



FUJITSU TECHNOLOGY SOLUTIONS

# PRIMEPOWER SERVERS

## AVAILABILITY *White Paper*

---

THE UNIX  
CHOICE WITH  
DATA CENTER  
CREDENTIALS

---

---

THE UNIX  
CHOICE WITH  
DATA CENTER  
CREDENTIALS

---

# PRIMEPOWER SERVERS

High-availability advantages make PRIMEPOWER servers enterprise-class, data-center-worthy platforms required for business-critical, customer-facing applications.

Powered by SPARC-64 processors, PRIMEPOWER servers run Solaris, the most popular UNIX operating environment in the world, giving customers a choice of more than 12,000 off-the-shelf applications and protecting their software investment. Scalable to 128 processors, PRIMEPOWER servers feature a high-bandwidth channel to memory and the I/O subsystem that maximizes performance and increases transaction speed.

---

April 2001

## CONTENTS

INTRODUCTION . . . . .	1
A BRIEF SYSTEM OVERVIEW . . . . .	2
DATA INTEGRITY AND AVAILABILITY GO TOGETHER. . . . .	2
End-to-End Data Integrity with ECC . . . . .	3
Address Arrays Protected by ECC and Parity . . . . .	3
Additional Protection for Memory . . . . .	4
Redundancy Contributes to High Availability . . . . .	4
Monitoring and Managing the System. . . . .	5
<i>System Control Feature</i> . . . . .	5
<i>Administration</i> . . . . .	5
Operating Isolation Through Partitioning. . . . .	5
Dynamic Reconfiguration . . . . .	6
Automatic System Reconfiguration . . . . .	6
Hot-Swap Capability. . . . .	7
PLATFORM STABILITY . . . . .	7
Stability for Solaris Applications with 100-Percent SPARC-Compliance. . . . .	8
SOFTWARE AND CONNECTIVITY CHOICES. . . . .	8
RELIANT CLUSTER . . . . .	10
SERVICE. . . . .	11

# PRIMEPOWER SERVERS

## AVAILABILITY WHITE PAPER

### INTRODUCTION

*This white paper reviews the characteristics of the Fujitsu® PRIMEPOWER® server that contribute to availability, the degree to which the server can process data when needed. For today's demanding data center environments, availability means the degree to which the server can run applications continuously. The essential points follow:*

- *Server availability.* The PRIMEPOWER server itself provides high availability and data integrity.
- *Platform stability.* PRIMEPOWER is conducive to a stable data center environment, which can contribute to availability.
- *Software and connectivity choices.* PRIMEPOWER can run so many applications and has such a wide range of connectivity options that it can easily be integrated with today's best enterprise, business-continuance, and availability solutions.
- *Clustering.* PRIMEPOWER servers can be integrated into a Reliant Cluster, an optional clustering solution from Fujitsu-Siemens Computers that combines high-availability clustering with application-management clustering in a single integrated product.
- *Service.* PRIMEPOWER is serviced by an organization oriented to and experienced in supporting mission-critical, multi-vendor environments.

Designed and manufactured by Fujitsu Limited, the third-largest computer company in the world, PRIMEPOWER benefits from Fujitsu's long experience providing highly available products for mission-critical mainframe applications. Mainframe standards of availability are increasingly being demanded in the UNIX® market for e-commerce, where UNIX is the dominant operating system, and by mainframe customers who are migrating applications to UNIX.

## A BRIEF SYSTEM OVERVIEW

PRIMEPOWER servers run the Solaris™ operating environment. Based on Fujitsu's SPARC® V.9, Level 2-compliant SPARC64-GP™ processor, PRIMEPOWER servers have the greatest scalability in the Solaris market, scaling from 1 to 128 processors.

