White paper
PRIMEQUEST 2800B2 Enterprise Server With Best Cost Efficiency and Highest Uptime

Business continuity and high performance for data access have become essential demands on IT platforms. Offering the best-blend of standard and high availability technologies, PRIMEQUEST 2800B2 is an open enterprise system platform that fully maximizes uptime and greatly imrpove database performance. This whitepaper explains the the features of the PRIMEQUEST 2800B2 that make it the best choice for enterprise operations.

Content

Introduction

PRIMEQUEST is a mission-critical server that supports up to eight Intel® Xeon® CPU chips and maximum 144 cores. By combining the cost efficiency of x86 servers and high availability, customers can build their solid business platform and achieve a high return on investment with PRIMEQUEST.
However, demands for high availability and cost efficiency are different on a customer-by-customer. To meet such various demands, Fujitsu provides two types of PRIMEQUEST models, one foused on high availability called Enterprise Model including model 2800E2 with 8 sockets and 2400E2 with 4 sockets, and another model focused on cost efficiency called Business Model including 2800B2.
Out of three models, this whitepaper focuses on PRIMEQUEST 2800B2.
The intention of this whitepaper is to convince the reader that PRIMEQUEST2800B2 can help you reduce operational costs without sacrificing the business demands for high availability and performance scalability. First, PRIMEQUEST 2800B2 can minimize planned and unplanned downtime because almost all the components are fully redundant and hot replaceable. Plus, the heart of the server – the CPUs and memory – is protected from failure by multi-level data protection mechanisms. Second, PRIMEQUEST 2800B2 has very good performance scalability up to a maximum 144 cores and 288 threads.

What are inside PRIMEQUEST

Fujitsu PRIMEQUEST 2800B2 is formed of components below.

- Systgem Boards, which are formed of CPU and memory, act as distinct systems
- Server management called Mangament Board (MMB) monitors, operates, and controlls server entirely
- Power supply units which efficiently use electric power
- Cooling fans to maximize performance

Management Board, the integrated server management, helps resolve system failures by identification of the exact point in failure.

- Problem detection including System boards, IO Units, Power Supply Units, and fans
- Detection of disallowed range of temperature and voltages in many points inside servers
- Preliminary detection of problems inr error-prone parts such as disks and memories.

Management Boards also controls startup & shutdown of server, and activation & deactivation of system resources. PRIMEQUEST maximizes electric power efficiency by control of supplied power and adjustment of the number of PSU operating according to power consumption.

Managment Board

Management Board controls server components to maximize the server uptime and cost efficiency.

- Efficient cooling so that server performance can be sustained
- Efficient power supply so that power supply loss is minimized
- Diagonosis based on feedback data from parts of server
- Server setup including Physical Partitions and Extended Partitions

Predictive Maintenance

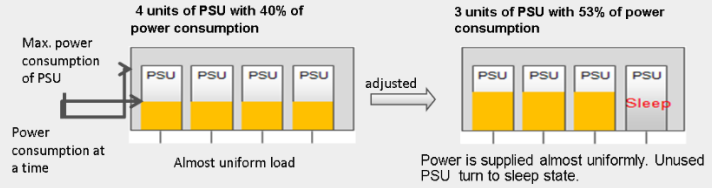
Predictive Maintenance for PRIMEQUEST 2000 helps take preventive measures for parts failures. This section focuses on internal disk drives, for which PRIMEQUEST 2000 can assure proper operations using statistical data called Self-Monitoring, Analysis, Reporting, Technology(S.M.A.R.T). Inter-working with ServerView Suite, PRIMEQUEST can detect problematic disc drives and store relevant statistical data to system trace. Report of the problems through e-mail or interfaces for system management software helps replace the problematic disk in early time.
PRIMEQUEST 2000 records error statistics including the number of correctable errors of CPU and memories to eliminate potential system problems.

Optimal Power Allocation

PRIMEQUEST 2000 controls power supply efficiently by adjustment of the number of Power Supply Units in operation according to power consumption of server.
Let's take an example for N+1 redundant PSU configuration. If four units of PSU operate with 40% of power consumption compared to max. value, PRIMEQUEST 2000 reduces the number of PSU to three with 53% of power consumption. As the result, 1 unit of PSU becomes nonoperational.

