Empowered by Innovation NEC

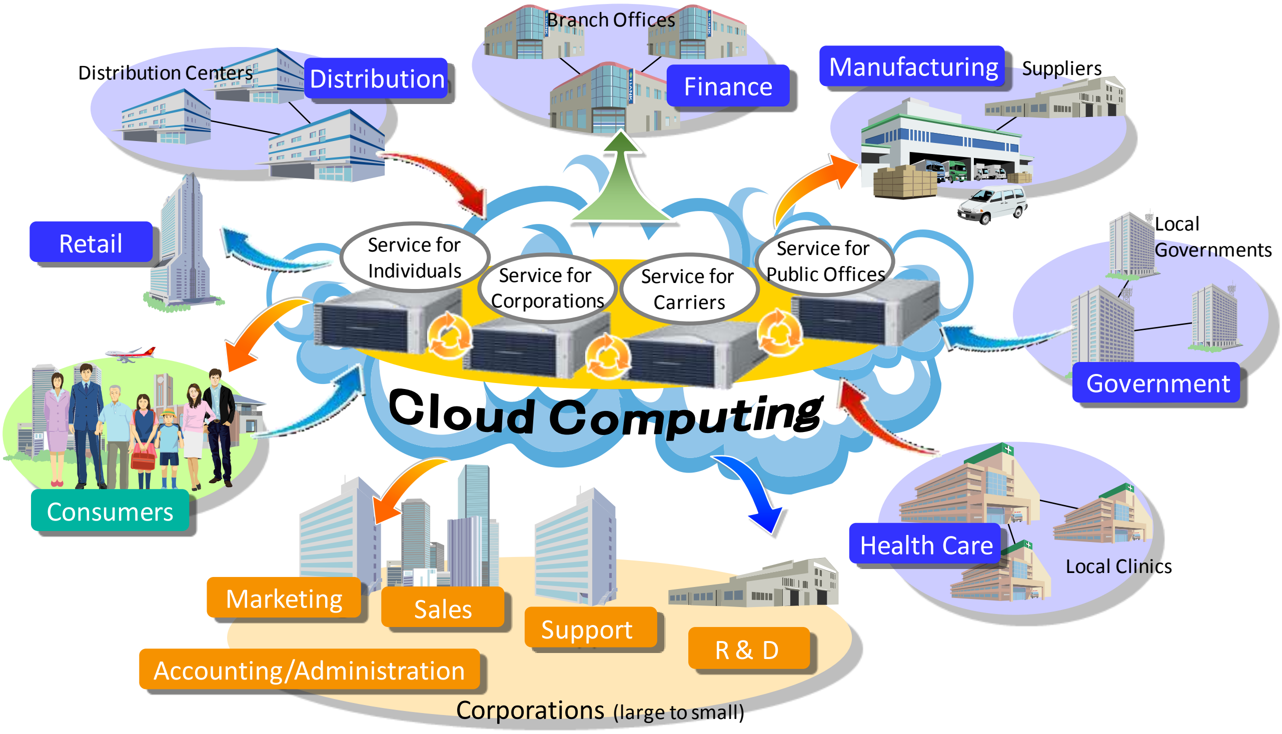
Fault Tolerant Server

White Paper

Table of Contents

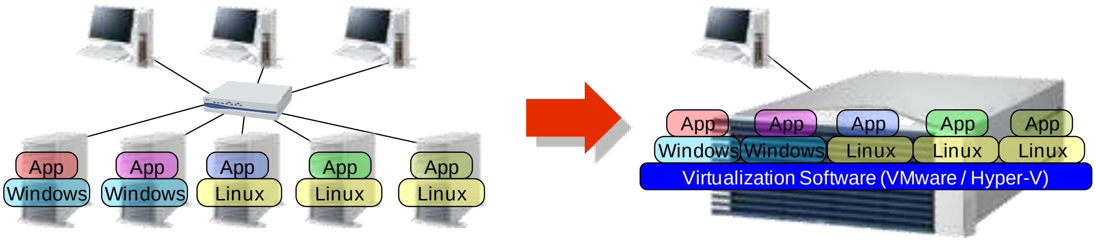
Introduction

Dramatic innovation in IT has brought significant changes to business practices and our way of life. Cloud services which leverage cloud computing are advancing with increasing speed, integrating all businesses and services through the network. Various IT devices and network infrastructure have become so prominent in our daily lives that we no longer can live without them. People demand for not only ease of use but also security and reliability in IT devices which are used to deliver lifeline services. Now essential to social infrastructure, the reliability of servers is becoming increasingly important.



Meanwhile, competition among server vendors has lowered server prices and is causing new issues. Affordable servers enabled companies and departments to readily configure new systems and add servers whenever necessary. As a result, many companies are now finding themselves owning a vast number of servers which incur excessive maintenance and management fees and drive up total cost of ownership (TCO).

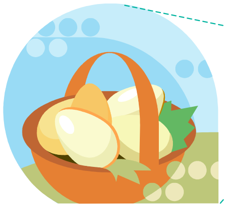
Aiming to address the situation, a solution now widely used is server consolidation leveraging virtualization technology. Virtualization tools, such as VMware® and Microsoft® Hyper-V™, enable a single server to accommodate more than one operating system each running multiple applications. This technology can simplify server management and internal control to help effectively reduce TCO.



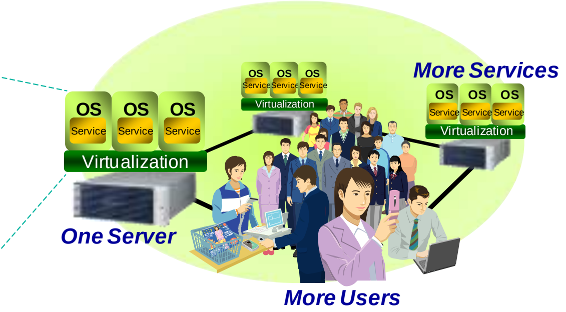
Unfortunately, server consolidation presents its own issues. Massive apps and large numbers of users relying on a single server significantly expand the scale of accidents and losses due to a single server failure. This implies that server consolidation goes against the old proverb about risk management—“Don’t put all your eggs in one basket.”

Don’t put all your eggs in one basket

This proverb reminds us that keeping your eggs in multiple baskets can save you from losing all your eggs if you drop one basket. Originally an English saying, it intimates the importance of investment diversification.



Eggs in one basket

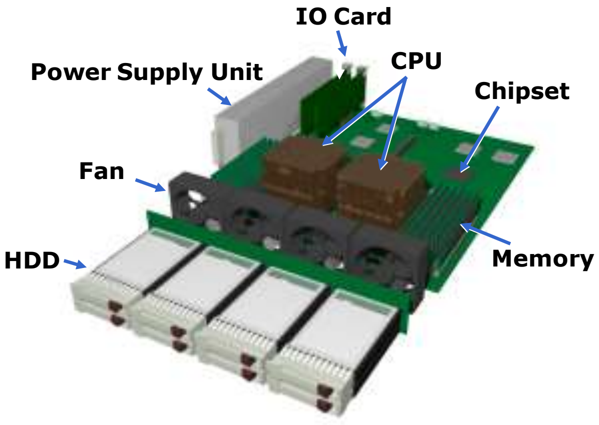


Therefore, with the spread of virtualization and server consolidation, there is a growing demand for high availability servers with superior failure resistance.

Reliability of Express5800 Server Series

Technical advances are increasing the processing power of general purpose servers. While this helps expand their presence in the business world, the heavy load inherent to high-level processing is increasing failure rates.

