Computing Surface

CS-2

ENTERPRISE SERVER

- A Technical Overview

meiko

CS-2 Introduces Enterprise Servers

Many large organisations lack the requisite support from their Information Technology because of the size, age and complexity of their mainframe systems. The provision of comprehensive support services, harnessing technologies such as large scale Relational Databases are paramount to supporting these needs, and requires an innovative architecture to support an enterprise-wide repository accessible to all users who need to process queries, transactions and models – the Enterprise Server.

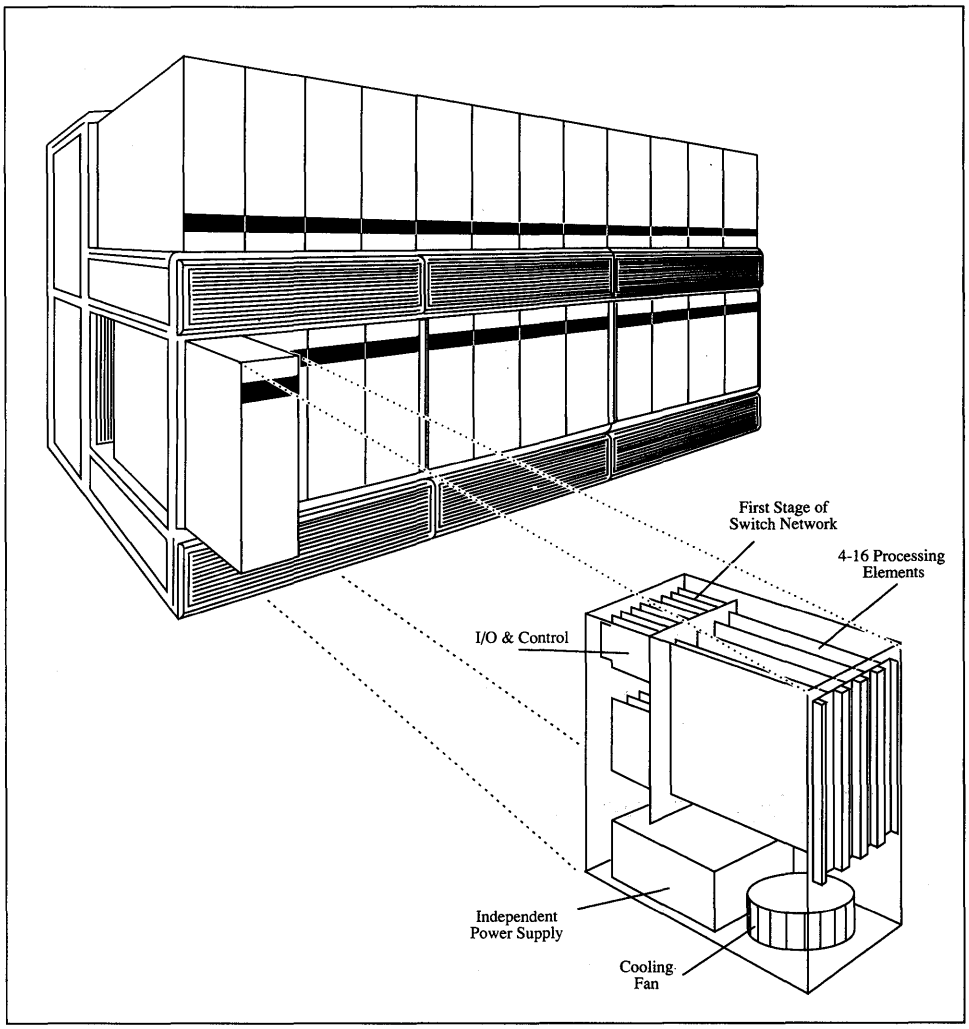
Meiko's CS-2 Enterprise Server is a Massively Parallel Processing (MPP) system designed to focus more power on commercial computing needs, at less cost than ever before. CS-2 is an MPP UNIX/RDBMS server supporting very large databases and user communities in open systems environments, more effectively than mainframes or cluster solutions.

