White paper
PRIMEQUEST 1000 series high availability realized by Fujitsu's quality assurance

PRIMEQUEST is an open enterprise server platform that fully maximizes uptime. This whitepaper explains how Fujitsu's expertise in high reliability and its long-term corporative policy for high quality help maximize PRIMEQUEST's business continuity.

Content

Introduction

Today's very dynamic business environments demand an increased need for non-stop, highly available, IT infrastructure. If a system failure occurs and data access stops, the result is not just a loss of productivity and business opportunity, but also the potential to jeopardize the brand image of the company. Whatever the outcome, such loss can create a serious impact on company profit. With increased globalization of business operations, and around-the-clock customer demands, comes a need for increased service levels and the highest in IT environment availability. As core components in the IT infrastructure, highly reliable systems are the key success factors in continued safe operation of all business applications.

PRIMEQUEST 1000 systems is Fujitsu's answer to such business continuity requests.The PRIMEQUEST expertise in high reliability has been fostered by the support of economic and social infrastructure. The result is a range of products and features focused on the delivery of high reliability and availability. One of Fujitsu's long-term corporate policies is "High Quality". This is reflected in the thorough quality assurance standards and the quality controlled supply of all Fujitsu products. In addition, Fujitsu has the capability to support the safe operation of all supplied customer systems, using optimized support structures that have been matured over extensive experience.
Fujitsu corporate values are detailed in its published corporate philosophy titled "The Fujitsu Way". There the value of "Quality" is defined as "We enhance the reputation of our customers and the reliability of social infrastructure."

This whitepaper describes the high availability features of PRIMEQUEST 1000 series together with Fujitsu's approach to quality assurance and control.

Hardware Redundancy

The PRIMEQUEST 1000 series is carefully designed with high resiliency to detect and recover problems at the earliest possible time and also to continuously operate even after failures. Single point of failures are thoroughly eliminated and almost all components including the power supply, cooling fan, and server Management Board (MMB) have redundant features.

1) QPI resend, degradation, and detouring

The PRIMEQUEST 1000 series uses the Intel® Quick Path Interconnect (QPI) high-speed system bus for communications between CPUs or between a CPU and a chipset. This protocol uses the Cyclic Redundancy Check (CRC) method to check transmitted data and perform resending if errors occur. In the event of a permanent bus failure, the system degradates to half or quarter of the bus width to avoid the fault location. Resending or bus-width degradation can be performed while operations continue. If a specific QPI bus becomes unusable because of failure, the protocol establishes a detour route around that bus and automatically performs a reboot. If a detour cannot be established because of configuration conditions, the degradation performs after reboot of the relevant System Board (SB).

2) Memory data protection

PRIMEQUEST 1000 series allows recover from failures in almost any part of memory and controller while minimizing memory throughput degradation. Double Device Data Correction (DDDC) allows continuous operations even double DRAM failures on the same DIMM at the same time. Scalable Memory Interface (SMI) assures memory access even in failure of CPU memory interface.
In addition, memory mirroring can free memory chip errors by duplicating units of memory and saving the same data to both units without sacrificing performance.