A typical general purpose server consists of components including hard disk drives, cooling fans, power supply units, I/O cards, CPUs, chipsets, and memory (see the figure on the right). At NEC, components with appropriate redundancy are used to meet the reliability target of each Express5800 server.

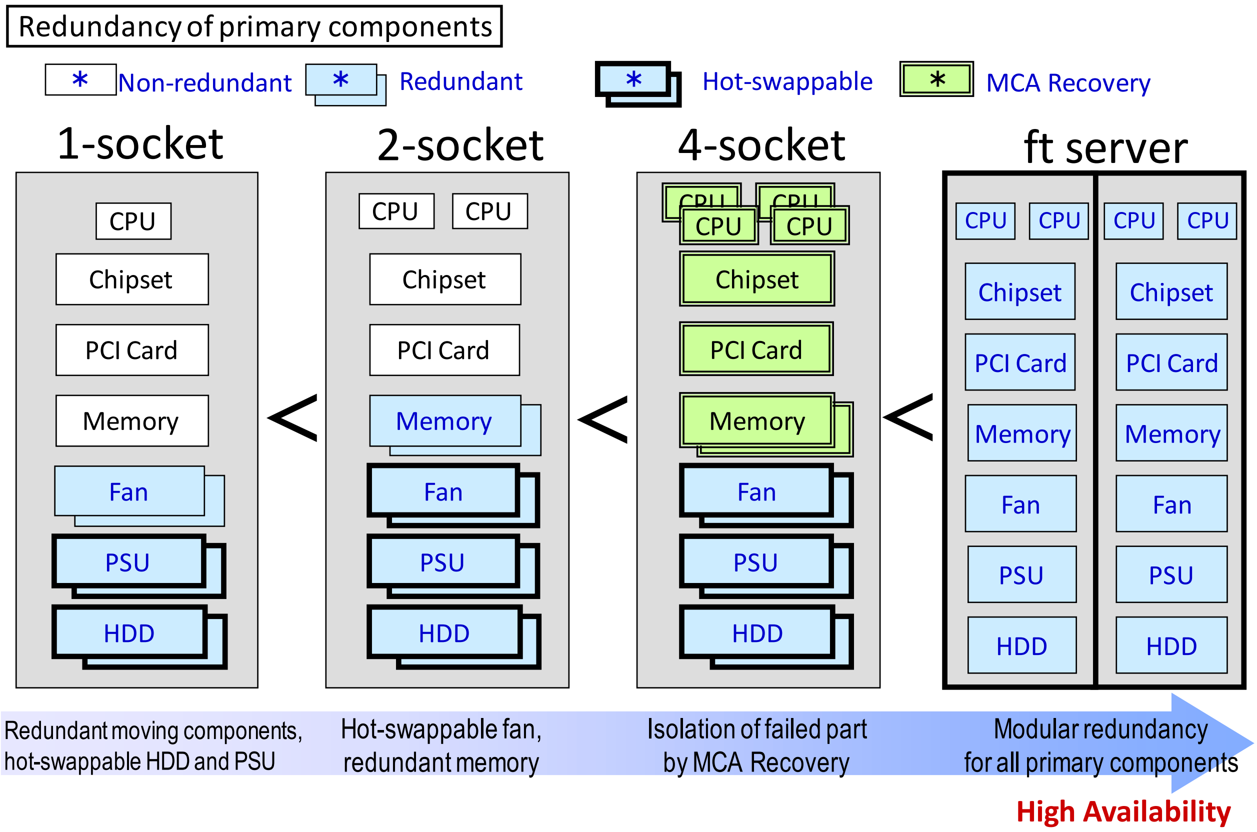


For example, in the most affordable single socket Express5800/R110 server rotating components such as hard disk drives (RAID), power supplies, and cooling fans are redundant. The hard disk drives and power supply units are also hot swappable. Still, the failure of other components can cause the system to go down.

The 2-socket Express5800/R120 server is more resistant to failures by adding redundant memory in addition to providing redundancy and hot swap capability for rotating components.

The 4-socket Express5800/R140 model maintains stable operation even when an uncorrectable hardware error occurs in CPUs, memory, and chipsets. The Machine Check Architecture (MCA) recovery function works with the operating system to tolerate failures that would be fatal to single- and dual-socket servers.

In terms of availability, the Express5800/ft series fault-tolerant servers are positioned at a higher level. The ft servers provide continuous availability through hardware redundancy in all primary components: CPU, memory, motherboards, I/O devices, hard disk drives, and cooling fans.


Redundancy of primary components
＊ Non-redundant
＊ Redundant
＊ Hot-swappable
＊ MCA Recovery
1-socket
2-socket
4-socket
ft server
CPU
Chipset
PCI Card
Memory
Fan
PSU
HDD
<
Redundant moving components, hot-swappable HDD and PSU
Hot-swappable fan, redundant memory
Isolation of failed part by MCA Recovery
Modular redundancy for all primary components
High Availability

Development of Express5800/ft Series Servers

Over the last two decades, mainframe and cluster computing systems have provided high availability for mission-critical systems for banking and utility systems. While there will always be a demand for such technologies, in our present networked society where every service and business is linked, there is a strong demand for high availability in more widely used IT devices. Unfortunately, costly mainframes and complex cluster systems do not best meet these needs. The high availability products of the current era must be affordable and easy-to-use for everyone.

Due to these pressures and in order to meet the high availability market needs for server products, in June 2001 NEC and Stratus Technologies co-launched the first fault-tolerant servers based on Intel® Architecture (IA) to deliver superior availability.

Concepts of the product include:

(1) Non-stop operation

Redundant hardware for continuous operation in case of component failures

(2) Non-disruptive maintenance

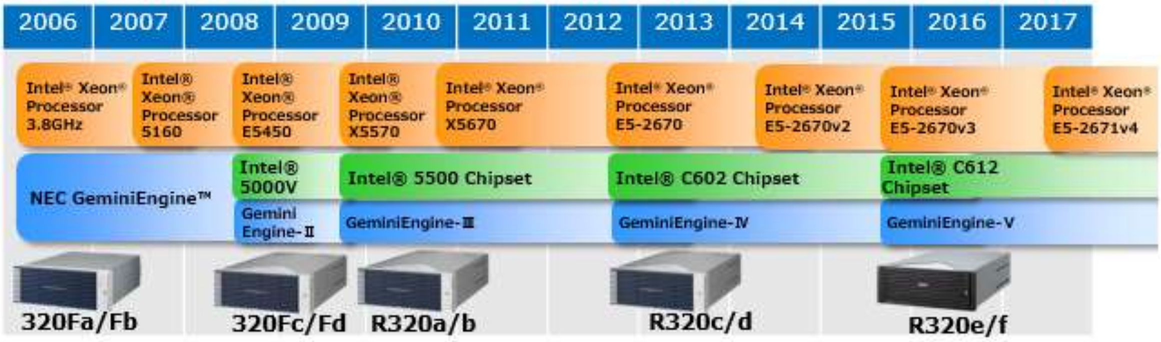
Hot-swappable components to enable replacements without interrupting operation

(3) General operating systems

General operating systems including Windows® / Linux® / VMware® to deliver the same operability as widely used servers

In 2003 as the ft series server gained greater acclaim in the marketplace, NEC started independent research and development for new ft series server technologies. While products up until then were limited within the scope of collaborative arrangements with Stratus Technologies, this new project aimed to flexibly incorporate the latest technologies, lower the product price, and meet the diverse needs of customers through leveraging NEC’s expertise in hardware development.

