



Fujitsu PRIMEQUEST: Itanium “Big Iron”

Quick Note

Thomas Deane
Gordon Haff

5 May 2005

Copyright © 2005
Illuminata, Inc. Licensed to
Fujitsu Limited for web
posting. Do not reproduce.

Big Iron systems are nothing new to Fujitsu. It's sold large multiprocessor Unix SPARC64 PRIMEPOWER and mainframe GlobalServer (GS) systems¹ for years. Fujitsu's no stranger to Intel chips either. Its current PRIMERGY line includes modest-sized Xeon² and Itanium 2³ systems with up to four sockets. But Fujitsu's recently unveiled PRIMEQUEST systems are both large *and* Itanium-based. Announced last month in San Francisco, these new systems complement large scale-up capability in Windows and Linux environments—where Fujitsu didn't previously have a product.

To produce PRIMEQUEST, Fujitsu plumbed its extensive in-house intellectual property and designs for large-scale, reliable systems—something that it's been doing for years, both in “Classic” Fujitsu and in Amdahl (which Fujitsu purchased in 1997). Fujitsu has also collaborated with Intel over the past two years; PRIMEQUEST uses off-the-shelf Itanium 2 processors, but Fujitsu developed a custom scalable chipset for the line. The first offerings in this new line, the PRIMEQUEST 440 and 480, are enterprise-class servers with a lot to offer.



Gladiator Games

PRIMEQUEST places Fujitsu in a gladiatorial coliseum with a considerable cast of competitors. But some are stronger and more directly matched with PRIMEQUEST.

1. Fujitsu doesn't sell the GS systems in the U.S.
2. Fujitsu's largest Xeon system is the PRIMERGY RX800, a 4U/4-processor 32-bit Xeon MP “building block” system, four of which can be grouped together into a 16-way system.
3. Fujitsu currently offers two PRIMERGY Itanium 2 servers—the 2-processor/2U RXI300 and the 4-processor/2U RXI600; neither are exactly Big Iron.

For its part, Fujitsu has explicitly declared battle with IBM and HP.⁴

In the absence of an Itanium entrant in its latest generation of scalable xSeries systems, IBM may not compete quite directly with PRIMEQUEST.⁵ But it's no less a competitor for that. IBM's eServer group has been on a roll, and certainly has Linux Big Iron in the form of both POWER5 and Xeon MP-based servers. The latter also run Windows.⁶ There's little doubt that Fujitsu will run into IBM plenty, notwithstanding the different processor architectures.

However, Fujitsu is most directly going after HP. To be sure, it can only fight HP for Linux and Windows deals; Fujitsu won't be licensing HP-UX on PRIMEQUEST as it does Solaris on PRIME-POWER. But, that important caveat aside, it's clear that as the dominant player in enterprise Itanium-based servers, HP is the most important and directly-matched competitor for Fujitsu's PRIMEQUEST. In much the same way that PRIME-POWER effectively gives customers a respected second source for SPARC systems, PRIMEQUEST is a Big Iron Itanium alternative to HP's Integrity Superdome systems.

To be sure, HP and IBM aren't the only competition. SGI and Unisys manufacture large Itanium systems, too. But SGI is exclusively focused on heavy-duty high-performance computing, and is most noted for selling large specialized systems to government organizations—NASA, for example.⁷ Fujitsu doesn't compete in that space. Nor does Fujitsu see Unisys and its ES7000⁸ as the main—or

even much—competition. That's not particularly surprising. The ES7000 is a relatively old design, and Unisys, as a company, is stepwise distancing itself from building computers in favor of offering services for and around them.

Fujitsu may run into secondary competitors such as these—as well as various purveyors of scale-up x64 gear, including partner Sun—from time to time. But PRIMEQUEST's success or failure will hinge on how it fares against HP and IBM.

Big and Bigger

First into the fray are the PRIMEQUEST 440 and 480. Developed in Japan by a project team of approximately 100 engineers, the systems can run either Linux or Windows.⁹ The 480 can house up to eight system boards, for a total of 32 sockets. Initially, the 480 will support 512 gigabytes of memory—one terabyte configurations are planned for later this year. The 440 is essentially a scaled-down version of the same system. It's half the size—four system boards, 16 sockets, and half a terabyte of memory. The two system models differ only in configurability; they have the same high-availability and other features.

