HP NonStop Gets Bladed

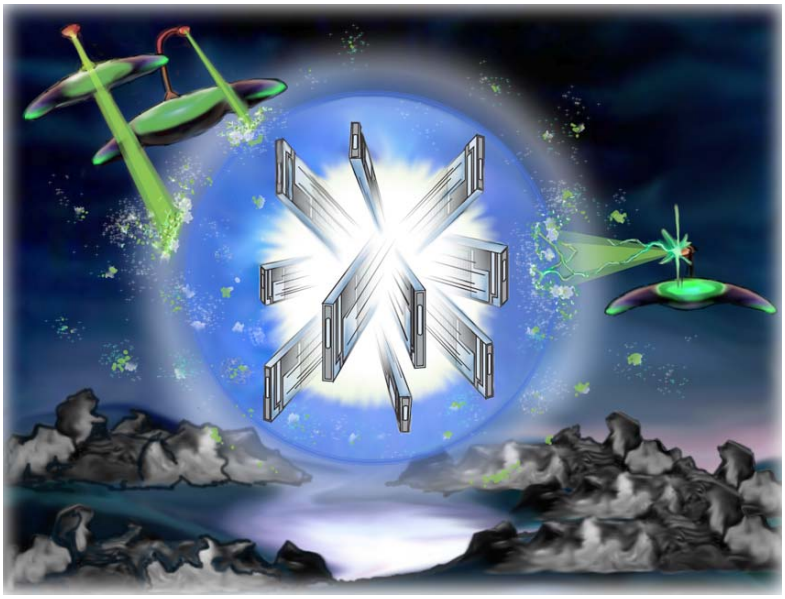
Product Insight

Gordon Haff
25 June 2008

In merging the Integrity NonStop line of fault tolerant servers into its blade server products, HP is bringing together a long-running product success with a relatively new breakout.

The NonStop architecture was developed at Tandem Computers, which was founded in 1974 by James Treybig and other Hewlett-Packard engineers. In 1997, Compaq bought, but never really absorbed, the company. When HP subsequently purchased Compaq, it acquired a lineup that was an integral component of many of the world's most critical computing infrastructures—which is to say infrastructures that move a lot of money around. Like Compaq, HP for a time seemed uncertain what to do with this marvelous asset other than to leave it mostly alone. However, over the past few years, HP has both modernized NonStop and leveraged its technology for a business intelligence play called HP Neoview.

Blades are a more recent development; as a technology approach they took a while to really hit their stride. In HP's case, it was the BladeSystem c-Class blade form factor and chassis design—introduced in 2006—that was the takeoff point; HP now holds the market share lead over the prior #1 vendor, IBM. HP's blade approach, like Dell's and IBM's, emphasizes the integration of processing, networking, and (in some cases) storage into a single chassis—in HP parlance, an "Adaptive Infrastructure in a Box."



When NonStop moved from its MIPS processor-based architecture to Itanium in 2005, it made a dramatic shift away from a design and components that were largely unique to ones that highly leveraged HP's other Integrity products.[1] But this latest marriage of the old and the new takes NonStop's mainstreaming to the next level.

[1] HP Integrity is HP's Itanium-based lineup that primarily runs HP-UX but, in many cases, also supports OpenVMS, Linux, and Windows. See our *The 'I' is for Integrity*. By amusing coincidence, the Integrity brand name came from Tandem which had originally used it for a separate line of FT systems that ran a Unix variant rather than the NonStop Kernel, or NSK (né Guardian) OS that ran on its primary lineup.

From Tandem to Integrity

When NonStop returned to HP by way of Compaq, it brought with it truly unique capabilities. NonStop runs many of the world's banking systems; HP estimates that it powers 75 percent of the 100 largest electronic fund transfer networks. It also handles the majority of all ATM and credit card transactions at organizations like Bank of America and Barclay's Bank. What is perhaps even more impressive, about 95 percent of the world's securities transactions take place on NonStop gear, at over 100 stock exchanges including the New York Stock Exchange, Chicago Mercantile Exchange, and Hong Kong Stock Exchange.

