White paper

PRIMEQUEST 2800E and 2400E Enterprise Server What's Inside The High Reliability Platform

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 2800E and 2400E are open enterprise system platform that fully maximizes uptime and greatly imrpove database performance. This whitepaper explains the the features of these enterprise servers 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 120 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 2800E with 8 sockets and 2400E with 4 sockets, and another model focused on cost efficiency called Business Model including 2800B.

Out of three models, this whitepaper focuses on PRIMEQUEST 2800E and 2400E.

The intention of this whitepaper is to convince the reader that PRIMEQUEST2800E and 2400E can help you maximize uptime and performance scalability.

What are inside PRIMEQUEST

Fujitsu PRIMEQUEST 2800E and 2400E are formed of components below.

- Physical Partitions, 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

Physical Partitions allows OS and applications operate without interference each other. Then, the core of system including database system in a Physical Partition can be protected from failures in other Physical Partitions.

Physical Partition is formed of components below.

- System Boards with CPUs and memories
- IO Units controling network and PCIe interfaces
- Disk units with disk storage space
- PCI Boxes extending PCIe interfaces

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 failure. 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.

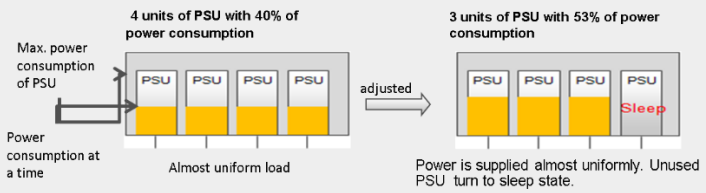
PRIMEQUEST 2000 records error statistics including the number of correctable errors of CPU and memories to eliminate potential system problems.

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.

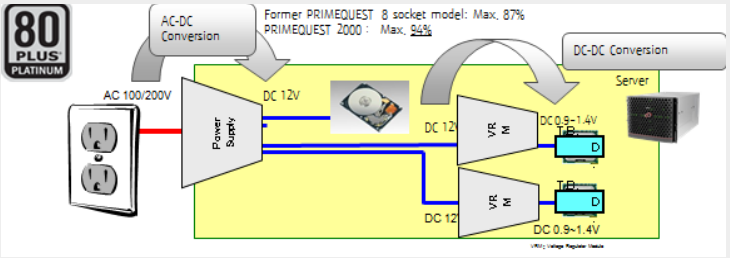
Figure 1. Optimized Power Allotment



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 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

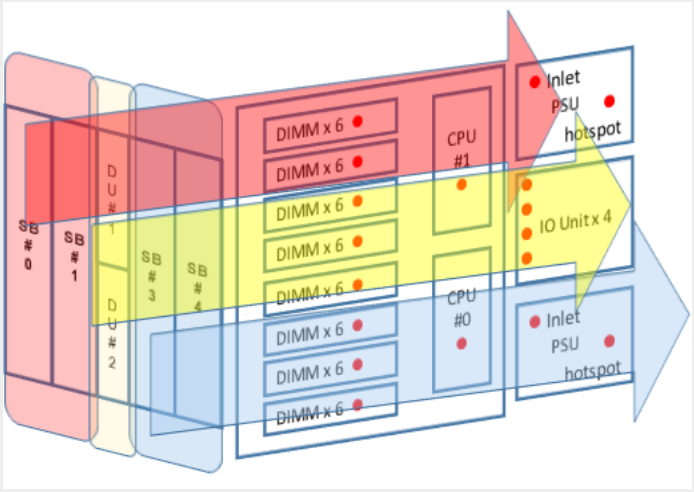
Figure 2. Power Conversion of PRIMEQUEST