For now, the four-socket system boards house one processor core per socket using Intel Itanium 2 "Madison" processors running at either 1.5 or 1.6 GHz.¹⁰ PRIMEQUEST, however, will also support Intel's forthcoming dual-core "Montecito"

4. Sun was conveniently not mentioned, likely because Sun and Fujitsu have ratcheted their historical development partnership up a notch and are now cross-selling each others' SPARC-architecture systems.
5. See Illuminata report "Xeon Zips with X3" (April 2005).
6. As well as in the form of its zSeries mainframes and the i5 iSeries systems that share common hardware with the POWER5-based p5 systems. See Illuminata reports "POWER5 Takes Off on pSeries" (July 2004), "The New iSeries" (May 2004), "The Mainframe Reloaded" (May 2003), and "IBM's Linux POWER-play" (November 2003).

7. Last year, NASA contracted with SGI for a cluster of 20 Altix supercomputers, each with 512 Itanium 2 processors. This Linux cluster with its 10,240 processors will be used by NASA to run its Space Exploration Simulator. See Illuminata report "SGI Brings Big Iron to Linux" (February 2003).
8. The Unisys ES7000 Orion 440 can contain up to 32 Itanium 2 processors.
9. Red Hat Enterprise Linux 4 is available now. Microsoft Windows Server 2003 and Novell's SUSE LINUX Enterprise Server are planned for September, 2005.
10. The 1.5 GHz Itanium 2 "Madison" processors shipped with PRIMEQUEST systems have 4 MB of Level 3 (L3) cache, whereas the 1.6 GHz Itanium 2 "Madison" processors shipped with PRIMEQUEST systems have 9 MB of Level 3 (L3) cache.

Itaniums when they start shipping in 2006¹¹ as well as the “Montvale” processor intended to follow. Fujitsu says that upgrading to future Itanium CPUs will require some changes—but is still working on exactly which upgrades it will offer and the other system changes that they will require.

Each system board also holds memory. Every processor in the system sees all the memory in the system as a unified whole, but a processor can access the memory on its own board faster than it can access memory on other system boards. That is, the PRIMEQUEST 480 is a non-uniform memory access (NUMA) system, as are most large-scale systems today.

Even though operating systems try to keep data physically close to the processor using it, they don’t always succeed. Thus, both local and remote access times can have a major effect on application-level performance—especially with volume operating systems like Windows and Linux that aren’t particularly tuned to any specific system architecture. The memory access times (based on the future Montecito processor) that Fujitsu shared with us are competitive with today’s Big Iron systems, especially for remote access,¹² although they trail some newer designs such as IBM’s X3.

Immunizing Against Faults

However, PRIMEQUEST isn’t just a generic big Itanium system. It’s got some unique features. Perhaps the most intriguing is its ability to split each system board into two separate segments, or system mirrors. These mirrors provide an additional level of “fault immunity”—to use Fujitsu’s parlance—at the cost of cutting the available processor and memory resources in half.

With this system mirror capability—the use of which is optional, POWERQUEST can guard

11. Although Intel plays up a “late 2005” date for initial availability, Montecito isn’t expected to ship in volume until 2006.

12. See Illuminata report “Latency Matters!” (September 2002).

against some hardware failures and their effects on the hosted OS and application. They would just not be seen.

In mirror mode, the address and data crossbars are put into lockstep pairs. When the software running on the system makes a memory request, that request gets sent out concurrently on both mirrored crossbars. If either mirrored crossbar fails, the system can continue running, because it retains the known state on the other crossbar. Similarly, if a fatal memory error occurs it can be intercepted and ignored, as the system has the known state on the other mirrored memory segment.

System mirror operations are all contained within individual system boards. Controlled by a custom Fujitsu chipset, there is a huge amount of redundancy and replication to safeguard these mirrors. There are multiple redundant address/data crossbar controllers—the dual address crossbars are mirrored and the four data crossbars are mirrored pairs. Memory is also “logically” mirrored—both pairs of mirrored Local Data Crossbars (LDX) are reading/writing the same contents from the mirrored address/data crossbars. Multiple mirrored system boards can also be joined together to form a single fault-immune partition.

PRIMEQUEST’s mirroring has certain similarities to the Linux and Windows fault tolerant (FT) systems that NEC and Stratus sell.¹³ Like those systems, PRIMEQUEST runs mirrored pairs of hardware. However, the processors on those mirrors do not run in lockstep with each other. As a result, although PRIMEQUEST can recover from many classes of failure that other systems cannot, it is not a fully fault-tolerant system that can ride through almost any hardware failure as can the NEC and Stratus designs—or the ultimately fault tolerant NonStop Himalaya systems from HP.

13. See Illuminata reports “Stratus Cuts the Cost of Fault Tolerance (Again)” (September 2004), “Stratus ftServers: Windows Fault Tolerance for Verticals” (September 2002), and “Linux FT, a la NEC” (January 2003).

