

# Description Paper

We make sure



## PRIMEQUEST Usage Patterns Usage template OLTP

Issue October 2007

Pages 6

### Contents

<b>Introduction</b>	<b>2</b>
<b>Definition and Target Market</b>	<b>2</b>
<b>Description</b>	<b>2</b>
Customer Problem/Challenges	2
PRIMEQUEST Business Value	2
Critical Server Attributes	2
<b>Solution Description and Requirements</b>	<b>4</b>
<b>Competitive Analysis</b>	<b>4</b>
Performance	4
Reliability	5
Flexibility	5
<b>Summary</b>	<b>5</b>
<b>Appendix A: Detailed Summary of PRIMEQUEST Attributes</b>	<b>6</b>

## Introduction

This paper describes the features of the Fujitsu Siemens Computers PRIMEQUEST family of servers and how they combine to provide an extremely powerful, reliable, and flexible platform for OLTP (Online Transaction Processing) applications.

It will be shown that the features of PRIMEQUEST servers, coupled with the Linux or Windows operating system, provide an easy to manage server environment offering robust performance and scalability.

## Definition and Target Market

OLTP (Online Transaction Processing) is the bread and butter of business operations. OLTP does stock inventory, performs order management, runs accounts payable, oversees shipping, etc: everything that is used to run a business. The OLTP host system typically supports hundreds/thousands of concurrent operations from multiple users, and the transactions range from very trivial (a customer lookup) to long-running multi-phase requests that run multiple minutes. A typical OLTP service may be multi-tiered, with a database backend surrounded by a tier of application servers, or may be embodied as a single large multi-purpose server.

OLTP is characterized by the Transaction Performance Processing Council (<http://www.tpc.org>) TPC-C benchmark, albeit in an artificially limited manner.

OLTP workloads typically do not co-exist with other workloads such as decision support or financial analysis unless the host system supports multiple partitions. In some instances, business hour OLTP processing can shift to decision support or data warehouse processor overnight, or in extreme instances, may enter into batch update operations for database updates that have been deferred during the day. This is because OLTP processing is characterized by significant write activity, where transactions are continually updating business data. These write operations must be synchronized between multiple transactions and users to insure data integrity, and place high loads on the memory sub-system and associated hardware.

## Description

The following sections describe the business and operational challenges of an OLTP host.

### Customer Problem/Challenges

Customers need to conduct business as close to real-time as possible, and this trend is only increasing as the IT industry moves forward with "business intelligence" initiatives and similar operations in the quest to improve business acuity and agility.

To support these requirements, customers host their business data and processes on large databases that support lots of concurrent users and transactions. The day to day business operations are conducted (buying and selling), inventory is updated, shipments are scheduled, etc, all in quasi real-time. Compounding the demanding nature of the workload is the increasing number of "data mining" or inquiry workloads, executed real time, that provide executive management with a snapshot of business activity for the day/week/month

In order for these business environments to be successful, transactions must be executed in a timely manner. Trivial transactions need to be consistently executed in one second or less, and transaction response times need to remain consistent during periods of peak daily usage. This requires a robust host system, and because of the write-intensive nature of OLTP business transactions, is best matched with a large single-image SMP server<sup>1</sup>.

Finally, since the business of the company is extensively based upon the database and its availability, the reliability of the system is paramount. If the system is down, the business is down.

### PRIMEQUEST Business Value

The Fujitsu Siemens Computers PRIMEQUEST server family represents best-in-class SMP capabilities for Itanium-based systems. From the ground up, the PRIMEQUEST server is designed to support robust and demanding OLTP workloads, providing excellent performance and reliability.

Examples of this are:

- High performance memory sub-system that provides the optimum host platform for demanding memory-intensive OLTP database workloads. These attributes are critical to the robust hosting of daily peak loads without degradations in transaction response times.
- Unparalleled hardware reliability with mirror-mode operations. For the most demanding business environments PRIMEQUEST can provide best-in-class availability.

Independent of the PRIMEQUEST hardware design, Fujitsu has worked with Microsoft and the Linux vendors, notably Red Hat and SuSE, to improve the reliability, availability, and robustness of the respective operating systems.

### Critical Server Attributes

The following sections describe the specific features of PRIMEQUEST servers within the context of Online Transaction Processing.