Every facet of the CS-2 server is scalable. Achieving true scalability requires that every aspect of the architecture scales with an increasing number of processors. CPU performance, memory bandwidth, inter-processor communication bandwidth and I/O system performance all scale, such that the same applications can be developed and run on a small development machine or a large scale production system. The entry level system starts as a deskside server and can be scaled to far beyond the capabilities of today's mainframes.

CS-2 Physical Structure

CS-2 systems are modular in construction, providing flexible configuration options and component redundancy. The basic building block is a module approximately 22 x 24 x 8 inches in size containing processor boards, switch network boards for inter-processor communications or mass storage devices. The processor module contains 4 processor boards of 1 to 4 Processing Elements (PE) each, and the first stage of the switch network. All systems, whatever their size, are constructed from the same processor and switch network boards.

CS-2 Physical Structure



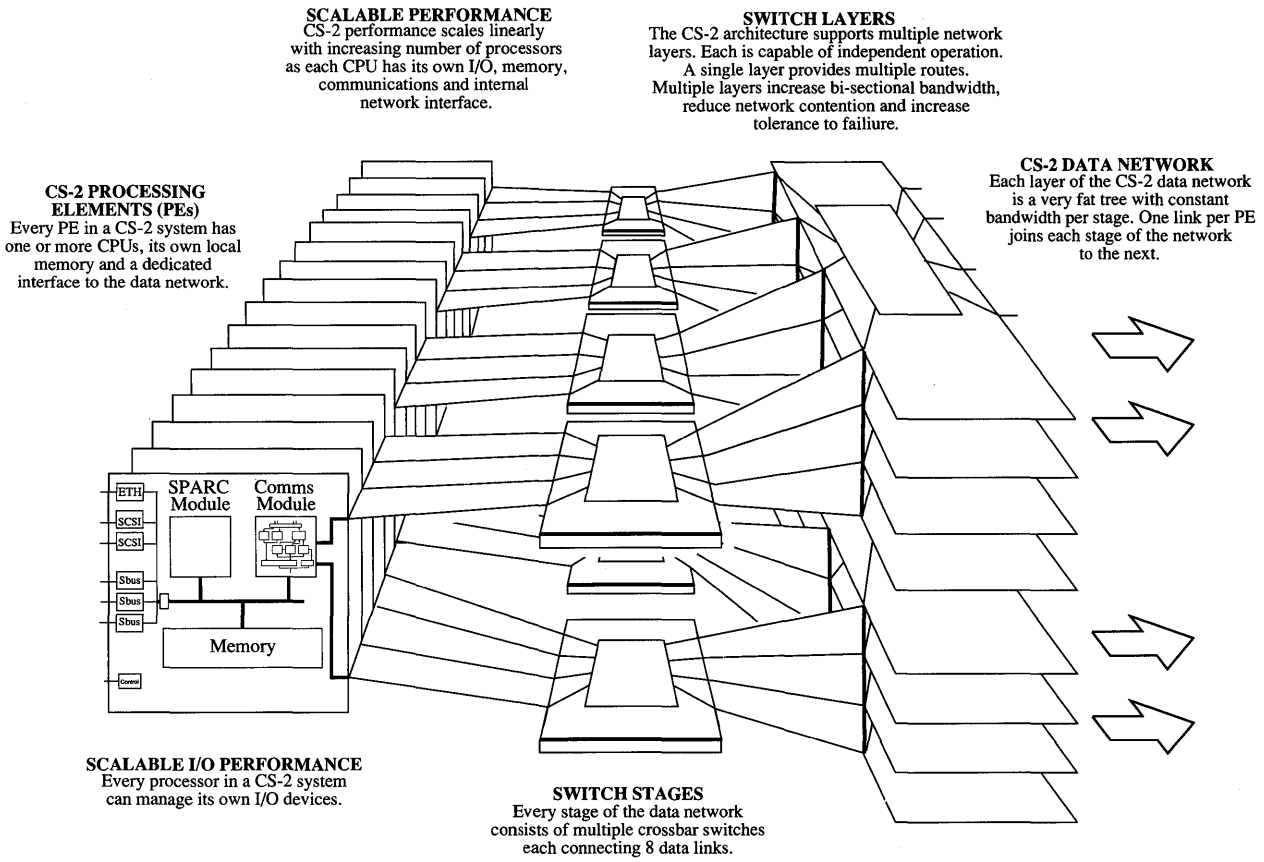
Modules are rack mounted and inter-connected in groups of 4. The 24 module system illustrated supports up to 256 superscalar SPARC PEs. Extension of this system is straightforward, with large systems constructed from multiple modules connected by a central switch.