Electric equipment distributes electric power to parts inside. This is similar to water system, which provides homes with water – aging or slack of water pipes causes water leak and disturbs efficient water

Figure 1. Optimized Power Allotment



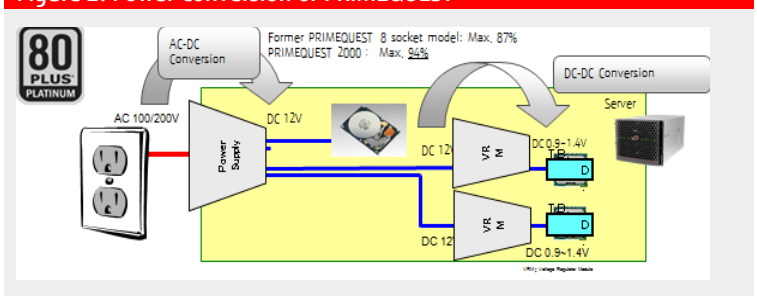
supply. For electric equipment, the deficiency of power consumption happens mainly in conversion or distribution of electric power. PRIMEQUEST 2000 has reduced loss of power conversion:

- Conversion of electric power from Alternating Current to Direct Current.
  Loss of electric power in this conversion has reduced to 6 per cent from 13 per cent
- Distribution of electric power to server
  Loss of electric power in this distribution has reduced to 12 per cent from 21 per cent

Cooling functions

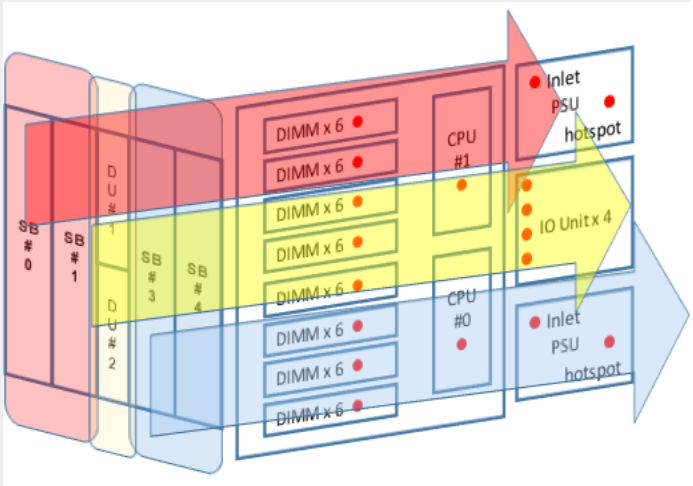
PRIMEQUEST 2000 maintains temperature of inside of servers as stable as possible to maximize performance and to reduce system disruption.

Figure 2. Power Conversion of PRIMEQUEST



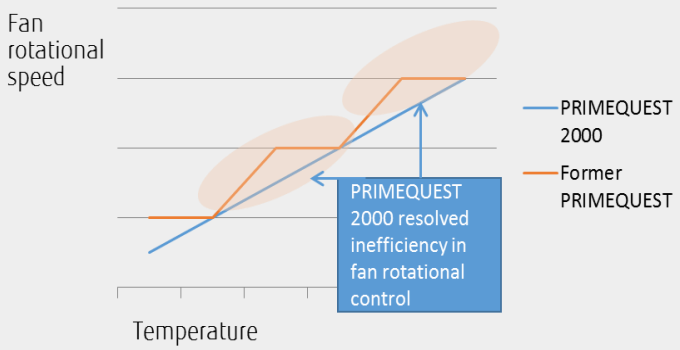
Fan rotational speed adjusts temperature changes. Integrating thermal data from thermo sensors attached in most of components, Maintenance Board can sense temperature rise in an area and blow out cooler air to the area included CPU and memory by speeding up fan rotation.

