HPE Superdome Flex server architecture and RAS

Unparalleled scale, flexibility, and reliability for in-memory workloads of any size
## Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction</td>
<td>3</td>
</tr>
<tr>
<td>Reliability, availability, serviceability</td>
<td>4</td>
</tr>
<tr>
<td>Manageability</td>
<td>4</td>
</tr>
<tr>
<td>System architecture</td>
<td>5</td>
</tr>
<tr>
<td>HPE Superdome Flex chassis—the modular building block</td>
<td>6</td>
</tr>
<tr>
<td>HPE Superdome Flex chassis management</td>
<td>11</td>
</tr>
<tr>
<td>HPE Superdome Flex Rack Management Controller</td>
<td>11</td>
</tr>
<tr>
<td>RAS</td>
<td>12</td>
</tr>
<tr>
<td>Fault management strategy</td>
<td>12</td>
</tr>
<tr>
<td>Firmware First</td>
<td>12</td>
</tr>
<tr>
<td>RAS differentiators</td>
<td>13</td>
</tr>
<tr>
<td>Management</td>
<td>18</td>
</tr>
<tr>
<td>Built-in management capabilities</td>
<td>18</td>
</tr>
<tr>
<td>Additional management resources</td>
<td>20</td>
</tr>
<tr>
<td>Conclusion</td>
<td>22</td>
</tr>
<tr>
<td>Resources</td>
<td>22</td>
</tr>
</tbody>
</table>
Introduction

HPE Superdome Flex server delivers a unique and unprecedented combination of modularity, flexibility, scale, and reliability to deliver in-memory computing for businesses of any size.

HPE Superdome Flex is a mission-critical platform flexible and powerful enough to handle the massive and growing amount of business data. Its in-memory design and unparalleled scale provide the ability to analyze data from the digital core to the Intelligent Edge in real time. And because the infrastructure is modular and cloud-ready, it’s the right fit for any business of any size.

Keep pace with your evolving in-memory computing demands through a unique modular design: start small and grow as needed, flexibly scaling up or out. With superior RAS and end-to-end security, HPE Superdome Flex safeguards vital workloads. Our broad services portfolio and partner ecosystem, and mission-critical expertise help turn data into insight, so you can turn insight into action, and action into success—with peace of mind that your business will be always on.

- Proven RAS capabilities not available on other standard platforms
- Best-in-class predictive fault-handling Error Analysis Engine, predicts hardware faults and initiates self-repair without operator assistance
- Firmware First approach to log analysis ensures error containment at the firmware level, including memory errors, before any interruption can occur at the OS layer
- Mission-critical resiliency from end-to-end implementation of processor RAS features, to redundancy of key system components to advanced system software
- Designed from the chipset up to provide up to five nines (99.999%) single-system availability

HPE Superdome Flex is a flexible in-memory computing solution with a unique modular design that scales easily and economically for businesses of any size. It blends trusted HPE Superdome reliability with a standard x86-based design. This document describes the HPE Superdome Flex architecture and explains its significant benefits in performance, manageability, and reliability for your mission-critical environment.

As shown in Figure 1, HPE Superdome Flex is designed around a 5U modular building block that can scale from 4-socket up to 32-sockets in 4-socket increments through a cabled crossbar interconnection fabric called the HPE Superdome Flex Grid. With this modular design flexibility, HPE Superdome Flex customers don’t pay for infrastructure costs (that is, power capacity, Flex Grid cables, and more) above and beyond the system scale they require, but can easily add capacity in the future as their application needs change. Service access for the system is limited to the front or rear of the enclosure by way of forward and rearward sliding rails, and components like Bulk Power Supplies (BPS), fan assemblies, boot drives can all be serviced while the mission-critical workloads and operating environments continue to run. Even HPE Superdome Flex Grid provides adaptive routing features designed specifically to enhance performance by routing traffic through the optimal latency path available, and provide superior uptime by automatically routing traffic around failed components.
**Reliability, availability, serviceability**

Reliability, availability, and serviceability, collectively known as “RAS,” are top reasons customers deploy mission-critical workloads on HPE Superdome servers as the system provide system capabilities that keep systems up and running despite component failures or recovering quickly and automatically from failures so those mission-critical workloads can resume quickly without waiting for a service person to show up and effect repairs. Robust RAS has always been a designed-in philosophy for HPE Superdome servers, and HPE Superdome Flex builds on that strategy with such RAS strengths as:

- A selective and extended implementation of Intel® Xeon® Scalable RAS capabilities through custom reference code modifications
  - For example, HPE Superdome Flex implementation of Adaptive DDDC utilizes more error correction regions and has a finer granularity than provided in Intel® standard reference code. This results in better annual service and replacement rates for memory than other vendors’ platforms can provide.

- Cabled HPE Superdome Flex Grid with adaptive routing:
  - Automatically chooses the optimum latency path through the crossbar fabric for best performance and even detects and routes traffic around failures without requiring a reboot.
  - HPE nPar support (coming soon) provides complete workload isolation and independent serviceability if cabled properly.

- Rack Management Controller (RMC) or embedded RMC (eRMC), for 8-socket or smaller systems, provides the server management command-line interface and includes an HPE Superdome Flex Error Analysis Engine to provide error handling/correction, self-healing, and system health monitoring.

- Hot-swappable power and cooling components.

A detailed discussion of RAS is provided in the RAS section of this document.