Cooling functions

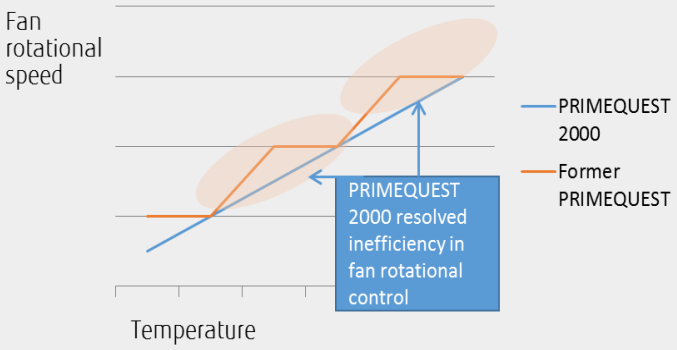
PRIMEQUEST 2000 maintains temperature of inside of servers as stable as possible to maximize performance and to reduce system disruption. Fan rotational speed adjusts temperature changes. Integrating thermal data from thermo sensors attached in most of components, Maintenance Board can sense temperature rise in an area and speed up fan rotation so that cooler air blows to the area included CPU and memory.

Figure 3. Air flows of PRIMEQUEST (side angle)



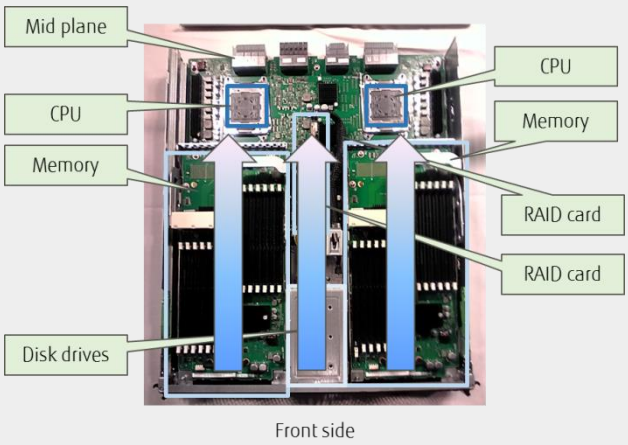
PRIMEQUEST 2000 controls fan rotational speed smoothly according to temperature changes. Former 8 socket PRIMEQUEST server controls three level of fan rotational speed. So, inefficiency of cooling was the problem because fan rotation speed tends to becomes too high, responding to a small rise of temperature. Fan control of PRIMEQUEST 2000 has much improved cooling efficiency because it respond 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.

CPU protection

Xeon E7 v2 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 in Intel Xeon E7 v2

High resilience of Xeon E7 v2 becomes clear if you compare its error recovery functions to Xeon E5 v2. To continue system operations, Xeon E7 v2 isolates the failed parts from system.

- CPU-CPU bus
  - E7 v2 can degrade failed buses. So, system can resume its operation by rebooting the system. But E5 v2 cannnot degrade the failed buses, so relevant CPUs must be replaced for resumption of system operation.
  - E7 v2 can fail over clock signal. But E5 v2 cannot fail over this.
- Memory controller
  - With E7 v2, multilple memory errors below can be recovered. With E5 v2, if such error happens, server operations must be stopped to replace memories.
    - (Xeon E7 v2) Two DRAM failures and one bit corruption can be recovered without system stoppage
  - Identification of failed DIMMs can reduce time for error recovery
    - E7 v2 can help identify failed DIMMs. E5 v2 cannot assure identification of failed DIMMs. With E7 v2, it is easier to recovery server from memory failures.
- CPU-memory bus
  - E7 v2 can degrade failed buses. So, system can resume its operation by rebooting the system. But E5 v2 cannnot degrade this, so relevant CPUs must be replaced for resumption of system operation.
  - E7 v2 can fail over clock signal. But E5 v2 cannot fail over this.

Table 1. Reliability comparison of Intel Xeon E7 v2 and E5 v2 component

Category	Items	Intel Xeon E7 v2	Intel Xeon E5 v2
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	Not supported
CPU-memory 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

*1 Two DRAM failures and one bit corruption can be recovered without system stoppage