In February 2006, two and a half years from the project launch, NEC released the Express5800/320Fa ft series server. This was the first ft series server equipped with GeminiEngine™—a NEC-engineered LSI chipset for fault tolerance control. Since that time, NEC has released ft series servers in line with Intel’s latest CPU and chipset design and the hardware is provided to Stratus Technologies[1]. As of 2017, ft series servers using the 5th generation GeminiEngine™ are shipping and the 6th generation LSI chipset is under development for successor models.


2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
Intel® Xeon® Processor 3.8GHz
Intel® Xeon® Processor 5160
Intel® Xeon® Processor E5450
Intel® Xeon® Processor X5570
Intel® Xeon® Processor X5670
Intel® Xeon® Processor E5-2670
Intel® Xeon® Processor E5-2670v2
Intel® Xeon® Processor E5-2670v3
Intel® Xeon® Processor E5-2671v4
NEC GeminiEngine™
Intel® 5000V
Gemini Engine-Ⅱ
Intel® 5500 Chipset
GeminiEngine-Ⅲ
Intel® C602 Chipset
GeminiEngine-Ⅳ
Intel® C612 Chipset
GeminiEngine-Ⅴ
320Fa/Fb
320Fc/Fd
R320a/b
R320c/d
R320e/f

GeminiEngine™

GeminiEngine™ is the NEC engineered ft-control LSI positioned as the core technology enabling hardware redundancy.

1st generation (Doppelgänger)
NEC engineered northbridge
Express5800/320Fa, Fb

2nd generation (Stratagem)
ft control chipset
Express5800/320Fc, Fd

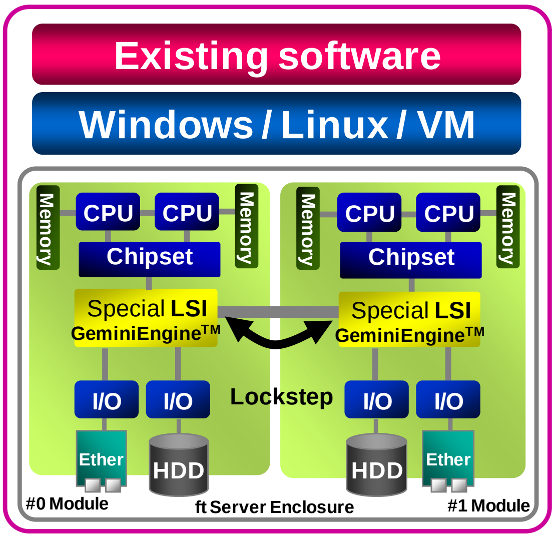
3rd generation (Argus-M/S) ft control chipset
Express5800/R320a, b

4th generation (Albireo-M/S) ft control chipset
Express5800/R320c, d

5th generation (Tarsus-M/S) ft control chipset
Express5800/R320e、f

[1] In the current collaboration in ft server development, NEC develops hardware while Stratus Technologies develops software.

Features of Express5800/ft Series Servers



To meet the requirements introduced in the previous section, ft series servers provide high availability through:

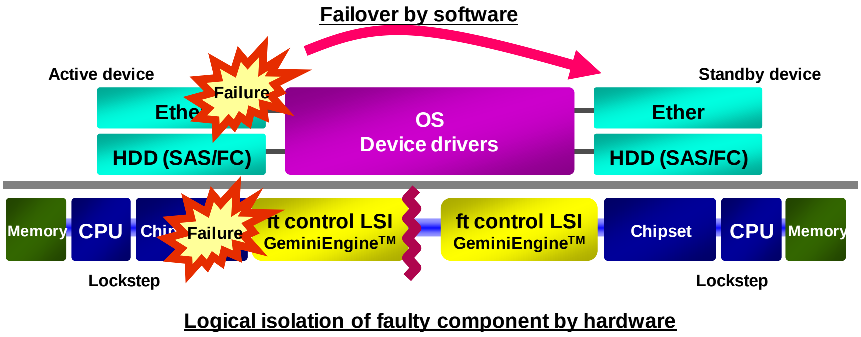
- Non-stop operation in the event of hardware failures
- Easy repair without interrupting operation
- Compatibility with existing operating systems and applications

Due to these features, ft series servers differ from normal servers in structure.

As the illustration on the right shows, ft series servers have two identical component groups called CPU/IO modules. Other than the special LSI in the center, each CPU/IO module consists of components much like those in typical general purpose servers and is capable of performing as a single server.

The special LSI, GeminiEngine^{TM}, is the key to the unique fault tolerant functionality. The combination of redundant hardware and redundancy control software enables ft series servers to operate non-stop. The following section elaborates on the non-stop features.

Non-stop in the event of hardware failures



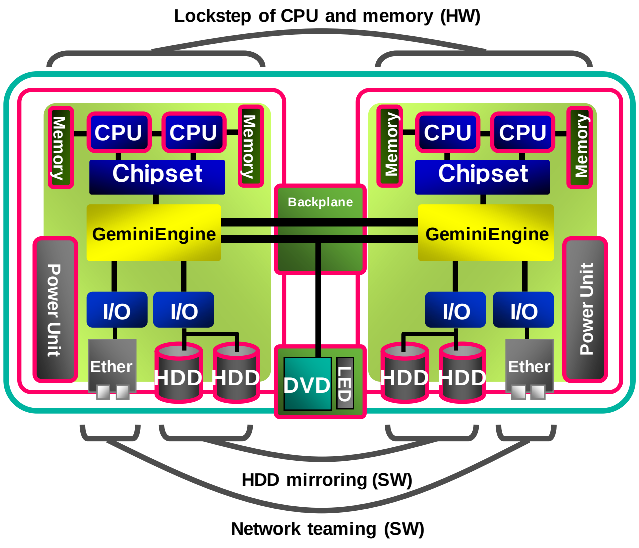
The illustration on the right shows how hardware and software work in the ft series server system.

Primary components such as the CPU and chipset create the system platform. On the platform, you find the operating system as well as I/O components such as Ethernet (Ether) and Serial Attached SCSI (SAS) hard disk drives or Fibre Channel hard disk drives. The operating system and device drivers control these I/O components.

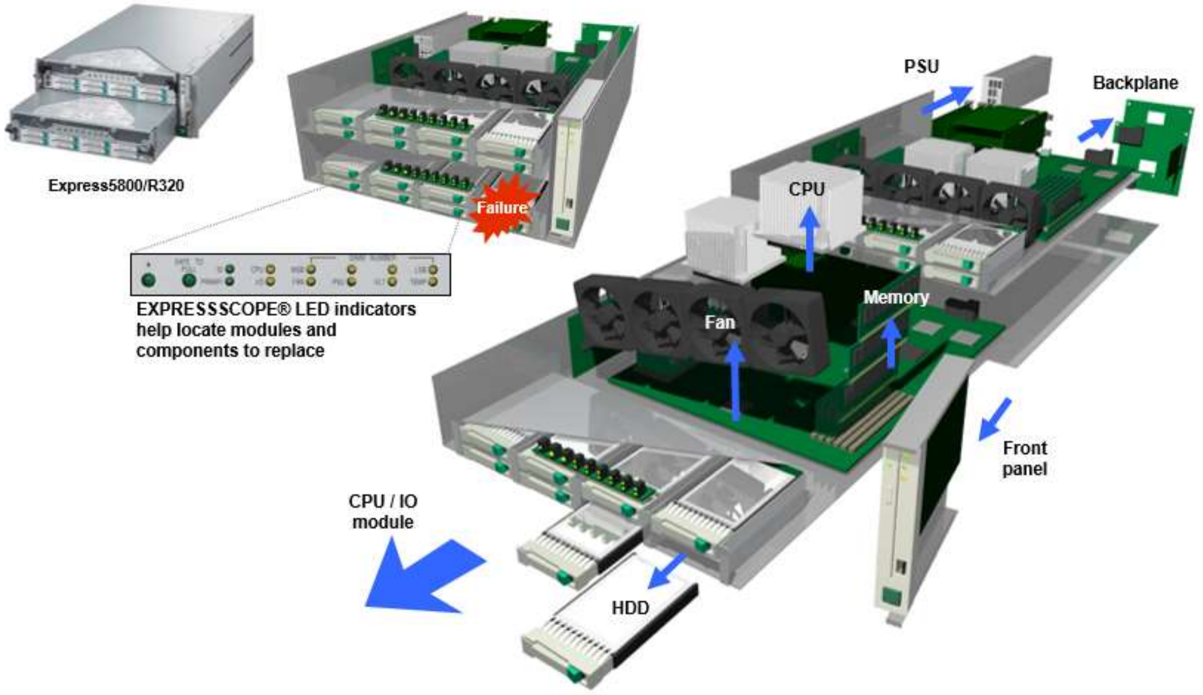
The CPU and memory which form the platform exist in pairs which work synchronously using lockstep technology (details to follow). The operating system does not recognize which set of hardware it is running on and the system sees the hardware as one OS instance.