**Manageability**

HPE Superdome Flex has an advanced manageability system that is always on, constantly monitoring and managing the system components, fabric, and infrastructure for mission-critical high availability. It combines the best of the HPE Integrity Superdome X RAS and management technology with the superior performance and modularity from HPE MC990 X platform to achieve unmatched scale, increased flexibility, extreme availability, and a simplified user experience. The design team has focused on management as a major development area in the system. It is designed to integrate seamlessly into both industry standard and the HPE suite of management products using industry-standard Redfish® APIs. In addition, an enhanced Error Analysis Engine similar to the one in HPE Superdome X provides improved self-diagnosis and automated recovery features. A comprehensive SSH command-line interface is available for administrators to get to the most detailed features in a repeatable (and scriptable) manner as well.

Major management components and resources include:

- HPE Superdome Flex Rack Management Controller (RMC)
- HPE OneView
- HPE Insight Remote Support
- Smart Update Manager

HPE Superdome Flex management subsystem integrates into data center solution management components using the industry-standard Redfish API. This enables standard solutions such as OpenStack®, as well as simple scripting to be used to get information and control the system.

A detailed discussion of HPE Superdome Flex management capabilities is provided in the Management section of this document.
System architecture

HPE Superdome Flex combines the best of Intel Xeon architecture and the new HPE Superdome Flex ASIC chipset to provide the most flexible x86-server solution available. The system is capable of delivering performance and scalability from 4-sockets to 32-sockets, in 4-socket increments, and is imbued with the necessary RAS features to operate in mission-critical environments where application availability is of paramount importance. Figure 2 shows the architecture of HPE Superdome Flex Base Chassis, which comprises the basic modular building block for the server solution. Each HPE Superdome Flex system consists of at least one Base Chassis plus as many as seven additional Expansion Chassis to provide the ability to scale up to 32-sockets or divide the system into hard partitions (HPE nPars) to isolate workloads and/or consolidate multiple workloads onto a single managed complex. HPE Superdome Flex ASIC provides the ability to connect Base and/or Expansion Chassis together via HPE Superdome Flex Grid cables, which provides many key benefits as listed here:

- **Adaptive routing:** Load balances the fabric and takes performance advantage by utilizing the available path with the most optimum latency and routing around any failed resources.

- **Configuration flexibility:** Cabling can be quickly reconfigured in the field to maximize performance whenever application needs change, whereas fixed backplane/midplane connections often leave ports unutilized in those smaller HPE nPar configurations.

- **Signal Integrity Advantages:** Cables have higher signal propagation velocity, which means less delay and lower latency, as well as, having better isolation between channels with lower crosstalk and a higher signal-to-noise ratio than can be achieved in copper traces across a printed circuit board.

- **Future-proof:** External copper cables today provide an infrastructure that can be more easily adapted to optical cabling technology in the future. When data transfer speeds exceed the capability of copper cables, and the cost to implement optical cabling becomes affordable, something as easy as plugging in a new version of HPE Superdome Flex ASIC assembly could convert HPE Flex Grid to optical fibers instead.

- **Lower entry costs:** With a modular chassis connect by cables, it is only necessary to buy the hardware, including power and cooling infrastructure, to fit the need, and not spend more securing expandability until that is needed.
HPE Superdome Flex chassis—the modular building block
Each HPE Superdome Flex chassis accommodates two Intel Xeon Scalable 81xx or 61xx series processors, which can provide as many as 28 cores per processor socket. With hyper-threading enabled, a 112-core HPE Superdome Flex chassis provides 224 logical processors. The four processors are connected together in a ring fashion via 10.4 GT/s Intel Ultra Path Interconnect (UPI) links. Each processor is also connected to one of two custom-designed HPE Superdome Flex ASICs via Intel UPI links to send remote targeted cache-coherent data traffic to external chassis in 8-socket or larger HPE nPars. HPE Superdome Flex architecture provides each processor with direct connections to 12 HPE DDR4 DIMM slots, which can accommodate 32 GB, 64 GB, or even larger DDR4 DIMMs, as well as, provide direct connections to PCIe 3.0 x8 and x16 stand-up card slots.
Also, the two HPE Superdome Flex ASICs are cross-connected to provide the system with the unique ability to support both 81xx Platinum and 61xx Gold series processors while still providing scalability from 4-sockets up to 32-sockets. The Intel Xeon Scalable processors offer the following advantages over previous generation processors:

- Up to 28 cores (56 logic processors with Intel Hyper-Threading Technology) per socket.
- New UPI links, running at 10.4 GT/s, provide increased bandwidth and performance over QPI, and include improved messaging efficiency and support multiple requests per packet.
- A rebalanced cache hierarchy to optimize and enlarge the mid-level cache, and still provide a last-level shared cache capacity up to 38.5 MB.
- Dual integrated and fully independent memory controllers break out into six fully independent (three per memory controller) that connect directly to 12 DDR DIMM slots running at a 2667 MT/s data transfer rate. With no scalable memory buffer in the path, latency and bandwidth are substantially improved.
- Fifty percent more PCIe 3.0 bandwidth available through 48 bifurcated (to easily provides x16 or x8 slot support) lanes, and new MCTP Scaling to provide native support for multisegment PCIe addressing.
- New Intel AVX-512 (S12 bit) instruction set, which provides up to 2X more peak floating point performance over AVX2, and makes this processor the best HPC engine available to date.

**HPE Superdome Flex ASIC**

HPE Superdome Flex ASIC is a new HPE custom-designed ASIC that interfaces directly to Intel Xeon Scalable processors and provides the system with the ability to connect up to eight HPE Superdome Flex chassis together in a cache-coherent fabric called HPE Superdome Flex Grid. HPE Flex ASIC provides two UPI links to interface with two Intel Xeon Scalable processors on one side, and provides sixteen Grid ports on the other side to interface with HPE Flex ASICs within the same chassis and to external chassis via HPE Flex Grid cables. The Flex ASIC maintains coherency by tracking cache line state and ownership across all the processor sockets within HPE nPar inside a directory cache built into the ASIC itself. This coherency scheme is a critical factor in the ability of HPE Superdome Flex to perform at near linear scaling from 4-sockets all the way up to 16-sockets and beyond, whereas, typical glueless architecture designs already see more limited performance scaling to as low as 4- to 8-sockets due to broadcast snooping.