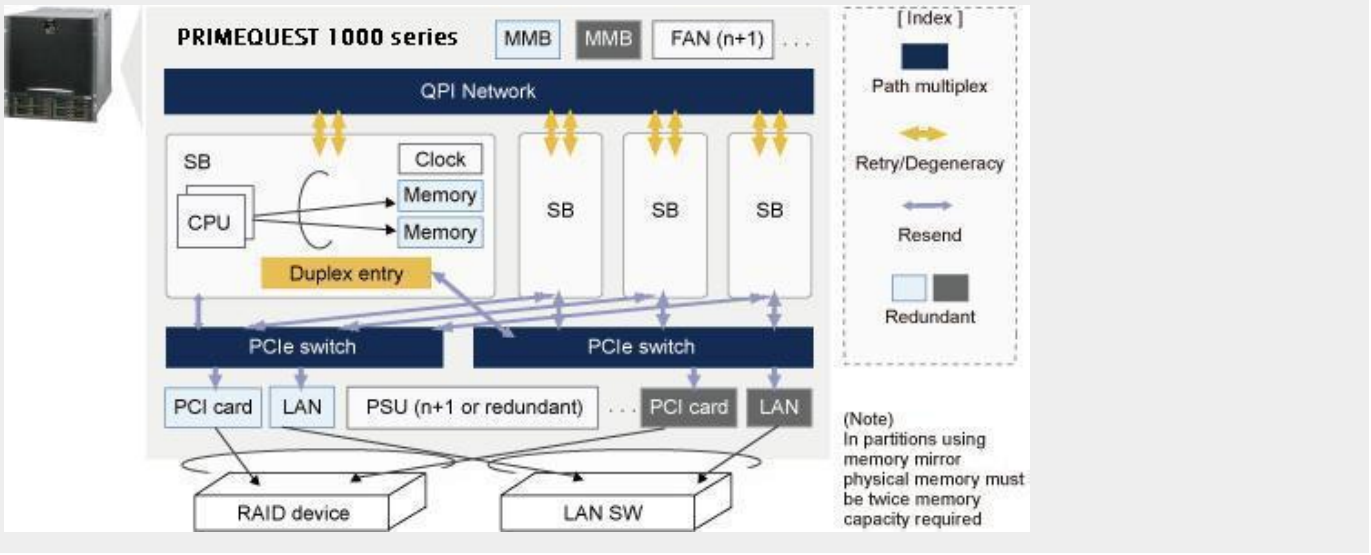
3) I/O data protection

I/O data is protected by data-send-retry mechanism in all of data paths between PCI Express, System Board, I/O Board (IOB), and individual I/O cards. Even if the error is found unrecoverable, the faulty part is automatically isolated.
Since a bus is configured in units of "lanes", the occurrence of a permanent bus error will result in degradation of the bus width by up to 1/2, 1/4, or 1/8. If the entire bus becomes unusable, devices under that bus are disconnected and a reboot is automatically performed. If a PCIe switch or I/O device fails, the failing components are disconnected and the system reboots.
All PCI's express slots including PCI Box slots that can be hot plugged. Using multipath SAN driver or LAN port teaming, PCI express card on failure can be replaced while the system is running if OS and Hypervisor support the PCI Hot Plug (PHP) mechanism.

4) Power supply units

The power supply can be made redundant by N+1 power supply units (PSUs) or N+N units.

Figure 1. Block diagram of PRIMEQUEST 1000 series



5) Fans

The cooling system is made redundant through the use of N+1 fans as standard.

6) Hardware RAID

The PRIMEQUEST 1000 series supports RAID 0, 1, 5, 6, 10, and 1E.

7) MMB duplication

Network ports and firmware codes are duplicated within the MMB, and the MMB itself can be duplicated. In the event of an MMB hardware failure, operation will automatically continue using the standby MMB.

Maintainability

PRIMEQUEST helps minimize down time in maintenance operation. Even if a faulty part must be replaced and the relevant hardware partition must be powered off, the other partitions can continue these operations.

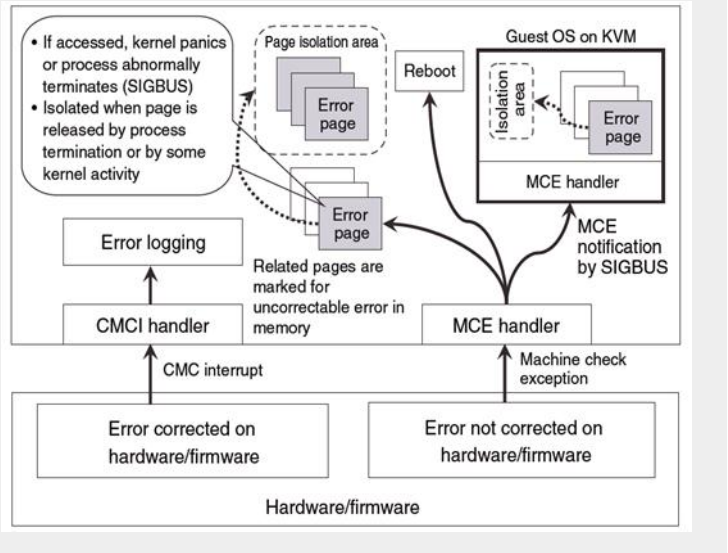
PRIMEQUEST makes maintenance operations as simple as possible by pull-out/push-in operations from the front or rear side of rack. The simplicity is enabled by a metal board called the midplane, which transmits electric signals between units without cables. PRIMEQUEST midplane does not contain any electronic components and consists of only connectors and wiring functions on printed circuit boards, so the failure probability is close to zero.