In a ft series server, all primary components are redundant. When an I/O component fails, the software initiates a failover to the redundant device to continue operation. The CPU and chipset achieve redundancy by hardware technology since the software itself resides on the platform. The CPU and chipset normally work in complete lockstep but in case of a failure, the faulty component is detected and logically isolated by the hardware without interrupting operation. With these functions, ft series servers achieve non-stop operation.

Repair and replacement without interrupting operation



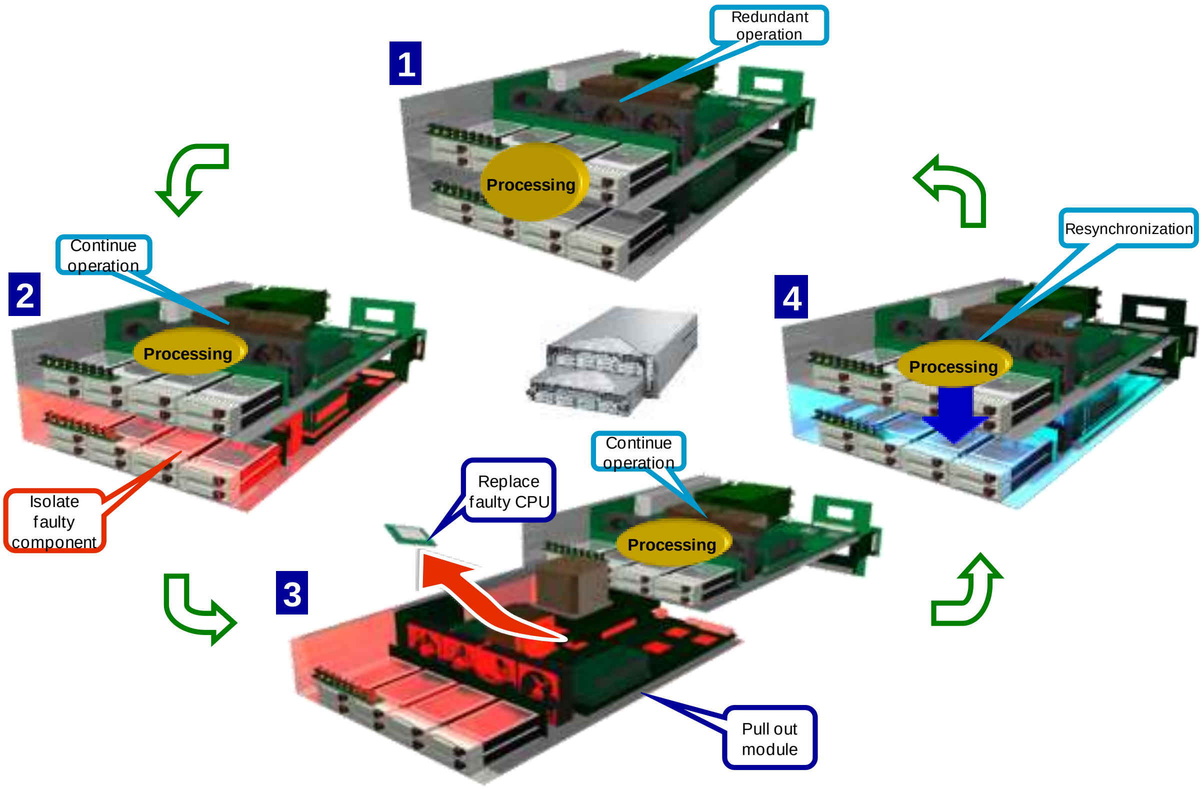
In the structural diagram of the Express5800/R320e/f ft series server on the right, the CPU/IO module, CPU, memory, power supply unit, HDD, backplane, and front panel have red frames indicating that the units are replaceable. With the exception of the backplane[2], their replacement does not interrupt system operation. The fully modular design, as shown in the next illustration, provides ease of replacement. In addition, EXPRESSSCOPE® LED indicators for the CPU/IO modules enable IT staff locate the faulty component at a glance to support swift identification, exchange, and recovery.



[2] As the backplane only includes connectors and wiring, there is small possibility of failure.

The following describes the module replacement process using the example of a CPU fault.

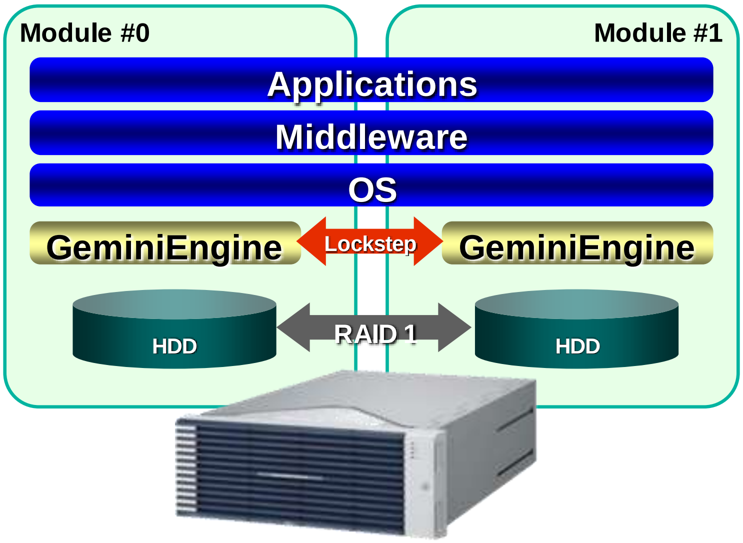
1. Normally, the two modules operate as one redundant server.
2. When a failure occurs (in this example, a CPU failure), the faulty module is isolated by the GeminiEngine™ while processing continues on the other module. (On the faulty module, the amber-lit LED on the EXPRESSSCOPE® indicates the failure.)
3. The faulty module is pulled out and replaced by IT staff while processing continues on the other module.
4. Following the replacement, GeminiEngine™ automatically resynchronizes the pair of modules and redundant operation resumes. (Back to 1)



Compatible with existing operating systems and applications

The first servers to pursue non-stop operation, developed by several manufacturers, provided extremely high availability but were dedicated machines based on special operating systems. Therefore, they were limited to specific areas of use. Following this, cluster systems using Windows® and Linux® were introduced. Since cluster systems covered both hardware and software failures, the technology could support highly available systems. However, the failover process relied on software. In addition, the systems had limited compatibility with applications. Not all applications could run on clusters because of the complexities involved in continuously running the applications on the backup server after failover. Applications not designed for cluster systems typically needed modifications to adapt to the system.

