or "EXCEED" warnings were present).

These tests show that, under heavy computational workloads, HP may significantly overheats its DIMMs. Blade buyers who plan to use virtualization software very heavily in order to drive maximum workloads on their blade systems need to be aware that cooked DIMMs represent a potential reliability/availability failure point.

Why HP Is Cooking Its DIMMs…

The big question that arises after conducting these tests is "why are HP's memory DIMMs running so hot?" And a closer look at HP's blade design reveals several quick and obvious possibilities. With respect to its memory complex design, it appears that HP may have:

- Jammed their memory too close together (DIMMS can use up to 6 Watts of power during peak loads — generating a lot of heat);

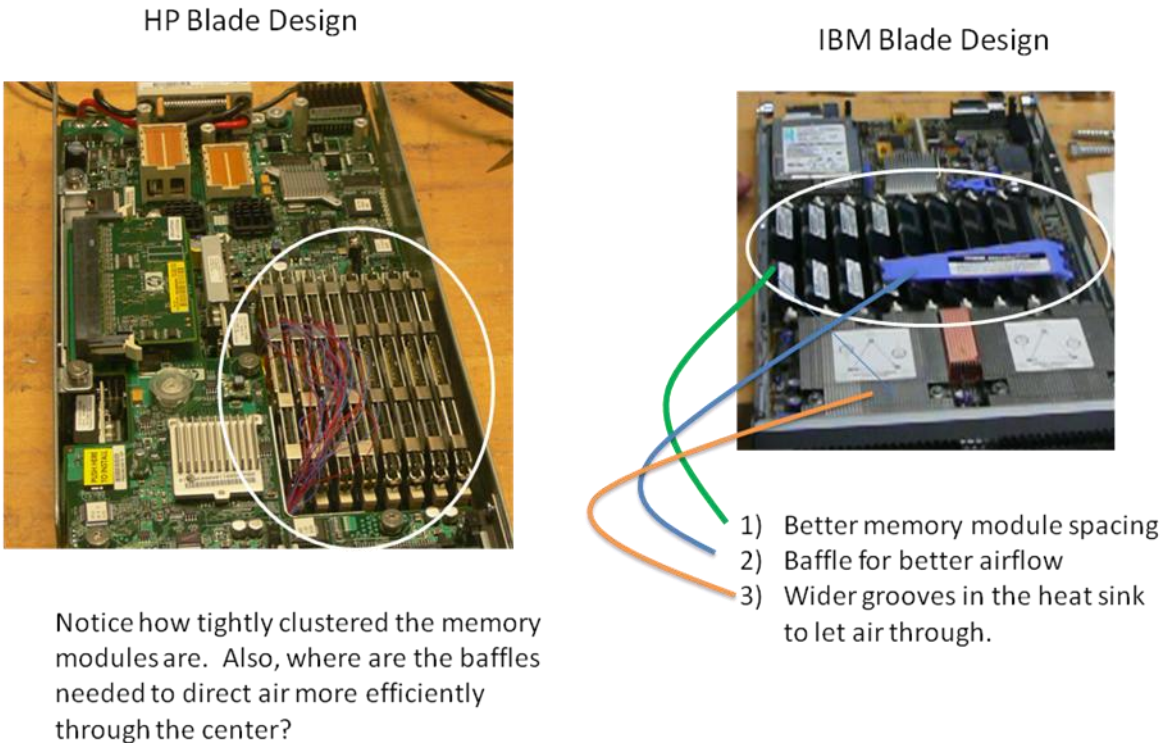
- Not baffled the airflow properly (it appears that HP is not feeding enough cooled air down the center of their memory complex); and
- Placed their memory modules directly behind a major heat source — the central processing units — but has very fine heat sinks that appear not to drive as much air through the center of the memory module complex as needed.

I believe that these design issues makes those memory modules in the center of HP's memory bank run hotter than they should.

Memory Jammed Too Close?

Figure 3 is a shot of HP's blade memory complex (on the left) as compared with IBM's blade complex (on the right). Note how HP has clustered its memory modules together, leaving about 10 millimeters distance between each DIMM — while IBM has left considerably more space between its memory modules and angled its DIMMS to enable better airflow and allow more separation between DIMMs (part of IBM's "calibrated vector cooling" design approach).

Figure 3 — Memory Jam vs. Memory Spacing



Source: Clabby Analytics, July, 2007

Baffles

A baffle can be used as a means to redirect air to potential hot spots. Notice that in the HP design baffles are not used — so, essentially, the center of HP's memory complex gets no special relief when it comes to cooling. IBM's design employs a baffle that flows air in a direction that cools its memory modules more efficiently.

