



PRIMEPOWER Server Architecture Excels in Scalability and Flexibility

September 2005

PREPARED FOR

Fujitsu

TABLE OF CONTENTS

Executive Summary	1
Performance.....	1
Scalability.....	2
Benchmark Evidence	4
Flexibility	5
RAS (Reliability, Availability, and Serviceability)	6
Review of Key Points	6

ABOUT THIS DOCUMENT

This paper offers an overview of the key architecture-related features of PRIMEPOWER servers. It is one in a series of six white papers that highlight PRIMEPOWER hardware and software capabilities from a customer-benefits perspective. Additional information concerning PRIMEPOWER's unique architectural features can be found on the Fujitsu websites: www.fujitsu.com/global/services/computing/server/unix/.

Executive Summary

Fujitsu's history of developing high-end computing systems stretches back over fifty years. Among the variety of systems Fujitsu continues to offer throughout the world, its SPARC64 V-based PRIMEPOWER series is the most widely known. Leveraging design expertise gained from its high-capacity mainframes and supercomputers, Fujitsu has developed high-end SPARC/Solaris servers with architectures that offer unique advantages in performance, scalability, flexibility, and RAS (reliability, availability, and serviceability).

Performance

Processor chips form the foundation of high-performance server designs. As outlined in the companion white paper covering processors, Fujitsu was a founding member of the non-profit organization that oversees the open-standard SPARC specification. Compliance with that architectural specification ensures that applications written for SPARC will run on the various implementations.

Even though the instruction set is well defined, processor designers retain the freedom to optimize their chip implementations. Fujitsu has leveraged the large-scale system expertise of its design teams to incorporate capabilities in its SPARC64 V processors that enhance the performance and RAS characteristics of its PRIMEPOWER servers.

Within the SPARC64 V processor, advanced out-of-order superscalar techniques accomplish more work per clock cycle than alternative implementations. Large on-chip caches help to keep the execution units supplied with data; and to keep those caches filled, the SPARC64 V provides a fast bus to memory. By using fast Double Data Rate (DDR) memory, PRIMEPOWER designers are able to satisfy demands for large amounts of data and achieve the high-performance potential of SPARC64 V processors.

To interconnect these powerful processors into a large SMP server, PRIMEPOWER engineers leveraged Fujitsu's supercomputer interconnect skills and devised a high-bandwidth, low-latency crossbar. This crossbar design offers scalable performance. That is, interconnect capacity grows as processors are added, so that performance is maintained from small servers all the way up the scale to very large server configurations.

This document is copyrighted © by Ideas International, Inc. (IDEAS) and is protected by U.S. and international copyright laws and conventions. This document may not be copied, reproduced, stored in a retrieval system, transmitted in any form, posted on a public or private website or bulletin board, or sublicensed to a third party without the written consent of IDEAS. No copyright may be obscured or removed from the paper. All trademarks and registered marks of products and companies referred to in this paper are protected.

This document was developed on the basis of information and sources believed to be reliable. This document is to be used "as is." IDEAS makes no guarantees or representations regarding, and shall have no liability for the accuracy of, data, subject matter, quality, or timeliness of the content. The data contained in this document are subject to change. IDEAS accepts no responsibility to inform the reader of changes in the data. In addition, IDEAS may change its view of the products, services, and companies described in this document.

IDEAS accepts no responsibility for decisions made on the basis of information contained herein, nor from the reader's attempts to duplicate performance results or other outcomes. Nor can the paper be used to predict future values or performance levels. This document may not be used to create an endorsement for products and services discussed in the paper or for other products and services offered by the vendors discussed.

From a customer perspective, processor performance and the number of processors in a server are key attributes. CIOs/CTOs and other IT executives view robust servers with large numbers of high-performance processors as allowing more work to be done with fewer systems. The reduced floor space requirements and lowered complexity save both facility and personnel costs in operations and support.

Scalability

PRIMEPOWER systems offer unmatched scalability, all the way to a 128-processor SMP. Competitive systems typically scale to about half as many processors, or degrade I/O performance by replacing PCI cards with auxiliary processors. PRIMEPOWER's balanced design, on the other hand, scales up the number of processors, memory capacity, and I/O capabilities without tradeoffs.