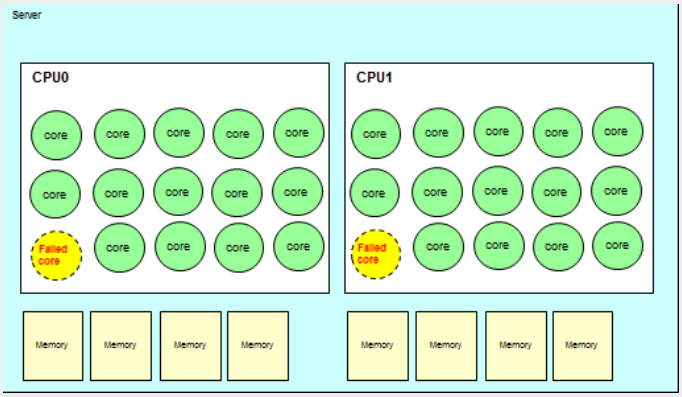
Mimization of CPU failure

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 2800E or 2400E because this high reliability server is designed to

minimize downtime. Even in CPU failure, this server isolates failed part of CPU then resume operations after rebooting the server.

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

Figure 6. Core degradation of PRIMEQUEST 2800B



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 v2 product family processors is able to recover 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).

- 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.

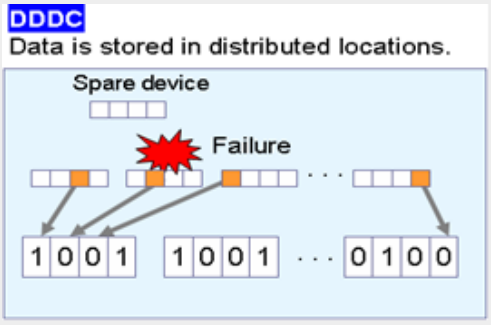
- Memory Scrubbing

  Memory Scrubbing detects a malfunctional 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.

Figure 7. Memory error correction by DDDC

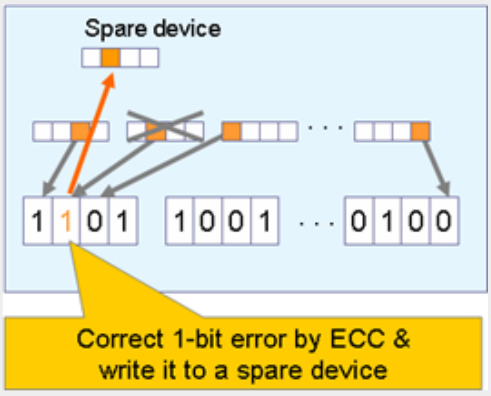
(1) First DRAM failure

Due to one DRAM failure, one-bit is corrupted.



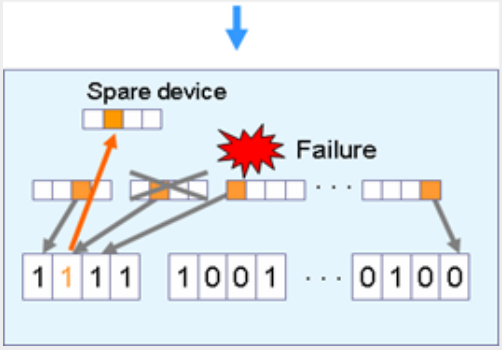
(2) Recovery of first DRAM failure

The corrupted bit is corrected by ECC.



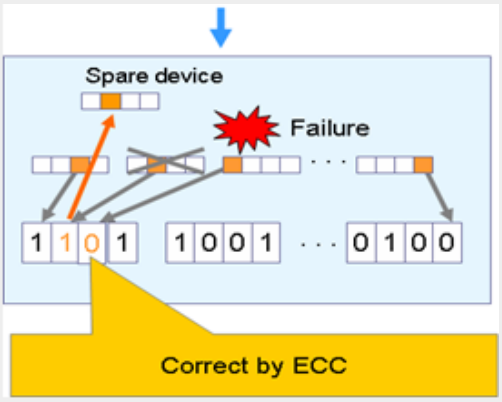
(3) Second DRAM failure

Due to another DRAM failure, one-bit is corrupted.



(4) Recovery of second DRAM failure

The corrupted bit is corrected by ECC.



Component Redundancy

The Figure 8. below shows that almost every component is redundant or can be used in multiplex configuration.