Placement Behind the Processors

The reddish-colored fins at the forefront of Figure 1 (on page 2) are heat sinks that sit on top of HP's blade processors. Cool air is drawn both under and over these processors in order to cool them. But a closer look at the fin design shows that these sinks have very fine channels through which air flows — presumably to accelerate airflow over the processors and increase the airflow pressure across the memory DIMMs located directly behind the processors.

> ***It is my belief that HP's fine-grained heat sinks are actually not directing cool air properly onto the memory modules behind them — and this is resulting in the memory modules located in the center of each the HP blade memory complex exceeding recommended heat specifications.***

It should be noted that IBM's blades also place the processors in front of the memory modules. But IBM's heat sinks have much wider grooves that appear to provide better airflow over its blades — and IBM memory modules are spaced further apart – allowing for better overall cooling of IBM's blade memory complex. Also, IBM spreads its memory components across the width of the blade — and tilts its memory at an angle — delivering a much better cooling impact.

Beyond the Blade: The BladeCenter Chassis Tear-down

As mentioned in the *Executive Summary*, as long as I was in Raleigh I requested an in-depth BladeCenter design review. To accommodate this request, IBM brought in several top-level senior technical staff members (STSMs), a distinguished engineer, as well as several blade design/testing engineers and program managers who conducted a thorough BladeCenter architectural tear-down.

In short, I learned that IBM's advantages in availability and reliability do not end at blade design:

- The IBM BladeCenter chassis provides dual power paths through the midplane. This redundant power plane design means that in the event that some sort of short incapacitates one plane, IBM BladeCenter can failover to a secondary plane. (HP's chassis does not provide a secondary power plane).
- IBM offers dual, solid-state disk drives that reside on its processor blades. These solid-state drives have no mechanical parts — and, accordingly, should fail far less often than mechanical drives. IBM is leading the industry with this new technology which provides IBM a distinct competitive edge over HP in disk drive reliability.
- IBM's Direct Attach disk subsystem is the first Blade server solution to combine the convenience and ease of use of direct attached storage with the benefit of disk consolidation external to the actual chassis; and
- IBM management products, including Open Fabric Manager and IBM Director, both enable switches and LAN settings to be preconfigured — and allow for the automatic failover of blades. The use of these products helps improve overall blade availability.

Each of these points deserves closer scrutiny.

Redundant Power Plane

IBM designed its BladeCenter chassis with a system backplane with a redundant power path such that, should a fault occur (such as a short in a power bus), another power path could be used in a failover mode in order to help deliver continuous power to blade servers. HP does not offer a redundant power plane in its BladeSystem.

When I pushed back on IBM engineers regarding the importance of a second power path, they explained the following:

> ***Without a second power path, any failure along a power plane can bring down several servers. Note that HP can stack 16 blades into their chassis. This means that potentially 16 servers could crash. Because IBM's chassis has a second power path, if one power backplane fails, the other can be used to deliver continuous power — helping result in none of IBM's 14 servers crashing. In high availability environments, this could prove to be a very big deal…***

Dual Solid-state Disks and Tightly Integrated External Storage

IBM offers an impressive array of storage products that can serve its blade architecture. These products include diskless blades, solid-state disks, flash drives, local hard drives, highly-integrated external hard drives, and even SIO (serial input/output) devices. But two of these devices/approaches were of particular interest to *Clabby Analytics*: the stateless solid state drives, and the work IBM has done to tightly integrate diskless blades with its DS 3200 dense storage family.

The reason I found these products so interesting is that they:

1. Help reduce power draw (and thereby reduce heat production) within blade enclosures; and they
2. Remove mechanical parts from within a blade chassis (reducing potential mechanical failures while helping increase overall availability and reliability).

My personal philosophy is that anything that can be done to reduce power draw from within a blade chassis — as well as anything that can be done to remove movable parts on a blade — represents "goodness". By reducing power draw, the amount of heat that a blade produces is reduced. And by reducing heat within a blade enclosure, blades cost less to cool and blade components can last longer. Further, by moving mechanical devices out of densely populated blades, maintenance is more easily accomplished. (Note: HP offers dual, hot swappable mechanical drives that are easily accessed and swapped within their blade environment. But I would argue that moving those drives to an external location is a better design for the above mentioned power/heat considerations).

A closer look at IBM's recently released dual-stacked solid-state drives (SSDs) reveals that IBM is able to put two 16GB flash SSDs into a custom carrier that can snap into and existing blade HDD (hard disk drive) casing (see Figure 4).

Figure 4 — IBM's Dual-stacked, Solid State Drive Designed for IBM Blades

Source: IBM, July, 2007

This type of stateless drive uses up to 93% less power than traditional mechanical drives (traditional 3.5" drives use 16 Watts of power — solid state uses 1 Watt or less); provides better meantime between failure (MTBF) due to no moving parts and "write wear leveling" technology; and is particularly good for fast operating system boots and random read intensive applications. Fast OS booting is becoming extremely important as information technology buyers continue to use virtualization and provisioning environments to build-up and tear-down blade environments — and these types of drives are ideal for such environments.

