the **Availability Digest**

HP's NonStop Blades

August 2008

HP has ported its fault-tolerant NonStop server to its HP c-Class BladeSystem. Named the HP Integrity NB50000c BladeSystem, a fully-configured system can contain up to sixteen processors, the same as HP's largest contemporary NonStop servers. Based on dual-core Itanium processors, the new multicore architecture is called NSMA, the NonStop Multicore Architecture. An NSMA system delivers twice the power of the HP NS16000, until recently HP's largest NonStop server, in half the footprint.

Existing applications can be ported seamlessly to the new bladed system. Using standard NonStop management facilities, NSMA nodes can be added to existing NonStop clusters comprising other Integrity (NS-series) and S-series NonStop servers.
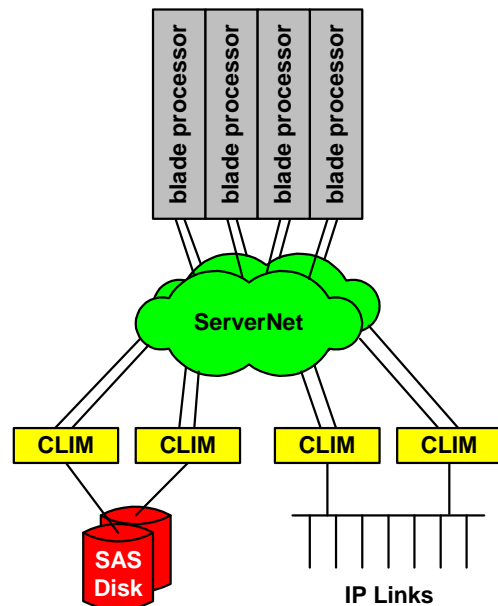
Overview

The NSMA system uses HP's c7000 c-Class blade processors running the NonStop operating system. An NSMA system comprises two to sixteen blade processors. Each blade is driven by an Intel dual-core Itanium microprocessor with up to 48 gigabytes of memory.

The standard NonStop fault-tolerant architecture has been ported to the BladeSystem. Processes may run as checkpointed process pairs or as persistent processes. For checkpointed pairs, one process is the active process. The other process, which is running in a separate processor, is the backup process, whose state is kept current via checkpointing for instant takeover. Should the active process or its processor fail, its backup will take over processing without the loss of any context.

Alternatively, processes can run as persistent processes under a checkpointed monitor that will restart a process in an operating processor should the process or the processor in which it was running fail.

A major modification to the NonStop operating system to support the NSMA system implements a new process scheduler that is multicore-aware. Its responsibility is to allocate processes to cores for proper load balancing.

NonStop Multicore Architecture

NSMA uses a new I/O subsystem called a CLIM (Cluster I/O Module). The CLIM is a duplexed-pair of Proliant servers that interfaces SAS disks and IP channels to the blade processors via a dual ServerNet fabric. The CLIM also supports XP storage and other NonStop disk and tape devices as well as other communication channels.

NSMA systems may join heterogeneous clusters comprising other NSMA systems, Integrity NonStop servers, and S-series NonStop servers. All are managed by the same system management facilities used by other NonStop systems.

NSMA Hardware Architecture

The Processor

NonStop Blade

The new NSMA system uses HP's standard c7000 c-Class blade processor. This blade fits vertically into a 10U enclosure[1] that can hold up to eight blades. Thus, a full sixteen-processor NSMA system comprises two enclosures.

A c7000 c-Class blade configuration includes:

- an Intel Itanium 9100 dual-core processor running at 1.66 gigahertz (each core is called an Instruction Processing Unit, or IPU).
- 18 megabytes of cache memory.
- 8 – 48 gigabytes of main memory, in increments of 8 gigabytes.