Modules are individually powered and cooled. Cooling is by forced air with no requirement for chilled water.

Modules are capable of independent operation and self-test. Each contains a control system which monitors the health and performance of its processing and network elements. CS-2 supports live module insertion, 'hot plugging', during operation without service interruption.

CS-2 Scalable Processing

CS-2 is a distributed global memory architecture. Every PE has one or more CPUs, its own local memory system and is capable of operating independently.



The distributed memory architecture guarantees a constant ratio of CPU performance, to memory and I/O bandwidth whatever the size of system giving scalable, balanced system performance.

CS-2 has been designed to track rapid technological development of its basic components, extending system lifetime and significantly reducing the cost of ownership.

Each PE utilises the SPARC International Mbus interface for plug-in daughterboard processor connection. Mbus is an asynchronous, cache coherent multiprocessor bus standard. All SPARC processors are provided on plug-in daughterboard Mbus modules allowing customers to upgrade PE performance while preserving investment in memory systems, infrastructure, peripherals and software. The flexibility to increase the number of processors or the power of individual processing elements permits selection of the optimal upgrade path.

A CS-2 system consists of multiple PEs, each with a common interface to the data network. This interface is designed for longevity, allowing new and more powerful PEs to be added to existing systems.

The inter-processor communications infrastructure has been designed with room for growth, both in terms of the link bandwidth, number of layers, and the functionality provided by the data network. This ensures that inter-processor communications performance will keep pace with advances in microprocessor power.

Timescales for major software projects are long in comparision with the evolutionary cycle of a parallel system. Strict adherence to standard application programming interfaces combined with Meiko's commitment to high performance implementations of these interfaces ensures that applications are readily portable from one generation of technology to the next.

Meiko systems integrate smoothly into an open systems environment. They provide a reliable, high availability computational facility suitable for both interactive development and production workloads.

CS-2 Scalar Processing Elements

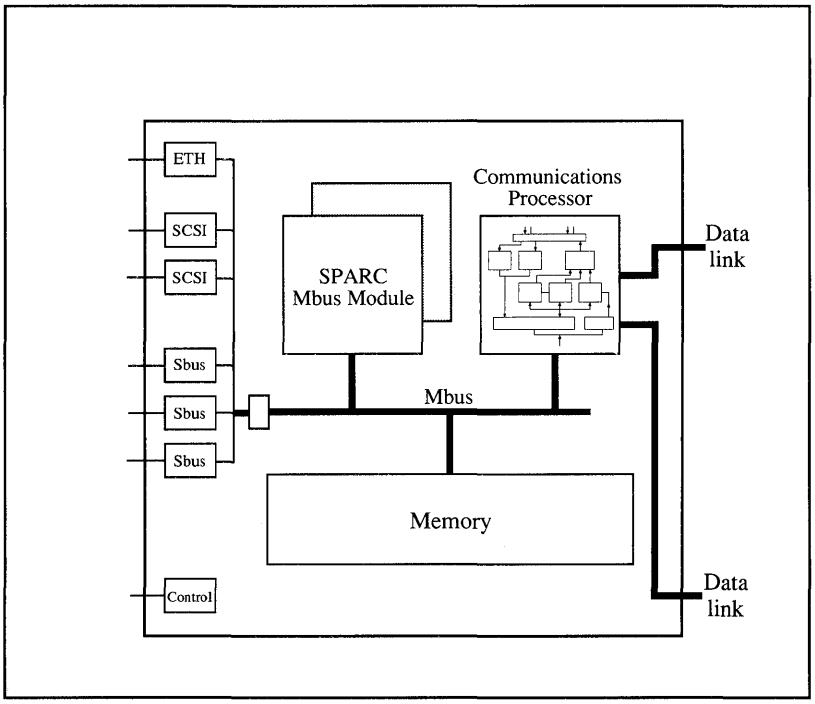
The CS-2 SPARC PE comprises a superscalar SPARC processor and a communications processor sharing a 32-512 MByte memory system. Two on-chip caches are provided: 20 KBytes of five-way associative instruction cache and 16 KBytes of four-way associative data cache. Processor modules with large (1 MByte) optional second-level caches are available.