Also worth mentioning is IBM's solution to deliver direct attach external storage for Blade booting and storage services. IBM densely packed DS 3000/EXP300 storage family subsystems has been tightly coupled with IBM's BladeCenter — creating an easy to use, direct attached storage environment that can serve blades transparently as a local drive. This configuration provides redundant, hot-swappable RAID controllers, power supplies, and cooling fans all within an external array — thus helping lighten the power draw/heat dissipation and maintenance problems that internal blade hard drives may cause. Also, placing storage external to the blade has other benefits including helping data be better protected, and made more available — as well as more easily available to other blades in a failover mode should the original blade fail.

Open Fabric Manager

An open fabric is a networking fabric that allows blade buyers great flexibility in choosing networking options for blade-to-blade, blade-to-other servers, blade-to-storage, and linkage to other devices. The ability to quickly and automatically link fabrics to servers or blades is a key determinant of overall application availability. IBM appears to have the broadest and deepest suite of network fabric offering in the blade market — with support for a variety of switches including 10 Gb Ethernet (the blade industry's first 10Gb solution), NPIV FC, and SAS switches from a number of vendors and supporting 5 I/O fabrics including Ethernet, iSCSI, Fibre Channel, Infiniband and SAS. IBM's entire x86 switch, blade and chassis portfolio is supported by BladeCenter Open Fabric. The BladeCenter Open Fabric Manager simplifies deployment and automates failover by pre-assigning LAN (MAC addresses for Ethernet) and SAN (worldwide names for Fibre Channel) connections and then letting the system manage them after that, regardless of the switch or passthru module being deployed. So, for example if a blade were to fail, when a new blade is added it can automatically connect to the required network and storage fabrics without any administrator intervention.

For blade buyers, Open Fabric Manager is important for two reasons:

- It makes the availability and interconnect of blades automated and easier to manage;
- It is OPEN — so it works across a range of vendor switches such as Cisco, Nortel, QLogic and Brocade (the leading suppliers of switched networking environments). HP's Virtual Connect, by contrast, works on HP switches — a major drawback for customers looking for choice in networking components; and,
- Since Open Fabric Manager is architected in the IBM Advanced Management Module, it is not dependent on the switch brand or design. This gives the added benefit of installed customers adding this function in the future as needed via software.

Summary Observations

Information technology (IT) buyers are moving multiple applications from tower servers and racks to blades to gain advantages such as reduced management cost, better overall system utilization, streamlined integration, power savings, improved density etcetera. And this move to blades creates new opportunities for improvements when done right.

Doing it right means paying close attention to the reliability/availability characteristics of a given blade environment. Little design flaws can have huge implications in blade environments where cooked memory can fry servers — or where lack of redundancy can bring down ten or fifteen servers in one fell swoop. The key message in this report is that *Clabby Analytics* finds IBM BladeCenter blades and blade chassis to be better designed for reliability and availability than HP's blade and enclosure design.

In this report I identified the following problems with HP's blades/BladeSystem:

- Clear heat dissipation issues around HP memory modules that can cause memory to operate much hotter than recommended by most memory manufacturers. This, in turn, leads to "memory cooking" — a source of potential component failure; and
- HP's blade design employs a single power plane that delivers power to processors, storage devices, fans, etcetera. IBM's blade design uses a redundant power delivery design. Again, a short or breach anywhere along HP's power plane could result in a system failure.

I also identified potential advantages in adopting IBM BladeCenter. These include:

- More reliable blade design;
- Redundant power backplane (for failover in case of power bus interruption);
- An innovative, integrated blade/storage subsystem package;
- Exciting new designs such as IBM's solid-state disk (which offer less heat in the blade and enclosure; no moving parts, hence high MTBF(meantime between failure); fast boot; etc.); and
- A more open communications fabric that facilitates easy management of blades (an availability benefit).

The HP BladeSystem did have a potential benefit for legacy environments as the company's blades offer two hot-swappable drives for availability (IBM's are not hot swappable). But this feature should be weighed carefully in that *Clabby Analytics* believes that a design goal for blades should be to move things that generate heat — and things with movable parts — off of blades and out of blade enclosures.

One last thought about HP blades keeps plaguing me. If blade processors or other components get any hotter, what sort of impact will that have on already-too-hot HP memory modules? I wonder if HP will need to undergo another blade redesign at some point in the near future to improve airflow and cooling…

Clabby Analytics
http://www.clabbyanalytics.com
Telephone: 001 (207) 846-0498

© 2007 Clabby Analytics
All rights reserved
August, 2007

Clabby Analytics is an independent technology research and analysis organization that specializes in consolidation, virtualization, and provisioning. Other research and analysis conducted by Clabby Analytics can be found at: www.clabbyanalytics.com.