The maximum size is not required at every site for all servers. The customer can choose the configuration that best fits planned requirements from the variety of PRIMEPOWER models offered. At the same time, the customer enjoys the assurance that the system can be expanded to meet demand. Larger SPARC64/Solaris systems are available as the computing workload increases.

Table 1 below highlights existing and newly introduced PRIMEPOWER models. When announced in 2000, the PRIMEPOWER family scaled from two-way models (Model 200) to the maximum 128-way Model 2000, and used SPARC64 GP chips that varied in clock rates from 250 MHz to 450 MHz. Since that time the processor clock speed has been improved a number of times, has led to the introduction of newer models with faster clock speeds and buses.

Table 1: PRIMEPOWER Models

Model No.	Max. No. of Processors	Announced 2003	Announced 2004	New in 2005
		SPARC64 V	SPARC64 V	SPARC64 V
250	2	1.1 GHz	1.32 GHz	1.87 GHz
450	4	1.1 GHz	1.32 GHz	1.87 GHz
650	8	1.08 GHz / 1.35 GHz	1.89 GHz	2.02 GHz
850	16	1.08 GHz / 1.35 GHz	1.89 GHz	2.02 GHz
900	16		1.89 GHz	2.16 GHz
1500	32		1.89 GHz	2.16 GHz
2500	64/128		1.82 GHz	2.08 GHz

In 2001, Fujitsu redesigned the system boards for its new models – the 650, 850, 900, 1500, and 2500 – to double the number of processors per board from four to eight. This density improvement permitted a smaller footprint. In addition, it provided a more balanced split of the interconnect traffic handled by each of the system boards and that handled by the backplane.

The rebalancing yielded a more efficient interconnect that supported faster processors in the newly introduced models compared with prior models. The newest PRIMEPOWER models reconfirm that designing for performance and scalability go hand in hand.

Building upon this implementation experience, Fujitsu has just released new PRIMEPOWER models that use a newly designed eight-processor board that supports the fast 2+ GHz SPARC64 V chips. Interconnecting these fast processors and memory is an on-board switch that transfers data at an impressive 540 MHz using a source-synchronous clocking technique. Although IT executives need not be concerned about the engineering details of the high-speed data transfer mechanism, they need to know that this technique will allow even faster rates in the future.

Since the 2+ GHz processors with enhanced instruction-level parallelism offer much more performance than the prior SPARC64 chips, more PCI slots are needed to balance the system properly. In fact, the newest models offer approximately twice the PCI slots per processor compared to the original PRIMEPOWER design. They provide up to 320 PCI slots in the top-of-the-line Model 2500. Table 2 below highlights some of the capacities of the new models.

By packaging eight processors per board rather than the previous four, the system configuration requires a smaller footprint. For the 128-way configuration, Fujitsu has also been able to cut the number of cabinets in half, slashing floor space requirements and eliminating the cabling that used to be required to join the sections of the interconnect switch. (Since cable connections are a source of errors, eliminating physical cables enhances overall system reliability.)

Table 2: Capacities of High-end PRIMEPOWER Models

Model	Processors	Memory	PCI Slots	Storage Bays
900	16	128 GB	36	16
1500	32	256 GB	72	32
2500	128	512 GB	320	128

By interconnecting more processors on each board, compared to prior implementations, the design of the inter-board backplane can be streamlined. Indeed, the reduced size of the redesigned backplane switch, coupled with some advanced clocking techniques, produces a very fast 520 MHz transfer rate for the PRIMEPOWER system interconnect used in the 2500 model. (PRIMEPOWER 900/1500 supports a 270 MHz system clock.)

IT executives appreciate high speeds and large capacities that help deliver high performance. However, the “bottom line” is whether such technology will allow them to get more of their computing workload done with less expense. Although the ultimate proof requires running the customer’s unique workloads in a true production environment, a good approximation of the performance potential can often be gained from industry-standard benchmarks.

Benchmark Evidence

Since benchmarks run on existing PRIMEPOWER models have demonstrated very impressive results, one can expect that the new models will ensure that PRIMEPOWER remains a leader in UNIX/RISC benchmarks. Table 3 offers a list of industry-standard benchmarks for which PRIMEPOWER set new performance records.