NonStop also underpins important parts of the world's network and telecommunications infrastructure (for example, at AT&T, British Telecom, and NTT). Among the wins it contributed at Compaq was the $100 million decision by Sabre to kick out well-entrenched mainframes; Sabre's also a poster child for a newer style of "hybrid" NonStop deployment in which distributed Linux front-ends feed NonStop transactional back-ends. NonStops are also found in healthcare, manufacturing, retail, and government—and they handle about half the US 911 emergency calls.

NonStop plays these important roles because the applications that run on it share two key characteristics. The first is that they demand absolute reliability and absolute transactional integrity. NonStop provides these guaranteed attributes through a combination of hardware and advanced software such as its NonStop operating system and NonStop SQL/MX DBMS. It is, in a phrase, fault tolerant (FT). They also require massive scale—and not just the sort of scale that hooking together a lot of x86 servers with Gigabit Ethernet can deliver. They need systems that can handle update-intensive transactional workloads where there's a lot of writing (requiring coordination across the system) as well as reading. The net is that there are few, if any, other systems on the planet that can do what NonStop can do: handle multi-terabyte data volumes, real-time data feeds, and on-the-fly transformations while simultaneously guaranteeing that the services will always be up and available.

But engineering the unique capabilities that made NonStop so attractive to all those organizations long meant highly bespoke design that little leveraged the work that went into more mainstream systems. And, as important, didn't benefit from the volume economics that came from mainstream ubiquity. And that cast a certain pall over NonStop's future, given that it filled an important niche, but a niche nonetheless.

By melding NonStop into its Integrity product set, HP sharply reduced the amount of unique hardware that it needed to develop in order to maintain the NonStop line. While many of the board- and module-level components were still unique to NonStop, they extensively leveraged the same processors, chipsets, and other parts of the hardware design that went into HP's mainstream Integrity Unix platform. This represented enormous savings—and allowed a faster pace of innovation—relative to having to largely build everything from scratch specifically for NonStop.

The integration with c-Class blades puts NonStop another step closer to the main current of computing—probably as close as possible for a system architecture with such unique characteristics.

c-Class Blades

When HP decided to shift its p-Class blades to c-Class in 2006, it took a calculated gamble.[2] Unlike the approach that IBM took with BladeCenter H,[3] the new c7000 chassis was not backwardly compatible with the components that went into the prior p-Class product. This meant that HP had to design and manufacture a whole new blade lineup for the new chassis—while also supporting existing p-Class customers for a reasonable interval. Furthermore, because a blade chassis in the Dell, HP, and IBM vein functions as a sort of integration

[2] HP's c-Class blades are discussed in our *HP Blades Go from p to c*, as is the evolution of blade servers from disaggregated computing to integration point.

[3] See our *A Resharpened BladeCenter*.

point for third-party network switches and the like, HP wasn't even the only company whose product designs were affected. Partners like Brocade, Cisco, and Nortel likewise had to come out with new products for the new chassis form factor.

However, changing form factors also gave HP the opportunity to start with "a clean sheet of paper" at a time when blade adoption was still fairly limited and the installed base correspondingly small. This allowed HP, for example, to revamp the power and cooling design of the chassis—a set of technologies and capabilities that it collectively refers to as Thermal Logic. One such component—which HP understandably loves to demo at every opportunity—is its Active Cool fans. Inspired by the high velocity propellers used in model airplanes, these HP-designed fans—on which it holds, or has applied for, over 20 patents—spin at fantastic RPMs. Equally important is their logic; they use a control algorithm to optimize operating parameters. They speed up and slow down in response to changes in load; there's also the potential for using extra fans in some configurations so that they run slower and, hence, quieter. HP estimates that these fans draw 66 percent less power than traditional ones.

In any event, c-Class blades—propelled by both their own capabilities and a general renaissance of overall business under Mark Hurd—have paid off handsomely for HP. Market share data from all the usual sources shows HP's having taken the blade lead away from IBM. This has mostly been an x86 story and doubtless reflects, in no small part, the overall strength of the ProLiant server product lineup, as backed by solid sales, marketing, and supply chain management.