In addition to this standard blade configuration, an NSMA blade carries one additional card – a ServerNet card for connecting to the dual ServerNet fabric There are two redundant ServerNet fabrics for a NonStop BladeSystem. The fabrics provide 12 or 24 duplexed I/O links per enclosure. Therefore, a full NSMA system with two enclosures can provide up to 48 ServerNet I/O connections. There is no additional cabinet height consumed by the ServerNet interconnect. The ServerNet switches are embedded inside the blade enclosure at the back and are contained in one Field Replaceable Unit (FRU).
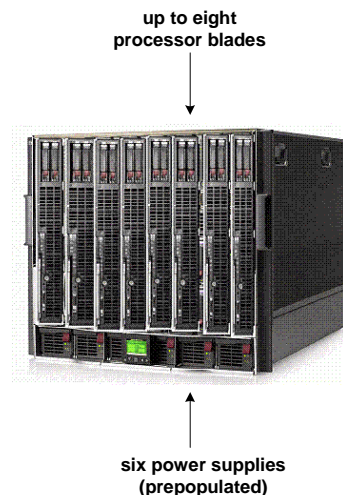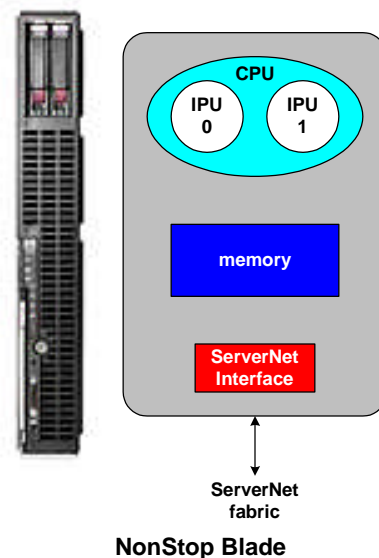

NonStop Blade

A NonStop blade comprises one logical processor in a NonStop BladeSystem. There can be up to sixteen logical processors (blades) in a system. Should one fail, its processing functions are taken over by one or more other logical processors.

Even though the Integrity microprocessor that drives the blade has two independent IPUs, the fault zone is the blade itself. If one of the IPUs fails, the blade fails. There is no attempt to continue to operate with only one IPU. Therefore, if a blade fails, the processes running in that blade will fail over to other blades.

Enclosure

NSMA uses the standard HP c7000 c-Class enclosure to house the blade processors. This is a 10U chassis that can hold up to

up to eight processor blades

six power supplies (prepopulated)

An NSMA Enclosure 2

---

[1] A "U" is a rack unit and is 1.75 inches high.

eight blade processors. It comes prepopulated with six power supplies (2,250 watts each) and 10 fans. Though only some of these are required to power and cool the enclosure, this provides spares and, more importantly, room for growth as quad- and eight-core chips become available.

The enclosure is designed to minimize single points of failure. The midplane interconnecting the processor blades and the ServerNet fabric is not active. Furthermore, only point-to-point links are used – there are no busses.

NSMA cabinets are 42U high. However, only one processor enclosure may be put into a cabinet because of power, cooling, and weight restrictions.

CLIM – The I/O System

The CLIM (Cluster I/O Module) is a newly-developed storage and communication interface for NSMA. It supports SAS (serial-attached SCSI) disks and IP interconnects. It also supports certain legacy disk and communication systems as well as HP's XP Storage.

CLIM

The CLIM platform is an HP Proliant DL385 rack-mounted server with a 2U height. It is driven by a 1.8 gigahertz dual-core Opteron processor with four gigabytes of memory and contains redundant power supplies and fans.

Since a NonStop BladeSystem can provide up to 48 ServerNet I/O links, it can support up to 44 CLIMs in addition to the two required storage CLIMs (which use two ServerNet ports each).