Two variants of SPARC PE are available, one optimized for general purpose and I/O intensive applications, the other for processor intensive database and application workloads. The general purpose variant includes an Ethernet interface, a pair of SCSI-2 disk controllers and three SBus slots per PE. The processor intensive variant is more densely packaged supporting four superscalar SPARC processors and their memory systems.

All CS-2 PEs correct single bit and detect double bit memory errors. All memory errors are logged by the operating system.

Effective co-operation between PEs is a crucial factor in determining the overall sustained performance of an MPP system. Maintaining effective inter-processor communication as a system scales in size is a vital aspect of preserving balance.

CS-2 Scalar Processing Element



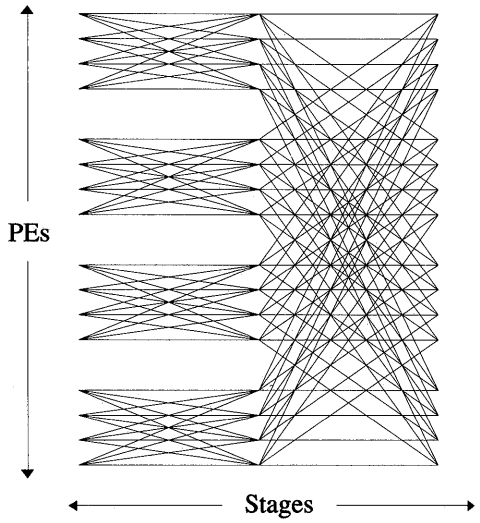
CS-2 Scalable Inter-Processor Communications

Processors share data using a highly efficient communications network. Each PE has its own interface to this network, allowing it to access data held anywhere in the system. The bi-sectional bandwidth of the CS-2 data network grows linearly with the number of PEs giving truly scalable network performance.

The CS-2 data network is a multi-stage switch network; a fat tree with constant bandwidth per stage. As the number of PEs grows, network stages are added to preserve bandwidth. The data network provides scalable inter-processor communications performance with only logarithmically increasing complexity and cost.

All CS-2 systems have at least 2 independent network layers, each a complete, independent, data network. The architecture supports up to 8 layers. Additional layers increase bi-sectional bandwidth, reduce network contention and increase tolerance to failure.

The longest path between any 2 PEs in a 256 element system is through 7 network switches.