The main server building block is the *system board*, each of which has

- Up to four SPARC64-GP processors
- Up to 16 GB of synchronous dynamic RAM (SDRAM)
- Up to six PCI cards, each with its own controller

The largest model, the Model 2000, can have up to 32 system boards, providing

- Up to 128 processors
- Up to 512 GB of SDRAM
- Up to 192 PCI controller cards

System boards and system board components are interconnected by the most advanced crossbar technology available with a bandwidth of up to 57.6 GB/sec, assuring near linear scalability.

The 563-MHz SPARC64-GP processor that is currently used in the Models 800, 1000, and 2000 is built with an 0.18 micron copper process. The processor will be implemented at progressively higher clock frequencies. The larger PRIMEPOWER servers, specifically the Models 800, 1000, and 2000, can be upgraded to these more advanced processors as they become available by hot-swapping system boards.

PRIMEPOWER has a combination of availability, performance, scalability, and Solaris capability that is currently unique in the UNIX market.

## DATA INTEGRITY AND AVAILABILITY GO TOGETHER

Mission-critical applications commonly require both data integrity and continuous system availability. Data integrity—the requirement that data should not be accidentally altered or lost by the processing system—has long been of paramount importance in mainframes, taking precedence even over availability. Fortunately, data integrity and availability can go hand in hand.

PRIMEPOWER provides a superior level of data integrity through the widespread use of error checking and correction (ECC). ECC provides a high degree of confidence that errors, however rare, will not slip through undetected, but rather will be contained and corrected. ECC takes the server from reliable operation to mission-critical levels of confidence. In contrast, many UNIX servers use only parity checking or even no checking at all except for error checking in memory.

---

## THE UNIX

CHOICE WITH

DATA CENTER

CREDENTIALS

---

ECC is superior to parity checking for two reasons.

- ECC provides a higher level of error detection, hence *better data integrity*, than parity. ECC detects both single-bit and double-bit errors, while correcting the single-bit errors. Parity checking offers awareness of errors, but not protection. Parity checking can detect single-bit errors, but lets double-bit errors pass undetected, possibly leading to unusual program failures and to the undetected spread of data corruption affecting database integrity.
- ECC provides markedly *higher availability*, because system failures due to single-bit errors are eliminated entirely in ECC-protected storage areas and buses. Using ECC in memory, for example, reduces unrecoverable failures by 70 to 90 percent over memory without ECC.

The data integrity features of PRIMEPOWER are comprehensive, providing the highest level of data integrity and availability in the Solaris market.

### **End-to-end Data Integrity with ECC**

To maximize data integrity and availability, data is protected by single-bit-correct, double-bit-detect ECC from the point of entrance to the server to the point of exit. All data storage areas are protected.

They include

- Memory
- Processor level 1 cache data
- Processor level 2 cache data
- Crossbar components

All data paths are also protected with single-bit-detect, double-bit-correct ECC. These include

- Level 1 crossbars that interconnect system board components, which are comparable to what in other systems would be called the system buses and level 2 cache buses.
- Level 2 crossbars that interconnect system boards
- PCI buses

### **Address Arrays Protected by ECC and Parity**

In addition to providing ECC for data, PRIMEPOWER provides error checking for addresses. Level 2 cache address tags are protected by ECC. Other processor address arrays, such as level 1 cache address tags and the translation lookaside buffer (TLB), are protected by parity. Parity errors on addresses can usually be retried successfully. The TLB, for example, can be reconstructed by retranslating virtual storage references. Consequently, errors on addresses will rarely cause a system outage.

### **Additional Protection for Memory**

For memory, PRIMEPOWER provides additional protection against uncorrectable errors.

- *Distributed memory modules.* In addition to single-bit error correction, PRIMEPOWER memory corrects for the catastrophic failure of an entire SDRAM module. The bits of a memory module are distributed one bit each to successive storage blocks. In other words, each byte contains one bit from eight different storage modules. If a storage module fails completely, it affects several locations, but all errors are correctable, if not compounded with other errors.
- *Memory scrubbing.* The storage controller uses background cycles to read through memory and clean up single-bit errors. Single-bit errors are corrected in the process of reading memory. The object is to prevent the coincidence of two transient single-bit memory errors compounding in the same block to produce an uncorrectable error.