---

<sup>1</sup> This is opposed to hosting a database on multiple clustered servers, ala Oracle Real Application Clusters (RAC). This is evident in the TPC-C benchmark results: while the clustered systems generate large aggregate numbers of transactions, their response times are longer and show considerably more deviation than the same workload hosted on an SMP.

## Scalability

Scalability is very important to the OLTP business environment. The ability of the host server to handle variable day-to-day transaction loads directly influences the transaction response times and how efficiently the business runs. This is true whether the workload demands are computational or I/O-intensive: both must scale to meet the demands of the business.

PRIMEQUEST servers offer unmatched scalability for Itanium-based servers. The combination of processor power, tied together into a true SMP system with the Fujitsu-designed memory and Flexible I/O sub-systems provide a robustly scalable platform for hosting OLTP database operations.

## Performance

Much like scalability, outright performance is critical to the OLTP environment. The speed at which a transaction executes is heavily dependent upon the overall system performance, ranging from processor speed to memory sub-system latency and throughput.

The PRIMEQUEST system is an ideal blend of processor performance, courtesy of the Intel Itanium2 design and implementation, and the Fujitsu-designed memory sub-system. Individual transactions will show reduced execution times relative to other competing platforms, and the excellent scalability will provide an ideal host platform for many concurrent users and transaction loads.

## Large Memory

In the OLTP benchmark environment represented by the TPC-C workload, enormous amounts of memory are allocated in order to squeeze the last bits of performance out of the benchmark. While these terabyte and above memory configurations are uncommon in commercial practice, the memory capability of the OLTP host system is critical to overall transaction performance.

One of these critical factors is the amount of memory that can be hosted per system board, and therefore per CPU. The PRIMEQUEST design hosts 256 GB per system board at the outset, or 32 GB per CPU core. This allows a generous pairing of memory with processors, even for systems that do not require the full complement of system boards.

## Flexible I/O

The ability to float I/O resources within partitions, either as the result of system component failure or system growth, is a critical requirement for OLTP workloads. Storage resources, and access to them, change or grow with the business, and insuring access to those resources for all components of the OLTP server is paramount. For example, the customer can add system boards to an existing partition, and using the PRIMEQUEST Flexible I/O facility, provide direct access to the I/O resources for the new board. Similarly, new I/O resources can be added to an existing partition, and be made available to all system board components.

## Partitioning

OLTP workloads tend to be database-centric, and the database image determines how the database is deployed. Commonly, a single large server is used to host the database, and in that context a system capable of offering multiple partitions is not critical.

The notable exception is the instance of a multi-tier OLTP/application service installation. In such a configuration PRIMEQUEST can support both the critical core database and the distributed applications, all on a shared platform running multiple partitions. This reduces the complexity of the multi-tiered OLTP application services.

## Dynamic Reconfiguration<sup>2</sup>

Dynamic reconfiguration (DR) is a desirable function for the host OLTP platform. This is because DR provides a means to:

1. Perform concurrent maintenance, where resources can be taken in and out of service to perform upgrades without taking the application service and database down, and
2. Increase the system configuration with no service intervention.

Both of these are important to a 24\*7 business environment.

## Operational Efficiency

OLTP environments are not typically concerned with operational efficiency, otherwise known as the ease of system administration. Most of the administrative challenges for businesses that rely on OLTP systems for day to day operation are related to the host database, not the host server.

However, the PRIMEQUEST servers offer a number of integrated services, starting with duplicated boot disks and redundant management LANs, and continuing with integrated Gbit switches and a KVM/USB interface. This latter aspect of the PRIMEQUEST design seems to be a trivial aspect of server operation, but it vastly simplifies the "jumper cable jungle" that is required to control multiple-partition servers and removes a source of potential operator error.

## Fault Immunity

24\*7 mission-critical OLTP environments have stringent reliability, availability, and serviceability requirements. Any downtime has an immediate and obvious effect on the business bottom line.

PRIMEQUEST servers provide the utmost reliability for hosting these demanding environments. The PRIMEQUEST servers are the only Itanium-based servers to provide mirror-mode operations, thereby providing extra assurance that if something did indeed happen to the hardware that its effect would be completely nullified. The level of redundancy in the PRIMEQUEST design is based upon years of experience with mission critical mainframe and large UNIX servers, and will provide a highly reliable host platform.