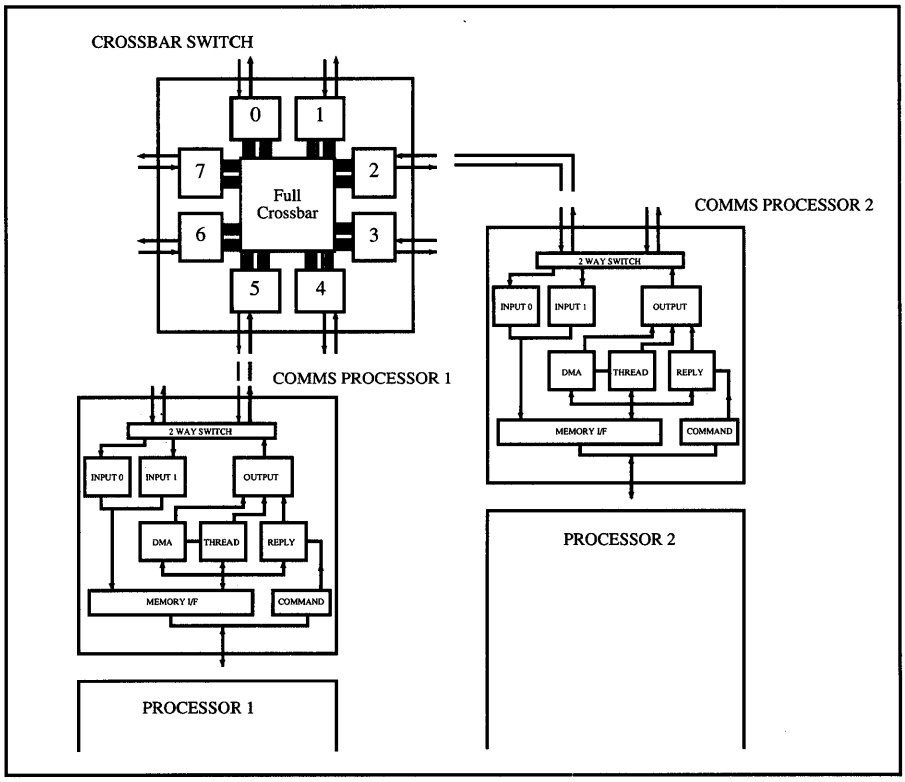
The CS-2 architecture supports up to 8 independent layers of this switching network, 1 layer being sufficient to achieve full connectivity with multiple routes. Current CS-2 systems use 2 of the 8 layers. They are engineered to expand to use all 8, giving a 256 element network a peak bi-sectional bandwidth of 102 GBytes/s.

This option for performance enhancement ensures that CS-2 inter-processor communications scale in line with anticipated increases in microcomputer processing power. Use of multiple layers is transparent to the application programmer.

In order to make the CS-2 generally applicable, interfaces to standard communications libraries have been provided. At the lowest level, the IP layer of the TCP/IP protocol has been implemented directly over the CS-2 data network. From this foundation, all of the higher level protocols that rely on this for transport services are directly available to applications and programmers. These include UNIX mechanisms such as Streams and TLI and extend up to all the important RPCs such as the Sun ONC+ RPC and the OSF/DCE RPC.

Every processing element in a CS-2 system has its own dedicated interface to the communications network; a Meiko designed communications processor. The communications processor has a SPARC shared-memory interface and 2 data links. Data links are connected by Meiko-designed 8 way cross-point switches. Each data link provides 50 MBytes/s in each direction of user bandwidth over a physical link operating at 0.6 Gbit/s in each direction.

CS-2 Inter-processor Communication



Latency is minimized in two ways. First, the communications processor manages all remote data accesses without the need for data copying, kernel intervention or main processor interrupts. Certain classes of job require this performance with perhaps the most demanding being OLTP. OLTP is characterised by a very large number of small communications requiring real time, deterministic response times. The architecture of the CS-2, with its extremely low inter-processor communications latency, is uniquely capable of supporting this load; representing a major advance over previous generations of commodity open systems.

CS-2 Scalable I/O

The CS-2 architecture provides a powerful file I/O system which is both flexible and scalable. Its flexibility derives from the fact that every PE is capable of managing its own independent I/O devices but with the data being accessible to any other processing element. The operating system permits a single large file or data table to be distributed across multiple controllers. It can be accessed concurrently at full bandwidth from large numbers of processors simultaneously delivering scalable I/O performance.

Systems are configured with a mix of devices appriopriate to their I/O requirements. Each PE can be directly connected to its own disk system or be served by remote controllers across the CS-2 network. Where concurrent I/O performance is important, for example, in large scale database applications, each processor can control its own array of fast disks.

User networks scale in the same way. Ethernet, Token Ring, ATM, X.25, FDDI, Fiberchannel and HiPPI interfaces can be added to as many processing elements as are necessary to support the load.

The CS-2 operating system is based on Solaris from SunSoft. Solaris, and strict conformance to the SPARC Software Compliance Definition 2, Application Binary Interface (ABI), provide a stable and familiar working environment giving access to the widest possible base of UNIX applications and software development tools. Solaris conforms to the X/Open Portability Guide 3, System Five Release 4 (SVR4), and POSIX P1003.1 (1990) standards.