### **Redundancy Contributes to High Availability**

The data integrity and availability of the processing units are complemented by redundancy in other units.

I/O connections and onboard disks:

- Each system board has six PCI buses, each connecting one PCI controller card. Thus, the server can have up to a maximum of 192 PCI buses, providing many possibilities in the I/O cabling plan for duplicate paths. Consequently, customers can configure PRIMEPOWER so that the failure of any single PCI bus will not cause a system failure.
- Similarly, the PCI controllers allow alternate paths to I/O devices.
- Disks installed in the cabinet support multiple RAID types and software mirroring.

Power and cooling:

- Front-end power supplies, which distribute power to the rest of the server, have N +1 redundancy.
- Dual power sources are supported, allowing connection, for example, to power sources from separate power grids or substations.
- Uninterruptible power supplies can also be attached.
- Cooling fan trays are duplicated, mounted one over the other. Each fan tray carries three high-performance dual-speed fans. When a fan tray fails or is removed for maintenance, the fans in the other fan tray increase their speed.

Server control and monitoring facilities:

- The System Control Feature (SCF), which is integral in monitoring and managing the system, can be configured with a redundant board.
- The network inside the cabinet connecting the SCF to the boards and components that it monitors is duplicated. If the primary network fails, the administrator can switch to the alternate.
- The connection to the system console can also be duplicated.
- The Models 800, 1000, and 2000 have a second system clock. Rebooting of the server is required to use the alternate clock.

## Monitoring and Managing the System

### System Control Feature

The SCF provides administrators and support personnel visibility into and control over server operation, allowing them to anticipate and respond to error conditions. The SCF monitors the functioning of processor, memory, and I/O components, fan rotation speed, power supply functioning, and environmental factors such as temperature and humidity. Through its Remote Cabinet Interface (RCI) it can also monitor and control external disk cabinets. The SCF collects and aggregates information about intermittent errors. When errors exceed a threshold, the SCF will alert the administrator, and the part can be scheduled for maintenance. With an integrated service processor, the SCF functions independently of the rest of the server, enabling support personnel to gather information even if the server is not operational.

### Administration

Simple, consistent, familiar administrative tools that integrate well with other tools make server administration less prone to error. PRIMEPOWER servers are administered through an easy-to-use Web-based tool, WebSysAdmin. Written in JAVATM with both a GUI and a simple command-line interface, the WebSysAdmin application can run on any PC with a browser. It has been qualified for use with both Microsoft Internet Explorer and Netscape Communicator. One WebSysAdmin PC can administer multiple servers and multiple partitions in a server.

WebSysAdmin is also used to administer related functions, such as Reliant Cluster, peripheral devices, and other systems and will be adapted to support future servers as well. WebSysAdmin can also be integrated with data center system management functions such as Tivoli, CA-Unicenter, and HP-OpenView.

## Operating Isolation Through Partitioning

PRIMEPOWER Models 800, 1000, and 2000 provide for the segregation of server resources into partitions, which are independent processing environments with their own dedicated processors, memory, and peripherals.

- Each partition runs its own copy of Solaris and is completely independent of other partitions—operations in one partition do not affect operations in others.
- Partitions are divided at system board boundaries—all of the processors, memory, and PCI controllers on the system board are assigned to the partition.
- Multiple partitions also allow *multiple generations* of SPARC64-GP processors to coexist in the same server. Within a given partition, all processors must run at the same clock frequency, but different partitions may operate at different frequencies.
- The Model 800 supports up to 4 partitions, the Model 1000 supports up to 8 partitions, and the Model 2000 supports up to 15.

A partition provides logical and physical *operating isolation* from other partitions, allowing mission-critical applications to be separated from less-critical applications. Separate partitions may be set up for unstable environments, such as development and testing, or administrator training, without impacting the availability of critical applications.

