Technical white paper

HP Integrity Superdome 2

The ultimate mission-critical platform

Table of contents

Click here to verify the latest version of this document

Introduction

For over 10 years, HP Integrity Superdome—our flagship, high-end Integrity server—has powered some of the world's most demanding mission-critical environments. For enterprise customers who require high availability, rich virtualization capabilities, and long-term investment protection, Integrity Superdome is the solution. With Integrity Superdome 2, HP pioneers a new category of modular, mission-critical systems, blending trusted Integrity Superdome reliability with HP Integrity BladeSystem efficiencies.

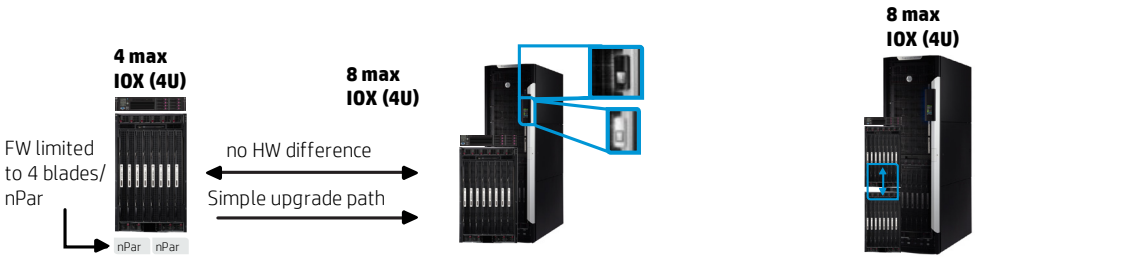
Integrity Superdome 2 is based on HP Mission-Critical Converged Infrastructure strategy, which attacks IT sprawl with standards-based, modular architectures. For Integrity Superdome 2, this means a common, bladed architecture as well as common components and common management environment across all HP servers. The result is an Integrity Superdome that is modular, modern, and easier to manage than ever before.

At the heart of Integrity Superdome 2 is the innovative sx3000 enterprise systems chipset, extending Integrity Superdome 2 reliability and availability to new levels. This paper provides an overview of the Integrity Superdome 2 architecture, its sx3000 chipset, and the host of innovative mission-critical features that make the Integrity Superdome for the next decade.

Product family

Using the design principles of a mission-critical converged infrastructure, Integrity Superdome 2 was designed to be more modular, to provide a lower cost of entry than the first generation of Integrity Superdome, and to leverage standards-based components throughout. As a result, Integrity Superdome 2 uses a 19" rack and includes a bladed design, with the basic building block being the Integrity Superdome 2–16s compute enclosure. The enclosure leverages HP BladeSystem c7000 enclosure and shares a common midplane, fans, and power supplies to give customers common, easy-to-service spares. It is 18U and supports up to eight Integrity Superdome 2 blades that contain compute, memory, and I/O resources. The architecture supports complexes with one or two compute enclosures to provide up to a 32-socket symmetric multiprocessing (SMP) system with 512 DIMM slots, and 64 on-blade NICs. For those customers needing more I/O capability, an Integrity Superdome 2 I/O Expansion Enclosure provides 12 PCIe slots in a 4U space. Up to eight I/O expansion enclosures can be supported in the Integrity Superdome 2–16s and Superdome 2–32s for an additional 96 PCIe I/O slots.

Figure 1. HP Integrity Superdome 2 product family



Superdome 2–8s
16s with 8 socket SMP
(18U, 19" rack)

16	Sockets (limited to 8 sockets per n Par)
128*	Cores
4 TB	Memory (16 GB DIMM)
32 Internal	10GbE ports
48 I/O slots	PCIe x8 Gen2

Superdome 2–16s
16s with 16 socket SMP
(18U, 19" rack)

16	Sockets
128*	Cores
4 TB	Memory (16 GB DIMM)
32 Internal	10GbE ports
96 I/O slots	PCIe x8 Gen2

Superdome 2–32s
32s with 32 socket SMP
(36U, 19" rack)

32	Sockets
256*	Cores
8 TB	Memory (16 GB DIMM)
64 Internal	10GbE ports
96 I/O slots	PCIe x8 Gen2

*Intel® Itanium® Processor 9500 series

Performance

The Integrity Superdome 2 Server has a well-balanced architecture that is designed to increase the performance of the Itanium® processors across a wide variety of commercial and technical applications. Integrity Superdome 2 provides greater than twice the local memory bandwidth and two to four times the I/O bandwidth of the sx2000 systems (see table 6). The sx3000 chipset also provides 128 MB-level four cache per blade, which caches data from memory on remote blades, significantly reducing average latencies. The memory controllers are integrated inside the Intel Itanium 2 processor CPU socket and provide more than twice the bandwidth (see table 6) with less latency than seen in sx2000-based Integrity Superdome Servers.

The sx3000-based Intel® Superdome 2 with 32 Intel Itanium processor 9300 and 9500 series sockets provides up to 816 GB/s of I/O bandwidth and an unprecedented 1248 GB/s of system fabric bandwidth. This results in excellent performance scaling up to 256 processor cores with the Intel Itanium Processor 9500 series and 128 I/O adapters, enabling customers to add capacity as their business needs grow.

The Integrity Superdome 2 Server leverages leading-edge, enterprise-ready technologies to enhance performance. In large part, this is accomplished by using many of the same infrastructure components as in the BladeSystem c-Class enclosure (that is, power supplies, fans, interconnect modules). In addition to using the industry-standard Itanium 2 processors, double data rate 3 (DDR3) DRAMs and PCIe I/O adapters are used.

Reliability and availability

Integrity Superdome 2 uses error-correcting, self-healing, and advanced diagnostics technologies to increase availability. At the heart of Integrity Superdome 2 is a fault-tolerant HP Crossbar Fabric that is enabled by passive midplanes, end-to-end retry, and link failover. The manageability subsystem, the clock distribution system, cooling, and the power system also have full redundancy. In addition, double-chip spare restores chip-spare protection after a DRAM has failed and protects against the vast majority of memory buffer and data bus failures.

Another key reliability feature is the HP Superdome 2 Analysis Engine, which moves platform diagnostic capabilities into the firmware level where it can drive self-healing actions and can report failures even when the OS is unable to boot. It is also able to correlate disparate errors across multiple partitions for simplified reporting.

Other key reliability and availability features of Integrity Superdome 2 include: ECC on all PCIe slots, nonshared PCIe busses to isolate failures to a single I/O card, redundant fans, end-to-end retry over SMP fabric links, L4 cache line replacement in the sx3000 chipset, and Intel Cache Safe Technology on CPU L2 caches, L3 caches, and directory cache.

Virtualization and manageability

Integrity Superdome 2 supports a wide array of partitioning and virtualization options to meet your needs. Hard partitions (nPars) provide full hardware isolation between partitions. Virtual Partitions (vPars) and HP Integrity Virtual Machines (VMs) allow hard partitions to be further subdivided to provide fine-grained control over assignment of resources. You can choose from HP-UX. We offer a tightly integrated suite of tools with HP Matrix Operating Environment (OE) for HP-UX. Matrix OE can significantly increase resource utilization with hard and soft partitions, virtual machines, and workload management to deliver mission-critical virtualization.

In addition, our orchestration capabilities support a tightly integrated environment that adapts to changing business needs. With HP Systems Insight Manager, common management tasks such as deploying or moving environments can be completed quickly and effectively. HP Global Workload Manager can increase the overall effectiveness of your environment by moving resources to the most critical tasks. That's integrated management and simplified automation.