HPE Superdome Flex ASIC provides the following features:

- Physical address support for up to 64 TB of main memory per cache coherence domain
- Very large directory cache to track cache line state and maintain coherency across all attached processor sockets
- Adaptive routing features provide fault resiliency and load balance the fabric
  - Performance gets optimized by choosing the optimal latency datapaths
  - Failed components get detected and routed around automatically
- Provides 16 Flex Grid ports, each capable of 13.3 GB/s data rates (80% faster than HPE MC990 X) for maximum Flex Grid bandwidth
  - More than 210 GB/s of bi-sectioned crossbar Grid bandwidth at 8-sockets with 4 Flex Grid link connections between each ASIC
  - More than 425 GB/s of bi-sectioned crossbar Grid bandwidth at 16-sockets with 2 Flex Grid link connections between each ASIC
  - More than 850 GB/s of bi-sectioned crossbar Grid bandwidth at 32-sockets with 1 Flex Grid link connection between each ASIC

**HPE Superdome Flex Grid—the crossbar fabric**

HPE Superdome Flex Grid is the fundamental differentiator for HPE scalable system design. This capability, enabled by the custom HPE Superdome Flex ASIC, is what allows HPE to offer a system with greater scalability than the 8-socket limit provided by the standard Intel reference architecture. The custom Flex ASIC and HPE Flex Grid cabling provides the low latency and high-bandwidth cache-coherent path between all the processors defined within each HPE nPar. And it is important to differentiate true hard partitioning (complete workload isolation) through the completely disablement of any cabled Flex Grid links between HPE nPars so no data traffic from one nPar can impact another. If HPE nPars are individually cabled for a specific HPE nPar size, as would be the case for a factory-integrated system, then each HPE nPar can be serviced independently and without interruption to other HPE nPars residing within the same complex.

By providing system scaling of up to 32-sockets, HPE Superdome Flex stands alone in providing the largest possible cache-coherent solution available today, thereby allowing customers to tackle workloads they couldn’t imagine doing before. An example of HPE Flex Grid interconnect scheme for a 32-socket HPE nPar is shown in Figure 3.
As Figure 3 shows, HPE Superdome Flex Grid links provide single-hop, direct interconnection between every ASIC to keep latency to a minimum and maximize performance. The cabled HPE Superdome Flex Grid is so flexible, that even above the 32-socket scale, where HPE Superdome Flex ASIC can no longer provide single-hop, direct interconnect, it is still conceivable to scale higher (64-sockets or even more) by allowing for multi-hop links. Such flexibility means that even special need customers may discover that a true cache-coherent system solution at a previously unimaginable scale can solve computing problems never before attempted.

**Memory subsystem**

Each HPE Superdome Flex chassis has 48 DDR4 DIMM slots that can accommodate 32 GB RDIMMs, 64 GB LRDIMMs, or even 128 GB TSV RDIMMs (future) for a maximum per-chassis capacity of 6 TB. This gives a fully scaled 32-socket HPE Superdome Flex a total memory capacity of 48 TB of main memory to support the most intensive in-memory applications. A diagram of the memory subsystem architecture is shown in Figure 4.
As illustrated in Figure 4, the memory subsystem includes the following features:

- Each Intel Xeon Scalable processor provides two fully independent integrated memory controllers (MCS), and the processors are connected together in a cache-coherent ring architecture providing three latency levels per chassis (local, direct attach, indirect attach).
- Each memory controller provides three fully independent memory channels.
- Each memory channel connects directly to two DDR4 DIMM slots.
- External chassis connections provided through (not shown).
Since these memory channels are fully independent, they can all run simultaneously at DRAM data transfer rates up to 2667 MT/s to provide each 4-socket HPE Superdome Flex chassis with >360 GB/s of local memory bandwidth (STREAM TRIAD), which scales linearly to 1.4 TB/s at the 16-socket maximum nPar size of HPE Superdome X. What’s more, with HPE Superdome Flex you are only at half capacity at 16-sockets, and you can count on that same linear memory bandwidth scaling all the way to 32-sockets where such incredible memory capacity and performance levels will be needed to keep as many as 896 cores of Intel Xeon Scalable processing power running the most demanding workloads.

I/O subsystem
As has long been a hallmark of HPE mission-critical server design, achieving breakthrough system performance means that maintaining balance between processing power, memory capacity/performance, crossbar interconnectivity, and system I/O capabilities is of paramount importance. Each HPE Superdome Flex chassis can be equipped with either a 16-slot or 12-slot I/O bulkhead to provide innumerable stand-up PCIe 3.0 card options and flexibility to maintain that vitally important system balance for any workload imaginable. The 16-slot I/O bulkhead provides nine low profile x8 and seven low profile x16 PCIe 3.0 card slots. The I/O bulkhead utilizes the available 48 PCIe lanes per processor to the maximum degree possible with as much as 110 GB/s per chassis of I/O bandwidth available. The 12-slot I/O bulkhead provides four full-height x8, four full-height x16, three low profile x8, and one low profile x16 PCIe 3.0 card slots. The I/O bulkhead provides sufficient extra power capacity to support ~300W full-height, double-width GPU cards for those high-demand HPC/HPTC workloads or for machine learning as the ultimate neural net training engine. With either I/O bulkhead selection, the I/O design provides direct connections between the processors and the card slots without need for bus repeaters or retimers that could add latency or reduce bandwidth. Hence, HPE Superdome Flex customers can rest assured they will get the best per card performance possible. Figure 5 shows the layouts for both the 16-slot and 12-slot I/O bulkhead options.