## Dynamic Reconfiguration

Two types of partitioning are supported. *Static partitions* are configured when the server is booted and the resource boundaries among partitions remain fixed. *Dynamic reconfiguration* allows the system administrator to add or remove system boards, with their processors, memory, and PCI controllers, from an active partition—Solaris does not have to be rebooted in the partition. With dynamic reconfiguration the resource boundaries of a partition are adjustable, creating *dynamic partitions*. Dynamic reconfiguration requires Solaris 8.

Dynamic reconfiguration makes it possible to make configuration changes without system outages, thus improving overall availability. It has the following typical uses:

- *Resource balancing.* System boards can be moved to partitions with applications requiring peak load resources temporarily and removed later.
- *Expanding the configuration.* A new system board can be hot-swapped in and added to a partition's configuration.
- *Upgrading a system board or online repair of failed components.* The administrator can idle a system board and remove it from its partition's configuration; a service representative hot-swaps the board out and a replacement board in; then the administrator can add it back to the partition's configuration.

## Automatic System Reconfiguration

Dynamic reconfiguration occurs at the request of the system administrator. Automatic System Reconfiguration (ASR) is a part of the boot sequence. When the server powers on, the power-on self-test (POST) automatically runs diagnostics on all of the server's active hardware components, including crossbars, I/O interfaces, and the components of each system board, as well as on auxiliary components such as fans, power supplies, and any attached uninterruptible power supplies. Similarly, when a partition is booted, diagnostics are run on that partition's own configuration.

If a hardware error causes a partition to fail—implying a kernel failure—the server will typically log information about the failure in its nonvolatile RAM. When the partition is automatically restarted or “bounced,” POST analyzes this information to determine which hardware component is the probable cause of the failure and supplements its routine *comprehensive* diagnostics with *exhaustive* diagnostics targeting that particular component.

If a component fails these diagnostics, ASR logically removes it from the active hardware configuration to avoid subsequent system failures. ASR also attempts to reconfigure around the loss of certain critical components. For example, it will employ an alternate path, if available, to circumvent the loss of a boot drive's SCSI I/O path. When ASR removes a component, Solaris runs without it until it is replaced. If the model supports dynamic reconfiguration, it may be possible to replace the defective component without rebooting the partition.

### Hot-Swap Capability

Hot-swapping provides the physical ability to alter the configuration and have the changes recognized while the remaining components of the server continue operation. The following system components can be hot-swapped:

- System boards (containing processors, memory, and PCI controllers)
- SCF boards
- Onboard disks (requires software mirroring if data is not to be lost)
- Built-in I/O: CD-ROM, floppy disk
- Environmental monitoring board
- Front-end power supplies
- Cooling fan trays
- Air filters

## PLATFORM STABILITY

PRIMEPOWER servers provide the most flexible upgrade opportunities and the largest range of configurations in the UNIX market with the following features:

- PRIMEPOWER servers can incorporate successive generations of processors.
  - Models 800, 1000, and 2000 can be upgraded to *faster processors* simply by hot-swapping installed system boards with boards mounting faster processors.
  - In the Models 800, 1000, and 2000, *multiple processor generations* can coexist in the same server in separate partitions by virtue of their partitioning features. Clock frequencies are the same within each system board and must also be the same within each partition. In the case of partitioned servers, some partitions can continue to operate while other partitions are upgraded.
- PRIMEPOWER servers also offer an unusually large range of scalability.
  - 1 to 128 processors
  - Up to 512 GB of memory
  - Up to 192 PCI controllers

These capabilities protect and enhance the customer's investment, but also provide for a stable long-term platform that can change by small increments. This stability in itself can contribute to availability.

### **Stability for Solaris Applications with 100-Percent SPARC-compliance**

The SPARC64-GP is 100-percent compliant with the specifications and interfaces of SPARC Architecture V.9, Level 2. The SPARC International Consortium, which defines SPARC specifications and certifies SPARC processors to those standards, has certified the SPARC64-GP in the same way as Sun's UltraSPARC processors. All applications that run on Sun's UltraSPARC processors will be binary-compatible with SPARC64-GP-based servers.