Investment protection

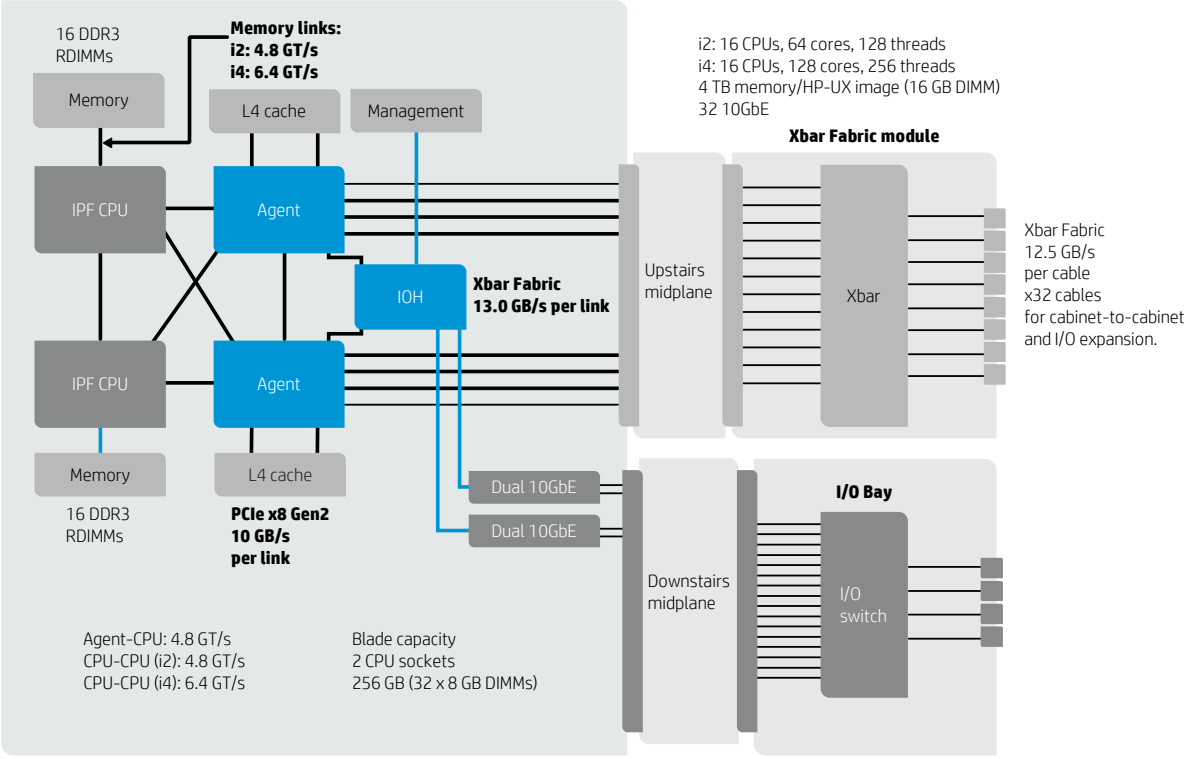
Integrity Superdome 2 continues the legacy of the superior investment protection offering of the first-generation Superdome servers. Integrity Superdome 2 Servers are designed to be upgradeable to keep up with your business needs. As part of the HP Mission-Critical Converged Infrastructure, a modular, building-block approach was used that provides both a lower cost of entry and scaling up to hundreds of CPU cores. You can add either new blades or I/O subsystems or both to the system to expand current partitions or to create new partitions. Integrity Superdome 2 blades can be quickly and easily removed to add memory. The Integrity Superdome 2 Servers were architected to simultaneously support nPars with future processors alongside nPars with Intel Itanium processor 9300 and 9500 series. In addition, applications running HP-UX 11i v3 run unchanged on the Integrity Superdome 2.

System topologies and sx3000 chipset

The Integrity Superdome 2 Servers combine the best of BladeSystem components and the unique sx3000 chipset to provide the next generation in mission-critical, scalable servers. These systems are based on building blocks called Superdome 2 blades. Each blade contains compute, memory, and I/O resources. These servers utilize a SMP computer architecture. Superdome 2 blades enable the deployment of hard partitions (also called nPars), which contain independent, hardware isolated OS instances, each with its own memory space. A hard partition consists of one or more blades within a server system. Multiple hard partitions can be deployed in a single complex to support server consolidation. Hard partitions running HP-UX can be subdivided into vPars to provide even finer grained partitions for server consolidation and virtualization.

The sx3000 chipset consists of three very large-scale integration (VLSI) component types: a blade controller (also known as Agent), a Crossbar switch (also known as Crossbar), and a PCIe system bus adapter (also known as IOH). The Crossbar switches reside on boards that plug into the passive system midplane and determine the overall topology of the system much the same as in legacy Integrity Superdome Servers. Placing the Crossbars on boards that plug into a passive midplane greatly increases the serviceability of these components as compared to previous-generation Integrity Superdome Servers where the Crossbars are attached to the midplane. As shown in figure 2, each Integrity Superdome 2 blade consists of two Agents, two processor sockets, 16 DIMM slots per processor socket, one I/O hub (IOH), and two dual-port 10GbE LOMs. The Agents communicate to the Crossbars through the Crossbar Fabric. External ports on the XFM can be used to provide interconnection either to another compute enclosure or Superdome 2 I/O Expansion Enclosures (IOXs), or both. The LOMs connect to the same passive midplane and interconnect modules used in the BladeSystem c7000 enclosures. A single 18U enclosure supports up to eight blades for a total 16 processor sockets, 256 DIMM slots, and 32 10GbE ports.

Figure 2. HP Integrity Superdome 2 Server blade



As in legacy Integrity Superdome Servers, the Crossbar switches determine the overall topology of the system and allow blades to be taken offline for upgrade or service without impacting the rest of the system. In contrast to legacy Superdome, the Crossbar switches provide not only the fabric between blades themselves, but also to the IOXs (see figure 4). This provides more flexibility because the IOXs are not connected directly to the blade, so the number of IOXs can grow independently of the number of blades. This also results in an improvement in availability because the IOXs can still be accessed even if a blade has experienced a fault. The Crossbar fabric is all new for Superdome 2 and is completely unique in the industry.

As previously mentioned, Integrity Superdome 2 reuses and leverages many components from the BladeSystem c7000 enclosure. As shown in figure 2, the downstairs midplane in Superdome 2 is the same midplane used in BladeSystem c7000 enclosures and connects to the same interconnect modules on the I/O bay side and same LOMs on the blade side. The upstairs midplane is unique to Integrity Superdome 2 and connects to the Agents on the blade side and Crossbar modules on the other side. This can also be seen from a figure of the rear of the Superdome 2 compute enclosure, which illustrates again the leverage of BladeSystem c7000 enclosure in the bottom 10U and Superdome 2 "value add" in the upper 8U.

There are several advantages to this type of architecture. With an Agent and Crossbar-based fabric, the full bandwidth is available regardless of the number of blades in the partition. Unlike directly connected topologies, the Superdome 2 fabric is fully redundant. IOXs can be added to a partition without adding blades, thus increasing the flexibility of both capacity and I/O bandwidth scaling. Finally, the availability of I/O in the IOX is not dependent on the blades themselves. Superdome 2 blades can be added and/or deleted from partitions without affecting accessibility to the IOX.

HP Integrity Superdome 2 blade subsystem

The Integrity Superdome 2 blade is constructed of processors, memory, and chipset VLSI (the sx3000 Agent and sx3000 IOH), as well as PCIe devices. The Agent chips extend the scalability of the processor and at the same time interface to I/O, interface to the Superdome 2 Server fabric, and manage a L4 cache for improving the latency of remote memory references.

