HP Superdome
Mainframe-class availability at one-eighth the total cost of ownership (TCO)
Business white paper
# Table of contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Executive summary</td>
<td>3</td>
</tr>
<tr>
<td>Introduction</td>
<td>3</td>
</tr>
<tr>
<td>HP Superdome: mainframe availability</td>
<td>4</td>
</tr>
<tr>
<td>Engineered for superior RAS</td>
<td>6</td>
</tr>
<tr>
<td>One-eighth the TCO of mainframes over three years</td>
<td>9</td>
</tr>
<tr>
<td>Conclusion</td>
<td>10</td>
</tr>
<tr>
<td>Appendix</td>
<td>11</td>
</tr>
</tbody>
</table>
Executive summary: For years, mainframe computers were considered almost the sole practical alternative for the world’s most mission-critical computing environments. They have long claimed the industry’s best reliability, availability, and serviceability (RAS). Today, with the HP Superdome, organizations have an alternative that features RAS and availability levels equivalent to mainframes, at approximately one-eighth the total cost of ownership (TCO) over a three-year period. This paper provides a detailed analysis of the availability HP Superdome systems deliver when compared to a recent IBM z10 mainframe—with potentially even more dramatic differences when comparing older mainframe models. It also touches on the superb reliability and serviceability of HP Superdome systems and provides a detailed business case comparing the TCO of the Superdome to a comparable mainframe system.

Introduction

The costs and risks associated with planned and unplanned downtime continue to escalate. According to the U.S. National Archives and Records Administration, 93% of companies that lost the use of their data center for 10 days or more filed for bankruptcy within one year of the disaster. Half of all businesses that found themselves without data management for this same time period filed for bankruptcy immediately.

Considering the dire consequences of downtime, many organizations—especially those in industries such as financial services, government, communications, transportation, and retail—have traditionally relied on mainframe systems to support their mission-critical computing needs. Mainframes have traditionally been known for being reliable and highly available. For instance, industry statistics for mainframes tout a mean time between failures (MTBF) of decades and near-perfect uptime. Mainframes also boast relatively easy serviceability.

The downside with mainframes, however, is their high cost—both upfront and on an ongoing basis. A baseline IBM z10 Model E40 with 40 active CPUs, for instance, has a baseline cost of $22.4 million, including initial purchase outlays and annual hardware and software maintenance over one year. With shrinking budgets and the need to do more with fewer IT resources, it is easy to see why so many organizations are looking for a computing alternative that offers RAS on par with mainframe systems—at a fraction of the cost.

The high availability of the Superdome is possible in part because of highly reliable cell board technology. One measure of reliability is mean time between failures (MTBF). Using actual field replacement rates, HP engineers calculated the Superdome cell board MTBF to be 57 years.
HP Superdome: mainframe availability

To counter the claim that no type of computer other than a mainframe can deliver mainframe-class availability, HP engineers conducted an in-depth study modeling the availability of HP Superdome systems. The models are based on actual field data gathered over 10 years of Superdome use in some of the world’s most demanding computing environments. The availability calculations were derived using continuous-parameter Markov chains. Markov chains are used to model the different states of degradation of each configuration outlined in this document. Failure and recovery rates between states are specified based on field failure data, field service data, and subsystem configurations. Once steady-state probabilities were computed, the system availability was calculated as the sum of the probabilities for the system states in which the hardware was operational.

While all HP Superdome servers offer superb availability, engineers chose two specific configurations for modeling purposes. Engineers ran two models, the first predicting the availability of a single HP Integrity Superdome sx2000 system with two hard partitions running HP Serviceguard, the high-availability clustering solution from HP, and Oracle Real Application Clusters (RAC). The second configuration featured two HP Superdome sx2000 systems, each with two hard partitions, running HP Serviceguard and Oracle RAC. (See the appendix for the full set of assumptions.)

The high availability of the Superdome is possible in part because of highly reliable cell board technology. One measure of reliability is mean time between failures (MTBF). Using actual field replacement rates, HP engineers calculated the Superdome cell board MTBF to be 57 years.
**Configuration #1:** single-server, 2-node Superdome cluster

<table>
<thead>
<tr>
<th>Configuration</th>
<th>Steady-state availability</th>
<th>Number of nines</th>
<th>Average annual unplanned downtime*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single-server, 2-node Superdome cluster running HP Serviceguard</td>
<td>99.9988%</td>
<td>Five nines –</td>
<td>6.3 minutes</td>
</tr>
<tr>
<td>Two-server, 4-node Superdome cluster running HP Serviceguard</td>
<td>99.9996%</td>
<td>Five nines +</td>
<td>2.1 minutes</td>
</tr>
</tbody>
</table>

*Due to hardware or operating system outages
Engineered for superior RAS

From redundant cell board components to double chip spare memory technology and hot swap I/O, the Superdome has been engineered since its inception for resiliency to prevent both planned and unplanned downtime. Over the years, these features have helped Superdome servers operate with near-perfect availability.