A CLIM contains eight PCI card slots. Five of these may be used for Host Bus Adapters (HBAs). The SAS disks and fibre channel links connect to the CLIM via the HBAs, which are described later. Via the HBAs, a storage CLIM can be configured to host four SAS ports or two SAS ports and two fiber channel ports. The communication CLIM can support up to 5 copper gigabit-per-second Ethernet ports or 3 copper and 2 fiber gigabit-per-second Ethernet ports

The CLIM functions are implemented via Linux. No customer application software can run on the CLIM. It appears to the outside world simply as a device controller.

CLIMs are usually configured in redundant load-sharing pairs. CLIM health is monitored by heartbeats sent to the NonStop blade processors. Should one CLIM fail, connectivity is maintained to the disk and communication devices by failing over all connections to surviving CLIMs.

Storage CLIM

The storage CLIM supports dual-ported SAS disks. The disks are a 2½ inch form factor, and 25 disks can fit into a 2U rack-mounted MSA70 enclosure.[2]
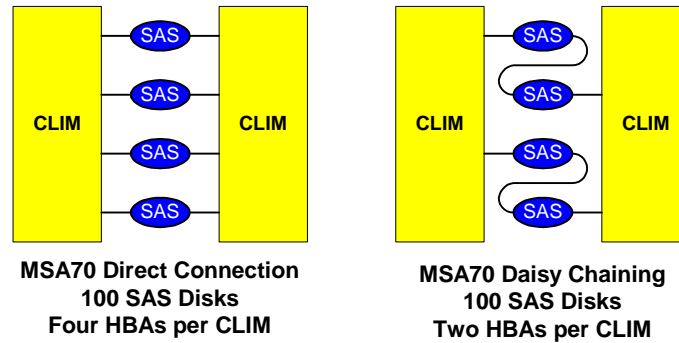

MSA70 SAS Disk Array

A storage CLIM pair can drive four MSA70 SAS disk storage arrays. Thus, a CLIM pair can control up to 100 SAS disks. The disks are usually configured as mirrored pairs, resulting in 50 disk mirrors per CLIM pair.

---

[2] As opposed to 14 disks in a 3U enclosure for fibre-channel connected disks.

The SAS disks can either be 72 gigabyte, 15,000 rpm disks or 146 gigabyte, 10,000 rpm disks.

The SAS disk arrays connect to the CLIM pair via host bus adapters. HBA slots can be used for disk-array connections using one of two options. Either each MSA70 disk array may be connected to each CLIM directly, or two MSA70 disk arrays may be daisy-chained and connected as a chain to CLIM HBAs. For a full complement of four disk arrays, direct connections require four HBAs of the allotted five on each CLIM. Daisy chaining requires only two HBA connections,

CLIM	SAS	CLIM
	SAS	
	SAS	
	SAS	

MSA70 Direct Connection
100 SAS Disks
Four HBAs per CLIM

CLIM	SAS	CLIM
	SAS	
	SAS	
	SAS	

MSA70 Daisy Chaining
100 SAS Disks
Two HBAs per CLIM

All HBA links in a CLIM may be active simultaneously, each carrying three gigabytes per second of traffic. Thus, a fully populated CLIM using direct HBA connections can transfer data at a rate of up to twelve gigabytes per second. To support this data rate, each storage CLIM by default uses two ServerNet ports. Thus, a fully-configured NonStop BladeSystem with 22 storage CLIMs can contain over 160 terabytes of mirrored storage.

The SAS disks are significantly faster than the fibre-channel connected disks used in the NonStop Integrity series. For one thing, each disk contains onboard cache that can be considered part of the storage device since it is mirrored. The Write Cache Enabled (WCE) option allows writes to complete to cache rather than having to be written to disk. Measured comparative performance improvements with Write Cache Enabled are tabulated below:

Read Sequential	70%
Read Random	20%
Write Sequential	500%
Write Random	35%

If WCE is used, in-cabinet UPS is required to prevent the loss of cached data following a power outage. The standard in-cabinet UPS system can provide power for about five minutes. This can optionally be extended to ten minutes.