On the other hand, the Express5800/ft series server performs as a single server running a single operating system[3]. This means the system can be configured as normal servers without special consideration for the redundant structure. This simplicity applies to middleware and applications as well—no special settings or modifications are required. Therefore, you can enhance the availability of your entire system by simply replacing your existing servers with Express5800/ft series servers.



Full compatibility with existing operating systems and applications

- ft series servers provide a single-server view on the network, free of any special consideration for the redundant structure.
- Applications do not need redundant settings. There are no limits to applications.
- Same operability as single servers enables easy and low-cost management.

99.999% system availability

Availability is the capability of a system to operate continuously. The terms availability and reliability are similar and often used synonymously but they actually are two discrete notions. Reliability refers to the likelihood of failures or the failure-free interval. Availability refers to the uptime provided by a system which is accessible to users. In general, if a system fails often, both reliability and availability are low. However, in a redundant system, high availability is achieved as long as failures do not affect the operation or the users.

It is often assumed that ft series servers do not fail, but this is not true. Because ft series servers contain the components for two servers, the failure rate is about twice the normal rate. The concept of ft series servers is continuous availability—not to eliminate failures but to enable continuous operation even in the face of failures.

Availability is described in % values.

Availability = Uptime / (Uptime + Downtime)

[3] Models supporting VMware® and Windows Server® 2016,2012R2,2008 R2 Hyper-V™ are capable of running multiple operating systems on a single server with virtualization.

Availability	Annual downtime
99.9999%	32sec
99.999%	5min, 15sec
99.99%	52min, 34sec
99.9%	8hrs, 46min
99%	3days, 15hrs, 36min

NEC Express5800/ft series servers deliver a remarkable 99.999% availability[4]. For a well managed system, this translates into just 5.25 minutes of downtime in a year.

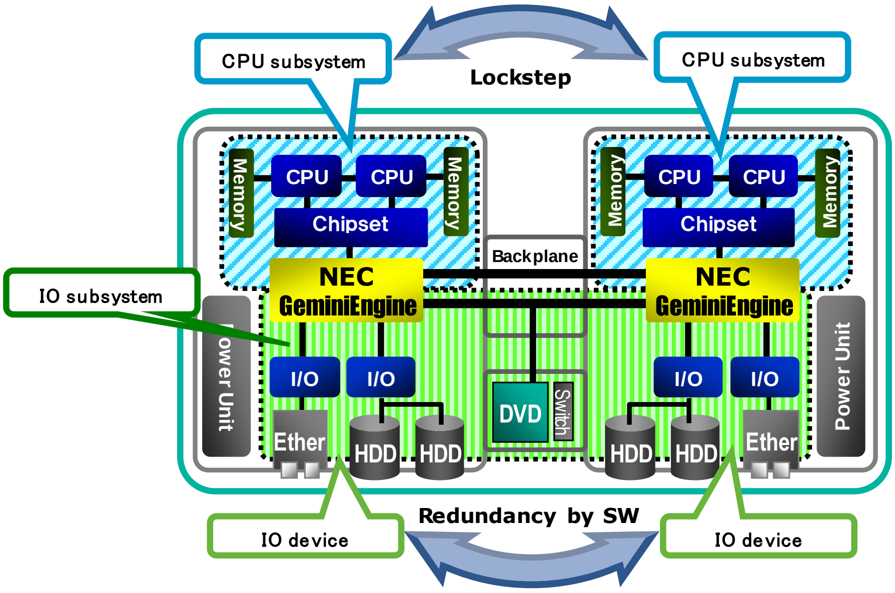
Basic Architecture of Express5800/ft Series Server

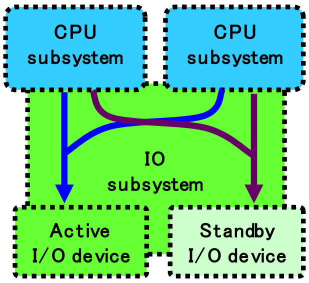
Although various functions are expected of ft series servers, the basic concept is simple—take two sets of hardware and if one set fails and stops, allow the other healthy set to continue without interruption. Among the many hardware and software functions required to realize this concept, the following are the three fundamentals:

- I/O failover
- Surprise removal and concealment of errors
- Lockstep

The CPU/IO modules of the ft series servers adopt different redundancy technologies for the two subsystems—the CPU subsystem including the CPU, memory and chipset, and the IO subsystem including the I/O devices.

The next illustration shows the scope of the two subsystems. The CPU subsystems work identically and the GeminiEngine™ processes the requests from the two subsystems as one request. Therefore, though redundant, it is as if there is only one CPU subsystem in operation. Meanwhile, the IO subsystems are linked via the backplane and I/O devices of





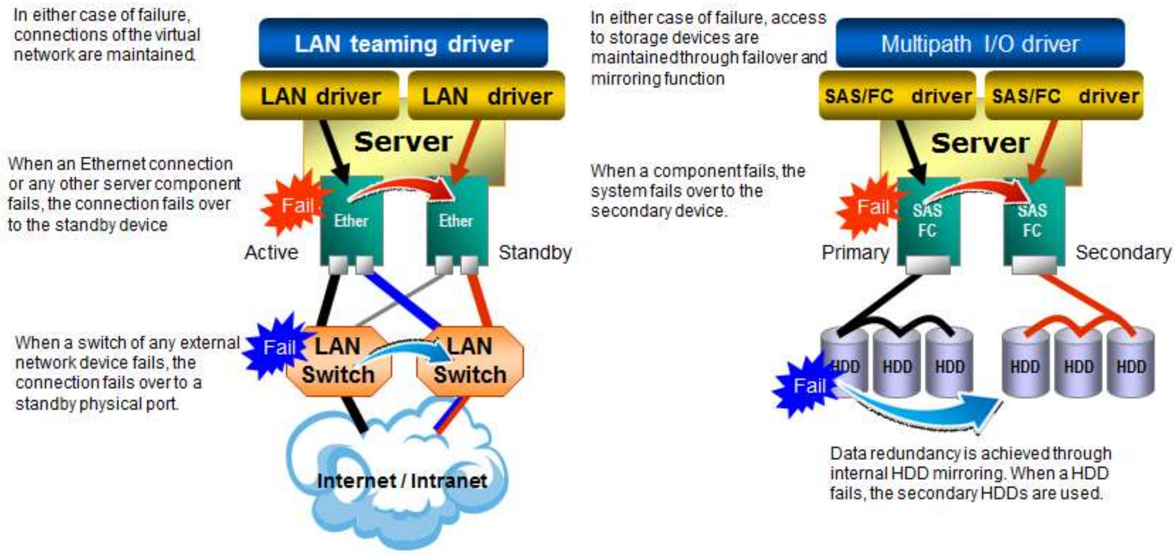
The CPU subsystem can see both I/O devices and therefore SW can control IO redundancy

[4] Theoretical value calculated by NEC. The actual availability of any particular system may differ.

the two modules are acknowledged by the system during normal operation[5]. The two IO subsystems are also acknowledged by the CPU subsystems to enable I/O redundancy control by software.

