Conclusion 6

Sun provides a comprehensive system solution for database needs. From the hardware design emphasizing modularity and performance to the finely tuned Solaris 2 operating system to the third-party vendor optimized database systems, the SPARCcenter 2000 and the SPARCserver 1000 provide the equivalent of a single vendor solution for databases. Sun delivers all that database customers need: support for large numbers of users, high rate of transaction throughput, reliability, scalability, and low cost -- a total solution all in a modular design. Sun has made it possible to have superior price performance at low cost providing a scalable, upgradeable architecture that protects a customer's investments.

Sun's on-going alliances with the major database vendors ensure a cooperative exchange of architectural, design and performance ideas and enhancements. Together, these alliances have resulted in finely tuned database servers: the Sun SPARCcenter 2000 and SPARCserver 1000 systems. Sun Microsystems is delivering the power for applications of today and tomorrow.

HA includes an agent toolkit to allow customization of additional services. HA for Sun is developed by OpenVision and works with Sybase, Oracle and Informix database systems.

system board includes a thermal sensor that issues a non-maskable interrupt if the local temperature exceeds the recommended operating range. If this occurs, the system will perform a shutdown to avoid damaging the machine.

The SPARCcenter and SPARCserver systems incorporate a power supply that is resistant to most kinds of power fluctuations. Additionally, provisions are made for incorporating a third-party uninterruptable power supply (UPS).

5.3 Automatic System Reconfiguration

The SPARCcenter and SPARCserver systems run boot time diagnostics to detect malfunctioning components. If a bad component is detected, many times the system will be able to bypass the component and rely instead on a redundant component. This effect may be a degradation in performance, but the system can re-configure and still provide service.[1]

5.4 OnLine: DiskSuite

Sun OnLine: DiskSuite is an unbundled product that offers better performance, greater capacity, easier administration and improved availability of disk storage on Sun SPARC systems. DiskSuite provides disk stripping (RAID 0) and multi-way disk mirroring (sometimes referred to as RAID 1). Both disk stripping and multi-way mirroring can improve performance. Data availability and reliability are also improved by multi-way mirroring. DiskSuite includes automatic replacement of failed components within a mirror using a "hot spare" facility. A hot spare is a disk that is not being used to store data, but is available to be to replace a failed disk in a mirrored set. The high availability and reliability gained from a hot spare can offset the cost of having redundant components.

5.5 High Availability for Sun

Where DiskSuite provides disk failover, High Availability for Sun (HA for Sun) provides automatic service restart after software or hardware failure is detected. Alternatively it provides automatic service "failover" to an alternate system. In either case, recovery of all data on the alternate machine is immediate since HA is integrated with standard dual-ported and dual-initiated mirrored disk technology. Furthermore,

1. For more detail on system reconfiguration, readers are referred to “SPARCcenter 2000 and SPARCserver 1000 Reliability, Availability, and Serviceability.”

High Availability Features and Solutions 5

Along with high capacity, high performance and efficient system software, the SPARCcenter and SPARCserver systems are designed with high availability in mind. The high availability features permit these systems to detect, avoid and recover from failures. These systems have data protection features formerly found only in more expensive traditional data center computers. In addition, partnerships with third-party vendors offer solutions to enhance further the high availability of SPARC systems.

5.1 Protected Data Paths

Memory reference integrity is enhanced by protecting system memory with an ECC mechanism. Single bit errors are corrected. Multibit errors are detected.

If any memory chip on a SIMM fails, the worst that happens is a single bit correctable error. The SIMM crossbar logic orders memory so that each bit of a 64-bit memory word comes from a different DRAM. With this organization, any errors due to a single chip failure will be corrected by the ECC hardware.

In addition, all bus traffic is parity checked for correctness -- detecting single-bit errors. The hardware system re-transmits in the event of an error.

