IMPROVEMENT IN HPC PERFORMANCE THROUGH HIPPI RAID STORAGE

Blake Homan
Maximum Strategy, Inc.
801 Buckeye Ct.
Milpitas, CA 95035
Tel: (408) 383-1600
Fax: (408) 383-1616
blakeh@maxstrat.com

RAID History

In 1986, RAID (Redundant Array of Inexpensive [or Independent] Disks) technology was introduced as a viable solution to the I/O bottleneck. A number of different RAID levels were defined, in 1987 by the Computer Science Division (EECS) University of California, Berkeley, each with specific advantages and disadvantages.

With multiple RAID options available, taking advantage of RAID technology required matching particular RAID levels with specific applications. It was not possible to use one RAID device to address all applications. Maximum Strategy's Gen 4 Storage Server addresses this issue with a new capability called *Programmable RAID Level Partitioning*. This capability enables users to have multiple RAID levels coexist on the same disks, thereby providing the versatility necessary for multiple concurrent applications.

Architecture

Gen 4 is essentially a parallel computer. Multiple CPUs work in parallel to facilitate the asynchronous data transfer to and from up to 20 IPI-2 channels and one or two HIPPI channels.

Gen 4 utilizes a Motorola 68040 microprocessor running a powerful real-time, multitasking operating system. The 68040 is the centralized task manager for all command and control.

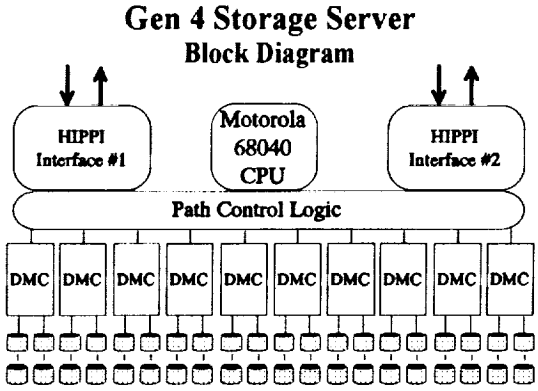
Each dual IPI-2 interface also uses a Motorola 68000 microprocessor running a real-time operating system and two independent microcontrollers to control the IPI-2 channels.

Gen 4 may be configured with one or two HIPPI channels, the open systems standard for high performance computing (HPC), and utilizes the IPI-3 command set.

Internally, the HIPPI interface controls data mapping between its high-speed buffers, and all IPI-2 channels. Additional dedicated hardware performs the functions required for RAID 3 or 5 data recovery.

Three additional interfaces, one Ethernet port and two RS-232 ports are available for external communication. The Ethernet is also capable of transferring data, however it is not well-suited for transferring at high-performance levels. The Ethernet port is best suited for third-party or complex data transfers where the command/status information is sent over Ethernet, and data only is sent over the HIPPI channel. This provides the capability to interface with various distributed computing solutions that are now becoming available.

Gen 4 may be managed and configured by the host using the IPI-3 command set, or from a system management console via RS-232 or Ethernet, allowing the operator to monitor real-time status of the system through a menu-driven interface.



Because Gen 4 is highly configurable, a user may start out with a minimum investment of five dual IPI-2 interfaces and one bank of disks. The system can grow to 10 dual IPI-2 interfaces and two banks of disks to increase the capacity and performance. Total formatted storage capacity ranges from 12 GB to 53 GB.

Programmable RAID Level Partitioning

Traditional RAID solutions suffer performance limitations and narrow applications viability because they are confined to one RAID level. By partitioning storage into different RAID levels, Gen 4 can be configured to handle all types of data, making it applicable to a variety of HPC applications.

Programmable RAID level partitioning enables simultaneous support of RAID levels 1, 3, and 5 on user-defined partitions of storage space. This ensures maximum performance is attained for both large and small I/O. Each RAID level has an advantage for different data types and applications.