![16-slot I/O bulkhead layout](image1)

![12-slot I/O bulkhead layout](image2)

**Figure 5.** HPE Superdome Flex I/O bulkhead options

In addition to having the choice to pick the best I/O bulkhead option to fit any application’s needs, the I/O subsystem design of HPE Superdome Flex system also provides Base I/O functionality. It also provides built-in, hot-swappable boot drives to ease deployment and minimize costs (Of course, a customer may also choose to boot from SAN, as is often recommend for SAP HANA®, or may even choose network boot options via PXE).
HPE Superdome Flex Base I/O, as illustrated in Figure 6, provides the Base Chassis of HPE Superdome Flex complex the following features:

- Built-in boot storage provided via four hot-swappable 2.5” HDD/SSDs
- One DVD-ROM/RW drive for OS/driver installation
- Two general-purpose 10GbE LAN ports
- Two general-purpose 1GbE LAN ports
- Four USB 3.0 ports
- OS Console serial, VGA, and KVM LAN ports
- One Rack Management Controller (RMC) LAN port for connection to the management LAN
- One Board Management Controller (BMC) console port for initial system setup, password recovery, and debug

Each HPE Superdome Flex system will have at least one Base or Partitionable Chassis to act as HPE nPar monarch, and may contain up to seven Expansion Chassis depending on customer configuration. If the system is to be carved up into multiple HPE nPars, then each HPE nPar will need at least one Base or Partitionable Chassis. However, selecting two to provide for fail-over purposes will increase overall system availability by allowing for automatic reconfiguration and reboot in the unlikely event of a chassis deconfiguration event.

HPE Superdome Flex chassis management
Each HPE Superdome Flex chassis includes a Platform Controller Hub (PCH) chip and baseboard management controller (BMC) to provide all the features required to tie the chassis into the HPE Superdome Flex server management's primary component, the Rack Management Controller (RMC). The PCH chip provides initial reset functionality and real-time clock functionality. The BMC provides the bulk of hard partitioning capabilities and error handling. The BMC hardware and firmware also provide remote server management capabilities over an Ethernet management network. The BMC of each chassis interface directly with the RMC to provide the processing power needed to manage a large and flexible system like HPE Superdome Flex.

HPE Superdome Flex Rack Management Controller
HPE Superdome Flex is managed through HPE Superdome Flex Rack Management Controller (RMC). HPE Superdome Flex RMC is based on HPE MC990 X RMC, with expanded functionality from HPE Superdome X product integrated into it. It provides the ability to manage the partitioning of the system and component inventory and health. While each chassis has its own Baseboard Management Controller (BMC), HPE Superdome Flex RMC collectively manages all chassis and the system fabric with the aid of each chassis BMC, avoiding the need to drill down when managing individual nodes.

HPE Superdome Flex RMC’s built-in Analysis Engine is constantly analyzing all hardware to detect faults, predict failures, and initiate automatic recovery actions as well as notifications to administrators and HPE Insight Remote Support and HPE OneView. HPE Insight Remote Support connects to HPE Superdome Flex RMC to monitor problems and troubleshoots. It works with the Analysis Engine and can connect to HPE back end for automatic notification to HPE support of any problem with the system.
RAS
HPE Superdome Flex servers offer RAS features in key hardware subsystems—processor, memory, and I/O—and provide the ideal foundation for mission-critical Linux®, and VMware® operating environments. HPE mission-critical Superdome Flex running the x86 operating environment reflects the growing emphasis on availability by ensuring that your business is always on and providing availability through a layered approach that offers application, file system, and operating system protection. HPE mission-critical Superdome Flex infrastructure and the x86 operating environment provide a comprehensive RAS strategy that covers all layers—from application to hardware.

Fault management strategy
HPE Superdome Flex servers fully realize HPE’s design strategy for systems handling mission-critical workloads, which is to implement, when applicable, a four-stage RAS strategy of detection, logging, analyzing, and repair (Figure 7).

![Figure 7. HPE hardware RAS strategy](image)

This strategy keeps customer workloads up and their data available even in the presence of faults. In the rare event of an unrecoverable fault, the strategy still provides detection and containment to protect corrupt data from reaching the network and permanent storage.

When faults do occur that require support action, accurate diagnosis of the fault is critical to determine what is wrong and how to fix it right the first time. The following are some of the design keys for diagnostic abilities that are built in to every HPE Superdome Flex server:

- Minimize time to repair
- Capture enough data to diagnose failures the first time
- Allow system to run after failure for complete error logging
- Ability to diagnose all system components (software, firmware, and hardware) via complete error logging
- Field-replaceable-unit (FRU)-level granularity for repair
- Component-level granularity for self-healing

Firmware First
Part of HPE Superdome Flex’s comprehensive strategy for fault management includes Firmware First problem diagnosis. With Firmware First, firmware with detailed knowledge of HPE Superdome Flex system is first on the scene of problems to quickly and accurately determine what’s wrong and how to fix it. Intel Xeon Scalable processors Enhanced Machine Check Architecture Gen 2 (EMCA2) allows firmware a first look at error logs so that firmware can diagnose problems and take appropriate actions for the platform before OS and higher-level software involvement. Firmware First covers correctable errors, uncorrectable errors, and gives firmware the ability to collect error data and diagnose faults even when the system processors have limited functionality. Firmware First enables many platform specific actions for faults including predictive fault analysis for system memory, CPU, I/O, and interconnect.
RAS differentiators

While features such as hot-swap N+1 power supplies and single-/multi-bit memory error correction have become common in the industry, a number of RAS differentiators set HPE Superdome Flex servers apart from other industry-standard servers. HPE Superdome Flex servers offer several types of RAS differentiators:

- Self-healing capabilities
- Processor RAS
- Memory RAS
- Platform RAS
- Application RAS
- OS RAS

Self-healing

When faults do occur, HPE Superdome Flex provides several mechanisms to react so that unplanned downtime is avoided. Primary means of downtime avoidance include disabling failed or failing components during boot and attempting recovery on failed or failing components during run time. Taking failed or failing hardware offline allows the system to remain running with healthy hardware until the system can be serviced. Such self-healing capabilities avoid unplanned system downtime.