5.2 Environmental Safeguards

A number of environmental sensors are provided to protect the system from hazardous conditions. For example, voltage is monitored; if voltage should rise or fall to points outside a recommended range, the system immediately shuts itself down. Also, each

SunSoft's Solaris 2 is a highly modular operating system, meaning only drivers and subsystems needed by the hardware configuration are used. It is a fully preemptable, real-time capable, symmetric multiprocessing (SMP) kernel that supports multiple threads of control. It also features user level thread support. With user level threads, software developers can take explicit advantage of SMP at the application level.

4.1.1 Solaris Benefits to Database Processing

To maximize the performance of database operations, the database architectures organize their operations so that they are doing many tasks simultaneously. To make this possible, the underlying operating system provides support for concurrent processing, such as multiple threads, shared memory and asynchronous I/O.

At the application level, the Solaris user level threads library provides a mechanism for true, i.e. simultaneous, multitasking activities. Using user level threads, applications can be decomposed into subtasks to take maximum advantage of multiple processors.

Most database systems make extensive use of shared memory for inter-process communications. Shared memory is the fastest and most scalable message passing scheme possible for Unix systems. Understanding this, Solaris 2 has provided the unique feature called Intimate Shared Memory. Processes sharing memory with ISM also share a set of page tables. By having one common set of page tables, context switch time is reduced. Memory requirements and swap activity are also reduced. This translates into faster response time and the ability to keep more database specific data in memory.

Solaris 2 provides three classes of priorities: multitasking, system, and real time. Real time processes have higher priority than system processes; system processes have higher priority than multitasking processes. While not really needing real time response times, database applications can take advantage of the scheduling privileges of real time processes. Multitasking processes are at the mercy of the kernel scheduler and can be preempted by system processes at any time. Ordinarily, this is a good thing because the kernel must run when it needs to. In multiprocessor systems, this requirement is not so critical because the kernel can run on alternate processors. With proper precautions, database processes can take over a processor and dedicate that compute power exclusively to database use.

Software Features 4

4.1 Solaris Software Features

While the SPARCcenter 2000 and SPARCserver 1000 systems offer many benefits, such as power, scalability, extensibility, it is the SunSoft Solaris operating system that enables these benefits. Solaris has been enhanced and highly optimized to increase performance of database systems.

In the past, it was necessary to purchase Sun DataBase Excelerator (SunDBE), a performance optimization package. SunDBE dramatically improved the performance of databases running on Sun's systems -- particularly with a large number of concurrent users. All of the SunDBE enhancements are now part of Solaris 2.2 making the SPARCcenter 2000 and SPARCserver 1000 systems even better choices as database servers. These enhancements include:

- Faster file writes
- Increased maximum number of users supported
- Efficient multi-threaded SunOS kernel
- Efficient page table manipulation
- Increased maximum number of file descriptors
- Highly tuned asynchronous I/O

requester arbitrates for the bus, sends a request packet that specifies the target address and then releases the bus to make it available for other activity. While the request is being serviced, the bus is free to perform other activities. All packets on the bus are tagged so that a request can be associated with its reply. Furthermore, there is an added efficiency due to the pair of XDBuses in the SPARCcenter 2000 servers. The effect is a single bus with twice the bandwidth.

The XDBus implementation operates at 40 MHz and has a peak bandwidth of 320 MB/sec. Due to the efficient arbitration scheme -- a packet is overlapped with actual transmission of the previous packet -- a sustained throughput of 250 MB/sec is realized.

3.3 SBus I/O Architecture

Databases require both I/O capacity and I/O bandwidth. For example, decision support requests read massive amounts of data, look at it briefly, and then read more data. The bus bandwidth of the I/O system is key to reducing the time to handle these requests. Bus bandwidth is important for doing backups as well. The system must be able to backup large databases without affecting concurrent transactions.