All of the results listed in Table 3 were record breaking at the time the measurements were published. For some of the application benchmarks, PRIMEPOWER still remains the top performer.¹ Note that PRIMEPOWER has upgraded its processors and the results in Table 3 do not reflect the latest 2+ GHz SPARC64 V chips.

Vendors prefer to avoid the time and expense of remeasuring benchmark tests whenever processors are upgraded. Thus, leadership on a particular benchmark will change as new systems are measured. PRIMEPOWER has held leadership for long periods, despite measurements that reflect the performance of older chips. The latest SPARC64 processors are much faster.

Table 3: Record-Setting PRIMEPOWER Benchmark Results

Benchmark	PRIMEPOWER No. of Processors (Speed)	Submitted Date	Comments
SAP SD 2-Tier	128, (2.08 GHz)	3/05	Still # 1
SAP ATO 2-Tier	128, (563 MHz)	5/01	Still # 1
SPECjbb 2000	120, (2.08 GHz) 128, (1.82 GHz)	1/05	Still # 1
SPECjAppServer 2002	56, (1.32 GHz) DB 9 x 4 (1.32 GHz) Client	9/04	Still # 1
SAP IS-U/CCS 2-Tier	16, (1.35 GHz)	3/03	Still # 1

Most of the benchmarks, such as the SAP variants and SPECjbb2000, represent general commercial computing applications and are familiar to IT executives. Each of the benchmarks stresses a different combination of processing power, memory, and I/O capability. The excellent showing of current PRIMEPOWER models indicates that the new, higher-performing PRIMEPOWER will once again attain leadership positions on industry-standard measures.

The PRIMEPOWER Model 2500 facilitates such high-end clustering by adding special-purpose instructions that help synchronize operations among clustered PRIMEPOWER systems. Up to 128 systems, each with 128 processors, can be linked into a truly massive computing powerhouse that can attack the largest scientific problems. Although most PRIMEPOWER users will not need such very large-scale performance clustering, this extension demonstrates how Fujitsu has leveraged its unique heritage of supercomputing and mainframe expertise throughout the PRIMEPOWER range.

¹ The benchmark rankings reflect results submitted as of August 2005. For the actual measured results, and any recent updates, refer to the respective benchmark websites, or the organization certifying those results.

Flexibility

The modular building blocks that help construct scalable systems can also offer configuration flexibility. Each of the PRIMEPOWER's system boards contains processors, memory, and connections to PCI I/O – essentially a self-contained server on each system board. Independent partitions can be created within PRIMEPOWER by instructing the backplane crossbar switch to isolate sets of system boards from each other. Each independent partition runs a different instance of the operating system and is electrically isolated from hardware or software instances in other partitions.

Note that system boards with different processor speeds can be joined into the same partition. For example, a PRIMEPOWER server could be procured with a subset of the maximum system board configuration to allow for future growth. By the time that server needs to be upgraded, faster speed processors may have become available. By allowing mixed processor speeds within a partition, Fujitsu provides investment protection. The original speed processors do not need to be replaced even when adding newer, faster processors to the server.

The ability to carve out partitions within a larger server is highly valued by IT organizations for server consolidation environments, software development, testing of new application release levels, and a variety of other uses. The Solaris Operating System (discussed in other white papers in this series) supports dynamic reconfiguration. Dynamic reconfiguration allows system boards to be moved from one partition to another without rebooting the entire PRIMEPOWER system.

Dynamically reconfigurable partitioning is an important PRIMEPOWER feature and is highly desired by IT organizations. However, reconfigurable system board partitioning is not unique to PRIMEPOWER. Similar capabilities can be found on other SPARC/Solaris implementations. Other RISC/UNIX vendors also have their own versions of partitioning that offer many of the same benefits. But the PRIMEPOWER models offer a unique feature: partitioning within each system board, which Fujitsu calls Extended Partitioning (XPAR).

Consider how partitions typically work. The electrically enforced isolation fundamental to hardware partitioning ensures that faults in one partition do not affect other partitions. However, the granularity of hardware partitioning has typically aligned along physical packaging boundaries, such as an entire system board with all its processors and memory.

Software partitioning by contrast, usually allows individual processors to be assigned to different partitions, but it depends on a software layer (not electronic isolation) to attempt fault containment. Customers trust the isolation provided by hardware partitioning, but often find that the granularity of a full system board is too coarse.