Figure 3. Air flows of PRIMEQUEST (side angle)



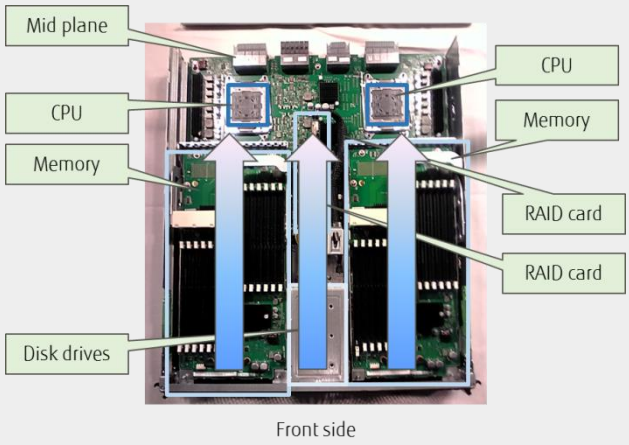
PRIMEQUEST 2000 controls fan rotational speed smoothly according to temperature changes. PRIMEQUEST 1000 server controls three level of fan rotational speed. So, inefficiency of cooling was the problem because fan rotation speed tends to becomes too high, responding to a small rise of temperature. Fan control of PRIMEQUEST 2000 has much improved cooling efficiency because it responds to temperature changes in smoother way. – Small change of rotation to small change of temperature.

Figure 4. Transition of fan rotational speed



Hot spots like CPUs and memories, temperature of which rises responsively to rise of power consumption, must be cooled down by flow of cool air. Other parts, temperature of which rises less responsively to rise of power consumption, can be cooled less intensively.
To adjust differences of temperature changes, PRIMEQUEST 2000 has three air flows. The two pass hot spots in System Boards, and the other one passes disk units.

Figure 5. Air flows of PRIMEQUEST (top angle)



High availability matched for mission critical systems

Maximizing business uptime is an axiom of mission critical server. However, to sustain server operation, even during a system failure, all components must be redundant. Especially the essential parts of the server such as CPU, memory, and system bus, must be especially assured as a failure of one of those components has the capability to cause an entire system shutdown. In addition maintenance operations such as component replacement, patch application, and testing, must be able to be executed while business applications continue to run, without interruption, or with the very minimum of downtime.

In-built high availability can slash mission critical server costs

Building mission critical systems could sacrifice cost efficiency. This is valid if the server being used lacks overall reliability. While clustering is a possible and practical solution with such servers, the need to double or triple the server count will bring extra costs – worse there are hidden costs that are easily overlooked.

In a multi-server cluster, maintenance costs will be more than doubled because administrators must apply the same patches to all servers in the cluster. They must also switch the cluster nodes before and after maintenance. License and support charges for some software may also double.

Even clustered systems are not totally safe from server failure. So, business losses by such downtime must be contemplated. In a High Availability (HA) cluster, for example, if failed nodes are switched to a stand-by node on failure, the cluster switching process requires a number of minutes to restart applications. If cluster parallelism such as Oracle Real Application Clusters (RAC) is used, performance deterioration will be unavoidable. For instance here, with a dual node Oracle RAC system, the performance would be halved during the time the failed server was offline.

PRIMEQUEST on the other hand embeds the equivalent of high-end UNIX server availability in every unit; this means the operational costs related to high availability are as low as those of UNIX servers.

CPU protection

Xeon E7 v3 processor family are designed to handle recoverable and unrecoverable errors.

- Recoverable errors
  Both data and tag fields in cache levels 1/2/3 can detect and correct bit errors. The data protection features of level 3 cache are described below.
  - Data array
    Up to three-bit errors can be detected and retried. Up to two-bit errors can be corrected.
  - Tag array, core valid array, and LRU (Least Recently Used)
    Up to two-bit errors can be detected and retried. One-bit errors can be corrected.
  - Registers, ALUs (Arithmetical and Logical Units), and TLBs (Translation Look-aside Buffer)
    One-bit errors are handled by each processor's circuits. They can detect and correct such errors.
- Unrecoverable errors
  If the above retry operations are successful, the application and operating system are not notified of the error. Only if the recovery is unsuccessful the application is stopped.