However, Itanium is part of the c-Class story, too. Introduced in 2005, the original BL60p blade was best thought of as a way of bringing HP-UX to HP's blade integration point—that is, a way to run HP-UX applications alongside those running on x86 blades in the same chassis.[4] The BL860c is the current dual-core "Montvale"-based flavor of Itanium blade that replaced the original BL60p.

Building Up a NonStop

Before delving into how NonStop works as part of a blades architecture, it's worth reviewing some basic NonStop concepts and how they're implemented in the current generation of Integrity NonStop systems, the H-Series.

The basic computing unit in a NonStop system is a "logical processor." The hardware components that make up a logical processor have varied among different generations of NonStop, but the common thread is that it's the smallest standalone unit of processing that's doing "real work."

A NonStop server is an aggregation of logical processors connected by ServerNet into a loosely-coupled, i.e. distributed memory or shared-nothing, computer system. Tandem introduced ServerNet together with its NonStop Himalaya S-Series in 1997. For a time, Tandem promoted ServerNet as a standard "system area network" in a vein similar to Dolphin's Scalable Interface (SCI) and InfiniBand. But it never saw wide use outside of Tandem (and, less so, Compaq after it acquired Tandem).

One also sees the term "node" used within the context of NonStop systems. The term isn't always used consistently, including within HP literature, but the official current meaning is to refer to a grouping of up to 16 logical processors that share a single system name. It's primarily a management concept that isn't normally visible to applications. For example, database schemas can span nodes out to the maximum system size—4,080 logical processors. This differs from the more conventional use of the "node" term to refer to the SMP building blocks within a cluster—that is, something akin to a logical processor in the NonStop case.

None of these basic concepts change in moving from the first generation of Integrity NonStop servers to the current one. However, at a detailed level, the system now implements fault tolerance in a conceptually quite different way.

[4] In other words, the BL60p wasn't really intended as a generic heavyweight processing blade for Linux high performance computing workloads and the like. See our *HP-UX Gets Bladed*.

Fault Tolerance

Before this move to blades, Integrity NonStop servers were based on what HP calls the NonStop Advanced Architecture (NSAA). In a DMR (dual modular redundant) NSAA configuration, two processors on two different physical modules carry out each computation.[5] The results are then compared; if the results aren't identical, a comparison error between the two boards is noted, and the system will try to determine the source of the error. If it can do so, it will disable the defective hardware and continue processing. Otherwise, it will disable both halves of the logical processor and transfer the job transparently to another logical processor using a NonStop software FT technique called "process-pair takeover." Logical Synchronization Units (LSU) handle the comparisons by looking at all the I/O traffic across the ServerNet interconnect. The LSUs compare output packets to ensure that an error or corruption hasn't occurred, and they convert I/O packets to and from ServerNet's format.[6]

NonStop has used conceptually similar techniques since abandoning its own processors in favor of those from MIPS in 1991. As you'll note from the previous paragraph, the NonStop software doesn't actually depend on the hardware to recover from a failure. If the hardware can transparently recover, all the better, but it isn't really necessary in most cases. Contrast this with the pure hardware FT approach used by NEC and Stratus. In this case, the whole idea is to mask any hardware failures from Windows or Linux (which, under normal circumstances, could no more recover from a CPU failure than automagically compose a sonnet).[7] Thus, while NonStop's approach is often lumped in with FT that depends on hardware-based lockstepping, it's actually quite different.

So, why bother with all the comparisons, then? In a word, detection.

In the NonStop architecture, the issue isn't so much recovering from a known error—the NonStop software knows how to do that—but detecting that such an error has occurred in the first place. If an undetected error does occur, the result can be worse than an outright failure because it can result in silent data corruption and other behaviors that just aren't acceptable for the sort of environments where NonStop plays. We're not talking high frequency events of course, but when you're aiming for "runs for years without interruption" even low probability glitches have to be eliminated.

In fact, explicit synchronization between processors was an approach that Tandem only introduced when it moved away from its own processors that had far more internal data checking mechanisms than their MIPS successor did. In other words, when the processor and associated hardware could no longer be depended upon to police its own actions, Tandem introduced a sort of "buddy system" to mitigate any deficiencies.