Deconfiguration of failed or failing components

HPE Superdome Flex provides the ability to deconfigure components so that any single hardware fault can be tolerated.

- Memory DIMM, I/O slot and CPU core deconfiguration: Reactive and predictive fault analysis allows for deconfiguration of failed or failing memory DIMMs, I/O slots and CPU cores so that the system can remain available with only healthy memory DIMMs and CPU cores in use.

Run-time deactivation of components before failure

Faults in many areas of HPE Superdome Flex servers result in run-time deactivation of resources to avoid continued usage of failing components. This level of self-healing provides zero system downtime and allows for repair actions at the next planned downtime event. System interconnects and the memory subsystem provide self-healing capabilities with deactivation of failing resources when needed:

- HPE Superdome Flex Grid link self-healing with link width reduction, online port deactivation, and alternate routing for fabric connections
- Adaptive Double Device Data Correction (ADDDC) to tolerate two failed devices on a DIMM
- MEMlog™ memory capability for high-performance applications

Processor RAS

HPE Superdome Flex servers use the Intel Xeon Scalable processors. These processors include extensive capabilities for detecting, correcting, and reporting hard and soft errors. Since these RAS capabilities require firmware support from the platform, they are often not supported in other industry-standard servers. HPE Superdome Flex implements RAS functionality provided in Xeon Scalable series processors including

- Corrupt data containment
- PCIe Live Error Recovery containment
- Poison error containment
- Processor interconnect fault resiliency
- Advanced MCA recovery
- Socket disable and deconfiguration
Corrupt data containment
HPE Superdome Flex servers with Intel Xeon Scalable processors enable corrupt data containment mode that provides the detection and possible recovery of uncorrectable errors. When corrupt data containment mode is enabled, the producer of the uncorrected data will not signal a Machine Check Exception. Instead, the corrupted data is flagged with an error containment bit. Once the consumer of the data receives the data with the error containment bit set, the error is signaled and handled by firmware and the operating system. Several recovery flows are possible including uncorrected no action (UCNA), software recovery action optional (SRAO), and software recovery action required (SRAR). The mission-critical HPE Superdome Flex infrastructure and the x86 operating environment support all of these. Corrupt Data error flows and provide end-to-end hardware, firmware, or software error recovery where possible.

PCle Live Error Recovery containment
Uncorrectable errors in a server’s PCIe subsystem can potentially propagate to other components, resulting in a crash of the partition—if not the entire server. To minimize this risk in HPE Superdome Flex servers, HPE implemented specific firmware features leveraging Intel’s Live Error Recovery (LER) mechanism that provides a means of trapping errors at a root port to prevent error propagation. LER containment allows the platform to detect a subset of Advanced Error Reporting (AER) and proprietary-based PCIe errors in the inbound and outbound PCIe path. When a PCIe error occurs, LER is able to contain the error by stopping I/O transfers to avoid corrupted data from reaching the network and/or permanent storage. LER containment also avoids the propagation of the error and an immediate crash of the machine. In parallel of this error containment, HPE firmware is informed and in turn, the OS and upper layer device drivers are made aware of the error. HPE contribution to the enhancement of the Advanced Error Reporting PCIe implementation allows Linux to better report the details of such errors in the Linux syslog files as well as cooperating with device drivers to resume from recoverable PCIe errors. HPE Superdome X and HPE Superdome Flex innovative solution for Live Error Recovery is not available on typical Xeon processor-based systems.

Poison error containment
HPE Superdome Flex servers further expand protection of customer data from corruption by tagging poison data in the processor and scalable server chipset. Poison data is prevented from going to or from I/O. Poison data can be read speculatively, but never-consumed Poison data will never be used by the processor or I/O.

Processor interconnect fault resiliency
All processor interconnects, including UPI, Memory Interconnect, and PCIe, have extensive cyclic redundancy checks (CRCs) to correct data communication errors on the respective busses. They also have self-healing mechanisms that allow for continued operation through a hard failure such as a failed link.

PCle links also support width reduction and bandwidth reduction when full width or full speed operation is not possible.

Advanced MCA recovery
Advanced MCA recovery is a technology that is a combination of processor, firmware, and operating system features. The technology allows for errors that can’t be corrected within the hardware alone to be optionally recovered by the operating system. Without MCA recovery, the system would be forced into a crash. With MCA recovery, the operating system examines the error, determines if it is contained to an application, a thread, or an OS instance. The OS then determines how it wants to react to that error.

Xeon Scalable processors expand upon previous Xeon E7 processor capabilities for advanced error recovery. Xeon Scalable processors now provide the ability to recover from uncorrectable memory errors in the instruction and data execution path (software recovery action required [SRAR]) in addition to handling nonexecution path uncorrectable memory errors (software recovery action optional [SRAO]). In expanding E7 processor memory error recovery including SAP HANA application recovery (Intel, 2011), HPE has done extensive development and testing of execution path recovery.

When certain uncorrectable errors are detected, the processor interrupts the OS or virtual machine and passes the address of the error to it. The OS resets the error condition and marks the defective location as bad, so it will not be used again and continues operation.