High resilience of Xeon E7 v3 becomes obvious if you compare its error recovery functions to Xeon E5 v3. To continue system operations, Xeon E7 v3 isolates the failed parts from system.

- CPU-CPU bus
  - E7 v3 can degrade failed buses. So, system can resume its operation by rebooting the system. But E5 v3 cannnot degrade the failed buses, so relevant CPUs must be replaced for resumption of system operation.
  - E7 v3 can fail over clock signal. But E5 v3 cannot fail over this.
- Memory controller
  - With E7 v3, multilple memory errors below can be recovered. With E5 v3, if such error happens, server operations must be stopped to replace memories.
    - (Xeon E7 v3) Two DRAM failures and one bit corruption can be recovered without system stoppage
- CPU-memory bus
  - E7 v3 can degrade failed buses. So, system can resume its operation by rebooting the system. But E5 v3 cannnot degrade this, so relevant CPUs must be replaced for resumption of system operation.

Table 1. Reliability comparison of Intel Xeon E7 v3 and E5 v3

Category	Items	Intel Xeon E7 v3	Intel Xeon E5 v3
CPU-CPU bus	Error detection using CRC and retrying	Supported	Supported
	Degradation of unrecovered-errored bus	Supported	Not supported
	Fail-over of clock signal	Supported	Not supported
Memory controll	Memory mirroring and memory sparing	Supported	Supported
	Memory-error recovery even in extreme case *1	Supported	Not supported
	Exact identification of DIMM in failure	Supported	Supported
CPU-memory bus	Error detection using CRC and retrying	Supported	Supported
	Degradation of unrecovered-errored bus	Supported	Not supported

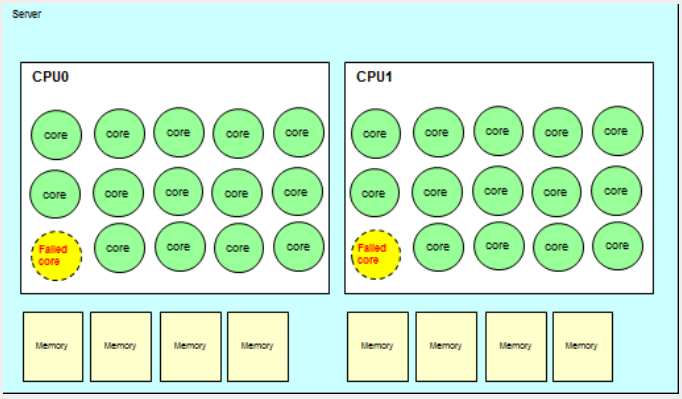
*1 Available with DDDC+1/SDDC+1

Mimization of CPU failure

If you face CPU failures, you have to give up using system and call a field engineer for system replacement. Until the repairment is completed, the system cannot be used. However, this is not the case for PRIMEQUEST 2800B2 because this high reliability server is designed to minimize downtime. Even in CPU failure, this server isolates failed part of CPU and resume operations. So, you can resume server operation after rebooting the server.

As shown in Figure 6, failed cores are isolated at system reboot.

Figure 6. Core degradation of PRIMEQUEST 2800B2



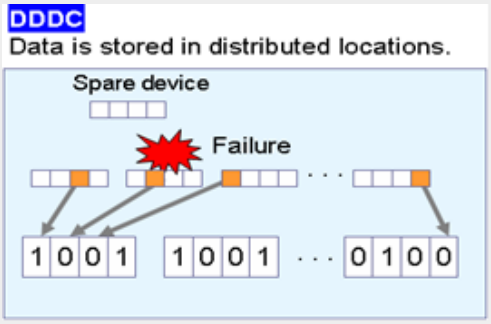
Memory protection

Memory chips and their interfaces to CPUs also have to be protected from errors. This is because memory is one of the most error-prone parts of the server and memory failures have the ability to cause an entire server stoppage.