I/O capacity is equally important, especially for OLTP systems. The more spindles the data can be spread across, the more transactions that can be supported simultaneously. In this case, both the number of separate spindles and the number of controllers are important. Furthermore with the new high data rate, complex transaction support -- such as voice and objects -- these requirements for I/O bandwidth and capacity will only grow.

The SPARCcenter and SPARCserver systems supports multiple full-featured SBus I/O subsystems. All peripheral devices are connected to the SBus. The SBus has a burst transfer speed of 80 MB/sec and a sustained stream-mode speed of 55 MB/sec for writes and 49 MB/sec for reads.

In the case of the SPARCcenter 2000, each system board has a SBus with four SBus slots. With a maximum of ten system boards in one system, customers can spread out their I/O capacity over as many as ten SBuses and forty different controllers. A SPARCcenter 2000 can be fully configured with over 500 GB of disk capacity.

In the SPARCserver 1000, each system board has an SBus with three SBus slots and an on board FSBE (Fast SCSI/Buffered Ethernet) SBus controller. Therefore, with a maximum of four system boards per system, customers can spread out their I/O capacity over as many as four SBuses and twelve different controllers. It can be fully configured to 100 GB of disk storage.

The SPARCcenter 2000 and SPARCserver1000 promote fast memory access by using two levels of parallelism. First, each system has multiple memory banks with bank interleaving. Second, a single bank consists of multiple SIMMs and DRAMs interconnected by a "crossbar" which distributes each memory access across a set of DRAMs. Chips are activated in parallel for each block accessed.

The interleaving not only reduces the latency for accesses to blocks of memory, but also minimizes the chance that back to back memory request will target the same bank. For sequential memory operations, the access pattern of all the memory system clients will be distributed uniformly across all the memory banks,

3.1.3 Non-Volatile RAM

An important performance consideration in database systems is the handling of the log file. All database systems use log files and they must be written synchronously, that is data written to the log file must be confirmed as being successfully written. In particular, a transaction cannot continue until the log record is safely on the disk. In many cases, log files are mirrored to avoid a single point of failure. Therefore, to complete a transaction, two writes must be completed (the same write to each mirrored disk). This wait time can be a significant portion of the total transaction time thereby making the database system "log bound". Recognizing this, Sun designed its SPARCcenter 2000 and SPARCserver 1000 systems with optional non-volatile RAM (NVRAM). NVRAM allows a synchronous write to be posted immediately -- the database system proceeds as if the write to disk had completed reducing total transaction time. Data committed to the battery-backed NVRAM is physically moved to the disk drive in the background. In addition, data committed to NVRAM will survive a system crash -- even one involving a power failure. On reboot, the data written to NVRAM is flushed to the disk.

3.2 Packet-Switched, Multiple Bus Architecture

The SPARCcenter 2000 and SPARCserver 1000 systems provide for multiple buses each handling specific high-speed buffering between the separate subsystems. The system XDBus is used to transfer data between CPU and memory, between memory and the I/O buses and to transmit interrupts. The XDBus is packet-switched to ensure the system remains balanced as the configuration is expanded.

The unique advantage of the XDBus is its "packet-switched" design. A packet-switched bus permits substantially greater overall throughput than comparable circuit-switched buses by separating bus requests from their corresponding replies. A

by, respectively, 2 MB and 1 MB unified second level caches. The SuperSPARC chip contains approximately 3.1 million transistors; the SuperCache chip contains approximately 2.2 million transistors.

With this superscalar design, the SuperSPARC can achieve a best case instruction completion rate of two integer and one floating point instructions per clock. Since database applications are integer compute intensive, this superscalar throughput is a verifiable competitive advantage.

3.1.2 Cache and Memory Subsystem

Multiple processors increase the likelihood of system resource contention as there are more processors competing for limited resources. Specifically, database processes are often busy manipulating buffer chains and scheduling client threads. As a result, database systems can benefit greatly from large amounts of memory and fast memory access. The memory subsystem performance affects how fast the database system can chain down its buffer lists. The SPARCcenter 2000 and SPARCserver 1000 systems have many features designed to eliminate memory contention. Each of these features speeds up the processing by reducing the likelihood that a processor will need to wait for one of its supporting functional components.