I/O failover

The two IO subsystems are common in structure and one serves as a standby during normal operations[6]. Software (device drivers) detects any malfunction in the active I/O devices and swiftly fails over operation to the standby subsystem. Failover methods are based on redundancy technologies developed for typical general purpose servers and have been greatly enhanced through the ft series server’s unique hot swappable module structure.



In either case of failure, connections of the virtual network are maintained.

When an Ethernet connection or any other server component fails, the connection fails over to the standby device

When a switch of any external network device fails, the connection fails over to a standby physical port.

In either case of failure, access to storage devices are maintained through failover and mirroring function

When a component fails, the system fails over to the secondary device.

Data redundancy is achieved through internal HDD mirroring. When a HDD fails, the secondary HDDs are used.

Network interfaces use a method called teaming (Windows®) or bonding (Linux®) where a virtual network port consists of multiple physical network ports. When there is a network failure, the connection fails over to another physical port without affecting the virtual network port.

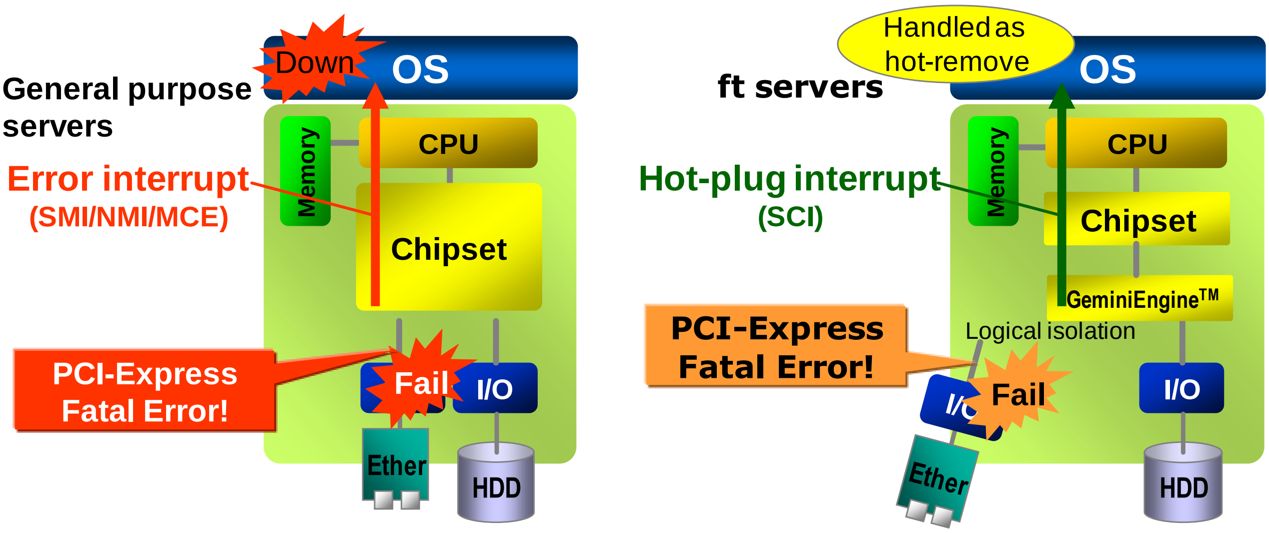
With Fibre Channel and SAS storage systems, I/O failover is achieved through the multipath I/O driver and by mirroring the hard disk drives.

[5] For the R320c,d,e,f models, several devices including the DVD and USB are not redundant to achieve better usability. Failures of these devices do not stop the system and the devices are replaceable without interrupting system operation.

[6] Changes in settings allow simultaneous use of the two I/O devices. However, as I/O performance differs between parallel use and single use, active and standby configuration is recommended.

Surprise removals and concealment of errors

Device failures can lead to unexpected problems and be fatal to the system. A typical failure would be an uncorrectable fatal error in the PCI Express connecting I/O devices and chipset (below on the left). With typical general purpose servers, this would be reported to the operating system as a critical hardware error and cause a system down.



In ft series servers all I/O devices are connected to the GeminiEngine™ and are fully monitored. The same fatal error in the PCI Express as described above is logically isolated by the GeminiEngine™ to hide the device from the system. Furthermore, instead of the true error report, a hot-plug[7] interrupt signal is generated to notify the operating system of a surprise removal meaning that a device was suddenly pulled out. This way the actual hardware error is hidden from the operating system avoiding a system shutdown.

Following this, the operating system notifies the related device driver to initiate I/O failover to the standby subsystem. While surprise removal support is essential to ft series servers, it is an optional hot plug feature of Windows® and Linux® and therefore not supported by every device driver. For this reason, I/O device support of ft series servers is limited when compared with typical general purpose servers.

Currently, supported I/O devices with surprise removal capability are the following:

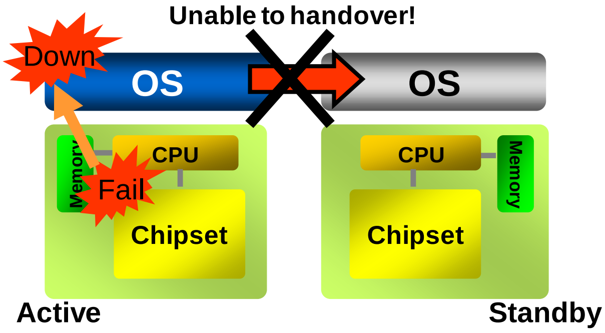
- Ethernet
- SCSI / SAS
- Fibre Channel
- Video display
- USB (Devices are treated as if were removed and reinserted after failover)

Unsuitable drivers without surprise removal support can cause incomplete I/O failover and lead to system failures. Attention is required not to use applications that have a filter to conceal drivers or directly access hardware. For details, please contact your sales representative.

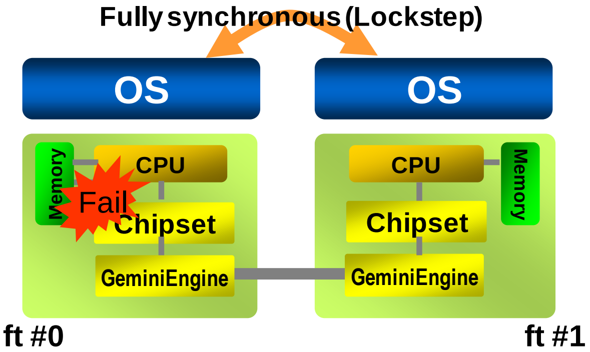
[7] Hot plug is a function that allows the devices as the PCI card and PCI Express card to be pulled out and inserted without shutting down the system.

Lockstep

Lockstep, unique technology developed by NEC, is critical to ft series servers.



In CPU subsystems, the operating system and control software reside on critical components such as CPU, chipset, and memory. Therefore, if a component in the subsystem fails, the operating system stops and all data of the subsystem are lost or ruined. Due to this structure, failover between active and standby devices, as applied to IO subsystems, is not available to CPU subsystems.



Instead, CPU subsystems in the two modules perform in a style called lockstep, working synchronously on a clock-cycle basis. Since the two subsystems perform the same instructions, a subsystem with a faulty component can be logically isolated while continuing operation on the other healthy subsystem. This means the concept of active and standby systems does not exist in CPU subsystems.

Lockstep is a combination of cutting-edge technologies. GeminiEngine™ adopts NEC's hardware developers' various unique and novel ideas to achieve lockstep on the latest hardware. Due to the complexity of the technology, NEC is the only company currently involved in hardware development of the lockstep type FT servers based on Intel® Architecture.