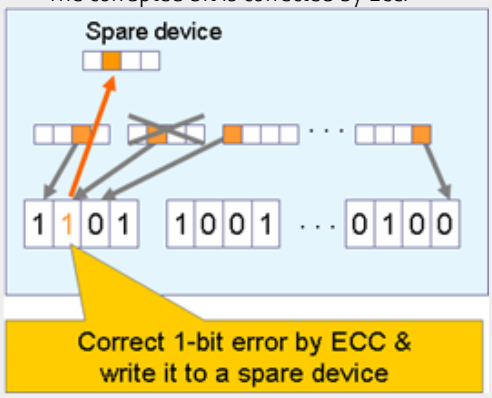
- Multi-bit error recovery
  Even with an error occurring in a DRAM module, the application can continue operating while the error is corrected.
  In DRAM 4-bit or 8-bit data chunks are typically assigned an additional DRAM bit. ECC (Error Check and Correct) uses this information to correct read errors so that CPU memory access can continue when an inconsistency is found.
  All models of PRIMEQUEST 2000 with Xeon E7 v3 product family processors is able to recover from dual DRAM failures using Dual Data Device Correction (DDDC) (Figure 7) and also able to recover the exreme condition that dual DRAM failures and one bit data corruption (DDDC+1).
- Bank SDDC/DDDC
  Bank SDDC/DDDC is expanded recovery from SDDC/DDDC recovery in units of DRAM to recovery in units of of bank of DRAM. Memory recovery in more granular level strengthens data protection of PRIMEQUEST – even if Maximum five banks fail all at once, read/write operations from.to memory can continue.
- Memory Mirroring
  Memory Mirroring is a memory redundancy function that allows each CPU to write to and read from a memory pair. This means CPU-memory access can continue even if a whole DRAM module fails, as the other available DRAM module still contains the correct data.
- Guaranteed read/write operations
  PRIMEQUEST 2000 detects and correct one-bit errors, detect two-bit errors and then performs retry operations using ECC.
  If an error occurs on one SMI2 (Scalable Memory Interconnect two) lane, which is an interface between processor and memory, memory access is able to continue using a spare lane.
- Memory Scrubbing
  Memory Scrubbing detects a malfunctioning memory chip before it is used. This is designed to ensure early detection and correction of memory errors using ECC. This includes Demand Scrubbing error checking at memory read time, and periodic error checking by Patrol Scrubbing.
- Multi Memory Rank Sparing
  This function allows reseves Memory Ranks of DIMM to replace Memory Ranks in correctable errors with the reserved ones without intervention of operation. Max. 4 memory ranks per DDR channel can be replaced automatically.

Figure 7. Memory error correction by DDDC

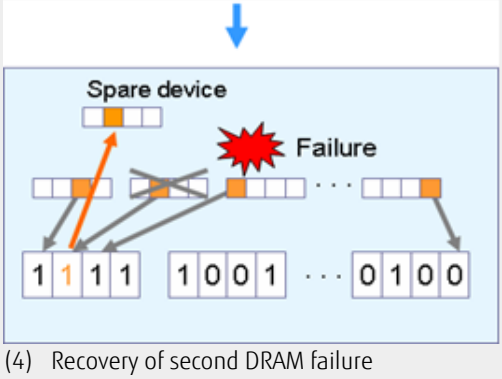
(1) First DRAM failure
Due to one DRAM failure, one-bit is corrupted.



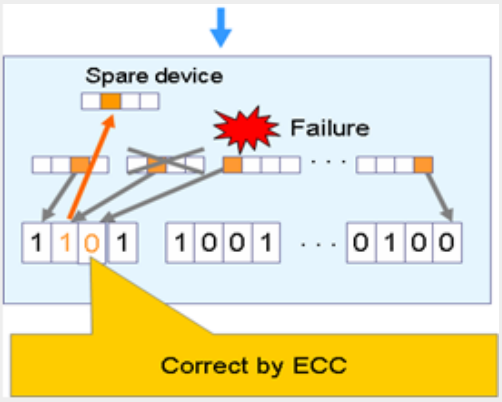
(2) Recovery of first DRAM failure
The corrupted bit is corrected by ECC.