Figure 2. PRIMEQUEST Midplane

Data Integrity

As the heart of enterprise-class servers, CPU and memory must equip multiple levels of recovery mechanisms. In PRIMEQUEST correctable errors are recovered by retrying to rectify data.
The error detection and correction mechanism inside PRIMEQUEST allows recover from CPU or memory errors.
Cache or circuit errors in processors such as registers, ALUs (Arithmetical and Logical Units) and TLBs (Translation-Lookaside Buffer) can be detected and operations retried based on a predetermined parameter.
Memory errors can be corrected by a similar method. Plus, Patrol Scrubber keeps memory in a healthy state by scanning the entire memory while incrementing memory addresses.
PRIMEQUEST can continue its operation in a non-disruptive way even if it contains a faulty memory. PRIMEQUEST 1800E2 with Xeon processor E7 family is able to recover dual DRAM failures using DDDC (Dual Data Device Correction).
Machine Check Architecture (MCA) in Intel® Xeon® enhances server uptime by letting the OS react properly on uncorrectable failures that are isolated safely by degradating failed pages. MCA can minimize the risk that the whole system becomes inoperable, because only applications and virtual machines access the relevant page failres.
For details, please read the article below.
"Fujitsu's Activities for Improving Linux as Primary OS for PRIMEQUEST"
http://www.fujitsu.com/downloads/MAG/vol47-2/paper17.pdf

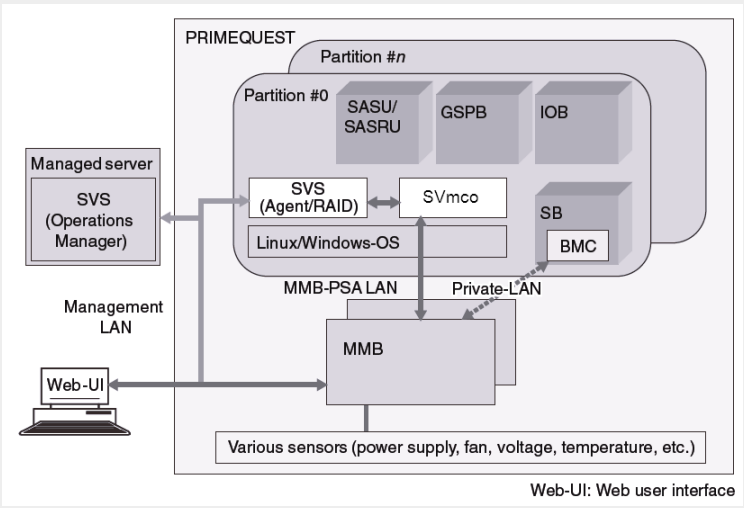
Figure 3. Machine Check Architecture (MCA)



Fault analysis and logging

The PRIMEQUEST series has three main types of firmware: Basic Input/Output System (BIOS), Baseboard Management Controller (BMC) firmware, and Management Board (MMB). If a hardware anomaly occurs, information collected by each type of firmware is integrated at the MMB to comprehensively determine the suspected fault location. Furthermore, event information associated with operational and environmental changes and error information as described above are recorded in the MMB system-event log and used in the analysis process whenever an anomaly occurs.

Figure 4. Server Management Mechanism

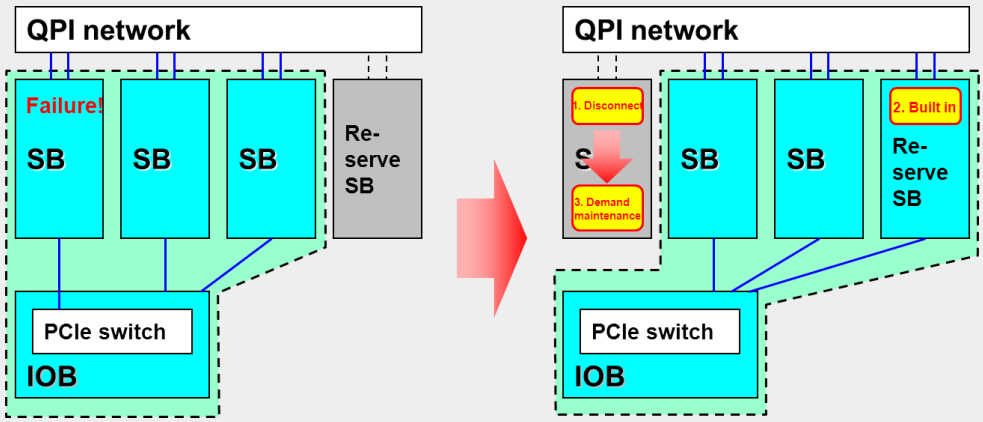


Server management functions are crucial for stable operation in a server system. The PRIMEQUEST 1000 series helps pinpoint and diagnose problems intuitively even in multiple server systems by MMB and SeverView Suite (SVS). MMB can integrate data required to identify problem, for instance, you can detect anomalies in PCI cards swiftly. SVS inter-works closely with the MMB through the internal LAN on Linux and Windows OS; it monitors the configuration and state of PCI cards. ServerView Mission Critical Operation (SVmco) can detect anomalies which are sent out by the SVS and can report them to the administrators by e-mail via the MMB.