The Solaris operating system has been agumented in four areas:

- Resource management.
- Parallel filesystem.
- Inter-processor communication.
- Parallel services for supporting the ORACLE RDBMS.

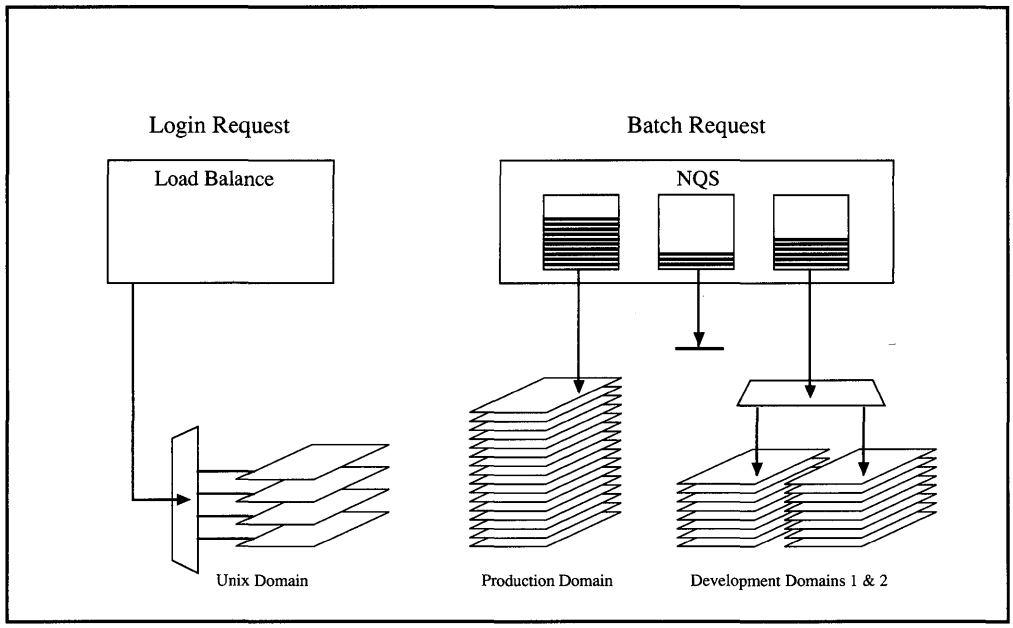
All standard features are identical to those of the market leading UNIX operating system. Unlike more traditional mainframe architectures, the CS-2 does not require specialist front-end processors. Each PE runs a full copy of the operating system and individual PEs within the system can be dedicated to the task of external data network support.

The parallel filesystem is implemented as a Solaris virtual filesystem striping the contents of its files over an arbitrary number of underlying partitions. It builds upon hardware striping used in individal devices such as RAID disk arrays.

These filesystems may be disk or network based. Therefore a single file in the parallel filesystem may be distributed over all or any of the disks and controllers available in a system. This removes the bottle-neck on seek performance and bandwidth imposed by filesystems backed up by only a single disk, or a single controller, and assures scalability.

File I/O to a single processor is at data rates up to the full inter-processor communications bandwidth as data distributed over the rest of the system converges on the requesting processor. For parallel applications with multiple channels to disk, file access rates scale accordingly.

CS-2 Resource Management



The CS-2 resource management suite extends standard UNIX to support production execution of parallel applications. It includes the access control, accounting, administration, batch processing and utilization tools necessary to manage a massively parallel system.

Systems resources, including processors, filesystems and network connections, are allocated to independently controllable groups called domains. This allocation can be changed dynamically, dedicating resources where needed. Scheduling, access control and accounting are on a per domain basis.

The system can be partitioned as required to concurrently support any application workload mix of OLTP, DSS, Batch and Modelling jobs. User logins and corrections are controllably distributed to appropriate processing resources.

The system administrator controls user access to domains and the distribution of resources between them.

The resource manager provides full control over administrative parallelism in a CS-2 system – the concurrent execution of large numbers of jobs. Its GUI controls the allocation of job queues and resources to domains, as well as providing constantly updated system status and performance information.