The SuperCache external cache on the SuperSPARC module enhances performance in two ways. First, it permits a SuperSPARC processor to run at speeds greater than the system clock. With this feature, new processors with higher clock rates can simply be plugged into the XDBus. Second, bus loading can become a problem as the number of installed processors increases. The SuperCache helps to alleviates this problem.

A processor can establish a working set of data in its cache memory and run from that. It does this by loading the cache from system memory. Once done, no system memory accesses are needed for relatively long periods (relative to processor clock speed). The likelihood of any one processor tying up the bus/memory complex is reduced; the chances that a different processor will be able to refresh its caches without having to wait on memory are increased. Processors work more efficiently and the system performance as a whole scales as processors are added. This design approach is a good example of what is meant by balanced performance.

Using currently available 16 Mbit DRAM chips, a fully equipped SPARCcenter 2000 can offers five GB of main memory. Using the 16 Mbit DRAM chips, a fully configured SPARCserver 1000 offers two GB of main memory. All memory can be reached at the same high performance level -- no matter how much memory is configured on the system.

This modularity will protect a customer's investment across several generations of processors.

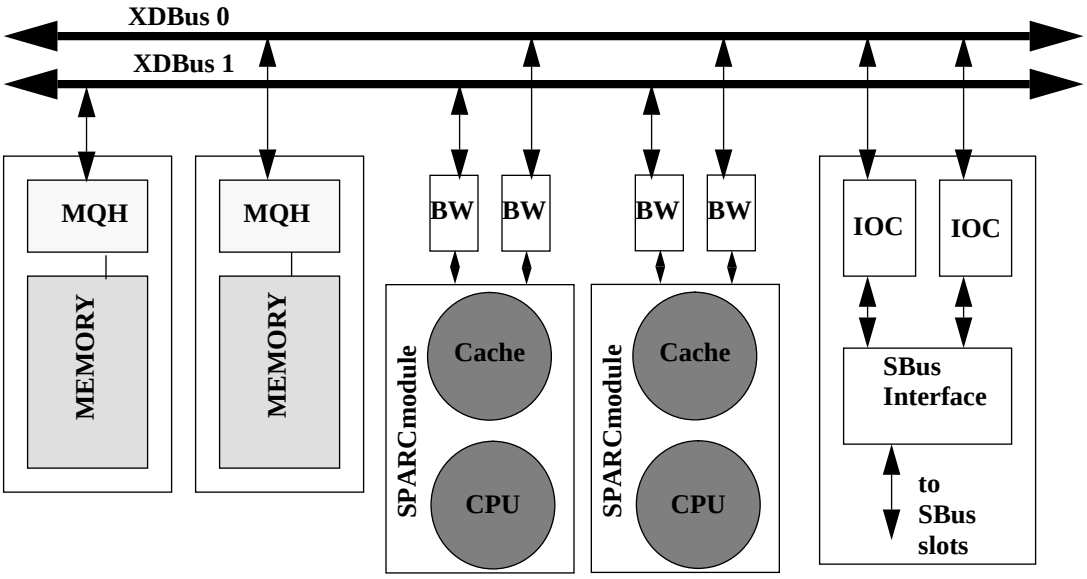

Figure 3-1 The logical organization of a SPARCcenter 2000 system board. Note all functional units are connected directly to the XDBus backplane. There are two memory units, each connected to one XDBus. The SPARCmodule contains the SuperCache and SuperSPARC CPU. The I/O Cache (IOC) provides buffers for the SBus devices. The SPARCserver 1000 has the same logical organization. However there is only one XDBus to which all functional units are connected.