GeminiEngine™ and Hardware Redundancy Technology

GeminiEngine™ is the LSI which enables the core technologies of the ft series server. The main features include:

- Ensure determinism and achieve lockstep
- Synchronize CPU context and memory
- Detect and isolate errors

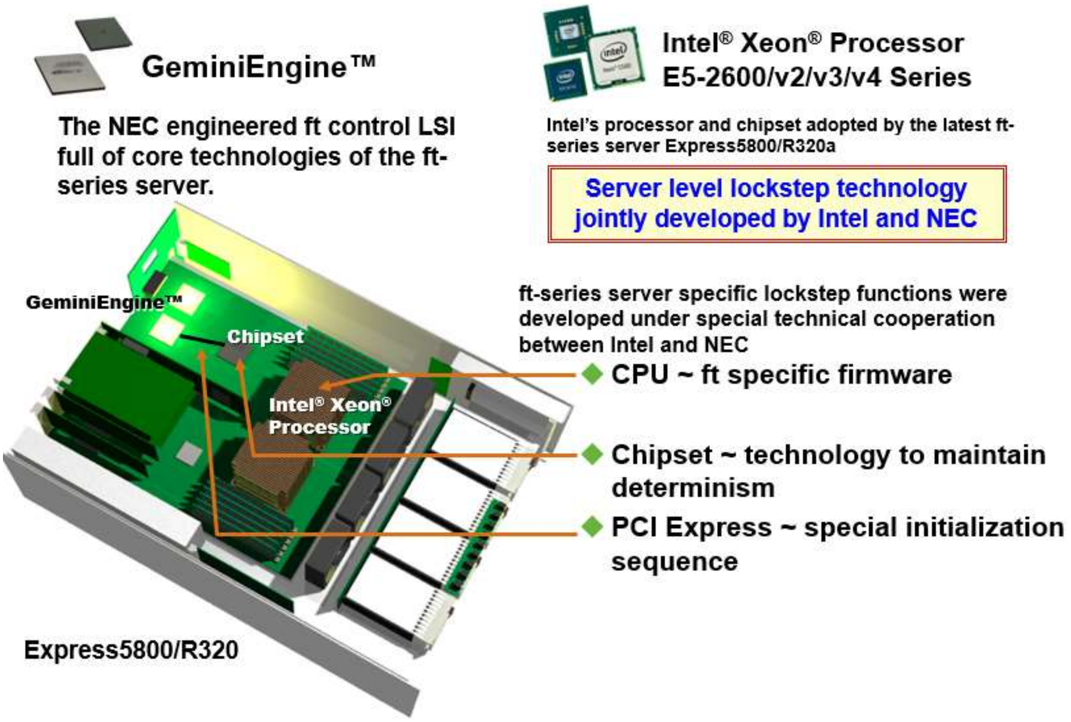
This section examines the details of these three main features of the GeminiEngine™.

Ensuring determinism and achieving lockstep

Lockstep is the key enabler of the ft series server. Here is how it works. Whenever the same clock signal is input and the reset signal is released at the same timing, the LSI's response is always the same; no matter how many times this is repeated. This trait is referred to as "determinism". In this context, when two identical LSI chipsets are applied the same clock signal and started at the same time, with determinism in effect, the two chipsets work synchronously. This status is referred to as lockstep.

In the past, many manufactures engaged in FT server development based on this lockstep principle. However, development of such determinism-based FT technologies has become extremely challenging as component and interface speeds have increased, and use of analog traits has expanded.

Challenging factors for determinism include, for example, high-speed CPU operating frequencies. Another is the analog elements involved in flexible control of CPU operating frequency and voltage to reflect temperatures and power consumption levels. In addition, the mainstream high-speed serial links—inter-chip interfaces such as Intel® QuickPath Interconnect (QPI) with 6.4GHz frequency, high-speed serial I/O interfaces such as PCI Express (5.0GHz)— increase the complexity of the problem as well. In such an environment, deviation of just a few hundred picoseconds[8] can be critical for lockstep. Therefore development of FT servers today calls for extremely advanced engineering skills.



The NEC engineered ft control LSI full of core technologies of the ft-series server.

Intel's processor and chipset adopted by the latest ft-series server Express5800/R320a

Server level lockstep technology jointly developed by Intel and NEC

ft-series server specific lockstep functions were developed under special technical cooperation between Intel and NEC

- CPU ~ ft specific firmware
- Chipset ~ technology to maintain determinism
- PCI Express ~ special initialization sequence

The most demanding lockstep technology for the CPU and chipset was co-developed by NEC and Intel. While ft series servers adopt the same Intel CPUs and chipsets as typical general purpose servers, they cannot function without a special ft series server specific mode in those components to enable lockstep. Therefore, special technology was required to enable lockstep in CPU, chipset, and PCI Express interface for the Express5800/ft series servers.

Clock technologies essential for lockstep which enable phase adjustments and redundant system clocks are being developed in close cooperation with clock chip vendors.

Such cutting-edge, lockstep-enabling technologies developed in collaboration with various component vendors, are incorporated in the GeminiEngine™. In the Express5800/R320e/f, GeminiEngine™ controls the 100 picosecond-level clock phase and reset timing, while ensuring the determinism of CPUs, chipsets and numerous LSI chipsets to achieve lockstep for fault tolerance.

[8] A picosecond is 10^{-12} or 0.000000000001 seconds and is usually abbreviated to psec.

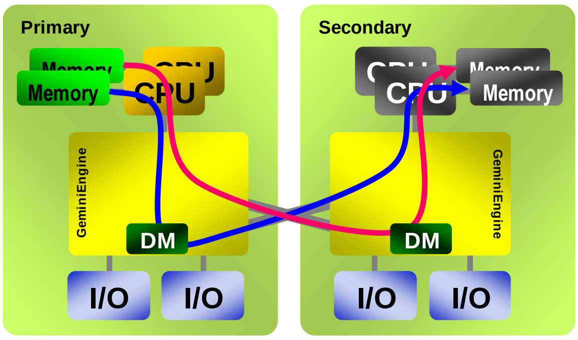
Synchronization of CPU context and memory

For the CPU subsystems to start redundant operation at startup or after board replacement, all memory contents are copied to the secondary module or the module returning to operation. Most of the copying is done in the brownout copy method without interrupting operation[9]. Only the last stage of the process requires blackout copy of the CPU context and cache. The system and services stop for such a short time that it does not affect the system.

Brownout Copy

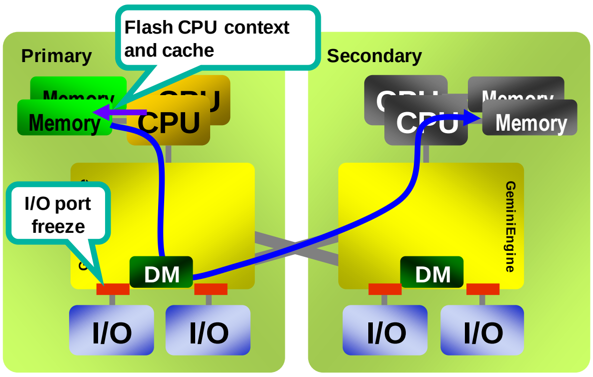
In the illustration, the system on the left is the active module and the right is the module returning to operation. The copy process is executed by the Data Mover (DM) inside the GeminiEngine™ and controlled by ft server control software. In the meantime, the memory contents constantly change as the CPU and I/O devices continue operation. Therefore, whenever a copied memory segment is renewed, the modified part is recopied.