The storage CLIM replaces FCSAs (Fibre Channel ServerNet Adapters) in an IOAME ((I/O Adapter Module Enclosure) used in HP's current Integrity NonStop servers, though these adapters are still supported by NSMA via ServerNet. Integrity and S-series databases can be migrated online to the NonStop BladeSystem.

Fibre-channel connected XP storage is directly supported by the CLIM, as are SAS and fibre-channel tape systems. Migrating an XP storage unit to a NonStop BladeSystem is simply a matter of reconnecting it to a CLIM pair.

Communication CLIM

As with storage CLIMs, communication CLIMs are normally configured as redundant load-sharing pairs for fault tolerance. Each communication CLIM can support five Gigabit Ethernet ports, either

4

as five copper ports or as three copper ports and two fibre-channel ports. One copper port is built into the CLIM. The other four ports are connected via Ethernet NICs.

A communication CLIM by default uses one ServerNet port, though two ports per CLIM may be configured for very high traffic. Thus, a fully configured NonStop BladeSystem with 44 communication CLIMs can theoretically support up to 220 Gigabit Ethernet channels.

The CLIM supports both IPv4 and IPv6 with IP Security (IPSec).

The communication CLIM also supports HP NonStop SWAN (ServerNet Wide Area Network) concentrators that handle bisync, async, X.25, and SDLC communication interfaces.

Power

The cabinet power and cooling are designed to support the systems of the future as quad-core and eight-core multiprocessors become available.

All cabinets in a c7000 system have a pair of power distribution units (PDUs) that supply power to the CLIMs and SAS disks. These PDUs can supply 8.6 kilowatts of I/O power. Each CLIM requires 250 watts, and each MSA70 SAS array requires 225 watts. Therefore, a fully populated I/O cabinet with 20 I/O units will consume about 5 kilowatts of power.

The processor blades are not powered by the in-cabinet PDUs. Rather, they derive their power from the six in-chassis power supplies, four of which are in standby mode ready to come into service if the power draw increases. These 2,250-watt power supplies are fed from an independent redundant pair of three-phase input feeds. Three-phase power is required because of the high power density created by the blade packaging.

A blade consumes about 350 watts. Therefore, a fully populated blade enclosure will consume 2.8 kilowatts. Fan power can potentially drive the enclosure power requirements beyond 3.8 kilowatts, which can be supported by just two of the six in-chassis power supplies. This power consumption will undoubtedly increase as more cores become available on the chips. A c7000 enclosure with redundant power supplies can supply 6.75 kilowatts of power to the enclosure.

In the event of an external power failure, an in-cabinet UPS can supply two three-phase output feeds for about five minutes. One feed powers the cabinet PDU, and one feed powers the c7000 enclosure. The UPS can optionally be configured to supply cabinet power for up to ten minutes.

NSMA Fault Tolerance

The fault-tolerant hardware characteristics of the NSMA architecture have been described above. Not only are there multiple processors in the system, but the CLIMs provide redundant access to network links and data storage devices. The redundant high-speed ServerNet fabric provides fault-tolerant communication between all devices. The system is architected so that it will survive any single hardware failure as well as some multiple failures.

This is hardware protection. But what about protecting the system and application processes[3] that are running on the fault-tolerant platform? What if a processor fails and takes down all of the processes in it? What if a process aborts?
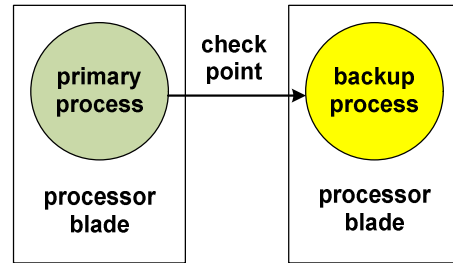
The NonStop server provides two primary mechanisms to protect processes against failure – *checkpointed process pairs* and *persistent processes*.

---

[3] A *process* is a *program* running in a *processor*. There may be several instances of a program running as different named processes.