---

<sup>2</sup> Dynamic Reconfiguration is planned to be supported from 2<sup>nd</sup> half of 2006

Supplementing the availability of PRIMEQUEST hardware is the Fujitsu Siemens Computers PRIMECLUSTER software. This mature and robust clustering software works at the level that System Mirror Mode cannot: application service availability and failover mechanisms. By deploying PRIMECLUSTER software on PRIMEQUEST hardware, the customer is configuring both fault immunity for the hardware, but also ensuring that their application services and host operating system are running.

### Mission Critical OS

Similar to the hardware reliability requirements, the host operating system for a 24\*7 OLTP environment needs to be resilient and robust.

Microsoft Windows Server, particularly Enterprise Edition and Datacenter Edition, has been proven to be a very suitable operating system for hosting the SQL Server database, which in turn is a robust and flexible database management product. Fujitsu has been working with Microsoft to enhance its capabilities, aiming for the "Longhorn" release.

Much of the work that Fujitsu has put into the Linux open source community has been for operating system "hardening". Examples are device driver enhancements, improved dump and diagnostic capabilities, and error recovery and retry. These modifications enhance the mission-critical aspects of the Linux operating system.

### Multiple OS Support

The essence of the OLTP workload environment is determined by the choice of database vendor. In the context that not all database products are available on all operating systems, the ability of PRIMEQUEST to support multiple operating systems is important.

The issue is whether or not these operating systems need to be concurrently supported on an OLTP platform. In the typical deployment, this is not the case: the database and host OS are typically the sole workload on the platform.

The notable exception is where a multi-tier OLTP database and application service is hosted on a common platform. Using PRIMEQUEST partitions, the central database and attendant transactions can be hosted in one partition using the appropriate operating system, and the application services can be hosted in other partitions using other operating systems.

## Solution Description and Requirements

Hosting a mission-critical 24\*7 OLTP environment requires a number of key contributions:

1. Optimal hardware platform, represented by the Fujitsu Siemens Computers PRIMEQUEST server.
2. Database software with proven capabilities. Fujitsu Siemens Computers recommends Oracle9i/10g as a class-leading relational database management system for Linux deployments, with proven scalability, reliability, and manageability. Similarly, Fujitsu Siemens Computers recommends SQL Server as a robust and flexible database management offering for Windows Server deployments.
3. Proven service and support. Fujitsu Siemens Computers and Oracle Professional Services can accommodate the most demanding customer application requirements, and with Fujitsu Siemens Computers' service organization, provide excellent ongoing support to guard against service outages.

The combination of top-flight hardware, software, and services and support provides the customer with the best solution for their OLTP requirements.

## Competitive Analysis

There are a number of vendors that offer products to compete against PRIMEQUEST, based both on Itanium and RISC processors.

Competing Itanium vendor in EMEA mainly is HP, offering Superdome Integrity servers. Competing RISC system vendor is IBM, offering systems based upon the Power technology.

PRIMEQUEST either matches or surpasses the performance of these respective platforms using several criteria: performance, reliability, and flexibility.

## Performance

### CPU Performance

At the heart of PRIMEQUEST servers are the Itanium 2 processors. The current incarnation, the "Montecito" design, incorporates a 24MB 3<sup>rd</sup>-level cache running on a 1.6 GHz processor. These capacities, coupled with the EPIC (Explicitly Parallel Instruction Computing) architecture, allow Itanium-based servers to out-perform servers using competitive vendor processors, such as IBM and AMD.

### Memory Performance

In commercial systems hosting databases or running high-performance computing (HPC) numerical equations, memory performance is the key to robust and sustainable **system** performance. PRIMEQUEST couples the Itanium 2 processors together using a custom-designed memory sub-system that offers three benefits:

1. Low latency for local (on-board) and remote (off-board) access. Memory latency is the defining attribute for many database applications, and consistent memory latency is the hallmark of a robust memory design.

2. Symmetrical multi-processor (SMP) memory design. Because of the low ratio of remote/local memory access, the PRIMEQUEST server offers an SMP environment for the hosted operating system and applications.
3. High bandwidth. Memory latency determines how fast a processor can get the first data element. Bandwidth determines how fast the processor can get the last. PRIMEQUEST offers class-leading memory bandwidth, both peak and sustained. Of the two values (see datasheets for details), sustained is the more realistic, in that it reflects how the memory sub-system supports the aggregate configuration of processors and I/O components.