RAID level 1 is best for smaller I/O, such as file system information, allowing the host to avoid the write penalty associated with RAID level 5 and the single block latency of RAID level 3. (Block sizes 4K, 8K, 32K, 64K)

RAID level 3 is well-suited for large I/O where high sustained bandwidth is required such as main memory swaps or large data sets. (Block sizes 256K and greater)

RAID level 5 can also provide high sustained bandwidth in addition to being uniquely capable of addressing I/O-intensive applications involving large or small amounts of data. (Block sizes 32K, 64K)

In a typical file system, a RAID 1 partition would be best suited for the superblock or metadata information, since this usually is a very small and frequently accessed portion of the storage. Additionally, other small, frequently used files could also be stored in this partition, however they would occupy a small portion of the available storage.

A RAID 5 partition would be best for mid-sized and large files. The RAID 5 partition provides high I/O rates for smaller files as well as high throughput for larger files. The number of files stored in this partition would be less than the RAID 1 partition, however they would require the majority of the available storage.

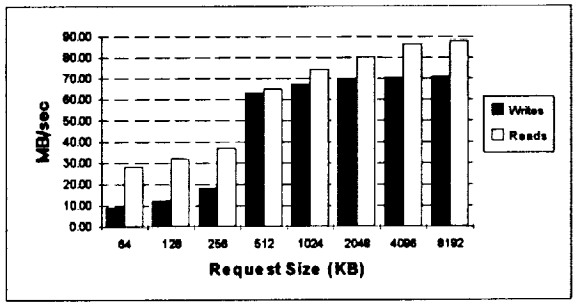
Performance

When addressing many independent I/O requests, Gen 4's large number of I/Os per second enables a sustained transfer rate of 30 MB/sec. When reading or writing large files, or when handling many sequential requests, Gen 4 supports sustained transfer rates in excess of 85 MB/sec. The following charts show the performance capabilities of the Gen 4:

Gen 4 System Performance
I/O Performance (64K Block Size)

RAID Level 5	RAID Level 1
560 reads per second	560 reads per second
180 writes per second	290 writes per second
375 4:1 mix	450 4:1 mix

Sustained Throughput



Reliability and Availability

Each RAID level incorporates fault-tolerance (automatic data-error correction). RAID levels 3 and 5 generate parity data which is stored along with user data. RAID level 1 duplicates or "mirrors" all user data in its entirety. In the event of an uncorrectable media error or disk failure, Gen 4 uses the parity or mirrored data to create the lost original data.

Another measure of reliability is data availability. Because disk failures do occur, Gen 4 features standby disks which are immediately available to replace failed disks. Data from the failed disk is reconstructed using the remaining disks and written to the standby. This reconstruction process takes place while the system remains operational.

Gen 4 also features hot replacement technology. Each disk in the array is packaged in a cartridge with its own power supply and cooling fan. This design allows users to remove an individual disk for maintenance, while the system remains on-line and functioning at full capacity.

Disk Interface

Because of the IPI-2 standard, the system is capable of supporting new disk technology as it becomes available. IPI-2 disks are currently the fastest and highest capacity disks available.

Investment Protection

The Gen 4 has been designed with the future in mind, utilizing industry standards. First, with the standardized HIPPI interface, users attaching a system directly to a host today, can move that same storage to a new host or to a distributed, network-attached storage architecture in the future.

Additionally, as higher performance standard channels and larger capacity disks become available, previously developed applications can be easily ported to take advantage of new storage capabilities.

Summary

RAID technology has become the accepted solution to the I/O bottleneck in the HPC community. As HPC becomes more mainstream, it is important that users have the flexibility to mix and match hosts with the highest performance peripherals. The HIPPI standard has been a major milestone in this process. Other fabrics such as Serial HIPPI and Fibre Channel are in the early stages of standardization.

Maximum Strategy is dedicated to improving the ability of the high-performance computing marketplace to provide solutions by increasing the transaction rates, throughput, and mean time to data loss of its storage solutions. This will allow researchers and businesses to solve problems much faster, save money and resources, and make HPC more interactive for multiple users.