Detecting Errors in Hardware

This latest NonStop generation, in a sense, comes full circle. It eliminates processor synchronization hardware in favor of a purely software-based approach. Programs running on each logical processor are "backed up" on another logical processor in the system. In this context, "backed up" means there's a copy of the running program and that context changes and other changes in the program status get copied to the backup over ServerNet. There's obviously some overhead in this, but the backup program isn't actually running and consuming CPU resources; it's just sitting in memory. HP estimates there's about five percent overhead associated with maintaining this copy.

What's made this possible is the substantial increase in error detecting and correcting features built into even relatively mainstream, if not exactly mass market, Unix servers. Thus, the zx2 chipset, developed by HP for its Itanium blade and other

---

5 A module is a hardware building block containing processors, I/O Bridge, zx1 memory and bus controller, and the memory complex. HP used the term "NonStop Blade Element" but I'm avoiding that terminology in the text to clearly differentiate from the standard blade chassis in this latest iteration.

6 See our *Itanium Goes NonStop at HP* for more on NSAA.

7 See our *Stratus Cuts the Cost of Fault Tolerance (Again)*, *Linux FT, à la NEC*, and *Stratus ftServers: Windows Fault Tolerance for Verticals*.

entry-level Itanium-based servers, includes any number of reliability and availability features. These include memory-related features such as page de-allocation, chip sparing, and DIMM address parity protection.

Intel has likewise been augmenting the Itanium processor itself with features such as Cache Safe (detects and shuts off malfunctioning areas of cache),[8] Enhanced Machine Check Architecture (enables multi-level error handling across hardware, firmware, and operating system), memory scrubbing, and ECC on the system bus. More features of this type are planned for "Tukwila," the next generation of Itanium processor planned for late-2008 or 2009. Reliability, Availability, and Serviceability (RAS) features that go above and beyond the norm are increasingly a big part of Intel's pitch for Itanium as a point of differentiation from 64-bit x86 processors from itself and AMD. For many classes of servers, one can reasonably ask how important these incremental functions are. NonStop, however, has an application environment and user base that gives even small reliability gains outsized significance. When "it must not go down!" is in effect, "detect and correct everything you possibly can!" becomes much more the rule of the road.

Bringing It All Together

The heart of the new NB50000c is a c7000 BladeSystem, a 10U-high chassis that houses from two to eight BL860c Integrity blades. Each blade incorporates one Itanium 9100 series dual-core processor ("Montvale") running at 1.66 GHz.[9] Blades can be configured with between 8 GB and 48 GB of DDR2 memory. Each blade functions as a logical processor within a NonStop system; it connects the other logical processors in the system using a ServerNet mezzanine card on each blade, in concert with redundant ServerNet switch modules that plug into the back of each blade chassis.[10]

One significant change that's enabled by the new J-Series NonStop Kernel is that the logical processors can make use of both physical cores of the Itanium processor. The NonStop system as a whole is still shared nothing/distributed memory/MPP—choose your term—but the individual logical processor can now be an SMP in which the two cores share the blade's memory. While this may seem straightforward, it's actually a significant change from historical NonStop logical processors in which only a single CPU actively executed code in each logical processor.[11]

This change is important because individual processor cores aren't getting much faster. Chip horsepower increasingly comes from having more engines, not more powerful individual ones. Therefore, benefitting from the increasing transistor densities predicted by Moore's Law pretty much requires making effective use of cores in the aggregate. To do otherwise is a recipe for stagnant performance.

The one downside of this approach is that, at least currently, it's incompatible with implementing an error-checking regimen that involves three copies of the program rather than two—a true belt-and-suspenders level of paranoia that goes by the acronym TMR (triple modular redundant). For now, this remains a feature of H-Series Integrity NonStops but not the new bladed flavor. HP estimates that TMR is of interest to about 10 to 15 percent of its base; it does expect to have a follow-on implementation of TMR that leverages blades by trading off some performance for the incremental availability.[12]

Storage and local area networks connect to the server through dedicated interfaces. In the NB50000c, the interface is provided by a

[8] Originally codenamed "Pellston."