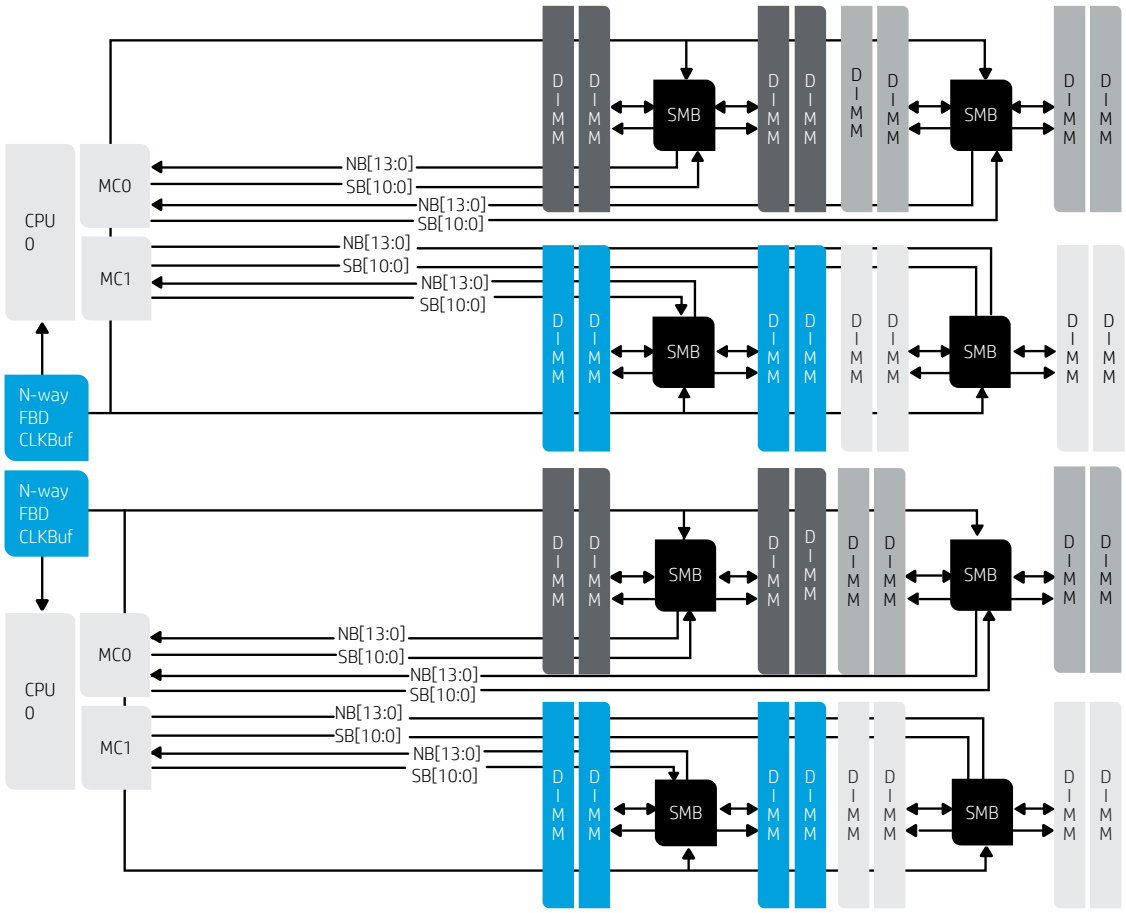
The two Agents on the blade provide four QPI links (19.2 GB/s each), six Superdome 2 fabric links that connect to system Crossbars (13.0 GB/s each), and two fabric links that connect to the on-blade I/O hub. Each processor module connects to four memory buffer chips, which in turn interface directly with DDR3 RDIMMs. This forms the core of the blade providing high bandwidth to memory, considerable memory capacity, and fabric to other blades and external I/O through the interconnect modules in the lower midplane. Superdome 2 processors have directly attached memory that provides low latency and high performance. The sx3000 architecture allows a processor core to reference this memory directly without consulting any other processor socket or the Agent chip, thereby preserving the advantage of on-die memory controllers. The processors supported by Superdome 2 are the Intel Itanium processor 9300 and 9500 series.

The new sx3000 chipsets maintain system cache coherence by using a Remote Ownership Tag Cache realized using on-chip SRAM that resides in the Agent chips. By tracking remote ownership for the cache memory that the blade hosts, scalability of the Intel Itanium processor 9300 and 9500 series can be extended from eight sockets to 32 sockets without compromising the latency of socket local memory references.

The Agent's L4 cache on the blade stores lines fetched from memory on remote blades. Once cached, future references do not have to cross the fabric, thereby reducing latency and increasing performance. The L4 cache is realized with low latency eDRAM chips that store the L4 cache data and on-chip SRAM that stores L4 cache tags. The L4 cache is 64 MB per processor socket and uses a write-back caching policy. The 64 MB is shared by all cores in the socket. When four or eight cores in one socket all cache a line, there is just one copy of the L4 cache. If one core has R/W access to a line and to another core on the same socket, the request goes all the way to the home agent responsible for the line. The L4 cache is not designed to perform ownership transfers.

A significant change in Integrity Superdome 2 from previous-generation Superdome Servers is that the memory controllers are integrated in the processor module itself. This reduces the latency time for memory accesses. There are two independent memory controllers per processor socket. Each memory controller communicates through a SMI link with two SMBs, which in turn communicate with DDR3 DIMMs as shown in figure 3. The cache line is split across two DIMMs that work in lockstep. Therefore, a minimum of eight DIMMs is required to ensure that all memory controllers on both CPU sockets are active. Memory off the processor socket can be accessed even if all four cores have been disabled.

Figure 3. HP Integrity Superdome 2 memory topology



The Agent supports advanced interleaving of memory. The blade design interleaves memory requests at a low level enabling the use of both Agents and all six Crossbar links when a processor core references a page of memory hosted by a remote blade. The sx3000 Agent and IOH support fine-grained interleaving of pages across multiple processor sockets to reduce the effects of memory hot-spots. The sx3000 architecture has the ability to support memory interleaving among as many as 32 sockets. In addition, Integrity Superdome 2 supports socket-local memory allocation for the lowest latency to local data.

Table 1. sx3000 features

Feature	Value
QPI links	5 per blade
Crossbar Fabric links	6 per blade
L4 cache size	64 MB per processor socket
Remote Ownership Tag Cache	100 MB per processor socket
Interleave	1 to 32 processor sockets

The I/O subsystem

The I/O subsystem connects the processors and memory to the I/O cards. An Integrity Superdome 2 Server has two types of I/O subsystems: the on-blade I/O subsystem (which is included on every blade) and an optional IOX for customers wanting to deploy systems with many I/O cards per partition.

On-blade I/O subsystem

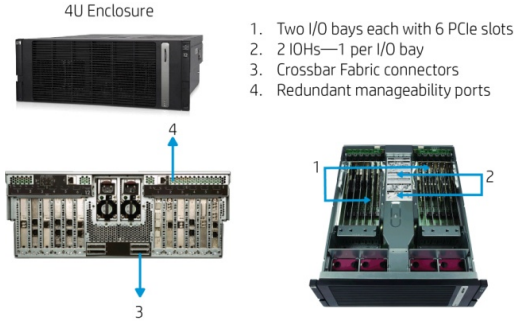
The on-blade I/O subsystem consists of the sx3000 IOH, two dual-port LOMs, and elements of the management subsystem including VGA, USB, and HP Integrity Integrated Lights-Out (iLO). The sx3000 IOH interfaces with the Agent chips through two Crossbar Fabric links each providing a bandwidth of 13.0 GB/s peak or 5.9 GB/s sustained. The IOH connects to the LOMs through two PCIe x8 Gen2 links. The LOMs (shown as dual 10GbE) are included on every blade and provide the blade redundant Ethernet connections to the Ethernet or pass-thru interconnect modules of the Integrity Superdome 2 (as shown in the figure 2).

External IOX

The IOX allows I/O card capacity to be extended without adding any additional blades, which is a cost-effective expansion approach. Using two sx3000 IOH chips, the Integrity Superdome 2 IOX extends the PCIe slot capacity by 12. Each PCIe slot supports industry-standard form factor cards and connects to the sx3000 IOH with a PCIe x8 Gen2 link, capable of running at 5.0 GT/s.

The IOX is a 4U rack-mounted enclosure that can be serviced while online.

Figure 4. HP Integrity Superdome 2 IOX



PCIe cards

The Integrity Superdome 2 can meet the most demanding I/O needs of today's computing workloads. For network connectivity, the Integrity Superdome 2 supports leading-edge 10GbE cards. To connect to storage, Integrity Superdome 2 supports 8 Gb Fibre Channel cards, InfiniBand cards, and SAS cards.

Fault-tolerant HP Crossbar Fabric

The Crossbar chip is the heart of the system fabric. It connects all of the blades in the system, providing a high-bandwidth, low-latency, coherent path among processors, memory, and I/O. This chip is a 20 ported nonblocking Crossbar, meaning it has 20 connections to and from blades, either IOX or other Crossbar chips, or both. There are eight of these chips in a 32-socket system. In addition to the scalability features, the sx3000 chipset is designed to tolerate the failure of one Crossbar chip while keeping the system running.

Table 2. sx3000 Crossbar features

Feature	Peak	Sustained
Bandwidth of each Crossbar link	12.0–13.0 GB/s	5.6–5.9 GB/s
Fabric bandwidth per blade	78.0 GB/s	36.0 GB/s

With this much fabric bandwidth, the sx3000 architecture provides room to grow so the full benefit can be seen with new applications and processors.