Key RAS features available on every HP Superdome include:

- **Intel® Itanium® 2 processors**: with a robust error-correction scheme and a world-class machine check architecture, these processors are specifically designed to enable mainframe-class availability in enterprise environments. Intel Itanium 2 processors, in combination with the advanced RAS capabilities of sx2000-based systems, support a level of availability that was previously only possible with high-end, proprietary platforms.

- **Intel Cache Safe Technology**: As cache sizes become larger, the likelihood of cache failures increases. Intel Itanium processors (starting with the dual-core Intel Itanium 2 “Montecito” processor) offset this increased likelihood of failures using Intel Cache Safe Technology (code-named “Pellston”). When there is a correctable error in the L3 cache, Cache Safe Technology tests the cache line and corrects the error. If the cache line is found to be defective, it is disabled.

- **Double chip spare**: An enhanced feature in HP systems is double chip-sparing technology. The hardware can permanently detect and correct an error in any given dynamic random-access memory chip (DRAM) and also detect and correct an additional memory error in any other memory location in the same code word. This detection and correction is done by firmware recognizing when the first DRAM has failed, and “erasing” its bits from the error-checking and correcting (ECC) correction calculations. This feature enables the ECC logic to correct for a second DRAM failure in the same ECC code word.

- **Electrical isolation of hard partitions**: Unique in the industry, electrical isolation of hard partitions enables true server consolidation. More than two generations of hard partitioning experience at HP are the customer’s assurance that a server divided into hard partitions is an excellent approximation of an array of smaller boxes, yet without all the system management and cost-of-ownership headaches.

- **Dynamic processor resiliency**: Dynamic processor resiliency (DPR) is the system’s ability to de-allocate (online) those CPUs that are exhibiting an unacceptable number of correctable errors. CPUs are de-allocated if the number of corrected cache errors reaches a specific configurable threshold. DPR is currently available on HP-UX and, to a limited extent, with the Windows® operating system. If HP Instant Capacity (iCAP) software is installed on the server, then an iCAP processor with no user downtime replaces the de-allocated processor.

**General Mills: a case in point**

Customers in real-world IT environments report exceptional availability and uptime in line with what the theoretical models predict. Leading consumer packaged goods company General Mills switched from a mainframe to Superdome for its mission-critical SAP R/3 environment. Application availability stays at mainframe rates—at a fraction of the cost of operating a mainframe.
• **Dynamic memory resiliency:** It is not uncommon to have single bits in a DRAM “go bad” (fail) during the life of a computer system. It is advantageous to map these locations out of main memory, because a persistent single-bit memory error can be a performance issue. Applying its high-availability design standard, HP has developed a unique solution to the single-bit hard-fail problem. If a location in memory is “bad,” that physical page (a 4Kb “chunk”) is de-allocated. This physical page de-allocation is persistent across reboots. Over the course of a system’s life, many memory locations can hard-fail, and the system can silently handle hundreds of these faults. These de-allocations occur dynamically without any operating system or application interruption. If a location in memory proves questionable (that is, exhibits persistent errors), that memory is de-allocated online, with no visible customer impact.

• **Dynamic multipathed I/O:** Multipathed I/O enables accessibility to a storage-device or networking end-node through multiple paths. The access can be simultaneous (in an active-active configuration) or streamlined (in an active-passive configuration). With this feature, points of failure between two end points can be eliminated. The system software can automatically detect network or storage link failures and can failover (online).

• **Advanced I/O error recovery:** The PCI error-handling feature enables an HP-UX 11i system to avoid a machine check abort (MCA) or a high priority machine check (HPMC) if a PCI error (for example, a parity error) occurs. Without the PCI error-handling feature installed, the PCI slots are set in hard-fail mode. If a PCI error occurs when a slot is in hard-fail mode, an MCA or HPMC occurs, then the partition crashes. When the PCI advanced error-handling feature is installed, the PCI slots containing I/O cards that support PCI error-handling will be set to soft-fail mode. If a PCI error occurs when a slot is in soft-fail mode, the slot is isolated from further I/O, then the corresponding device driver reports the error and the driver is suspended. PCI online addition and replacement (OLAR) commands can be used to recover online, restoring the slot, card, and driver to a usable state. PCI advanced error-handling, coupled with multipathing, is expected to remove upwards of 90% of I/O error causes from system downtime.