High reliability in virtualized server

A hardware failure is fatal for high-dense consolidation because it causes disruption to all applications running on the hardware. Adding to the highest reliability possible at the component level, rapid recovery within one hour through the Reserve SB and flexible I/O provides optimum uptime without additional costs.

Figure 5. Reserved System Board Operation



Quality Assurance System for High Availability

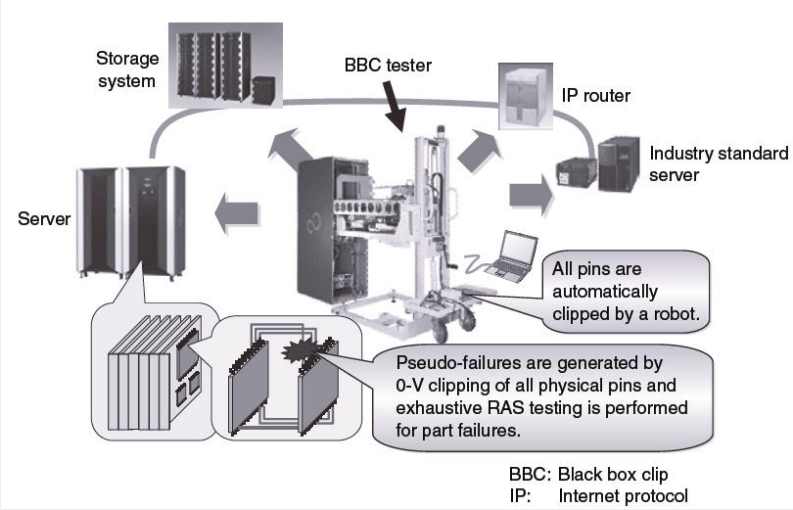
Fujitsu's high quality standard covers the entire life cycle of PRIMEQUEST, from system design, quality assurance tests, manufacturing, and customer support. To keep the quality at high level, Fujitsu monitors quality statistics to ensure PRIMEQUEST's failure rates do not exceed the pre-determined target for Annual Failure Rates (*1). Failure analysis of the information enables Fujitsu to improve its product reliability and quality assurance processes proactively.

*1 AFR: Fujitsu Annual Failure Rate, is the highest level quality target in the industry.

RAS Testing

Fujitsu has developed a test tool to ensure both recovery from hardware errors and operational continuity. BBC (Black Box Clip) tester: this tester generates pseudo hardware errors automatically. It does so by deactivating signals at each point on all the wiring patterns of each Printed Circuit Board in the PRIMEQUEST 1000 series. In addition, pseudo hardware errors can be generated repeatedly and automatically by this tool, varying the error occurrence times and intervals at each location.

Figure 6. BBC Tester



Extensive Experience in IT Delivery

Fujitsu has 30 years of experience in building and supporting solutions for mission-critical environments. This rich knowledge has also optimized the structure for very effective customer support. This support capability is yet another important factor in realizing the high reliability of PRIMEQUEST 1000 series.

Alliances

PRIMEQUEST 1000 series with Intel® Xeon® processors and Red Hat Enterprise Linux is the fruit of long standing cooperation with Intel and Red Hat. Its high value of cost efficiency and reliability are based on Intel Xeon's RAS function and proven reliability of Red Hat Enterprise Linux.

Conclusion

PRIMEQUEST 1000 series has been designed as optimized hardware and software for minimum downtime. The surrounding service infrastructure also enables monitoring services and constant failure analysis that help maximize availability and reliability. Over the years Fujitsu has developed its ability to support mission-critical environments as an expert IT vendor.

Contact
FUJITSU Limited
Address: Shiodome City Center, 5-2, Higashi-shimbashi 1-chome, Minato-ku, Tokyo 105-7123 Japan
Phone: +81-3-6252-2605
Website: www.fujitsu.com