At a given system level, Solaris copies are identical for all Solaris customers. When Solaris is built, the model-independent Solaris functions are combined with model-dependent code for all platforms on which Solaris runs, including those for PRIMEPOWER, then tested for compatibility and quality as a single integrated product, and shipped to all Solaris customers regardless of their specific hardware platforms.

Consequently, Solaris applications can be transferred to PRIMEPOWER *without porting*, avoiding the instability that can result from modifying applications. Furthermore, PRIMEPOWER performance is available to these applications without recompiling or optimizing. Not only is the SPARC64-GP designed specifically for the Solaris environment, but the SPARC64-GP also provides, in effect, dynamic code optimization with its deep out-of-order execution engine that can keep up to 63 instructions pending while they wait for their operands.

## **SOFTWARE AND CONNECTIVITY CHOICES**

By virtue of Solaris compatibility, more than 12,000 commercial and industrial applications are available for PRIMEPOWER customers. PRIMEPOWER customers may choose those applications best suited to their availability needs from Baan, BEA, BMC Software, Computer Associates, EMC, Informix, JDEdwards, Oracle, SAP, Veritas, and many other vendors.

PRIMEPOWER offers a similar scale of connectivity choices, with both a large number of I/O connections and a wide variety of interconnection protocols.

- PRIMEPOWER servers provide connection through industry-standard PCI cards, rather than through a proprietary bus.
- I/O connectivity scales up with the number of processors, reaching 192 PCI controller cards in the largest model.
- Each system board has six PCI buses, allowing up to 192 PCI buses.

---

**THE UNIX  
CHOICE WITH  
DATA CENTER  
CREDENTIALS**

---

- PCI controllers are available for the following I/O protocols:
  - Fibre Channel, both switched and arbitrated loop
  - Single channel differential UltraSCSI
  - Dual channel differential UltraSCSI
  - Fast Ethernet
  - Quad Fast Ethernet
  - Gigabit Ethernet 2.0
  - FDDI/P Dual Attach 2.0
  - ATM-155/Mfiber
  - ATM-155/UTP
  - ATM-622/Mfiber
  - High Speed Serial Interface 2.0
  - Serial Asynchronous Interface

These connectivity options enable customers to complement PRIMEPOWER's own high availability with high-availability storage and communications networks. A true storage area network (SAN) can be built using switched Fibre Channel attachments as well as hybrid SANs with SCSI, Fibre Channel, and SCSI/Fibre Channel bridges. The most highly rated backup and archiving solutions are available, including Networker backup software and a wide range of tape devices and robotics libraries, as well as FMWORM archiving software and the PXM2 Optical Disk Library family. PRIMEPOWER customers can also incorporate EMC products such as Timefinder and Symmetrix Remote Data Facility (SRDF).

With this range of application and connectivity options, PRIMEPOWER customers have access to today's best enterprise, business continuance, and availability solutions.

## RELIANT CLUSTER

The high availability that PRIMEPOWER servers bring to applications can be further enhanced with a high-caliber clustering system—Reliant Cluster (RC), an optional clustering solution from Fujitsu-Siemens Computers. RC is both a high availability (HA) clustering solution and a performance-oriented application-management clustering solution in an integrated product.

RC is a versatile and robust HA cluster solution.