Checkpointed Process Pairs

Critical system processes are implemented as checkpointed process pairs. With this technique, a failed process can be recovered in milliseconds.

Two copies of the process are spawned, each in a different processor blade. One process is designated the primary process, and the other process is its backup. The primary process keeps its backup process synchronized via checkpointing. Whenever the state of the primary process changes, that state change is sent to the backup process via a checkpoint message. The backup process uses the checkpoint information to update its state. In this way, the backup process can immediately take over processing from the exact point at which the primary process failed.

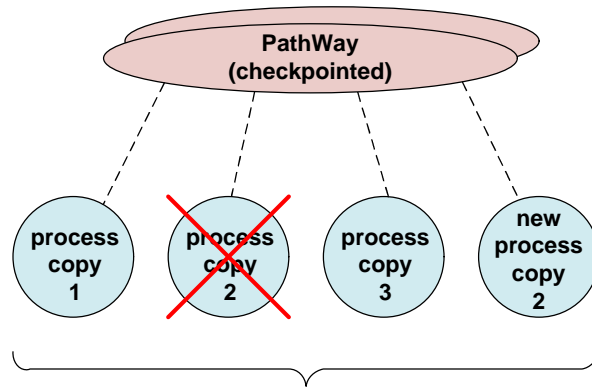**Checkpointed Process Pair**

If the processor in which the primary process is running fails, or if the primary process aborts, the NonStop operating system will immediately activate the primary process' backup. Processing continues uninterrupted. The mechanism exists for a backup process that has just been promoted to primary to start its own backup process in a surviving processor, though this is usually not done.

Critical operating system processes such as disk processes, communication processes, and process monitors (used for process persistence, as described next) are generally implemented as checkpointed process pairs. Many third-party products use process pairs for their critical processes.

Persistent Processes

Developing proper checkpointed process pairs is a difficult task and is therefore not usually used for application fault tolerance. It is used mainly at the system level.

To provide a simple-to-use process recovery mechanism for application developers, the NonStop operating system supports persistent processes. Process persistence is the responsibility of the NonStop checkpointed process monitor, PathWay. Persistent processes recover in seconds from a hardware or software failure.

Persistent Processes

Any application can be run under PathWay. All that this requires is the specification of some configuration parameters. PathWay is responsible for spawning the application's processes and then to monitor their health. Should a process abort, PathWay restarts it. Should a processor fail, PathWay restarts all the processes that had been running in that processor in surviving processors.

PathWay can also spawn server classes. A server class is multiple copies of the same process (or server) distributed across several processors. PathWay distributes transactions to processes within a server class to balance the load across all processes. If the load increases, PathWay can spawn more servers. If the load diminishes, PathWay can terminate unused servers.

NSMA Software

NonStop blades run substantially the same NonStop operating system as do all other contemporary NonStop systems. However, some modifications were necessary to create what is now the J-Series of blade operating systems:

Memory Allocation

Both IPUs share the same memory as well as the operating system image, locks, and other shared resources. However, a small amount of main memory is allocated to each IPU. This private memory of 64 kilobytes holds such structures as the IPU's live register set, its ready list, and its data cache and instruction cache.

Process Scheduler

A new process scheduler was implemented for NSMA to take advantage of the multicore architecture. The role of the process scheduler is to assign processes to specific IPUs.
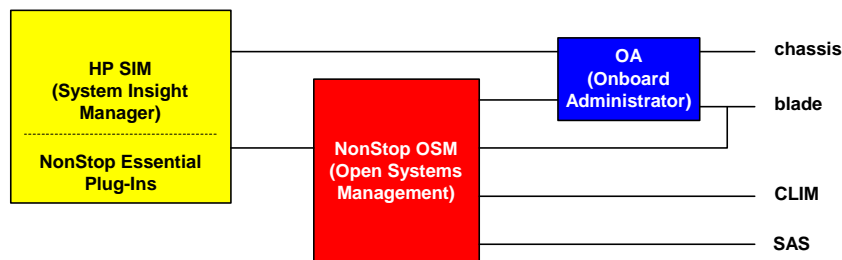
Rather than running the IPUs off a common task queue as is often done in SMP (symmetric multiprocessing) systems, the NSMA process scheduler assigns processes to a particular IPU. This is because moving processes between IPUs is very expensive. For user processes, this assignment is done at process-creation time and generally is fixed for the life of the process.