The links

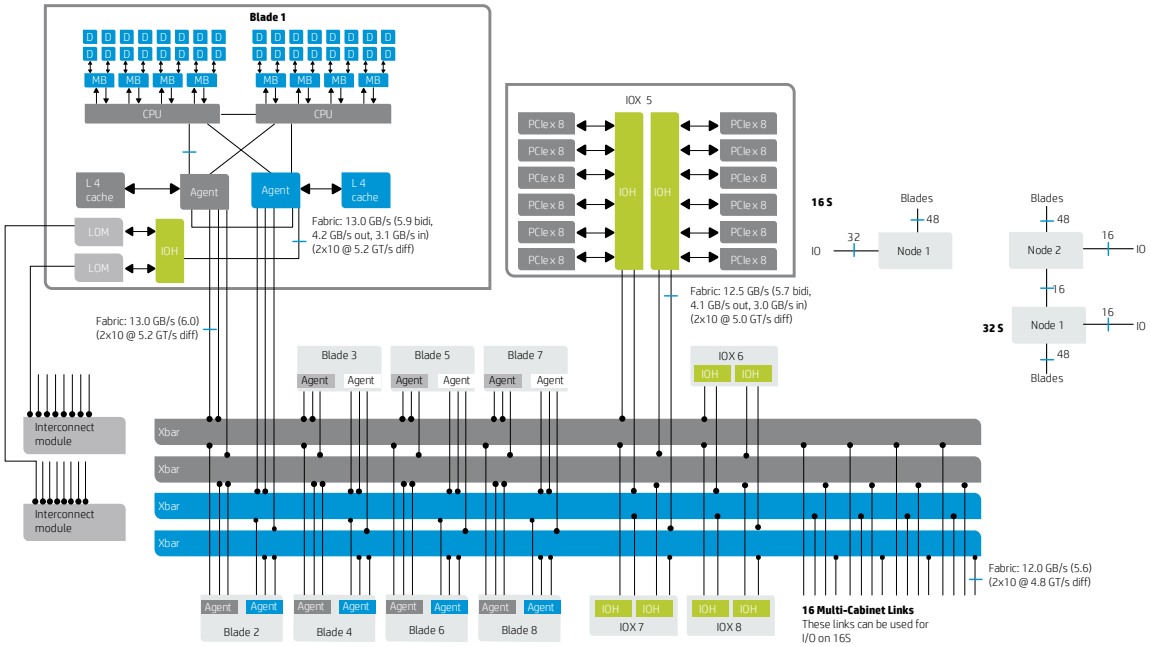
The sx3000 links take advantage of high-speed SERDES technology and have been designed to support transmission both through cables and in board traces. The same link technology is used for blade-to-blade interconnection and for

blade to I/O subsystem interconnection, both within a single enclosure or among multiple enclosures. In addition to providing higher bandwidth, the high-speed links have been designed to provide availability and reliability. For example, reliability is achieved by tolerating transient errors. When an error is detected on a high-speed link, any transactions in flight will be re-transmitted with the correct data. Should the link error prove to be persistent, the sx3000 chipset automatically re-routes data across an alternative path, without firmware intervention. This innovative end-to-end retry capability is unique to HP and prevents the loss of any single link from causing loss of data or a partition crash.

System topologies

In Integrity Superdome 2, the blades are connected using system fabric Crossbars to build "nodes." Each node (also known as "compute enclosure") contains up to eight blades, 16 processor sockets, 256 memory DIMMs, and I/O. These nodes are connected to one another and to IOX enclosures via cabled links to build systems as large as 32-processor sockets and 36 IOHs (16 on blades, 16 in IOXs). The following diagram is a higher topology diagram. Fabric link speeds are shown as peak (sustained).

Figure 5. HP Integrity Superdome 2 system topology



Capabilities of Integrity Superdome 2 and the sx3000 chipset

The Integrity Superdome 2 of HP system was designed to continue supporting the mission-critical features of legacy Superdome Servers, while extending many of the manageability and serviceability features in the BladeSystem c-Class enclosures. The sx3000 chipset was designed to provide exceptional performance and reliable operations even in the presence of errors. In addition, Intel Itanium processor 9300 and 9500 series have extended and enhanced the reliability, availability, and serviceability (RAS) features implemented in previous-generation processors. This section describes how these benefits are delivered with Integrity Superdome 2 Server of HP family.

Reliability and availability

Systems with the new sx3000 chipset have numerous innovative self-healing, error-detection, and error-correction features to provide the higher levels of reliability and availability.

Several new hardware features contribute to the increased reliability of the Integrity Superdome 2 infrastructure. This includes a fully redundant clock distribution starting with the clock source and continuing through the distribution to the Integrity Superdome 2 blade itself. All midplanes in these servers are completely passive, unlike legacy Superdome where the Crossbar switches are attached onto the midplane. The fabric reliability has been increased many-fold over legacy Superdome through a combination of link-level retry, link-width reduction, and end-to-end retry. A list of reliability features is shown in table 3.

Table 3. Superdome 2 system reliability and availability features

Location	Features	Customer experience
Memory system	DRAM protection (ECC, SDDC, DDDC) double device data correction in memory Memory scrub (patrol and demand) Memory channel protection (retry, reset, and lane failover) Can distinguish SMI link CRC error from Memory ECC error	17x fewer DIMM replacements than with traditional chip spare. Extreme levels of availability with no compromise of system performance or any added hardware cost. Risk of memory data corruption is drastically reduced to near zero with HP DIMM enhancements.
Processor	Cache error detection/correction Self-healing L2, L3, and directory caches Soft error (SE) hardened latches Core logic ECC and parity protection Dynamic processor resiliency Core deconfiguration Advanced machine check architecture with new CMCI support MCA Error Recovery with assistance from HP-UX QPI Interconnect path detection/correction (CRC, retry, reset, and lane failover)	Covers all cache errors and the majority (70%) of the CPU core errors resulting in much better error coverage and data integrity than can be expected with x86 CPUs: enterprise-class reliability for enterprise customers.
Blade, I/O, and fabric links	Link-level retry Link-width reduction End-to-end retry IOX attached to XFMs	Fault-resilient links mean partitions that stay up This feature eliminates errors due to environmental glitches and latent manufacturing imperfections, common causes of field server failures. Able to service links without bringing the system down.
Crossbar/System fabric	Redundant links to blades Explicit support for hard partitioning	A key enabler of leadership partitioning strategy of HP.
I/O slots	Error detection/correction PCI failure isolation to a single slot enhanced I/O error recovery Multipathing PCI card OLARD	Moves I/O errors from one of the major contributors of system. The ability to online repair further enhances the fault avoidance capabilities.
Chipset	Internal data path error detection/correction Hardened latches L4 cache line sparing	HP value-added chipset puts performance and availability above all else.
Partitioning/System infrastructure	nPars—hardware and software isolation between partitions Redundant and hot-swap clock Fully redundant clock distribution Automatic failover and hot swap manageability models (OA & GPSM) Redundant packet-based management fabric with automatic failover Ease of service—hardware can be repaired without bringing down multiple partitions 2N power and power grid redundancy Redundant fans Passive midplanes HP Integrity Superdome 2 Analysis Engine	Integrity Superdome 2 enables true server consolation. With a measured infrastructure[1] MTBF of greater than 300 years, combined with two generations of hard partitioning experience of HP, a customer can be confident that an Integrity Superdome 2 Server broken up into hard partitions is an excellent approximation of an array of smaller boxes, but without all the system management, reliability, and cost of ownership headaches.

[1] Infrastructure includes the enclosure power distribution, cooling, and passive midplanes.

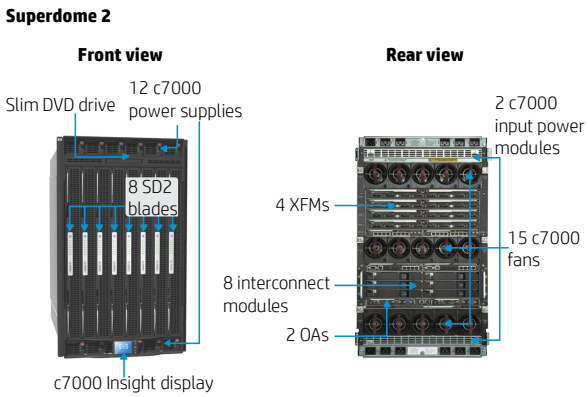
A key new feature of these servers is the industry-unique Integrity Superdome 2 Analysis Engine. In many systems, a large percentage of errors are handled and reported at the OS level. With virtualization and partitioning, this can be an issue, because a shared piece of hardware can trigger messages to multiple partitions. With the Analysis Engine, the diagnostics capabilities moved from OS-based agents to the manageability firmware. This allows for correlation of disparate failures, reduces issues of inconsistent responses between different OSs, and ensures that errors in shared resources are reported only once rather than multiple times. As the Analysis Engine is a part of the firmware, it also means that error-handling rules need to be updated in only one location, it is available even when online diagnostics are not loaded on the system, and errors can be analyzed even if a partition cannot boot. The Integrity Superdome 2 Analysis Engine also provides a single CLI for reporting the health of the server, including the replacement history of parts in the server. These automatic actions lower the impact of failures and help restore the server as quickly as possible.