(3) Second DRAM failure
Due to another DRAM failure, one-bit is corrupted.



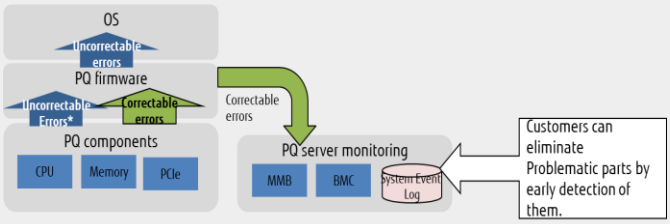
(4) Recovery of second DRAM failure
The corrupted bit is corrected by ECC.



Problem Prevention

PRIMEQUEST 2000 helps eliminate server problems including correctable errors at early stage. To do so, it records server problems including correctable errors of CPU and memory to System Event Log for system administrators to diagnose server problems and to take the best measures at earlier stage. This mechanism called eMCA Gen2 allows uninterrupted system operations after recording of error information to the trace.

Figure 8. Problem Prevention

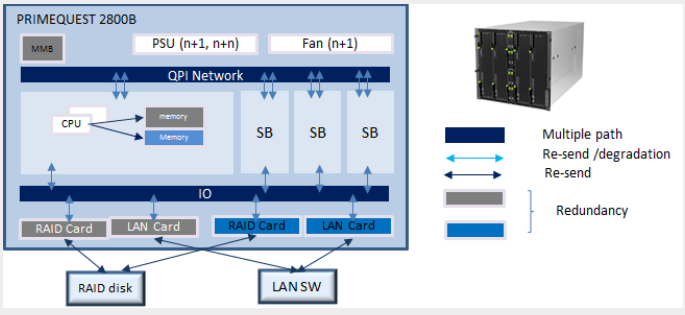


Component Redundancy

The Figure 9. below shows that almost every component is redundant or can be used in multiplex configuration.

- Redundant components
  Memory, PCI cards, standard LAN ports, fans, HDDs
- Path multiplex
  Interconnections between System boards and PCI switches, CPUs and other System board components.

Figure 9. PRIMEQUEST component diagram



.

Hot Replacement

All main components are hot-replaceable.

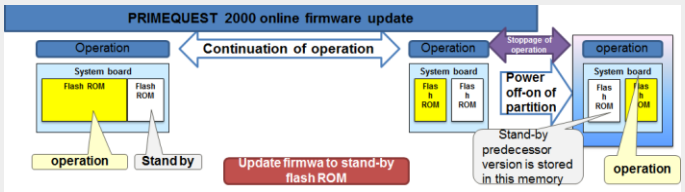
- Power supplies, fans, disk drives, PCI cards, service processors, and DVD drive

Online firmware update

Online Firmware Update can minimize time to apply firmware. Management board called MMB holds firmware to be updated in its memory area called flash ROM. At power-off operation, the updated firmware is applied to server. In predecessor model of PRIMEQUEST 2800B2, system must be powered off before application of firmware update and system can be restarted after completion of the application. So Online Firmware Update of PRIMEQUEST 2800B2 can eliminate downtime which was necessary for the application in the old model.

To do so, MMB have two flash ROMs and each ROM is assigned status :"operation, and "stand-by". The firmware update is stored to the stand-by flash ROM. At the power off of the server, the flash ROM mode is changed to "operation" status (Figure 10).

Figure 10. Online firmware update



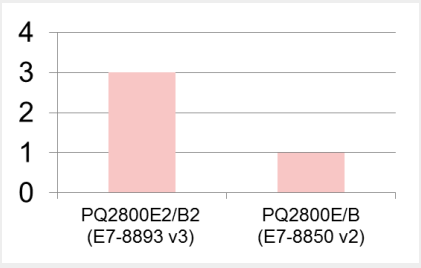
Improvement of performance per cost

Reduction of Oracle Database license

With improvement in data throughput and in per-core performance, PRIMEQUEST 2800B2 much improved performance per cost.