- Redundant components

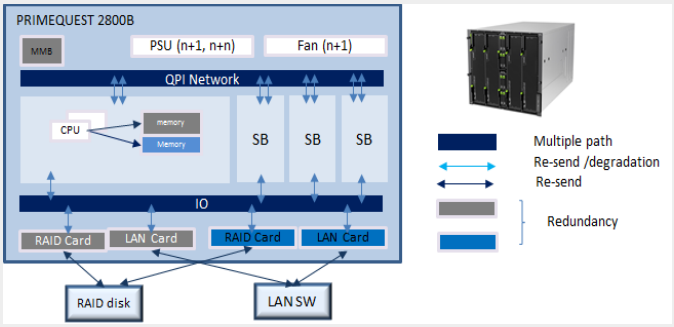
  Memory, PCI cards, standard LAN ports, fans, HDDs

- Path multiplex

  Interconnections between System Boards and PCI switches, CPUs and other System board components.

  .

Figure 8. PRIMEQUEST component diagram



Hot Replacement

All main components are hot-replaceable.

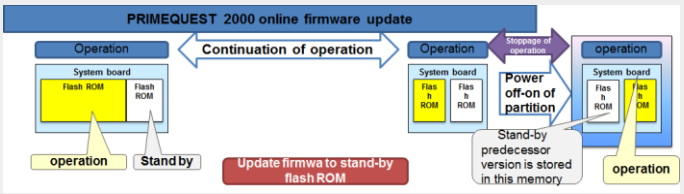
- Power supplies, fans, disk drives, PCI cards, service processors, and DVD drive

Online firmware update

Online Firmware Update can minimize time to apply firmware. Management board called MMB holds firmware update in its memory area called flash ROM. At power-off operation of relevant Physical Partitions, the updated firmware is applied to the Physical Partitions. The other Physical Partitions can keep operating. In predecessor model of PRIMEQUEST 2000, entire server 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 2800E and 2400E can eliminate system 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 9).

Figure 9. Online firmware update



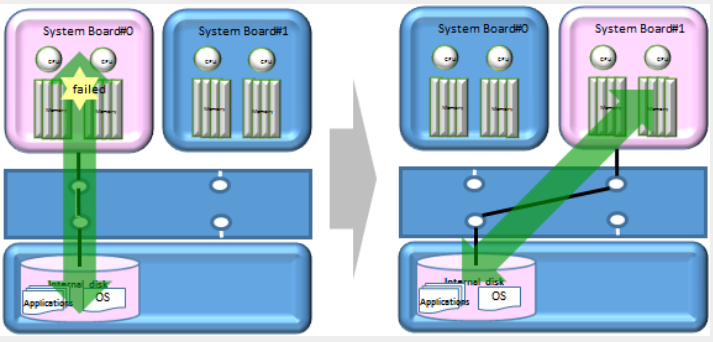
Physical Partitioning

To reduce operational costs of server, Physical Partitiioning is the easiest and fool-proof method because no knowledge about virtualization is needed to configure server consolidaiton. Physical Partitions can work independetly without interaction but under integrated administration by Managment Board, PRIMEQUEST 2800E can have max. four Physical Partitions. PRIMEQUEST 2400E can have max, two Physical Partitions. Each Physical Partition is formed of System Boards, IO Units, Disk Unit, and PCI Boxes under integrated power and cooling control by Management Board.

Reserved System Board

Failed server can be recovered swiftly without maintenance operations in a short time. Automatic replacement of System Board called Reserved System Board switches the failed System Board with reserved System Board, then re-connects connetions to disk storage which contains binary of applications and OS. With this unique function, uptime of customers system can be much improved.

Figure 10. Reserved System Board

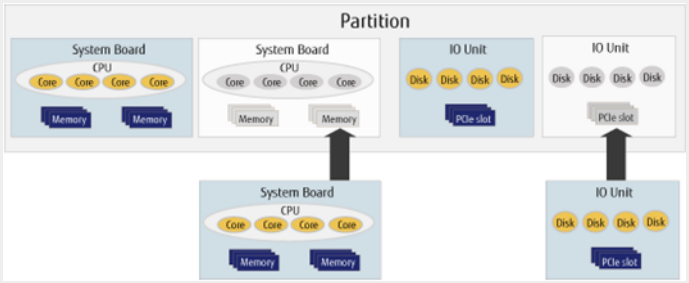


Dynamic Reconfiguration

To reduce system downtime, you have to replace the failed parts in a shor time and you have to add resources in system upgrade without system stoppage. For the system recovery, Reserved System Board mentioned in the last section is the best method.