Par-ti-tion-ing

Partitioning is another technique that's often used to improve system uptime. A fault in one partition won't bring the whole system down. Today's Big Iron systems are also routinely partitioned into smaller chunks to keep individual workloads from getting in each other's way. Partitioning can be very useful to an IT manager interested in reducing costs and efficiently using system resources. Partitioning a large server into exactly the right size chunks to satisfy a particular workload is a cost-effective way to optimize overall utilization. How finely the partitions can be divided, and how quickly and easily they can be reconfigured or resized on-the-fly are two key features.

In some cases, partitions are based in hardware and are fairly static—but very effectively isolate against just about any sort of fault, including the physical failure of components. In others, the partitions are based in software and can be dynamically changed without application service disruption—but primarily the partitions just isolate against software failures and interference.

PRIMEQUEST systems go the hardware partitioning route, just as their PRIMEPOWER relations do. A partition can span multiple boards, up to the size of the system. It can also be as small as one system board containing four processors. PRIMEQUEST systems cannot currently be sub-partitioned smaller than one system board. Therefore, a PRIMEQUEST 480 with 8 system boards today could be configured into a maximum of 8 partitions. Fujitsu has indicated, however, that future PRIMEQUEST servers will allow subdividing each system board in half, for a maximum of 16 partitions in a fully-populated 480.

Compared to, say, IBM's POWER LPARs or HP's vPars, this is relatively coarse partition granularity. That's the downside of physical partitioning approaches. The upside is that, in addition to better protecting against hardware faults, physical partitioning doesn't depend on software and hardware's working closely together. Thus, it's a good match

for operating systems like Windows that don't have partitioning smarts built in yet.

Neither can Fujitsu currently modify partition resources and sizes on-the-fly without application service disruption. Such dynamic partition reconfiguration (DPR) requires an operating system working in close concert with the hardware. Windows, Red Hat Linux, and Novell's SUSE LINUX—the 3 OSs used on PRIMEQUEST—do not currently have this capability. Fujitsu is currently working with the Open Source Development Laboratory (OSDL) to implement DPR in Linux; it expects this functionality to be available in the latter half of 2006. Fujitsu is also working with Microsoft to include DPR in the "Longhorn" version of Windows Server—when-ever it finally arrives.

PRIMEQUEST also includes a "Flexible I/O" feature by which each system board can dynamically reconfigure its I/O connections through software. As a result, it's not necessary to manually re-configure to make I/O changes or to allocate I/O to each partition in proportion to the number of processors that it has. Thus a partition running an I/O-intensive workload can be allocated extra I/O capacity while another running a compute-intensive job gets fewer I/O slots and adapters.

PRIMEQUEST's hardware-based partitions may not be the most flexible, or have the finest granularity, but they are still small and plentiful enough to handle many of today's workloads and server consolidation scenarios. And today, when partitions are used in a production environment, the reality is that they tend to be fairly static—they do not get changed or moved around all that often. PRIMEQUEST meets a lot of IT requirements well—and promises more flexibility and granularity in the future as the support moves into Linux and Windows.

Conclusion

PRIMEQUEST's high scale, advanced high-availability features, processor upgradeability, and flex-

ible I/O capability make it an entirely credible Big Iron platform for Linux and Windows. Indeed, the question isn't so much around the suitability of PRIMEQUEST to run the biggest Linux and Windows jobs, but how many such jobs there are—given that both OSs tilt far more toward horizontally-scaled environments than vertically-scaled ones. Although HP and IBM both likewise offer very large Linux and Windows servers, in practice, most of their biggest systems run Unix as well—just as Fujitsu does on its PRIMEPOWER line. As a result, in spite of PRIMEQUEST's high scale points, we'd expect far more midrange—or even low-end—servers to sell—a common happenstance in servers covering such a broad range.

Although not as well known in North America as IBM and HP, Fujitsu is nonetheless a major global IT provider that's very well known in Europe and

Asia. It also has a history of delivering large, rock-solid systems. It has capitalized on its mainframe expertise to produce PRIMEQUEST, an impressive Itanium server.

Although the new PRIMEQUEST line will be available globally, Fujitsu strategically chose San Francisco as its announcement venue. Over the next few years, Fujitsu's goal is to dramatically increase North American revenue,¹⁴ and it hopes PRIMEQUEST will make a significant contribution to this goal. Making the announcement in North America signals this desire. It has also declared it will be doing battle with the Titans in the gladiatorial arena—HP and IBM. Let the games begin.

14. According to Fujitsu's 2003—the latest available—yearly financials, Fujitsu's North America revenue is \$3 billion—only about 7 percent of its total global revenue of \$45 billion.



Through subscription research, advisory services, speaking engagements, strategic planning, product selection assistance, and custom research, Illuminata helps enterprises and service providers establish successful infrastructure in five key areas: Server Technologies, Information Logistics, Application Strategies, Enterprise Management, and Pervasive Automation.