These attributes combine to provide class-leading memory-system performance in the PRIMEQUEST servers. Competitive designs from HP suffer from ccNUMA (cache coherent Non-Uniform Memory Architecture) effects, where memory performance degrades as the system gets physically larger, i.e., more processors are added to the server.

## I/O Performance

PRIMEQUEST servers are well-balanced systems, with robust Itanium 2 processors coupled together with an SMP memory sub-system to provide excellent computational and database processing power. The PRIMEQUEST server extends these capabilities by providing a flexible, robust, and powerful I/O sub-system.

Key to the PRIMEQUEST I/O capabilities is the Fujitsu-designed "Southbridge" implementation. It combines a custom ASIC that interfaces directly to the PRIMEQUEST memory cross-bars, that in turn interfaces to a flexible high-performance PCI Express fabric. These interfaces, four-lanes wide for an aggregate bi-directional bandwidth of 2GB/sec, are externally extended to PCIe and PCI-X host bus adapters and network interface cards.

With 128 PCI-X (or 64 PCIe as well as a mix of both) interfaces the connectivity to PRIMEQUEST servers is excellent. Adding to this capability is the large number of integrated services, where PRIMEQUEST servers offer up to 32 Gbit interfaces, 32 Ultra320 SCSI drives, and integrated Management LANs and KVM interfaces., none of which requires the customer to consume valuable PCI-X slots.

## Reliability

PRIMEQUEST servers offer unparalleled reliability. Not only are all major components redundant and hot-pluggable, PRIMEQUEST servers are the only systems to offer System Mirror mode.

System Mirror mode takes the two Global Address crossbars (GAC) and runs them in lock-step. It also takes the four Global Data crossbars (GXD) and mirrors them into two steps. Finally, System Mirror mode mirrors banks of memory on systems boards.

The benefit of System Mirror mode is that a failure of any of the components will not take the system down. This stands in unique contrast to systems that, despite offering significant redundancy, will "bounce" if one of their redundant components fails. PRIMEQUEST servers offer that last bit of insurance for application availability, by providing a platform that is, or all intents and purposes, immune to faults.

## Flexibility

PRIMEQUEST servers can support up to sixteen partitions, each hosting an independent operating system instance. While this is higher than competitive products such as the HP Superdome, PRIMEQUEST flexibility goes a step further by allowing customers to separately configure processor/memory resources from I/O resources.

Fujitsu calls this Flexible I/O, and it offers several benefits. For one, partition configurations are determined by the application requirements. Both System Boards and I/O Boards can each be physically split into two extended partitions (XPAR), or can be combined into larger SMP partitions hosting larger applications. Thus, if the application requires a high processor count with relatively little I/O, then multiple System Boards can be configured with a single I/O Board. Inversely, if the application requires a high degree of connectivity but not too many processors, multiple I/O Boards can be combined with a single System Board running XPAR.

If there is a System Board failure, a hot-standby System Board can be deployed to take its place. Using Flexible I/O, all of the resources available to the previous System Board become transparently available to the replacement system board.

## Summary

The family of PRIMEQUEST servers offers "game changing" platforms for hosting advanced Linux and Windows applications. Engine for engine, they offer class-leading performance and scalability, unmatched reliability, and superior flexibility.

The robustness of the Itanium 2 processors, coupled with the Fujitsu-designed memory and I/O sub-systems, offers industry-leading performance for database and high performance computing platforms. The System Mirror mode places PRIMEQUEST servers into the rarified plateau of fault-tolerant servers, where component failures are not even noticed by the hosted operating systems. Finally, PRIMEQUEST XPAR capabilities, coupled with Flexible I/O, offer a high degree of freedom for configuring physical components to best match application requirements.