However, the process scheduler has several options, called *affinities*, built in for later expansion, some of which are used in the initial implementation. These scheduling affinities include:

- Dynamic – The IPU is picked when the process is ready to run (used for interrupt processes).
- Hard – The process is locked to the IPU (used for Measure performance-measuring processes).
- Group – All of the processes in a group always run on a single IPU and are moved as a group. DP2 processes are currently the only process group.
- Soft – The IPU is selected by the scheduler. In the initial release, a process is assigned to an IPU at process-creation time. In later versions, processes may be moved between IPUs for dynamic load balancing.

NSMA Management

The management of NonStop blades is integrated into the wider HP system-management products, especially HP's Systems Insight Manager (SIM), which provides management of heterogeneous systems across the entire enterprise environment. All NonStop management tools and third-party products continue to provide their existing functionalities. Full centralized monitoring and management of remote resources ensures Integrity Lights-Out operation (iLO).

New to the NonStop world is the Onboard Administrator (OA). Implemented as a pair of redundant modules within the blade enclosure, the OA manages enclosure power and cooling and monitors the health of the processor blades and the ServerNet interconnects, generating alerts when necessary. It is accessible via a Web interface and interfaces with NonStop OSM and HP SIM.

The major NonStop serviceability application for bladed systems remains the Open System Management (OSM) facility. OSM is a browser-based system management tool for NonStop systems. It has been upgraded to support the OA and to monitor CLIMs and SAS disk arrays.

HP Systems Insight Manager is HP's management facility for managing heterogeneous systems across the enterprise. Several plug-ins called NonStop Essentials have been developed for it to manage NonStop systems. They include:

- Cluster Essentials for managing clusters of NonStop systems.
- I/O Essentials for managing CLIMs and SAS arrays.
- Performance Essentials for monitoring the performance of heterogeneous clusters of Linux and NonStop servers.

SIM runs on a Linux system. However, the interface provided by the NonStop Essential plug-ins require no knowledge of the syntax of NonStop or Linux commands. The interface is an intuitive GUI.

NSMA Performance

Performance tests of NonStop BladeSystems using HP's Order Entry benchmark (which is TPC-C compliant) show that the processing power of an IPU is substantially that of an NS16000 processor. Therefore, since there are two IPUs, the throughput of a NonStop blade is roughly twice that of an NS16000 processor. Order-Entry benchmark tests show the following per-logical processor capacities:

NS16000	165 tps
NonStop Blade	349 tps

Since an IPU is equivalent to an NS16000 processor, the response times are roughly equivalent. However, it should be noted that because a blade has twice the processing capacity, it may require twice the memory.
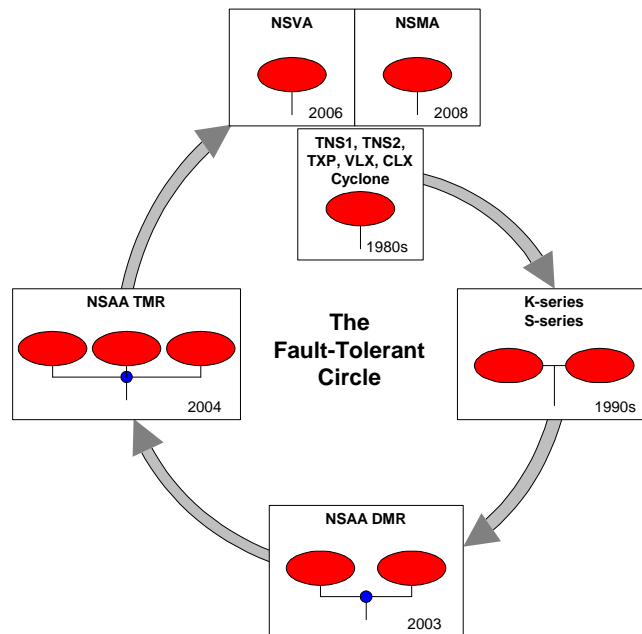
An interesting observation is that a blade processor is slowed down somewhat due to the multicore scheduling overhead and memory contention. However, this is offset by the overhead caused by the LSU (logical synchronization unit), which is the voting mechanism for the multiple physical processors that comprise a logical NS16000 processor.

