CM-5 I/O System

CM5

Introduction

The CM-5 I/O system meets the challenge of supporting the high performance, scalable computing of the CM-5. Taking full advantage of the CM-5's highly scalable architecture, it works in concert with the partitions of processing nodes that comprise the CM-5 to achieve the highest performance, and to create the most balanced and cost-effective supercomputer system available today. Users of the CM-5 are given seamless access to their data, display devices, and other computers while sustaining overall machine throughput and response. Because it is built on top of the scalable architecture of the CM-5, the I/O system can accommodate the widest possible range of application requirements.

True I/O scalability is achieved by connecting I/O devices directly to the CM-5 Data Network. By this means:

- Any I/O device, or collection of I/O devices, can be given the necessary bandwidth to achieve high throughput.
- A system's I/O capacity can be determined strictly from a site's I/O requirements, independent of the system's computational capacity.

In addition to scalable bandwidth, the CM-5 I/O system supports industry software standards, such as UNIX file system access and TCP/IP networking with other machines. The CM-5 Operating System (CMOST) provides straightforward, consistent methods for performing I/O, and accommodates a wide range of I/O devices, such as the Scalable Disk Array, Integrated Tape System, HIPPI, and FDDI network connections.

Functional Highlights

- Scalable I/O supports a virtually unrestricted choice of performance and capacity, allowing balanced combinations of:
  - High bandwidth disk systems
  - High capacity (lowest cost/Mbyte) disk systems
  - Tape storage
- UNIX I/O. All I/O operations are idealized as generic operations on file systems. These operations include open(), close(), read(), and write(); they apply to disk files, pipes, and sockets. A new file system, the Connection Machine Scalable File System, is added to support parallel disk devices and files larger than 2 Gbytes.
- I/O devices are shared resources. Any partition or any computer connected to the CM-5 by a LAN may access all CM-5 I/O devices. Additionally, I/O devices may communicate directly with one another. This allows, for instance, direct disk-to-tape copies without the use of partition resources.
- Data is always preserved in serial order and is therefore accessible both by CM-5 partitions of different sizes and by serial machines.
- I/O from one partition does not affect the performance of other partitions. Simultaneous I/Os from several partitions see minimal interaction if not using the same device.
- The CM-5 supports standard LAN connections (Ethernet, FDDI, HIPPI) using appropriate software standards (TCP/IP, NFS).

Performance without Pain

The CM-5 I/O system has three paramount features:

1. Performance and capacity are scalable independent of processing power.
2. The system's full performance can be brought to bear on a single application.
3. Applications have seamless access at all times to data throughout the system.

High Performance, Scalable I/O

On the CM-5, I/O devices do not connect to a special I/O bus, but instead connect directly to the CM-5 Data Network. If a device needs more bandwidth than one connection can supply, it is given multiple connections. As a result, I/O capacity is expanded merely by adding Data Network connections. Since the bandwidth of the CM-5 Data Network expands linearly with the number of connections, the performance of the Data Network expands to meet the additional needs of I/O devices.

Bringing Performance to Bear

This ability to increase the aggregate I/O bandwidth in an unlimited way is a significant achievement, but is not sufficient. Parallel computers that can run n applications on each of n processors do not meet the needs of today's users. The real challenge is to deliver a system that can apply thousands of processors to a single task in a coordinated way. An I/O system that cannot similarly deliver scalable I/O performance to a single application is not meeting the challenge. The CM-5 I/O system, however, can.

(over)

CM-5 I/O System (continued)

For example, consider the following Fortran code fragment, where a partition of processors reads data into the parallel array, HugeArray, from a file named array.dat, which is stored on the CM-5 Scalable Disk Array:

```
Real HugeArray (1000000)
Open (1, file='array.dat', form='unformatted')
Read (1) HugeArray
Close (1)
```

All disks in the parallel file system act together to transfer data to the partition. A parallel file system of 50 disks transfers at 50 times the speed of an individual disk; a system with 100 disks transfers 100 times faster. The operating system knows how data in the array HugeArray is spread across the processors, and automatically spreads the data coming from the Scalable Disk Array to the correct location. From the application's perspective, the Scalable Disk Array appears to be a single, high capacity, high performance disk. This is true scalability.

Seamless Access of Data

The heart of the CM-5 I/O system is its seamlessness.

- Users can access all data, on any device in the system, using familiar, simple, intuitive methods.
- System software sees partitions and I/O devices as similar entities. (In fact, the operating system handles partition-to-partition transactions using standard I/O software mechanisms.)
- The CM-5 Operating System provides a file system accessible via NFS, and communication via UNIX sockets. This use of standards allows both CM-5 processors and external serial machines to access data in a uniform and consistent manner.

In parallel programming, the elements of an array are typically spread across many processors. Depending upon choices made automatically by the compiler or under the direction of the applications programmer, the layout of two identically sized arrays may not be the same. Additionally, the actual layout varies with partition size. Serial machines view their arrays in the order in which they exist in memory (generally row- or column-major order). For different sized partitions to share files, or for serial machines to share data with a parallel program, the data must be stored in an order that both a parallel program and a serial program can understand.

The CM-5 provides a seamless mechanism for moving data between the CM-5 processors, CM-5 I/O devices, other serial machines, and I/O devices connected to other machines. Special-purpose hardware in the I/O devices assists the operating system in handling data-ordering issues. This combination of hardware and software support removes the burden of dealing with data-ordering issues from the application programmer.

Consider again the code fragment:

```
Real HugeArray (1000000)
Open (1, file='array.dat', form='unformatted')
Read (1) HugeArray
Close (1)
```

Notice that the application specifies nothing regarding special characteristics of the disk storing the file array.dat. If the file is on Disk Storage Nodes, the operating system gets the data from there. If the file is on a disk connected to a Control Processor, the operating system gets it from there. If the data is on a server attached via a LAN to the CM-5, the operating system uses NFS to retrieve the data. This property of the CM-5 system sets the standard for seamless data access.

I/O Products

The following CM-5 I/O products are available:

- Disks (Scalable Disk Array—SDA)
- Networks
  - HIPPI
  - FDDI
  - Ethernet
- Tapes (Integrated Tape System—ITS)
  - 3480/3490-compatible
  - EXABYTE 8500

The following CM-2 I/O products are supported on the CM-5:

- CM-2 DataVault (60 Gbyte Disk Drive)
- CM-2 HIPPI Interface
- CM-2 IOP (Tape Controller)
- CM-2 VMEIO Adaptor