For instance, PRIMEQUEST 2800B2 with 8 sockets of Intel® Xeon® E7-8893 v3 4 cores triples the database transaction perfomance per Oracle license compared to PRIMEQUEST 2800B with 8 sockets of E7 8850 v2 12 cores. Because two servers have similar performance and the former server PRIMEQIUEST 2800B2 has the number of CPU cores just one-third of PRIMEQUEST 2800B, software license charged in units of CPU cores like Oracle database can be much reduced.

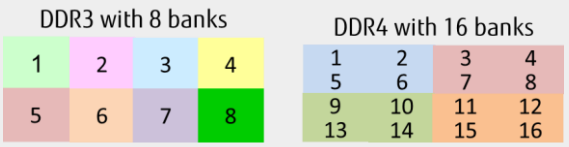
Figure 11. improvement of performance per cost



Improvement of data access performance

PRIMEQUEST 2800B2 with high-dense and high throughput memory, can improve data access performance. With DDR4 memory, formed of doubled the number of banks per DRAM compared to DDR3 memory, PRIMEQUEST 2800B2 shortens the time to reach the memory location for read/write operations.

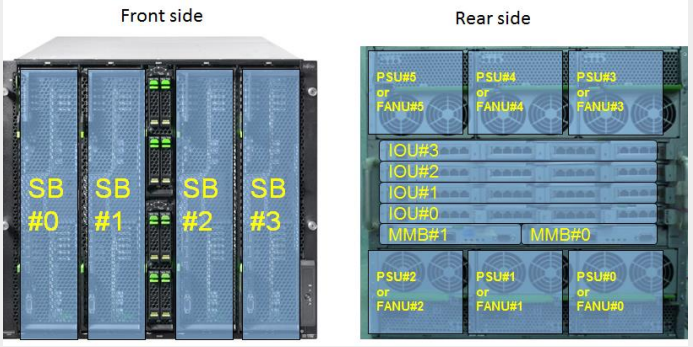
Figure 12. High density of DDR4 memory



Simple maintenance

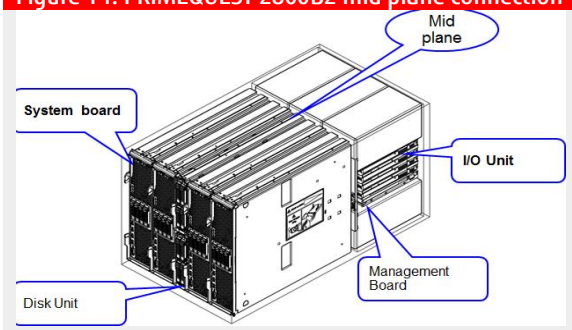
PRIMEQUEST much simplifies maintenance operations because most of components can be removed and mounted from/to front or rear side (Figure 13.). In most cases of component replacement, you do not need to lift down servers from rack. Most of components including System Board, PSU, fan, IO Board, System Management Board can be removed and mounted from front or rear side.

Figure 13. PRIMEQUEST 2800B2 front and rear side



PRIMEQUEST 2000 much reduces the quantity of cables because components are linked by metal board called mid plane. This means maintenance operations for PRIMEQUEST does not require cabling inside chassis. If taking off and taking on of cables were necessay, power-off of PRIMEQUEST chassis would have been necessary. So, this PRIMEQUEST design helps reduce downtime for maintenance (Figure 14)

Figure 14. PRIMEQUEST 2800B2 mid plane connection



Conclusion

With full fledged error detection and correction and high reliability design, PRIMEQUEST 2800B2 can maxmize uptime. This server can much reduce downtime necessary for maintenance operations such as replacement of failed parts and firmware update. Plus, componets can be replaced through front/rear side without lifting down equipments from racks due to its simple design where most of components are linked through mid plane almost eliminates cables.,

Looking to cost-efficiency of PRIMEQUEST, you can find the model 2800B2 can tripple performance per cost related to Oracle database license.

With elimination of planned and unplanned downtime and tripled performance per database cost, PRIMEQUEST 2800B2 is the best partner for your businesses.

Contact
FUJITSU LIMITED
Website: www.fujitsu.com
2015-5-5-WW-EN