Memory copy by Data Mover



Blackout Copy

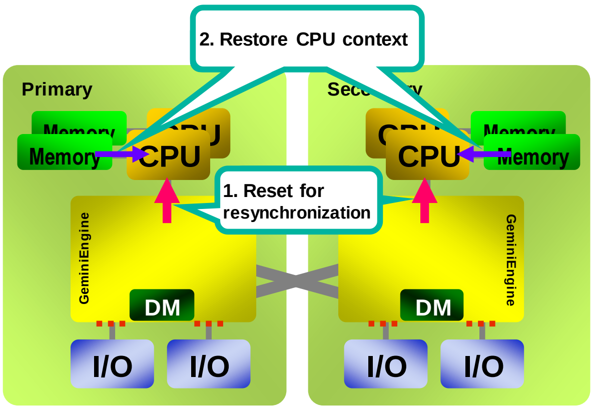
After the brownout copy of the whole memory area is completed, all I/O devices and operating systems are stopped at the hardware level. Then, once the CPU context and cache is flashed to the memory by the ft control firmware, the Data Mover copies the necessary memory content to the module on the right.


Flash CPU context and cache

I/O port freeze

For the modules to operate in lockstep, both CPUs are reset synchronously. This enables the two CPUs to work identically moving forward. Finally, after the CPU context prior to the stoppage is restored, the I/O functions and the operating system return to operation.

The blackout is so short that it does not affect the running services.

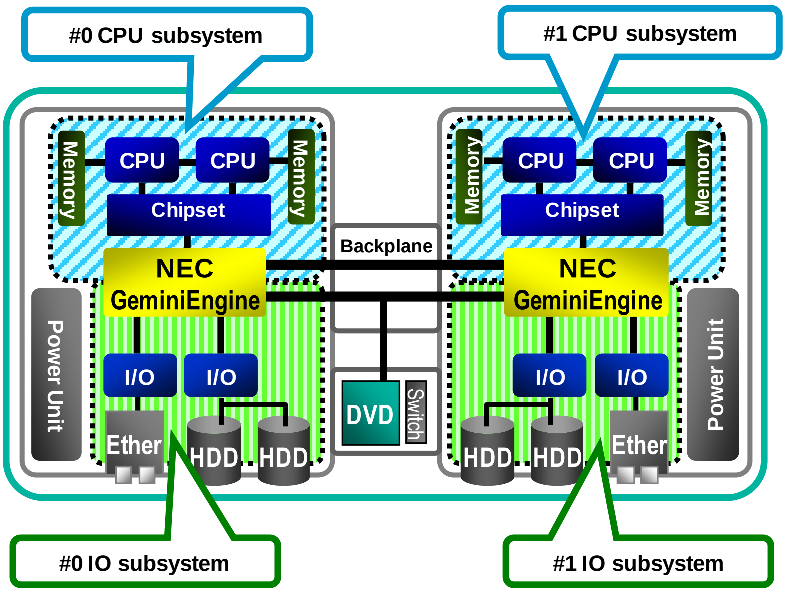

2. Restore CPU context

1. Reset for resynchronization

[9] Windows Server® 2016, 2012R2, 2008 R2 Hyper-V$^{\text{TM}}$ requires blackout copy for the whole memory copy process.

Error detection and isolation

The enhancement of capabilities to detect and locate errors is very important for ft series servers. Furthermore, technology to logically isolate errors is also essential to minimize damage and achieve continuous operation. For this reason, NEC Express5800/ft series servers consist of four subsystems that can be cut off separately when an error is detected.



The GeminiEngine™, positioned across the subsystems, monitors all system transactions and error signals from the chipset. Upon detecting a hardware error, the GeminiEngine™ immediately executes logical isolation of the subsystem. However, this does not indicate that the module with the malfunction needs immediate repair. There are various reasons for a hardware error and not all errors are failures. Primary causes of errors include:

1. Errors caused by faulty components
2. External electrical noise causing transient errors
3. Cosmic rays and other sources of radiation causing memory corruption

While the second largely depends on the operating environment and the third occurs in normal conditions, both are transient errors that are repeated at a certain rate and do not require replacement. Since such errors may happen, ft series servers are designed to return to redundant operation after diagnostic tests are executed on the erroneous subsystem and find no apparent fault. Nevertheless, the possibility of an undetectable failure cannot be eliminated. Therefore, the errors are analyzed for each subsystem and when the count reaches the threshold, the module is taken offline for replacement. Administrators are notified and replacement is suggested by the EXPRESSSCOPE® LED indicators and in error reports.

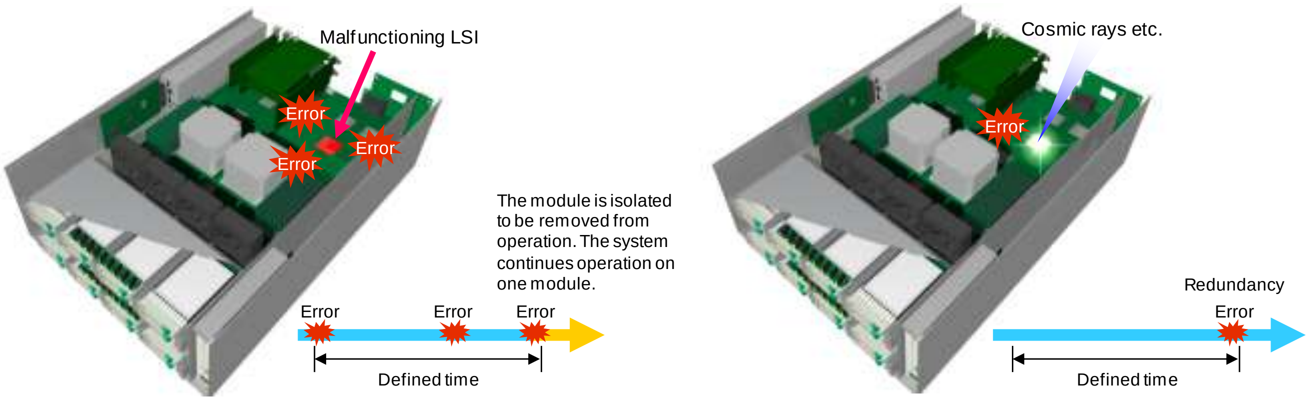
This mechanism is based on the Mean Time Between Failures (MTBF) which is the average elapsed time between failures of a component during operation. For example, if a system's MTBF is 100,000 hours, the system is estimated to fail once in 100,000 hours (about 11.5 years). In general, the MTBF is derived from the calculations made for all components.

In addition to the MTBF factor, the Express5800/ft series server also incorporates unique methods that reflect error frequencies to control the isolation and resumption of modules.

Through such ingenious approaches, NEC's ft series servers are designed to differentiate transient errors from critical errors to maximize the time of redundant operation for higher availability.

Overview of isolation and restoration control

In both cases, the lower module is running non-stop.



The module is isolated to be removed from operation. The system continues operation on one module.

When the number of errors reach the threshold within a given time, the system regards the error as a failure and isolates the module. Meanwhile, operation continues on the other module alone.

When the number of errors do not reach the threshold within a given time, the system regards the error as a transient one and resumes redundant operation.

Conclusion

This paper introduced fault tolerant technologies and the features of the GeminiEngine™ LSI, which form the core of the ft series servers which provide high-availability platforms essential for cloud computing and virtualized server consolidation.

NEC is one of the few companies with expertise in two high availability technologies—clustering technology, which is highly effective for increasing availability of typical general purpose servers, and fault tolerant server technology. NEC is committed to continue leveraging both technologies to provide innovative solutions to address the diverse needs of our customers.

Contents of this material are as of Aug. 2017.