To eliminate planned downtime which you have to ensure in system upgrade, PRIMEQUEST 2800E and 2400E provide you with excellent resolution called Dynamic Reconfiguration.

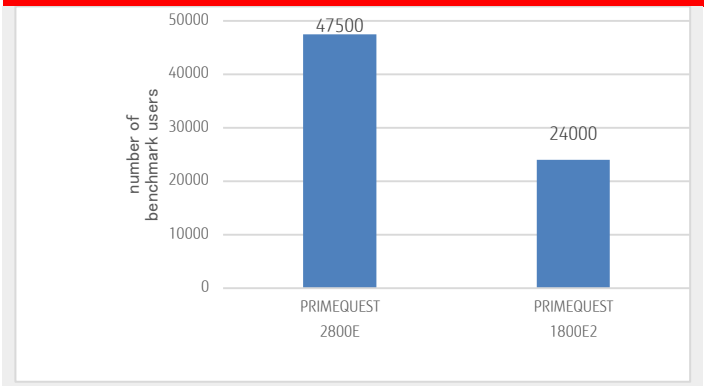
Figure 11. Dynamic Reconfiguration



Reasons of high performance

PRIMEQUEST 2000 can double the performance compared to predecessor 8 sockets server called PRIMEQUEST 1800E2. This performance improvement is attributed to higher parallelism of core and threads, improvement of memory & IO accesses. Due to these functions, SAP SD 2-tier benchmark with PRIMEQUEST 2800E is doubled compared to model 1800E2 (Figure 12).

Figure 12. SAP SD 2-tier performance *1



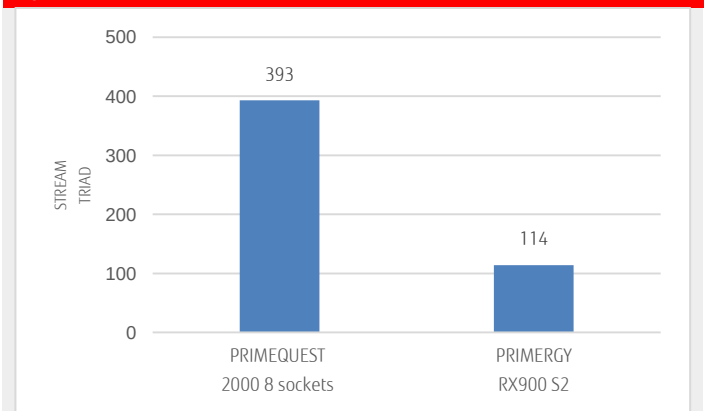
*1 For details refer to peformance reports.

- PRIMEQUEST Performance report
- PRIMEQUEST 1800E2 SAP benchmark result

It would be fair consideration that huge improvement of memory access performance contributes to the doubled performance.

According to STREAM benchmark results, the memory access throughput of PRIMEQUEST 2800E is 3.4 times that of PRIMERGY RX900 S2 with 8 sockets of Intel® Xeon® E7 (Figure 13). High throughput for data access helps much improve performance for enterprise applications. So, this benchmark results prove PRIMEQUEST 2800E can be best platform for enterprise applications.