- RC supports up to 64 cluster nodes, far more than other clustering solutions. Clustering solutions that use a disk drive as a quorum device to determine which server has control face an upper bound on the number nodes that is set by the number of available SCSI IDs. RC avoids this limitation by using an Ethernet connection to determine which server has control. RC also avoids the “split brain” problem that can result when using a disk to mediate which server has control.
- RC is currently the only cluster solution that allows failover to another partition on the same server.
- RC supports failover for individual applications as well as systems.
- RC supports multidirectional failover—applications can failover to multiple servers.
- Server configurations need not be identical. Larger servers can failover to smaller servers. Customers may determine whether all applications will continue to run at reduced performance, or whether only the more critical applications will continue to function.
- Failover is supported for heterogeneous servers. Applications can failover to Solaris, Linux, or Reliant UNIX environments. In practice, the key consideration for failover to a heterogeneous server is one of data compatibility between the two systems.
- Failover is supported to distant locations, making RC a candidate for some disaster recovery strategies. Failover to distant servers is essentially a matter of how far apart the disks can be. PRIMEPOWER’s I/O connection options give users many ways to configure for failover. In particular, Fibre Channel allows disk or SAN connections up to 10 km. Failover clusters have been constructed using RC at even greater distances using other connection strategies.
- RC is certified to run Oracle Parallel Server.
- RC failover has been proven in hundreds of installations—RC is widely used in Europe.

As an application management tool, RC provides a way to start and stop applications consistently. This consistency results in fewer operational problems and is especially needed in today’s increasingly operator-less environments. RC allows users to manage applications in the way that they wish, allowing standardization on customer procedures rather than preset vendor preferences.

RC manages application dependencies using hierarchical trees of resources—for example, to run, an application needs a disk, a file system, a mount point, etc. RC uses software “detectors” to monitor application resources (including PRIMEPOWER system components), switching the application to a different node if any of these resources becomes unavailable. By distributing individual applications among other nodes (in response to component failure or administrative directive), RC promotes the efficient use of the total cluster resources.

RC is administered through the same WebSysAdmin tool as the PRIMEPOWER server. WebSysAdmin can indicate the status of monitored resources, thereby providing early detection of a fault.

---

**THE UNIX  
CHOICE WITH  
DATA CENTER  
CREDENTIALS**

---

## **SERVICE**

Fujitsu Technology Solutions supplements the outstanding availability of PRIMEPOWER servers with a genuine understanding of the computing enterprise. Fujitsu and its subsidiaries have had long experience with business continuance issues from the data center perspective and many years of experience installing and servicing high-end Solaris servers internationally. Cross-trained on PRIMEPOWER, Solaris software and administration, clustering, and many peripheral products, the PRIMEPOWER service organization is both used to and prepared for multi-vendor environments. Fujitsu Technology Solutions offers a proactive approach to customer issues, responses to problems that are suitable to mission-critical applications, and management of problems through to resolution.

Multiple levels of service offerings allow customers to tailor service to their needs cost-effectively. Consulting services are also offered in association with internationally known consulting firms.



FUJITSU TECHNOLOGY SOLUTIONS

**Headquarters**

Fujitsu Technology Solutions, Inc.

1250 East Arques Avenue

P.O. Box 3470

Sunnyvale, CA 94088-3470

United States of America

**Tel:** 877 213 6674

**Sales:** 877 905 3644

**Fax:** 408 746 6595

**Internet:** [www.fujitsu-technology.com](http://www.fujitsu-technology.com)

**Canada**

Tel: 416 510 3111

Fax: 416 510 3353

Fujitsu, the Fujitsu logo, and PRIMEPOWER are trademarks or registered trademarks of Fujitsu, Ltd. UNIX is a registered trademark in the U.S. and other countries, licensed exclusively through X/Open Company Limited. Solaris and JAVA are trademarks or registered trademarks of Sun Microsystems, Inc. SPARC is a registered trademark and SPARC64-GP is a trademark of SPARC International, Inc. Products bearing the SPARC trademark are based on Architecture developed by Sun Microsystems, Inc. All other trademarks and product names are the property of their respective owners.

The information in this document may be superseded by subsequent documents. For details regarding delivery of specific products, features, and services, contact your local sales representative.

© 2001 Fujitsu Technology Solutions, Inc.

All rights reserved. Printed in the U.S.A.

MM003047-US-002 [1:35] 4/01