Availability also refers to the ability to add, replace, or delete components while the system is running. Just as in the c-Class enclosures, everything seen can be physically removed and replaced while partitions continue to run including fans, power supplies, OA modules, interconnect modules, and GPSMs.

Serviceability

The Integrity Superdome 2 has been designed to be highly serviceable. Many components have been leveraged from the BladeSystem c-Class enclosures. Service repairs can be done quickly and efficiently and most are without tools.

Figure 6. Front and rear view of HP Integrity Superdome 2

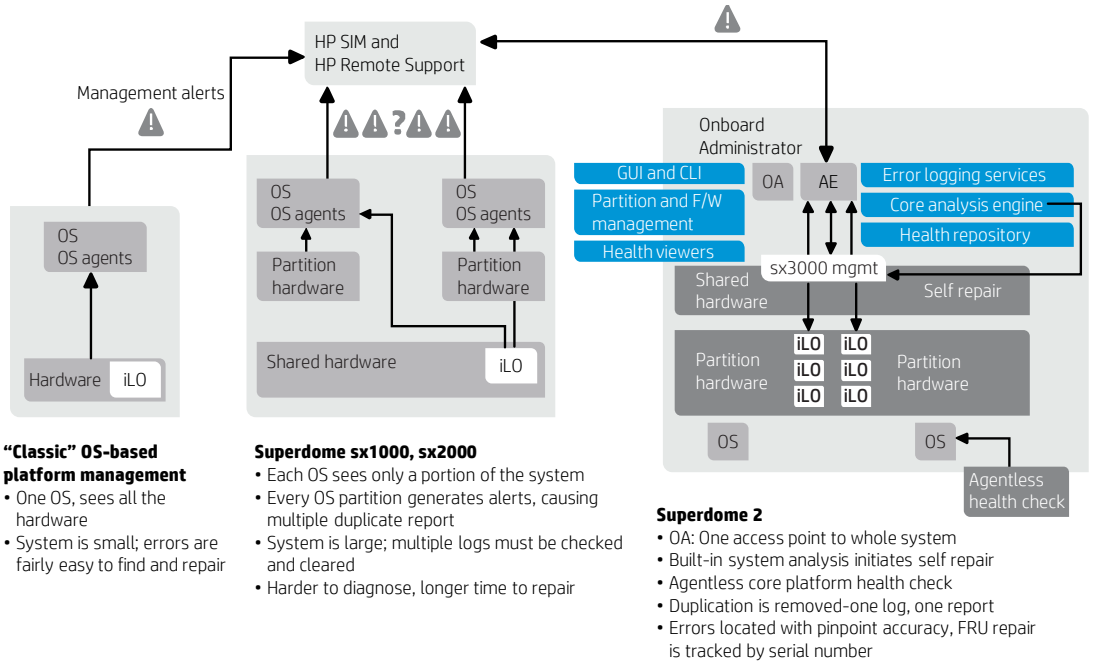


In addition, many of the software tools used to service and manage the system are built directly into the Integrity Superdome 2. Figure 6 compares platform management of HP Integrity Superdome. The “classic” image describes a common method used today in most x86 systems and smaller Integrity servers. Basic hardware configuration is done through system firmware tools such as BIOS or Extensible Firmware Interface (EFI). The platform includes a management processor (shown here as HP iLO), which provides remote administrator access and also monitors basic hardware “health” requirements such as voltage, temperature, power supply redundancy, and fans. In the “classic” design, if the management processor detects a change that needs an administrator’s attention, such as a rise in temperature, the management processor signals software “agents” that are running the OS. These server health agents then alert the administrator through protocols such as SNMP or WEBM. The classic OS-based platform management works very well for small servers that are fairly easy to diagnose and repair.

The middle, Superdome sx1000, and sx2000, part of the figure shows how HP took the “classic” picture and applied it to our cell-based partitionable servers: HP Integrity Superdome Server, HP Integrity rx8640 Server, and HP Integrity rx7640 Server. In these systems, there is a management processor, or a set of management processors, monitoring the shared system hardware. In addition, there are separate components monitoring the partition-specific hardware. Because these servers contain multiple OS partitions, every OS partition must be notified should the management processors detect a changed condition in the shared hardware. For example, if a chassis power supply fails, every OS partition must be notified. Consequently, every OS partition sends an administrator alert, and this can cause redundant error messages. In other situations, if an error is found in one partition’s hardware that information is not shared with monitoring components in other partitions or with the main management processor. This means operators much check multiple separate health logs to get complete system information. This architecture works, but it could be redesigned to be more efficient.

The final, Superdome 2 shows the new management architecture custom designed for Integrity Superdome 2. In Integrity Superdome 2, the core platform OS agents have been removed and replaced with analysis tools, which run in the management processor subsystem. Administrative alerts now come directly from the innovative Integrity Superdome 2 Analysis Engine, not from each OS partition, which removes duplicate reports. In this paper, it is referred to as "agentless health checking." But that's only a piece of the story. The Analysis Engine is doing much more than just generating alerts—it centrally collects and correlates all health data into one report. Then it analyzes this data and can automatically initiate self-repair without any operator assistance or manually-run diagnostics tools. Analysis is core to the new Integrity Superdome 2 platform management architecture.

Figure 7. Platform management comparison of HP Integrity Superdome



"Classic" OS-based platform management
- One OS, sees all the hardware
- System is small; errors are fairly easy to find and repair

Superdome sx1000, sx2000
- Each OS sees only a portion of the system
- Every OS partition generates alerts, causing multiple duplicate report
- System is large; multiple logs must be checked and cleared
- Harder to diagnose, longer time to repair

Superdome 2
- OA: One access point to whole system
- Built-in system analysis initiates self repair
- Agentless core platform health check
- Duplication is removed-one log, one report
- Errors located with pinpoint accuracy, FRU repair is tracked by serial number

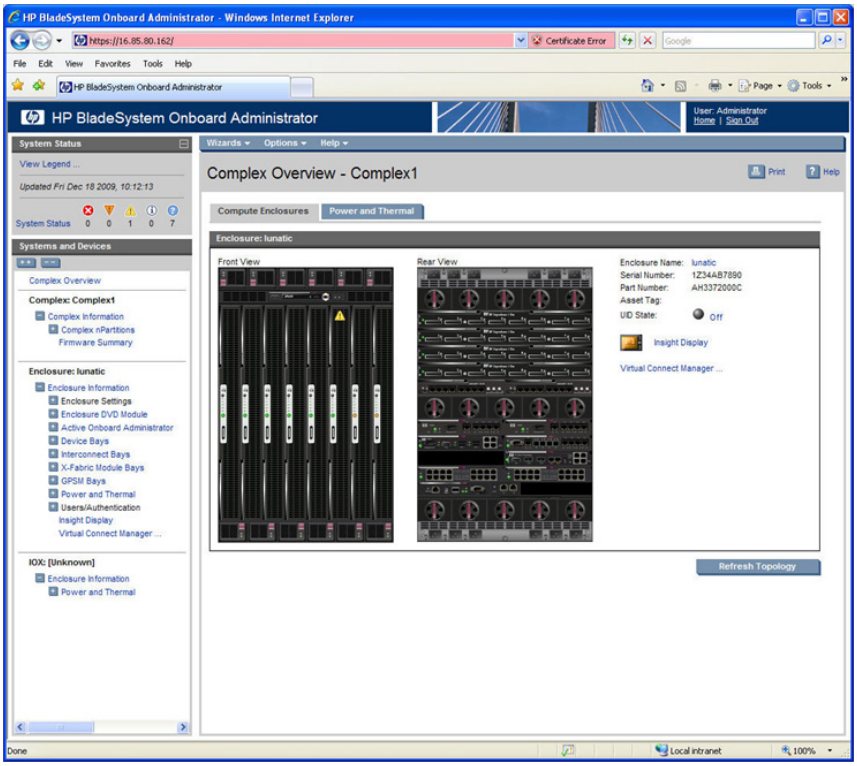
In this new architecture, every blade has a full iLO built into it. That means every partition automatically has an iLO in it, no matter how partitions are built. The entire system and all the iLOs are managed through the Integrity Superdome 2 Onboard Administrator (OA) with no need to drill down into individual iLOs. The iLOs are shown in gray in this figure 6 to indicate that they are working silently.