3.1.1 SuperSPARC Superscalar Pipelined Architecture

The SuperSPARC microprocessor is a highly integrated superscalar SPARC microprocessor, fully compatible with the SPARC Version 8 architecture. It contains on a single chip an integer unit, a double precision floating point unit, fully consistent instruction and data caches, and a SPARC Reference MMU. The SuperSPARC modules provided with the SPARCcenter 2000 and SPARCserver 1000 are supported

SPARCcenter 2000 and SPARCserver 1000 Hardware Features

3

In order for a computer system to benefit from an SMP architecture, it needs to reduce the potential for bottlenecks due to the increased number of processors requesting information. In other words, the architecture must be balanced. As more processors are added, other aspects of the system must also be able to expand to accommodate the additions and avoid bottlenecks. These counterparts include the memory subsystem, bus, peripherals, networking capabilities and supporting software.

The SPARCcenter 2000 architecture allows up to 20 SuperSPARC processors per machine.; the SPARCserver 1000 up to eight. Processors are assured balanced access to system resources such as memory and I/O by providing high bus bandwidth over one (SS1000) or two (SC2000) system XDBuses, up to 5 GB (SC2000) of memory, and space for up to 40 SBus peripheral controllers (SC2000). The logical organization is seen in Figure 3.1 below. Sun has demonstrated that scalability for databases improves as processors are added.

3.1 SuperSPARC Architecture and Memory Subsystem

System performance is ultimately delivered by one or more of the SuperSPARC superscalar RISC processors. [1]Because of the modular design of the SuperSPARC both SPARCcenter 2000 and SPARCserver 1000 systems will easily accommodate the higher speed SPARC modules expected in 1993 and beyond.

1. The SPARCcenter 2000 is configured with 50 MHz SuperSPARCs and 2 MB external cache per Super SPARC. SPARCserver 1000s are configured with 50 MHz SuperSPARC processors with 1 MB of external cache per SuperSPARC.

In today's marketplace of standards, it is easy to build a system with hardware and software from a variety of vendors. Difficulties arise if the various pieces do not work well together. For database customers, a complete solution is the most important requirement for their mission-critical applications. This complete solution begins with a well tuned, well-balanced platform and includes a close working relationship and commitment from both the system vendor and the database vendor.

2.2 Sun's Partnership with Leading Database Vendors

Sun works with database vendors to ensure a complete solution for its customers. To the customers of Sun systems and database products, these alliances achieve the same benefits as a single vendor solution.

Each of the major database vendors: Oracle, Sybase, Informix and Ingres, is a Premier Partner in the Catalyst program. The program offers a wide variety of technical, marketing and communications services to support these vendors as they develop, port and sell their products on Sun SPARC systems. Some of these services are seminar series, cross-training to the field organizations, and technical information to help customers tune their databases.

Sun and each of the major database vendors are alpha sites for each other's products. Sun is given early access to database products and, conversely, each database vendor has early access to Sun products. This ensures a high level of integration between the products. Furthermore, the engineering groups work closely together. They have on-going performance reviews, share data and plans, and exchange source code. The source code exchange allows each vendor to make minor changes to the other's system to see what performance benefits may result. This working relationship encourages the exchange of ideas for feature improvement and enhancement.

Database systems, especially relational database management systems (RDBMS), are CPU-intensive. This means a significant portion of the execution time is spent processing algorithms and information, rather than disk I/O. Advances in RISC technology have supplied the horse-power required to speed up compute-intensive applications. Multiprocessing RISC systems boost the processing capacity further.

Since performance is a key factor in database use, server architecture should accommodate gigabytes of fast memory with low latency access. Databases typically read and write large amounts of disk data. To reduce the overhead of this I/O, they create data caches in memory. It follows that performance will increase with additional high speed memory.

Even with a large data cache, at some point the system will have to access disk data. Therefore, efficient disk I/O is also a requirement. This is especially true for decision support requests, such as strategic planning, decision making and reporting which require massive amounts of data and quick response. For both decision support and large databases, the system must be amply equipped with storage capacity and provide sufficient I/O bandwidth for increased disk access rate. Efficient, tuned buffering and caching algorithms are also a must.