Socket disable and deconfiguration
During run time, if there is a CPU socket failure, the partition will log errors and crash. The Analysis Engine detects a socket error indication, a service notification will be sent to Insight Remote Support along with the CPU physical location. The system should reboot and come back up after system logs are collected.

If a CPU fails during boot time, BIOS will call it out, a service notification gets sent to the IRS. BIOS disabled the CPU group containing the faulty socket. A CPU group can be two sockets or four sockets depending on the CPU type. The rest of chassis will continue booting up and be part of the partition.
Memory RAS
Main memory failures have been a significant cause of hardware downtime. HPE Superdome Flex servers use several technologies for enhancing the reliability of memory such as proactive memory scrubbing and adaptive double device data correction (ADDDC). Additionally, HPE Memory DIMMs are qualified to provide both performance and quality. HPC applications may take advantage of the MEMlog memory RAS handling.

Proactive memory scrubbing
To better protect memory, systems such as HPE Integrity Superdome 2, HPE Superdome X, and HPE Superdome Flex implement a memory patrol scrubber. The memory scrubber actively scans through memory looking for errors. When an error is discovered, the scrubber rewrites the correct data back into memory. This scrubbing, combined with ECC prevents multi-bit, transient errors from accumulating. However, if the error is persistent, then the memory is still at risk for multi-bit errors. Accumulated memory DIMM errors can result in multi-bit errors that cannot be corrected and can result in data corruption. Proactive memory scrubbing is a hardware function included in HPE Superdome X servers that finds memory errors before they accumulate.

Data correction
The industry standard for memory protection is single error correction and double error detection (SECDED) of data errors. Additionally, many servers on the market provide single device data correction also known as chip sparing or chipkill.

Adaptive double device data correction (ADDDC) in HPE Superdome Flex servers addresses this problem. ADDDC technology determines when the first DRAM in a rank has failed, corrects the data, and maps that DRAM out of use by moving its data to spare bits in the rank. Once this is done, Single device data correction is still available for the corrected rank. Thus, a total of two entire DRAMs in a rank of dual in-line memory modules (DIMMs) can fail and the memory is still protected with ECC. This amounts to the system essentially being tolerant of a DRAM failure on every DIMM. After the second defective DRAM is detected, the OS will shut down to prevent any data corruption.

ADDC drastically improves system uptime, as fewer failed DIMMs need to be replaced. This technology delivers up to a 17X improvement in the number of DIMM replacements versus those systems that use only Single-chip sparing technologies. Furthermore, ADDDC significantly reduces the chances of memory-related crashes compared to systems that only have Single-chip sparing capabilities.

Although ADDDC is based upon an Intel Xeon processor E7 processor feature, HPE Superdome Flex has enhanced the feature with specific firmware and hardware algorithms. ADDDC provides a memory RAS improvement over Intel base code and reduces memory outage rates by 33% to 95% over standard x86 offerings.

After two DRAMs have failed, the memory contents are at risk. The firmware will initiate an OS shutdown to prevent data corruption.

New memory RAS introduced with Intel Xeon Scalable processor
Intel Xeon Scalable processors and their DDR4 memory subsystem provides a new memory RAS feature and retains two memory RAS features not available in previous E7 versions. These new features are described as follows.

Multiple Rank Sparing
Failures may span DRAM ranks. This feature provides the ability to move data from a faulty rank. This is a new feature for ADDDC over DDDC.

DRAM Bank Sparing
To better target the most likely memory failure modes at the DRAM level, DRAM Bank Sparing provides the ability to move data away from a faulty bank. DRAM Bank Sparing is automatically enabled as part of ADDDC and provides up to 33% more error resiliency compared to E7 v2 enhanced DDDC.

DDR4 Command/Address Parity Error Retry
DDR4 Command/Address bus is parity protected and the E7 v4 and v3 integrated memory controller and memory buffer provide detection and logging of parity errors. In previous E7 platforms, all Command/Address bus parity errors were fatal events, which caused an OS crash. Command/Address Parity Error Retry, ADDDC provides resiliency to errors across all memory interfaces and components.

Platform RAS
HPE Superdome Flex offers built-in RAS features, including System Fabric RAS and Fault-Tolerant RAS.

System Fabric RAS
HPE Superdome Flex Grid is a new and improved interconnect scheme, similar in look and feel to HPE MC990 X, which provides a flexible solution with adaptive routing capabilities. The system not only routes traffic down the optimal latency path for performance reasons, but also provides the ability to route traffic around failed component in the Grid and continue running in the event of most fabric failures. HPE innovative scalable enterprise system chipset includes extensive self-healing, error-detection, and error correction capabilities.
Designed with the goal of achieving a fault-tolerant fabric

HPE Superdome Flex Grid has been designed to achieve fault-tolerant fabric resiliency. The basics of the fabric are high-bandwidth links providing multiple paths and a packet-based transport layer that guarantees delivery of packets through the fabric. The physical links contain adaptive routing features to dynamically route traffic around failed components without requiring downtime. Strong CRCs are used to guarantee data integrity. The cabled fabric interconnect itself has no active components to fail, and it is conceivable that cables themselves may be serviceable in the future, including replacement and link re-activation without an HPE nPar reboot.

Partitioning and error isolation

Resiliency is a prerequisite for true hard partitions. HPE nPars are hard partition technology providing complete workload isolation, enabling you to configure a multichassis server complex as one large server or as multiple, smaller, independent servers. Each HPE nPar has its own independent processors, memory, and I/O resources of the blades that make up the partition. If the Grid is cabled appropriately, resources may be removed from one partition and added to another by using commands that are part of the system management interface, without having to manipulate the hardware physically. HPE nPars offer optimal performance with a recabling for up to 16 socket nPars.