The entire set of server health and configuration is managed through the Integrity Superdome 2 OA, removing the need for an SMS type of external management station. The Integrity Superdome 2 OA contains a full Superdome toolbox including partition tools that used to run in an OS partition. This new Integrity Superdome toolbox is always available regardless of the state of the system (up, down, rebooting, or OS not even yet loaded). These features are built into the Integrity Superdome 2 OA and sx3000 chipset and are only available in Superdome 2.

Manageability

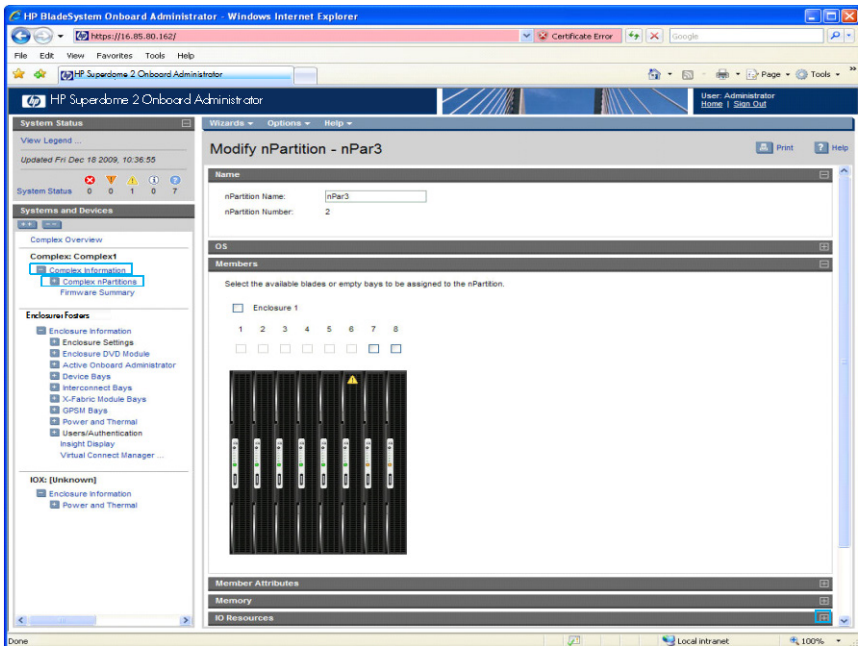
As part of the HP Mission-Critical Converged Infrastructure, a key benefit of leveraging the BladeSystem elements into Integrity Superdome 2 is simplified management across the wide range of HP servers. In addition to the wide range of resources offered through HP System Insight Manager (HP SIM), additional manageability features are available through the Superdome 2 OA. This allows Superdome and c-Class servers to share a common management "look and feel" including health alerts and monitoring, automated power and cooling management, role-based security, and other features. Compare this to other solutions that require different management user interfaces that make up a patchwork of tools, each using a different interface and process flow for completing tasks.

Figure 8. HP BladeSystem Onboard Administrator



Integrity Superdome 2 also has simplified partition setup and configuration. Unlike legacy systems where the nPar tools run on an external management station or from within a booted nPar, in Superdome 2 the nPar and vPar management tools reside within the firmware on the Superdome 2 OA. This means users can log into the Superdome 2 OA through a Web browser and run the partition management GUI and CLI to create nPars and or run CLI to create vPars.

Figure 9. HP Integrity Superdome 2 Onboard Administrator



HP Integrity Superdome 2 door display

Integrity Superdome 2 has the same Insight Display as c-Class systems to help manage and monitor the health of individual enclosures. However, Superdome 2–16s and -32s systems have an additional LCD display on the outside rack door as well. This enables customers to view the complex name, overall complex health, power, and ambient air temperature in a single glance without opening the rack door. In addition, user ID (UID) functionality enables easy location of a particular system within a crowded computer floor.

Figure 10. Superdome 2 LCD door display

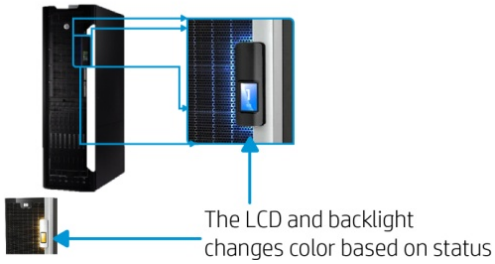


Figure 11. Superdome 2 LCD door display screen

Flexible virtualization and partitioning

Enterprise computing is more than providing massive computing power. It is providing a high quality of service and leadership RAS capabilities in a form that can be easily applied to the existing software architecture of the enterprise. The performance and RAS needs of the individual enterprise applications must be met while maintaining the customer's service level agreements. Optimize your resource pools with virtualized features to satisfy your dynamic enterprise class workloads.

Integrity Superdome 2 Servers provide a great deal of flexibility in how systems are managed. They support multiple methods of virtualization, multiple types of processors, and multiple operating environments.

These servers based on the sx3000 chipset support a wide array of virtualization options. Each virtualization strategy splits up the resources in the server (CPUs, memory, I/O) into independent servers that can each run an OS instance. The following virtualization strategies are available:

- nPars
- vPars
- HP Integrity VM
- HP-UX Containers

nPars are electrically isolated hard partitions with security provided in the hardware. Secure firmware configures the sx3000 fabric to isolate resources in an nPar from the rest of the system. This creates a hardware firewall to prevent other OS instances from disrupting that partition. The firewall also minimizes the chance that a single failure (in hardware or software) can take down multiple partitions. This is a required feature in order to perform hardware maintenance on one partition while the other partitions continue to operate. The size of an nPar can range from a single blade to the entire system.

vPars and Integrity VM, run within an nPar. vPars offer partitioning granularity down to the CPU core, memory slice level, with near-native performance. Integrity VM offers sub-CPU-core granularity and offers good security from un-trusted guest operating environments, supporting both HP-UX 11i v2 and v3 guests.

Table 4. sx3000 partitioning capabilities

	sx3000 chipset
nPars (Hard partitions)	• Offers electrically isolated hard partitions • Reduces failures that can crash multiple partitions • Provides improved hardware firewalls • Delivers granularity improved to 2s • Enables I/O card failure in one partition, which does not impact other partitions • Enables no hypervisor fails
vPars (Virtual partitioning)	• Offers CPU core-level partitioning granularity • Delivers minimal overhead to coordinate among the guest operating environments • Runs within a hard partition • Is offered on HP-UX 11i
Integrity VM (Sub CPU partitioning)	• Offers sub-CPU-core partitioning granularity • Offers good security • Is run within a hard partition • Is offered on HP-UX 11i • Provides low overhead I/O through direct I/O • Offers shared I/O • Enables dynamic resource allocation

Performance

A high-level summary of Integrity Superdome 2 Server level performance enhancement as shown in table 4. The sx3000 chipset contributes to these system-level performance in a number of ways. Support with the Intel Itanium processor 9300 and 9500 series is where we start. However, without an enhanced system architecture, those processors would not deliver their full performance potential. The sx3000 can keep the Intel Itanium processor 9300 and 9500 series fully utilized. The sx3000 chipset delivers increased bandwidth on every interface, allowing the processors to do the maximum amount of work. These bandwidths are shown in table 5.

Table 5. sx3000 partitioning capabilities

Feature	sx3000 chipset peak (sustained)
Memory subsystem/blade	90.0 (48.0) GB/s
Fabric links/blade	78.0 (36.0) GB/s
I/O/blade	26.0 (11.8 duplex, 8.4 outbound, and 6.2 inbound) GB/s
I/O/expander	50.0 (22.8 duplex, 16.4 outbound, and 12.0 inbound) GB/s
I/O single slot	10.0 GB/s (8.0 GB/s)

This increased bandwidth allows communication-intensive or large data-set applications to run faster, whether it is transferring data to/from memory, communicating among the processors running the application, or transferring data to/from an I/O device.