• **PCI computer bus online addition, replacement, and deletion:** The system hardware uses per-slot power control combined with operating-system support for the PCI card online addition (OLA) feature to enable the addition of a new card without affecting other components or requiring a reboot. This feature enhances the overall high-availability solution for the customer because the system can remain active while an I/O adapter is being added.
• **Link-level retry and link retraining**: When a link error is detected, the HP sx2000 chipset communicates the error event, retrains and retests the link, determines which communication packets have been transmitted successfully, and returns to normal operation. These recovery events occur without affecting system operation.

• **Spare channel: spare wire**: Each system communication link is composed of 20 parallel channels, of which only 19 are needed to transmit data and parity at any given time. Channel 20 is a spare, but it is used during normal operation to carry additional parity information. During link training (or retraining), if a channel is recognized as unconnected, the spare channel is used so that the bad channel can be left unused. This enables each link to remain functional through a single-channel failure.

• **Clock redundancy**: The system clocks are powered by two fully redundant and hot-pluggable hardware reference oscillators (high-stability oscillators [HSOs]) that support automatic, “glitch-free” failover and reconfiguration, and are hot-pluggable under all system operating conditions.

• **Power redundancy**: Nearly all DC–DC converters in Superdome sx2000 systems are fully redundant, significantly reducing downtime associated with power conversion.

• **Redundant and hot-plug fans and power supplies**: All fans in the system (that is, the system blowers and the fans for the I/O card cages) are fully redundant and hot-swappable, as are power supplies. These elements can be serviced without affecting any partition in the system.

All of these features substantially reduce planned and unplanned downtime. HP provides a host of additional features that further prevent planned downtime. HP Serviceguard high-availability clustering software, for example, avoids planned downtime and streamlines routine maintenance by enabling rolling upgrades. Previously, for a system with many components, the typical scenario was to bring down the entire system, upgrade every node to the new version of the software, and then restart the application on all the affected nodes. For large systems, this could result in lengthy downtime. An alternative is to enable a rolling upgrade. A rolling upgrade rolls out the new software in a phased approach by upgrading only one component at a time.

Other ways that Superdome helps prevent planned downtime include:

• **Dynamic Root Disk**: allows the system to be patched online. Installation and verification can be conducted during normal business hours; dual-bootable system images can be easily maintained and activated to change system personalities, product versions, etc., so that downtime is reduced to an efficient single-boot recovery.

• **Online addition, replacement, or deletion of I/O**

• **Dynamic partitions**: the ability to add or delete hardware resources without rebooting.

---

**Philippine National Bank: a case in point**

Philippine National Bank (PNB) had three mainframe systems spread across several sites. According to John Howard Medina, director of the core banking transformation project, IT costs were becoming a major issue. By moving to HP Superdome, the bank replaced the mainframe systems with a single HP Superdome and saved millions by retiring its power-hungry iron.
One-eighth the TCO of mainframes over three years

While Superdome delivers mainframe-class RAS, including up to “five and a half nines” availability, depending upon the configuration, its TCO is a fraction of a comparable mainframe. HP conducted an in-depth business analysis comparing the cost of the single HP Superdome sx2000 configuration used for the Markov chain model outlined above with a comparable IBM z10 mainframe system. The systems were comparably sized, with a 64-core Superdome system capable of 18,687 MIPS (million instructions per second), based on variable workloads. The Superdome was compared with a similar-sized IBM mainframe: the IBM z10 E35 capable of processing 18,586 MIPS. Note that although IBM often uses the term CPU, Processor Unit, or processor to describe the capacity of their mainframe systems, what they are referring to is a processing core. Based on a comparison of these two similar-sized systems, the business analysis concludes that the HP Superdome has a TCO one-eighth as much as the mainframe system’s over a three-year period.