CS-2 Fault Tolerance

Provision of unprecedented levels of system availability is foremost in the design and implementation of the CS-2. Single points of failure have been eliminated. Errors are detected and corrected automatically where possible, detected and reported where correction is not possible.

CS-2 fault tolerance is based on guaranteeing availability in the presence of component failure. This approach extends throughout the system, from individual memory systems to whole processor modules and network layers. When combined with appropriate redundant resources, the likelihood of system failure is dramatically reduced, from the probability of an error occurring, to the probability of a second error occurring in the time taken to correct the first. Availability is increased further by the addition of multiple redundant modules.

Failures in the communications network are detected in hardware in its link layer protocol using a Cyclic Redundancy Check (CRC) data integrity check. Failed network transactions are not committed to memory but generate errors on the communications processor which cause data to be resent. The network supports multiple routes between processors, allowing data to be re-routed around failed links if necessary.

The MTBF of a modern disk drive is approximately 250,000 hours, sufficiently high for a small system. However, in a large system, and when data integrity is of vital importance, CS-2 systems use dual ported RAID disk sub-systems. Each RAID sub-system of between 5 and 20 drives provides between 2.4 and 16 GBytes of storage capacity. Drives, controllers and power supplies are all redunded, and hot pluggable in the event of failure.

The CS-2 architecture includes a control and diagnostics network which is completely independent of the data network. This network is used to monitor network performance, diagnose errors that may occur in the switch network, extract diagnostic information from individual PEs and to monitor and control power supplies and cooling systems in each module.

This network is distributed throughout the system, at board and module level. It has sufficient embedded processing power to make local decisions, issuing warnings for non-urgent classes of error and initiating module shutdown for immediate, high priority faults.

From the system administrator's point of view there maybe three types of error. Those that the system corrects itself, those that require operator intervention but do not alter the functionality of the system, and those that require replacement of a module.

The third class of serious error may cause affected jobs to be terminated, but cannot disable the operating system, which runs on at reduced capacity. In the event of such a failure a hot spare can be allocated to the domain and the job restarted. Dynamic reconfiguration does not require a system reboot. CS-2 systems can guarantee a given level of availability by redunding modules that are subject to this class of failure modes.

CS-2 Client/Server Platform

Each PE within a CS-2 is an independent computer in its own right with its own full function operating system, connected by a scalable, high performance internal network. A CS-2 can be viewed as a distributed network but within a single system. PEs can be configured to run processes that are clients or servers, at the System Administrators discretion. These client and server applications can then communicate using the CS-2 data network for transport services or with clients connected to the CS-2 by external networks. The configuration of the CS-2 in this mode is completely flexible and can be changed to service dynamic workload requirements.

To further enhance the utility of the CS-2, Meiko provides certain support services to the ORACLE Parallel Server RDBMS. Working together, these allow the CS-2 to provide a seamlessly scalable database platform utilising the true parallel architecture of the machine. All of this though is hidden from the user, and the ORACLE instance that a user or application connects to is indistinguishable from an ORACLE instance running on a conventional SPARC workstation.

Conclusion

Many large organisations are aiming to become more proactive and cooperative; to sustain their own internal growth, and to improve the services provided to the customer. Traditional mainframe technology, constrained by prohibitive cost and complex integration, cannot provide the infrastructure to support the systems for this enterprise-wide growth.

Meiko's CS-2 Enterprise Server, is a new generation of IT infrastructure, offering an Open Systems solution to the most demanding query and transaction activities. With an architecture designed for growth and flexibility, CS-2 systems bring the most effective technologies and price-performance to the datacentre, supporting the current and emerging needs of the most demanding enterprise.

For further information about Meiko and the Computing Surface CS-2, please contact Meiko at:

Meiko Scientific Corporation
1601 Trapelo Road
Waltham
MA 02154
USA

Tel: +1 617 890 7676
Fax: +1 617 890 5042

Meiko Limited
650 Aztec West
Bristol
BS12 4SD
UK

Tel: +44 (0)454 616171
Fax: +44 (0)454 618188