Scalability is an important factor in building a balanced system. The best way to achieve scalability is to design the system to use modular components. Effort can be put into building a module with high performance and a high throughput, uniform interface. Interconnections are then built to support multiple modules. Modular components can be manufactured economically and lend themselves to easy upgrade. With the modular approach, a customer need only purchase enough to get started. As demand grows, additional components can be added to the system to expand its capabilities. For example, a multiprocessor machine can start with two or four processors and then add additional processors as demand requires.

Another advantage of modular design is that it allows customers to protect their investment. A significant cost of any database system is in its chassis and peripheral equipment. When, for example, faster processors or higher capacity memory becomes available these components can be easily swapped into the current system -- rather than writing off the current system and replacing it with a new one.

High availability and fault tolerant features are requirements for mission critical database computing. This is especially true as corporations right-size their organizations by moving from large mainframe computers to spread the work across a network of systems.

Database Server Requirements 2

2.1 Requirements Of A Database Server

Each of the varieties of database system implementations tends to be specialized and achieves these goals in different ways. However, they all need the same kind of support from the underlying hardware and software. This includes CPU power, low latency high-performance memory, optimized I/O throughput, high availability features and adaptation of the operating system to its underlying hardware. These features combine to provide a well tuned platform for database systems and can be summarized as follows:

- High transaction processing rate
- Rapid response to a large number of clients
- Room for growth and expandability.
- High availability

Above all, a database server must provide balanced access to shared system resources such as memory, buses, and peripherals. As processor power increases, the demands they make on system resources also increase. A balanced system meets these demands by assuring that there are sufficient resources to go around and that they can be used efficiently.

more performance, capacity, and connectivity is needed. The open systems philosophy has given vendors and users the flexibility they need to choose a configuration that best fits their needs.

These systems embody a broad mix of the underlying functionality that DBMS vendors rely on when building their products. For users, this translates into the freedom to choose among database systems running on a unified, scalable product line without sacrificing reliability, high performance, or low cost.

Executive Overview 1

In terms of performance, capacity, and connectivity, Database Management Systems (DBMSs) are among the most demanding applications offered. In today's computing landscape there are few computers distinguished by their ability to deliver exceptional database performance. By nature, they are costly resources. It is imperative that they be utilized fully by the enterprise wide community depending on them. Until recently, only mainframe class systems have been able to provide this specialized mix of performance and functionality. Now, as a result of innovative design and packaging technology, server class systems can offer the same or better performance and at a fraction of the cost of mainframe solutions.

Important as the price/performance point is for DBMS platforms, it is not the whole story. There are several variations of database engines available. Each variation may employ a different set of techniques to achieve performance and other goals. Trying to graft these specialized applications onto an inappropriate platform can defeat any anticipated price/performance gains. A far more satisfying answer to this problem is to port to a system designed with DBMS requirements in mind.

Sun Microsystems and the leading database vendors have worked closely together to develop hardware and software necessary to support optimized database architectures. The SPARCcenter 2000 and SPARCserver 1000 have benefitted greatly from this collaboration. They represent design features such as multiprocessor hardware and a symmetrical multiprocessing operating system which provide an excellent environment for the many simultaneous activities characteristic of DBMSs. Modular packaging has reduced costs and increased reliability. It has also made for an easy upgrade path when

Contents

Please
Recycle

Sun Microsystems Computer Corporation
2550 Garcia Avenue
Mountain View, CA 94043
U.S.A.

SPARCcenter2000 & SPARCserver1000 as Database Application Servers

Sun Microsystems Computer Corporation
2550 Garcia Avenue
Mountain View, CA 94043
U.S.A.

Part No: 8xx-xxxx-xx
Revision X, Month 1992