The Fault-Tolerant Circle

NSMA has gone full circle to return to the original Tandem architecture. The architecture of Tandem's first product, the TNS1,[4] carries through to the HP NonStop systems today. Each system can have up to sixteen logical processors. Should a processor fail, the processes it had been running will instantly fail over to surviving processors (though process failover in the early days was all through checkpointing since persistent processes had not yet arrived on the scene).

---

[4] The Tandem NonStop 1 used a 0.7 MIPS processor with up to one megabyte of memory. We have come a long way!

The early systems (TNS1, TNS2, TXP, VLX, CLX, Cyclone), were each powered by custom-designed logical processors. In the early 1990s, the logical processors were redesigned with commodity RISC (Reduced Instruction Set Computer) microprocessors. These logical processors were used in the K-series and S-series systems. Since the RISC processors did not have much in the way of internal error checking, each logical processor used two RISC chips running in lockstep at the memory-access level. The logical processor would fail if there were a mismatch, thus providing *fast-fail* to prevent data corruption.

Under HP, the NonStop logical processors evolved to use dual lock-stepped Itanium microprocessors.[5] Because these microprocessors were not deterministic, memory lockstepping could no longer be used. Therefore, the microprocessors were lock-stepped at the I/O level (any packet delivered to the interconnecting ServerNet fabric). Since each microprocessor was now an independent processor, the logical processor survived even if one of its microprocessors failed. This dual modular redundancy (DMR) architecture was named NSAA (NonStop Advanced Architecture). Subsequently, a TMR (triple modular redundancy) option for a third microprocessor was offered to provide extreme reliabilities.

However, the Integrity microprocessors were so reliable in terms of their internal error checking that fast-fail became not so important for many applications. In recognition of this, HP introduced a simplex version of its Integrity series that it called the NonStop Value Architecture. NSVA completed a full circle back to the early days of Tandem systems, in which each logical processor was a single processor with no lockstepped, fast-fail protection. After all, if this architecture was good enough in the '70s to give Tandem the niche edge in fault-tolerant computing, it is good enough now for many applications.

The NSMA architecture follows in the footsteps of NSVA, offering a single microprocessor-based logical processor in a fault-tolerant configuration and utilizing software failover to recover from processor failures.

---

[5] R. Buckle, W. Highleyman, The New NonStop Advanced Architecture: A Massive Jump in Processor Reliability, the *Connection*; July/August, 2003.

Summary

The NonStop BladeSystem is a major advancement in NonStop fault-tolerant technology. It provides twice the power in half the footprint of current NonStop servers. It can be integrated as a cluster member with other NonStop Integrity and S-series servers, and it is managed as seamlessly as any of the other servers.

Perhaps equally important is that NSMA leverages existing HP technology. Except for the ServerNet fabric so important to tying a NonStop system together, all of the hardware in a NonStop BladeSystem is standard hardware used in other HP products. This includes the c7000-class blades and enclosures, the Proliant CLIM I/O servers, and the SAS disk arrays.

A "Small" NonStop Blade System