With XPAs, Fujitsu has retained the rigorous isolation of hardware partitioning but has also allowed a reconfigurable granularity of less than a full system board. Just as the backplane crossbar switch can be designed to isolate the different system boards connected to the backplane, so too can the on-board crossbar interconnect be designed to isolate the different processors that connect to that switch. On the

PRIMEPOWER models, each eight-processor-capable system board is subdivided into four XPARs (PRIMEPOWER900/1500) or two XPARs (PRIMEPOWER 2500).

By extending hardware isolation to the system board interconnect, Fujitsu has blended the best characteristics of hardware and software partitioning. As implemented in the PRIMEPOWER 900 and 1500 models, XPAR allows hardware partitions to consist of two processors and 1 GB of memory. (Actually, XPAR on a fully populated 900/1500 system board results in partitions with two processors. XPAR does not require that both processors be installed. So, on a partially populated system board an XPAR could consist of a single processor.) For PRIMEPOWER 2500, an XPAR partition must contain at least two processors and two GB of memory. Eventually, other vendors may copy Fujitsu's technical achievement, but for now, PRIMEPOWER still offers the smallest reconfigurable granularity with hardware-enforced partition isolation.

RAS (Reliability, Availability, and Serviceability)

As discussed in the companion white paper on processors, Fujitsu leveraged its mainframe design knowledge to add hardware instruction retry and extensive error checking to the SPARC64 V processor. Complementing those unique processor features, the PRIMEPOWER system design includes dual-power supplies and redundant fans that ensure the system continues to run without interruption in the event of a failure.

The backplane crossbar is actually constructed in two sections. If one section fails, no connectivity is lost – all system boards can continue to communicate with each other, albeit with reduced data bandwidth. The Error Correcting Code (ECC) and the layout of memory chips allows recovery not only from the typical situation – single-bit errors – but also from a group of errors that come from the failure of an entire memory chip.

The collection of reliability and availability features in both chip and system will minimize system crashes caused by failures. Should an unrecoverable error occur, the dynamic partitioning capabilities allow the failing components to be isolated in their own partition for further diagnosis or to be hot-swap replaced.

PRIMEPOWER's overall management is controlled by the System Control Facility (SCF), which itself can be duplexed in the PRIMEPOWER 2500 to guarantee continued operation. (Further details on the SCF can be found in the companion white paper addressing system management functions.)

Review of Key Points

This white paper has briefly described some of the key features of the various models of PRIMEPOWER servers. Although benchmark evidence is not yet available to confirm the performance superiority of the latest PRIMEPOWER models, the strong performance of existing PRIMEPOWER systems suggests that the faster processors and improved interconnect will result in the new models being extremely competitive.

The PRIMEPOWER design relies on a modular system board that provides superior scalability while maintaining a balance of processors, memory, and I/O. The modular system board is also the fundamental unit of hardware partitioning. With XPARs, the new PRIMEPOWER models have enhanced the flexibility of hardware partitioning by introducing a reconfigurable granularity that is a subset of the system board. Since many IT organizations have already embraced partitioning enthusiastically, the enhanced flexibility of XPARs should prove attractive to IT executives looking for an operational edge.

Complementing the performance, scalability, and flexibility attributes, Fujitsu designers have engineered mainframe-like RAS characteristics into both the SPARC64 V chip and the PRIMEPOWER system design. These enterprise-level characteristics, combined with the broad application portfolio available for the Solaris Operating System, make PRIMEPOWER a must-consider candidate for IT organizations seeking a robust UNIX server or UNIX cluster. (Additional information concerning PRIMEPOWER's unique architectural features can be found on the Fujitsu websites.)

Americas

Ideas International, Inc.
800 Westchester Avenue
Suite S620
Rye Brook, NY 10573-1330
USA
Tel + 1 914 937 4302
Fax +1 914 937 2485

Asia/Pacific and Worldwide Headquarters

Ideas International Limited
Level 3
20 George Street
Hornsby, NSW, 2077
Australia
Tel +61 2 9472 7777
Fax +61 2 9472 7788

Europe, Middle East, Africa

Ideas International Europe
1 Deanes Close
Steventon
Oxon OX13 6SZ
United Kingdom
Tel +44 (0) 1235 437 850
Fax +44 (0) 1235 437 851

www.ideasinternational.com