[9] The BL860c has two processor sockets and can normally be configured with either one or two Itaniums. In the NonStop application, only a single dual-core processor per blade is used.

[10] Although the ServerNet mezzanine card and switches are specific to NonStop, they plug in the same way that Ethernet or Fibre Channel devices do.

[11] That the two CPUs in this case reside on the same processor die doesn't change the fact that they're architecturally independent CPUs as far as the software is concerned.

[12] The difference here is somewhere in the five-nines versus seven-nines range; the difference between 99.999 percent uptime and 99.99999 percent uptime translates to about 4 minutes a year.

ServerNet-connected Cluster I/O Module (CLIM) that is just software running atop a standard ProLiant rackmount server (a DL385 G5); the prior generation used a purpose-built device called an I/O Adapter Module Enclosure (IOAME) together with other specialized hardware.[13]

As with other NonStop hardware, CLIMs are typically configured redundantly. Thus, if an entire IP CLIM (i.e., the CLIM used for networking) fails, all the Ethernet interfaces on that CLIM are failed over to its preconfigured failover destination on a different CLIM. When individual links to an external network fail, interface resources can be switched within a CLIM using a feature called bonding interfaces.

The Storage CLIM connects to SAS disks within an MSA70 enclosure. This, too, is standard HP hardware, except that it's configured with a second I/O Module for increased redundancy. The NB50000 can also connect to various FibreChannel-based HP StorageWorks XP arrays.

The rack is rounded out with an in-cabinet UPS and a slide-out management console. The new system makes use of standard HP management tools. HP NonStop Cluster Essentials integrate with HP Systems Insight Manager (SIM), which can also monitor and manage an entire bladed infrastructure: the HP SIM Blade plug-in. The Onboard Administrator provides management specifically for the blade infrastructure.[14] Alternatively, NonStop management tools and applications remain available for deep-dive tasks that are specific to the NonStop architecture or operating environment.

Conclusion

Ebullient with the success of its c-Class blades, HP has, of late, been loudly touting a strategy that goes by "Blade Everything" or "Blades Unbound" (depending upon who is doing the talking). In practice, the strategy is a more practical "Blade Everything It Makes Sense To." HP will happily continue to ship rackmount, and even tower, servers to those who prefer them—whether in SMB, emerging markets, or even enterprises that haven't quite seen the light yet. Furthermore, while some degree of modular design is doubtless appropriate for even large SMP servers, that's not the same as saying that everything must be shoehorned into a single standardized blade chassis.

That said, the NB50000c provides a data point that demonstrates that HP is serious about going further in this direction, more quickly than any other major vendor. Perhaps SGI comes closest, but with a far more specialized product lineup. Certainly, none of the other companies typically considered Tier 1 OEMs are currently pushing blades as broadly as HP despite past leanings in this direction—such as IBM's (largely x86-centric) modular computing push of a few years back.

None of this should be taken as a suggestion that NonStop is Just Another Blade. Notwithstanding ServerNet I/O cards and switches, the NB50000c is best thought of as largely co-opting hardware developed for other purposes, rather than gracefully co-existing on the same set of hardware. Thus, although future HP plans call for "Adaptive Infrastructure in a Box" blade solutions in which NonStop works alongside Linux, Windows, or HP-UX blades, for now we're talking about dedicated racks of NonStop gear that just—by no means incidentally—take standard gear, add ServerNet, and sprinkle a liberal dose of NonStop software pixie dust. Think of these as NonStop systems that happen to leverage the BladeSystem design, rather than standalone blade servers.

But that's OK. HP has (if belatedly) recognized the incredible intellectual property it has in NonStop, and started to apply it to problems and customer sets that fall outside its traditional market. NonStop will never be a truly mainstream product if, by mainstream, you mean something that's applied to pedestrian problems. Rather, the opportunity here is when "good enough" isn't enough even when it doesn't truly reach the most insane levels of the most ultimate scale and ultimate reliability.

[13] IOAMEs are also supported for compatibility.

[14] The HP SIM Integrated Blades plug-in manages the chassis and server blades through the Onboard Administrator.