### Table 2. Total cost of ownership (TCO) comparison

<table>
<thead>
<tr>
<th></th>
<th>HP Superdome (64c)</th>
<th>IBM z10 mainframe 35c*</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Hardware (one-time charge)</strong></td>
<td>$2,495,932</td>
<td>$16,500,000</td>
</tr>
<tr>
<td><strong>Annual hardware maintenance</strong></td>
<td>$232,239</td>
<td>$1,200,000</td>
</tr>
<tr>
<td><strong>Operating system</strong></td>
<td>HP-UX 11i</td>
<td>z/OS**</td>
</tr>
<tr>
<td>(one-time charge and annual software maintenance)</td>
<td>$549,379</td>
<td>$4,700,000</td>
</tr>
<tr>
<td><strong>Start-up services (one-time charge)</strong></td>
<td>$107,140</td>
<td></td>
</tr>
<tr>
<td><strong>1-year cost</strong></td>
<td>$3,384,690</td>
<td>$22,400,000</td>
</tr>
<tr>
<td><strong>3-year cost</strong></td>
<td>$4,025,936</td>
<td>$34,200,000</td>
</tr>
</tbody>
</table>

* Sizing assumes 115 transactions per minute per MIPS.

** z/OS pricing assumes 2,366 million service units (MSUs) with Parallel Sysplex License Charges (PSLC) pricing.

**Chart. TCO comparison: HP Superdome sx2000 (64c) and IBM z10 mainframe 35c**
Conclusion

After more than a decade of operation in some of the world’s most demanding computing environments, the HP Superdome has proven itself as an outstanding platform for mission-critical operations. Superdome features mainframe-class availability, as well as mainframe-level reliability and serviceability. At the same time, it has one-eighth the TCO of mainframes over a three-year period.

The HP Superdome paved the way for some of the most important trends in computing today, including partitioning, virtualization, and on-demand computing. And today, the evolution of Superdome continues to stay a step ahead of the industry.

Looking ahead to the next generation of Superdome, HP will continue to deliver features and attributes that will make the Superdome an even better alternative to mainframes. New Superdome systems will be even more modular, scalable, virtualized, and space- and energy-efficient. They will also offer even higher utilization rates and more exceptional availability. Especially in light of today’s economic climate, in which IT genuinely must do more with less and help enable business innovation, Superdome servers make more sense than ever before as a superior alternative to mainframes.
Appendix

Markov chain model assumptions

• In the study, “available” is defined as the ability for end users to access the database after possible failover and automated reconnection to a different partition, which implies that one or more partitions are up and the Oracle database is also up, so that failover and reconnection can occur.

• The modeled server configuration has two 16-socket/256 GB hard partitions, each running HP-UX in a 32-socket/512GB sx2000 Superdome server with 32 dual-core “Montvale” processors; active/active Serviceguard clustering software is assumed to be operating.

• Oracle RAC database software is assumed to be operating.

• All hard partitions within each server have redundant links to dedicated, mirrored external disks where the system/root files reside.

• The external shared storage arrays are not included in the modeling results.

• Server hardware crash rates are based on the latest field data for HP Integrity sx2000 Superdome with Intel “Montvale” processors.

• Average database recovery time after server partition hardware or OS failure = 0 (assumes Oracle RAC running on top of Serviceguard, called Serviceguard Extension for RAC [SGeRAC]).

• Average application restart time = 0 (application running on healthy node will allow users to reconnect without having to restart the application).

• OS crash rate is 20% for each OS instance, one instance for each hard partition.

• Average Serviceguard failover time is 30 seconds.

• Average end-user reconnect time is 10 seconds via an automated reconnection script.

• 24x7 operation is assumed.

• CE travel time is 2 hours.

• Memory dump time after OS crash is 10 minutes.

• Reboot time is 15 minutes.

References for Markov chain model


• Oracle Real Application Clusters on Extended Distance Clusters, Updated for Oracle RAC 10g Release 2, Oracle white paper, October 2006, (see p. 15, HP Extended Serviceguard clusters).


MTBF calculation

• The reported MTBF is based on actual field replacement rate results.

• Note that “faults” that occurred on the cell board, such as correctable memory errors that did not crash the server, were not included in the cell board failure rate calculation.

• The cell board failure rate was then expressed in estimated failures per operating year, which was then inverted to obtain MTBF in years.

• Rounded to the nearest year, the MTBF result was 57 years.

TCO analysis assumptions

The line items are grouped into one-time costs and annual costs to determine a simplified 3-year cost of ownership. Assumptions are as follows:

• The HP Superdome 64 core = 2,149,000 transactions per minute (tpm)

• 115 tpm per MIPS, based on variable workload

• Similar-sized IBM mainframe: IBM z10 E35 (18,586 MIPS)

• Only costs included for mainframe are hardware and z/OS

• z/OS pricing uses 2,366 million service units (MSU), an IBM mainframe software pricing metric that varies by system size, and PSLC pricing methodology

• Source of IBM mainframe MIPS, MSU and hardware pricing: http://www.tech-news.com/publib/pl2097.html

• The 3-year cost is determined by adding one-time costs to three times the annual support charges.