Figure 13. STREAM TRIAD benchmark result *2



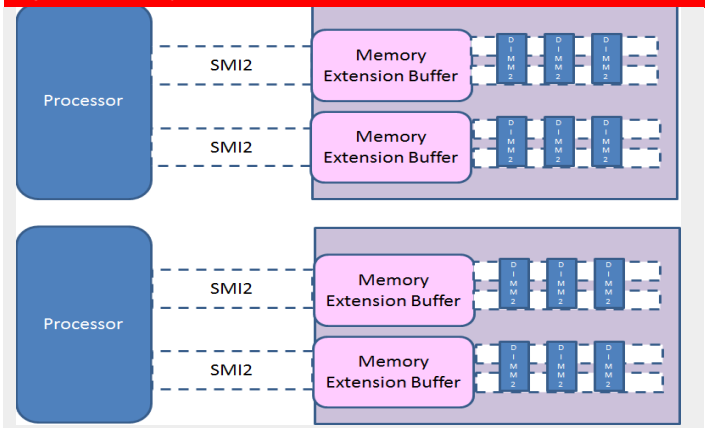
*2 For details refer to peformance reports.

- PRIMEQUEST Performance report
- PRIMERGY RX900 S2 Performance report

What inside PRIMEQUEST can improve memory access performance?

- Bandwidth of CPU-memory routes is much improved
  - Routebetween CPU and Memory Extension Buffer improves its throughput by 25%
  - Routebetweenf Memory Extension Buffer and memory chipS (DIMM2) improves its throughput by 56%
- Memory access parallelism
  - Intel C102/C104 Scalable Memory Buffer inside PRIMEQUEST 2000 can double its data rate of DRAM bus if 2:1 Independent Memory Channel Mode is used.

Figure 14. Memory controll architcture

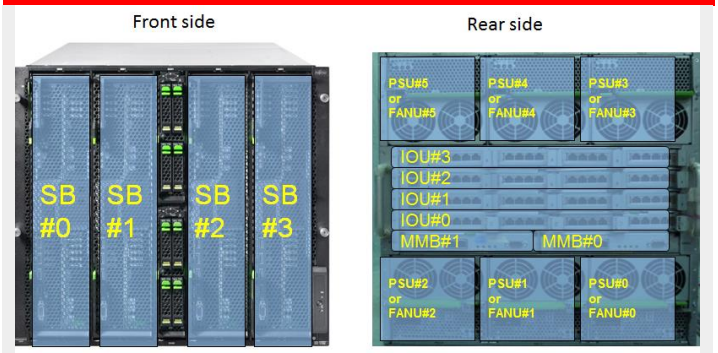


Reduction of IO access performance also contributes to this doubled application performance. Intel Xeon E7 v2 can improves IO access performance because this incorporating PCIe controllers called Internal IO(IIO), directly linked to data access parts of CPU. IIO is directly linked to CPU cache, so path length of data transmission to/from PCIe cards can be reduced.

Simple maintenance

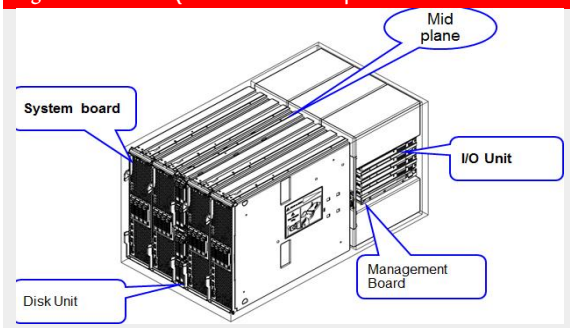
PRIMEQUEST much simplifies maintenance operations because most of components can be removed and mounted from/to front or rear side (Figure 15.). In most cases of component replacement, you do not need to pull out servers from rack. Most of components including System Board, PSU, fan, IO Board, System Management Board can be removed and mounted from front or rear side.

Figure 15. PRIMEQUEST 2800E front and rear side



PRIMEQUEST 2000 much reduces 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 16)

Figure 16. PRIMEQUEST 2800E mid plane connection



Conclusion

With full fledged error detection and correction and high reliability design PRIMEQUEST 2800E/2400E 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.,

With much improvent of memory access throughput and IO access throughput, PRIMEQUEST 2800E/2400E doubles application performance compared to its predecessor model.

With elimination of planned and unplanned downtime and doubled performance, PRIMEQUEST is the best partner for your businesses.

Contact
FUJITSU LIMITED
Website: www.fujitsu.com
2014-09-10-WW-EN