In some systems, higher bandwidth means longer latency; however, the sx3000 has decreased system latencies. This means processors and I/O devices have faster access to memory, improving application performance. The sx3000 chipset provides a 128 MB 4th level cache on each blade, that caches data from memory on remote blades. This significantly reduces the latency to shared global data such as program code and page tables, as well as to unshared data that has not been localized to the blade via application tuning. The L4 cache particularly helps "real-world" applications where a significant tuning effort is not practical.

The Intel Itanium processor 9300 and 9500 series and the sx3000 chipset also add the ability to transfer unmodified data directly between processors, rather than going from one processor to memory, and then to the destination processor. When a processor requests data that a second processor has recently created, modified, or read and is still in its cache, that data can be delivered through a cache-to-cache (C2C) transfer to the requesting processor with a maximum of two Crossbar hops, no matter where those two processors are in the system. In addition, the Intel Itanium processor 9300 series CPU cores each have a dedicated 6 MB L3 cache, carrying forward the Itanium 2's use of a dedicated L3 cache per core to reduce the chances of the workload on one core interfering with that on another core. The Intel Itanium processor 9500 series CPU cores have the same dedicated L1 and L2 cache per CPU core, but have a large, shared L3 CPU cache that provides significantly lower C2C latency for data shared among multiple threads that reside on the same socket. The sx3000 system latencies are shown in table 6.

Table 6. sx3000 system latencies

	Intel Itanium processor 9500 series		Intel Itanium processor 9300 series	
Memory latency (load to use):	**L4 hit**	**L4 miss**	**L4 hit**	**L4 miss**
Local socket	N/A	145 ns	N/A	176 ns
Remote socket, same blade	N/A	202 ns	N/A	245 ns
Remote blade, same chassis	147 ns	395 ns	158 ns	426 ns
Remote blade, remote chassis	147 ns	453 ns	158 ns	486 ns
Cache-to-cache latency (farthest sockets):				
Local socket	45 ns (8 cores)		115 ns (4 cores)	
Remote socket, same blade	160 ns (16 cores)		186 ns (8 cores)	
Remote blade, same chassis	392 ns (128 cores)		408 ns (64 cores)	
Remote blade, remote chassis	451 ns (256 cores)		467 ns (128 cores)	

The sx3000 provides up to six full-speed PCIe x8 Gen2 slots per I/O adapter chip, to fully support the bandwidth requirements of next-generation PCIe Gen2 cards.

Scalability

HP Integrity Superdome 2 Servers using the sx3000 chipset are inherently scalable systems. These systems are designed to support up to 32 processor sockets, all accessible to each other through the low-latency, high-bandwidth system fabric. Each of these processor sockets support the Intel Itanium processor 9300 and 9500 series. Investment protection is provided by setting the foundation for the next generation of Itanium processors. These systems support up to a maximum of 512 GB of memory per blade, using 16 GB DRAMs. In addition to up to 96 I/O cards supported in eight IOXs, there is onboard I/O capability to provide more flexibility in meeting your needs.

Green IT

Green IT in the Integrity Superdome 2 is a combination of features at levels from the most basic of components to the total data center solution.

Component level

Two versions of power supplies are available. The standard power supplies are rated Gold by 80 Plus with an efficiency that exceeds 88 percent at all loads and exceeds 92 percent at 50 percent load.[2] An optional power supply that is Platinum rated (exceeds 90 percent at idle, 91 percent at full and 94 percent at 50 percent) is available. With the Dynamic Power Saver mode enabled, power supplies are continuously monitored and kept as close to their optimum efficiency load level as possible by enabling and disabling supplies as needed.

The Intel Itanium processor 9500 series has a new layout with twice the number of cores than the Intel Itanium processor 9300 series, yet uses less power—more than 80 watts less at full load. The Intel Itanium processor 9300 and 9500 series and Integrity Superdome 2 also utilize "Green Active" and "Green Idle" which are OS-supported power reductions. Green Active is a user selectable operating system function which utilizes the processor's multiple power states (called P-states which are changes in the processor's voltage and frequency) to match the processing power and performance to the current task. Three P-state steps are available (P0-P2) with the largest reduction in power occurring in P2. Green Idle is another OS supported function which reduces the power used by the processor when specific idle conditions are met. This function places the processors in the C1E state which both invokes the P2 state and stops issuing instructions. The Intel Itanium processor 9500 series saves 75 watts over the Intel Itanium processor 9300 series in idle state even before C1E is enabled and C1E saves an additional 40 watts.

Integrity Superdome 2 leverages many innovative features from the HP BladeSystem c-Class enclosure, one of which is the fans. These fans are very capable when necessary and Integrity Superdome 2 uses sophisticated speed control to save energy in normal operation.

Component-level savings are of particular interest due to the additional savings realized at each stage in the delivery of electrical energy. Studies by Emerson Network Power (conducted in 2007) showed that for every watt saved at the component level, approximately 2.8 watts are saved at the utility connection to the data center. Using this metric, a 16-socket system saves approximately $1000/year USD over the same configuration using the last generation processor.

System level

Enhanced ducting and automatic blocking of the airflow for partially loaded systems insures minimum overall airflow at any temperature.

Control of the fans themselves is a significant improvement over legacy Superdome. Not only are the fans controlled to compensate for environmental changes, the algorithm used is a non-linear mapping of fan speed to temperature such that the airflow more closely matches the demand of the system. When the temperature in the data center is in the ASHRAE recommended range, Integrity Superdome 2 reduces the speed of the fans saving over 1200 watts of power from their highest speed setting. In the unlikely event that the temperature of the data center increases, Integrity Superdome 2 uses the capacity of the fans to keep running well above the ASHRAE recommended maximum, typically allowing critical operations to continue uninterrupted.

Integrity Superdome 2 continues the cell-based architecture of its predecessor, which in turn enables independent partition power up and down. In conjunction with HP-UX 11i Workload Management (gWLM/WLM), it is possible to maintain a higher average load per processor than the common practice of provisioning for the peak load on every system, thereby saving significant power when partitions are powered down. The use of HP Instant Capacity also depends on this feature set and provides the customer the ability to have unpowered resources installed rather than consuming power.

[2] 80 Plus is an electric utility-funded incentive program to integrate more energy-efficient power supplies into desktop computers and servers. 5 ratings for server power supplies are available: Base, Bronze, Silver, Gold, and Platinum.

Data center level

Aligning with data center best practices, Integrity Superdome 2 is cooled via front to rear airflow and fits in a standard 19-inch rack. This allows more placement flexibility; not only in traditional data center hot aisle/cold aisle layouts but in support of more cutting-edge solutions such as containment, rack-level cooling, and the HP Performance Optimized Datacenter (POD). In addition, the fan control as previously described results in a much higher exhaust air temperature under normal conditions. This in turn allows a much more capable and efficient operation point to be attained by the entire set of cooling equipment in the data center. This is independent of the cooling architecture (under-floor, rack, row, and so on) and results in the ability to cool more load with existing cooling infrastructure.

Benefits of HP Integrity Superdome 2 Servers for mission-critical application environments

The Integrity Superdome 2 Servers based on the sx3000 chipset are ideally suited to a wide variety of applications and market segments. In this section, we will describe several key applications in different market segments, and show how these systems meet the specific needs of each of these applications.

Online transaction processing

Online transaction processing (OLTP) applications handle the day-to-day activities of running a business, such as receiving and processing orders, tracking inventory in warehouses, and maintaining customer information. OLTP applications generally consist of a database server running a DBMS package such as Oracle or SQL Server, along with several application servers or clients. The database maintains its information in tables, which reside on disk and are typically cached in main memory to improve performance. These tables are updated frequently, which requires frequent communication and synchronization between the processors.

The Integrity Superdome 2 Servers handle the demands of OLTP workloads with ease. The Integrity Superdome 2 Servers scale to handle even the most demanding OLTP applications. These systems support up to 8 TB of main memory 512 GB per blade with 16 GB DIMMs, enabling a large portion of the database tables to be cached in memory for improved performance. The sx3000 based systems also includes 128 MB of L4 cache on each blade. This cache serves to reduce the latency of access to remote data structures stored on other blades. Superdome 2 offers improved multi-threaded performance and the ability to scale to 2X the number of cores as the sx2000.