Many systems use a shared backplane, where all blades are competing for the same electrical resources, and this raises the potential for a number of shared failure modes. For example, high queuing delays and saturation of shared crossbar resources may limit performance scaling, or enclosure power failures in systems with shared power may cause multiple partitions to fail at the same time. On HPE Superdome Flex system, subsystems are directly connected via cables, and each chassis provides its own bulk power conversion for a more flexible, reliable and scalable system.

Application-level RAS

HPE Serviceguard for Linux (SGLX) monitors the availability and accessibility of your critical IT services including databases, standard applications, and custom applications. Those applications—and everything they rely upon to do their job—are meticulously monitored for any fault in hardware, software, operating system, virtualization, storage, or network. When a failure or threshold violation is detected, HPE SGLX automatically and transparently resumes your normal operations in mere seconds by restarting the service in the right way and in the right place to enable improved performance.

HPE Serviceguard Storage Management Suite (SMS) also allows you to use a clustered file system to achieve the highest levels of availability and outstanding performance for applications benefitting from improved manageability and scalability. Furthermore, you can extend the comprehensive protection of HPE Serviceguard for Linux beyond the walls of your data center. HPE Serviceguard Metrocluster for Linux and HPE Serviceguard Continentalclusters for Linux offer robust recovery mechanisms for geographically dispersed clusters and enable your business to remain online even after a catastrophic event with disaster recovery solutions.

OS-level RAS

HPE mission-critical Superdome Flex environment provides an unparalleled set of features to detect and recover from faults. Many years of collaboration between the processor, firmware, OS, and application design teams has led to the delivery of several advanced error-recovery capabilities. Specifics of OS error recovery for memory and PCIe faults are described in the RAS differentiators section.

RAS feature summary

Chassis-level features

- Firmware First error handling
- Redundant, hot-swappable power supplies (N+N or N+1) and fans
- HPE Superdome Flex Grid link failover, link-level retry, dynamic link tuning, and bandwidth negotiation
- Adaptive routing finds bad fabric links and routes traffic around failures
- Cyclic redundancy check (CRC) protection per micro-packet and fast retry for transient errors
- Systemic transient errors, triggering retry or recovery attempt
- HPE Superdome Flex Grid link failover
- Socket disable at boot subject to limitations
- Chassis deconfigure at boot
- HPE nPars
Processor coverage
- EMCA2 architecture and recovery
- Integer pipeline or instruction pipeline retry capability
- Error-correcting code (ECC) coverage on all internal caches and cache tags
- Register or TLB parity protection
- Improved error (viral) containment, aiding system survivability
- UPI link-level retry, restart, or recalibrate
- UPI rolling CRC check for transient errors
- Core disable, indictment and deconfiguration at boot time and core-level corrupt data containment
- Data containment (Poison)

Memory features
- Proactive memory (patrol and demand) scrubbing
- Adaptive DDDC for mission-critical and HPC needs (with HPE Superdome X thresholds)
- Address or Command parity error resiliency
- DIMM indictment and deconfiguration
- Memory error logging or history in management firmware
- OS-level page deallocation with MEMlog
- Memory error storm response
- DRAM post package repair (coming soon)
- Rank and bank sparing

I/O capabilities
- PCIe Live Error Recovery (LER); PCIe root port containment and card error recovery
- PCIe Stop and Scream; PCIe root port corrupt data containment
- PCIe end-to-end CRC checking
- PCIe corrupt data containment (data poisoning)
- PCIe link CRC error check and retry
- PCIe link retraining and recovery
- Root port and card-level deconfigure

Management
- Analysis Engine
- Onboard analyzer
- Virtual KVM
- IRS and Proactive Care support
- SUM (coming soon)
- HPE OneView (coming soon)
- Onboard error logging services
HPE Superdome Flex: key areas of RAS superiority over standard x86

- Firmware First
- Automatic error logging
- Auto self-healing (Analysis Engine)
- Disabling and deconfiguration of failed FRUs
- Onboard fault analyzer
- Automatic restart
- Advanced processor error handling (EMCA2)
- Advanced memory resiliency (ADDDC)
- Memory error storm response
- Enhanced fabric resiliency (Flex Grid adaptive routing)
- Advanced PCIe error recovery (LER)
- Hard partitions (HPE nPars)

Management

HPE Superdome Flex offers extensive management capabilities through both built-in management capabilities and additional management resources.

Built-in management capabilities

HPE Superdome Flex management offers the following built-in management components:

- HPE Superdome Flex Rack Management Controller
- HPE Superdome Flex BMC

HPE Superdome Flex Rack Management Controller (RMC or eRMC)

The main component in the management subsystem is the Rack Management Controller (RMC), which connects to all the system chassis via a physically secure private LAN (see Figure 8, for N chassis system). Also, each chassis is managed by a Baseboard Management Controller (BMC), which configures and manages the hardware in that chassis as well as providing vMedia and vKVM features.
There is an option for one and two chassis configurations of HPE Superdome Flex to run the RMC functionality on one of the BMC management processors, connecting the two chassis together. This is called the embedded RMC or eRMC configuration. Functionality is identical to the larger configuration, but the extra 1U RMC appliance is not needed. For three or more chassis, or if there are plans to grow to that scale, use the RMC appliance.