## Appendix A: Detailed Summary of PRIMEQUEST Attributes

	PRIMEQUEST 580	PRIMEQUEST 540	PRIMEQUEST 520
Type	Floor-standing cabinet	Floor-standing cabinet	Rackmount Chassis: 12U
<b>CPU</b>			
Processor	Itanium® 2 Montecito dual-core Clock speed: 1.4 GHz (12MB L3 cache) or 1.6 GHz (24MB L3 cache)		
<b>System</b>			
Interconnect	Point-to-point crossbars: 2 Address, 4 Data		2 Address, 2 Data
	Bandwidth: 136.5 GB/sec sustained	68.2 GB/sec sustained	34.1 GB/sec sustained
	Latency: 208-276 ns	208 - 256 ns	208 - 256 ns
System Boards	8	4	2
Main Memory	2TB (256 x 8GB)	1TB (128 x 8GB)	256GB (64 x 4GB)
Internal HDD	Max 4.7TB (32 x 147GB)	Max. 2.35TB (16 x 147GB)	Max 584GB (8 x 73GB)
Expansion Slots	Maximum 128 PCI slots: IOU x 8 x 4 = 32 internal + PCI Box external: 8 x 12 PCI-X or 8 x 6 PCI-X plus 8 x 4 PCIe IOU: 64bit/133MHz Short x 16 64bit/100Mhz Short x 16  PCI Boxes max: PCI-X: 64bit/133MHz Long x 32 PCI-X: 64bit/100MHz Long x 64 PCI Express 4lane Short x 32	Maximum 64 PCI-slots: IOU x 4 x 4 = 16 internal + PCI Box external: 4 x 12 PCI-X or 4 x 6 PCI-X plus 4 x 4 PCIe IOU: 64bit/133MHz Short x 8 64bit/100Mhz Short x 8  PCI Boxes max: PCI-X: 64bit/133MHz Long x 16 PCI-X: 64bit/100MHz Long x 32 PCI Express 4lane Short x16	Maximum 16 PCI slots: IOU = 4 internal + PCI Box external: 1 x 12 PCI-X or 1 x 6 PCI-X plus 1 x 4 PCIe IOU: 64bit/133MHz Long x 2 64bit/100Mhz Long x 2 PCIe: 8lane Short x 4 PCI Boxes max: PCI-X: 64bit/133MHz Long x 4  PCI Express 4lane Short x 4
I/O Interfaces	10GBASE-LR: maximum 4 1000BASE-T: maximum 32 100BASE-Tx: 2 External SCSI Port: 16 Video (Analog RGB): 1 Serial (Dsub 9pin): 16 USB1.1: (KB, Mouse): 4 USB2.0: 32 DVD ROM: 1 Integrated KVM	10GBASE-LR: maximum 4 1000BASE-T: maximum 16 100BASE-Tx: 2 External SCSI Port: 8 Video (Analog RGB): 1 Serial (Dsub 9pin): 8 USB1.1: (KB, Mouse): 4 USB2.0: 16 DVD ROM: 1 Integrated KVM	1000BASE-T: 4 100BASE-Tx: 2 External SCSI Port: none Video (Analog RGB): 1 Serial (Dsub 9pin): 4 USB2.0: (KB, Mouse): 2  DVD ROM: 1 Integrated KVM
<b>Reliability</b>			
Redundancy	Disks, Power units, Fans, Cross-bars, Server Management processor, Gbit switch		Disks, Power units, Fans, Cross-bars
Hot Swap	PCI cards, Disks, Power units, Fans, Server Management processor		
Partitions	Maximum 16	Maximum 8	Maximum 4
<b>Installation Details</b>			
Dimensions	BASE Chassis: W x D x H 740 x 1,100 x 1800 mm (29.1 x 43.3 x 70.9 inches)	BASE Chassis: W x D x H 740 x 1,100 x 1800 mm (29.1 x 43.3 x 70.9 inches)	Rackmount Chassis: 12U 482 x 820 x 530 mm (17.3 x 32.3 x 20.9 inches)
Weight	Maximum 720 kg	Maximum 600 kg	Maximum 120 kg
Power Requirements	AC200-240 +/-10% Single Phase 50/60 Hz (+2/-4%)		
Power Consumption	10,900W	6,300W	3,320W
Operating Systems	Red Hat Enterprise Linux AS 4.0 , AP 5.0 SUSE Linux Enterprise Server 9, 10 Windows Server 2003, Enterprise Edition for Itanium-based Systems Windows Server 2003, Datacenter Edition for Itanium-based Systems		