Your OLTP systems are often mission-critical, with outages resulting in loss of revenue and your customer goodwill. The sx3000 systems were designed with multiple self-healing technologies to reduce both scheduled and unscheduled downtime. Double device data correction removes DRAM failures as a source of system failure, plus allows replacement of the DIMM to be deferred until the next regularly scheduled maintenance period. In the midplane and I/O fabrics, the sx3000 chipset incorporates end-to-end retry and enables the ability to route around failed links. The upper and lower midplanes in Integrity Superdome 2 are passive. This virtually removes the need to ever repair the midplanes. ECC on all I/O busses enables correction of intermittent failures. These and other technologies make the Integrity Superdome 2 Servers an ideal choice for mission-critical OLTP databases.

Business intelligence/Decision support

Business intelligence (BI) applications analyze large amounts of historical data to help answer business planning questions. BI applications typically stream large data files from disk into memory, while simultaneously using multiple processors to analyze the data. These applications are often highly parallel, with performance scaling well with the number of processors.

Integrity Superdome 2 is ideal for running BI applications. Users may use from 2 to 32 high-performance Intel Itanium processor 9300 or 9500 series to tackle BI problems that range from modest to mammoth. All Integrity Superdome 2 Servers provide 6.2 GB/s of sustainable inbound I/O bandwidth per cell (40 percent more than the sx2000), which removes I/O bandwidth bottlenecks and enables the processors to run at maximum performance. The sx3000 offers four times the I/O bandwidth per I/O slot as the sx2000. Each IOX enclosure offers an additional 2X the I/O bandwidth of the I/O built onto the blade. A single blade can be configured to operate with multiple IOXs. The Integrity Superdome 2 is the clear leader in I/O flexibility.

Server consolidation

Server consolidation is a practice that can be applied to market segments like those described above to reduce IT costs. Server consolidation enables a user to replace a large number of discrete systems with a smaller number of partitioned servers. Benefits of this include reduced hardware costs, reduced application and OS licensing fees, fewer OS instances to manage, and reduced facilities costs. Partitionable servers make it easier and faster to implement a new application as compared to a stand-alone server. This is because the provisioning of a new stand-alone server usually requires making a purchase, or at least allocating a resource, whereas adding a new partition to an existing system is very fast.

The Integrity Superdome 2 Servers are the ideal target for server consolidation initiatives. Each blade, which is the building block of nPars, is a fully isolated computing unit. As compared to competitors' systems, where cache coherency traffic continues to flow between partitions thereby creating potential single points of failure and unwanted performance interactions, Integrity Superdome 2 Servers keep nPar local coherency traffic entirely within the blades of the nPar. In addition to significantly reducing the risk of a hardware failure affecting more than one nPar, this increases the performance of a partitioned system. To handle a user's changing computing needs, blades and IOXs for new nPars can be easily added to a running system.

Virtualization

Integrity Superdome 2 Servers also allow users to further virtualize the resources allocated to an nPar. In Integrity Superdome 2 Servers, vPars allow multiple instances of HP-UX to execute in parallel without the overhead of hypervisors. In addition Integrity VM can be used to provide virtual partitioning of the server allowing multiple guests to run on a core.

As more servers are consolidated onto single systems, the high-availability features of Integrity Superdome 2 reduce the likelihood of a fault causing an outage that would affect more than one virtualized guest.

Conclusion

HP Integrity Superdome 2 Server has continued to build upon the success of previous-generation Superdome products, boosting RAS and performance. Now, based on HP Mission-Critical Converged Infrastructure, Integrity Superdome 2 Server shares common components, power, cooling and management with HP BladeSystem. Our Integrity Superdome 2 Servers running business-critical and research applications such as online transaction processing business intelligence, or systems being utilized to consolidate the IT environment benefit from the advancements in the new sx3000 chipset and the next generation of Intel Itanium processors.

Glossary

Term	Definition
Agent	Chip in sx3000 chipset that provides interface to processors and HP proprietary links to I/O and other Integrity Superdome 2 blades for scaling.
CMCI	Corrected Machine Check Interrupt. Allows the CPU to trigger interrupts on corrected machine check events, which allows faster reaction to them than traditional polling timers.
CLI	Command-line interface. Means of interaction with a computer program where the user issues commands to a program in the form of successive lines of command lines.
Crossbar Fabric	HP proprietary fabric used to interconnect Agents to Crossbars, Agents to IOH, and Crossbars to IOH.
DRAMs/DIMMs	Dynamic RAM/Dual In-line Memory Module. The memory in a computer system uses DRAM chips to store the data. DIMMs hold multiple DRAM chips on a memory "card."
DDDC	Double Device Data Correction. Advanced technology that uses ECC code words to restore chip-spare even after a DRAM has failed. This technology keeps the memory system fully operational even if two DRAM chips fail.
ECC	Error Correcting Code. ECC consists of extra bits that travel with data to protect the data from errors. ECC can detect when the data has been corrupted. ECC used in today's computer systems will typically correct single-bit errors and detect double-bit errors.
GPSM	Global Partition Services Module. Integrity Superdome 2 utility board that contains global clock source and resources to monitor fans and power supplies in the upper half of the compute enclosure. There are two GPSMs (for redundancy) per Superdome 2 compute enclosure.
IOH	I/O Hub
IOX	I/O Expansion Enclosure
Load-to-use latency	The latency seen by the CPU from when it requests data from memory to when it receives the data. The faster the data can be supplied, the more work the CPU can do.
LOM	LAN On Motherboard. Network connections embedded directly on the motherboard or blade.
MP-SPOF	Multipartition single point of failure
MTBF	Mean time between failure
OA	Onboard Administrator. Same manageability module used in c7000. There are two OAs (for redundancy) per Superdome 2 compute enclosure.
OLARD	Online Addition, Replacement, and Deletion. This term refers to swapping, adding, or removing components such as blades or PCI cards while the partition is running.
PCI/PCI-X/ PCI Express	Industry-standard buses/links for I/O devices. PCI-X mode 2 is double the data rate from PCI-X 133 MHz, and it supports approximately 2 GB/s. PCI Express is based on links instead of buses, and will start replacing PCI/PCI-X in the future.
QPI	QuickPath Interconnect. Point-to-point processor interconnect developed by Intel. Replaces the Front Side Bus (FSB) used in previous Itanium processors.
SMB	Scalable Memory Buffer. Intel chip that acts as a buffer and router between the high-speed FBD2 channel on the memory controllers and the DDR3 busses at the RDIMMs.
SMI	Scalable Memory Interconnect. Interface between processor and memory subsystem.
Superdome 2 CB900s i2	Official name for Integrity Superdome 2 blade with Intel Itanium processor 9300 series.
Superdome 2 CB900s i4	Official name for Integrity Superdome 2 blade with Intel Itanium processor 9500 series
sx3000	Chipset used in design of Integrity Superdome 2 for use with the latest Intel Itanium processors.
Crossbar	Fabric chip in sx3000 chipset. Resides on XFM. Provides fabric between multiple Superdome 2 blades and to external I/O Expansion Enclosure (IOX).
XFM	Crossbar Fabric Module. Assembly that contains the Crossbar chip.

Resources

HP Integrity Servers: hp.com/go/integrity

HP Partitioning Continuum: hp.com/go/partitions

Green IT: http://h18004.www1.hp.com/products/blades/thermal-logic/index.html

HP Matrix Operating Environment: hp.com/go/insightdynamics/integrity

HP Mission-Critical Converged Infrastructure: hp.com/go/integritynow

Learn more at
hp.com/go/superdome2

Sign up for updates
hp.com/go/getupdated

Share with colleagues

Rate this document

© Copyright 2010, 2012–2013 Hewlett-Packard Development Company, L.P. The information contained herein is subject to change without notice. The only warranties for HP products and services are set forth in the express warranty statements accompanying such products and services. Nothing herein should be construed as constituting an additional warranty. HP shall not be liable for technical or editorial errors or omissions contained herein.

Intel, Itanium, and Intel Itanium are trademarks of Intel Corporation in the U.S. and other countries.
Oracle is a registered trademark of Oracle Corporation and/or its affiliates.

4AA1-7762ENW, July 2013, Rev. 4