HPE Superdome Flex RMC provides the following key features:

- Error Analysis Engine
- Firmware manager
- Partition management
- SSH-based command-line interface
- A console for each HPE nPar
- Redfish interface

The Error Analysis Engine is constantly analyzing all hardware for faults. Based on detected errors, the Analysis Engine can predict failures and initiate automatic recovery actions as well as notifications to administrators and to HPE OneView and HPE Insight Remote Support.
Onboard Firmware Manager can scan a partition and report components with incompatible firmware versions. Mismatched firmware (due to parts replacements or system upgrades) of a single HPE nPars or an entire HPE Superdome Flex can have firmware updated to a consistent level with just a click of a button. Partitions with consistent firmware levels are fully validated for the most reliable operation by the system developers. HPE Superdome Flex firmware is managed as one version of complex firmware, which includes a version of HPE nPar firmware, similar to server BIOS. The complex firmware is all the infrastructure components of the enclosure including the RMC, the BMCs, FPGAs, and CPLDs in the hardware, and BIOS images for the system. Having a single installation and version that is upgradable greatly simplifies firmware management and enhances platform reliability.

Partition management is implemented entirely in firmware. There are no dependencies on additional software tools and no need for an external management station or special hypervisor to build your desired partition configuration. The result is faster and easier partition configuration and partition start or stop. HPE nPars are fully electrically isolated from each other and run independently. All are controlled and monitored from the RMC. Support for HPE nPars is coming soon to the HPE Superdome Flex product, but the base architecture for it is already in place—the system is designed with HPE nPar composability in mind from the start.

The RMC SSH-based command-line interface provides access to all capabilities of the RMC, Analysis Engine logs, system control, as well as access to the console for each HPE nPar in the system.

The Redfish interface from the RMC enables the system to be managed by both HPE tools as well as being easy to script for using simple Python, curl, or other methods, in a secure, modern RESTful interface over https. Because HPE has been working with the industry on the Redfish standard from its inception, the standard already supports large, partitionable systems managed by a single aggregated controller like HPE Superdome Flex RMC. So the ability to present and control a system with many chassis, multiple HPE nPars, and more, is all comprehended by the standard, making it easier to integrate with software such as OpenStack.

HPE Superdome Flex BMC
Each chassis is managed by a Baseboard Management Controller (BMC), which configures and manages the hardware in that chassis as well as providing virtual media and virtual keyboard, video, or mouse (KVM) features. System control and inventory remains with the RMC in order to ensure coordination across HPE nPars and the entire system. The BMCs carry out instructions from the RMC and monitor each chassis independently, reporting any problems to the RMC. The virtual media (vMedia) and vKVM functions are served directly by one BMC for an HPE nPar, from the chassis that has the active Base I/O.

Additional management resources
Additional management resources such as HPE Insight Remote Support, HPE Insight Online, HPE OneView, and Smart Update Manager offer efficient and comprehensive monitoring and control of the HPE Superdome Flex from virtually anywhere.

HPE Insight Remote Support
HPE Superdome Flex Analysis Engine has been upgraded to work directly with HPE Insight Remote Support. No connection to the OS LAN is required on HPE Superdome Flex for monitoring or inventory collection by Remote Support (RS). Monitoring and troubleshooting is done entirely through the RMC, as well as inventory collection. A software package in HPE Superdome Flex Foundation Software called DCD supplements the analysis and inventory coverage of the RMC by monitoring for a few errors, such on I/O and disks that the firmware cannot detect as easily. DCD only communicates inside the system to the RMC, it does not require any configuration or connection out of the OS. The Analysis Engine utilizes all the information from DCD, the BMCs, UEFI, and other sensors throughout the system to determine if a problem has occurred or is about to occur on the system. It will send service events and periodic assessment reports to HPE Insight Remote Support, which can connect to HPE back end for automatic notification to HPE Pointnext of any problem with the system. Various support contract levels are available for HPE Superdome Flex. For more information, go to hpe.com/services/getconnected.

HPE Insight Online
HPE Insight Online provides one stop, secure access to the information you need to support HPE Superdome Flex with standard warranty and contract services. Through the HPE Pointnext, HPE Insight Online can automatically display devices remotely monitored by HPE. It provides the ability to easily track service events and support cases, view device configurations and proactively monitor your HPE contracts and warranties. This allows your staff or HPE authorized services partner to be more efficient in supporting your HPE environment. What’s more, they have the ability to do all this from anywhere and at any time. HPE Insight Online also provides online access to reports provided by HPE Proactive Care services. The embedded management capabilities built into HPE Superdome Flex server have been designed to seamlessly integrate with HPE Insight Online and HPE Insight Remote Support.
HPE OneView
Integration with HPE OneView provides a GUI look and feel of HPE Superdome Flex system, including detailed inventory of its components, and the ability to receive alerts and health information for up-to-date status. Future releases of firmware and HPE OneView will add features to control all aspects of the system, in a manner similar to all products supported by HPE OneView, as well as supporting advanced features and capabilities only found on HPE Superdome Flex system.

Figure 9. HPE OneView screenshot

Smart Update Manager
Smart Update Manager (SUM) is HPE’s firmware management and update tool for enterprise environments. It can remotely update all HPE firmware as well as firmware from other HPE products. SUM gives recommendations for firmware that needs updating, and has an easy-to-use web user interface providing reporting capabilities, dependency checking, and installing updates in the correct order through CLI and/or GUI.
Conclusion
Through our very close association with Intel during processor development, we have been able to have HPE Superdome Flex fully exploit the performance and RAS functionality built into the CPU. Our Platinum membership of the Linux Foundation yields a high degree of kernel enablement to ensure scalability, reliability, and performance at the OS level. This results in the groundbreaking performance, robust RAS, and flexible manageability that sets HPE Superdome Flex apart as the x86 scale-up solution for mission-critical environments.

Resources
HPE Superdome Flex information
hpe.com/servers/superdomeflex
HPE Insight Remote Support
hpe.com/services/getconnected
Smart Update Manager
hpe.com/info/hpsum

Learn more at
hpe.com/superdome