RAID Disk Arrays for High Bandwidth Applications

Bill Moren
Ciprico, Inc.
2800 Campus Drive
Plymouth MN 55441 .
bmoren@ciprico.com
612-551-4000
FAX: 612-551-4002

Introduction

High bandwidth applications require large amounts of data transferred to/from storage devices at extremely high data rates. Further, these applications often are 'real-time', in which access to the storage device must take place on the schedule of the data source, not the storage. A good example is a satellite downlink - the volume of data is quite large and the date rates quite high (dozens of MB/sec, typically). Further, a telemetry downlink must take place while the satellite is overhead; once it passes over the horizon the telemetry is lost forever.

A storage technology which is ideally suited to these types of applications is RAID (Redundant Arrays of Independent Disks). The concepts of RAID were presented in an academic paper from the University of California's Berkeley campus in the mid-1980s. This paper (often referred to as the 'RAID paper') offered five different architectures, colloquially referred to as the 'RAID levels'. Each RAID level, numbered one through five, defined a different methodology for using multiple disks grouped together to improve performance and offer redundancy. Each of the levels had distinct strengths and weaknesses. It is a fallacy to believe the RAID levels with higher numbers (e.g. RAID-4 versus RAID-2) are superior; the ideal RAID level for an application varies with applications - one application may find RAID-1 best suited, RAID-5 for another, and yet another application's best choice may be RAID-3.

RAID Levels

RAID-1 is classic disk mirroring, in which every disk has a mirror image of its data stored on another disk. This level was the frame of reference in the RAID paper. Mirroring has been around for some time, primarily in mainframe computing. Its strengths are redundancy and performance. Any single drive in any given data pair may fail and the disk system will remain accessible, though at a reduced performance level. Because there are two disks for any given piece of data, read performance is quite good as any two arbitrary requests for a single logical disk can be serviced simultaneously on two physical disks. However, the cost for mirroring is quite high - essentially a 100%

premium since every disk is duplicated. The power, cooling, and packaging costs are also quite high. Reliability is also halved because of the duplication of disks.

RAID-2 and -3 stripe user data across a group of data drives (typically four or eight drives per group). Every block of user data is striped, typically a byte at a time, resulting in all the data disks servicing every user data request in parallel. This results in extremely high data transfer rates, since multiple disks are transferring data simultaneously. RAID-2 and -3 differ in their redundancy methodologies. RAID-2 uses multiple disks to implement a Hamming error detection and correction code. The codes stored on a RAID-2's redundant disks were generated from the data on the data disks. RAID-3 uses a single redundant disk to store a error correction code generated by calculating the logical exclusive-or of the data on the data disks. Because RAID controller technology doesn't require the use of a Hamming code to detect a failed drive, RAID-2 hasn't found commercial acceptance as it is more costly than RAID-3.

RAID-4 and RAID-5 also stripe user data across a group of data drives. However, instead of striping every block of data across all drives, each block (or sometimes groups of blocks) is stored entirely on an individual disk. This results in good transaction performance as each disk in the group can service separate requests for individual blocks, simultaneously. RAID-4 and -5 differ in the methodology used for storing the error correction codes. Both use the exclusive-or code as used in RAID-3. RAID-4 dedicates one drive for the error correction codes while RAID-5 rotates the codes throughout all drives in the array. RAID-5 has better write performance because of this rotation as there is less contention for access to the redundant codes.

The Right RAID Level for High Bandwidth Applications

Real-time, high bandwidth applications require the following from disk storage: high sustained data transfer rate under all normal operating conditions. Of all the RAID levels, only RAID-3 fits the profile.

RAID-4 and RAID-5 don't fit because their performance characteristics are designed for delivering a large number of independent requests (high I/Os per second). These RAID levels operate best when each disk is servicing a separate request. However, high bandwidth applications are characterized by large sequentially stored data sets. For such data sets, transfer rate (measured in MB/sec) is the important metric, not I/Os per second. Also, both RAID-4 and RAID-5 have severe performance degradations after a drive failure, which is considered a normal operating condition in RAID disk arrays. For real-time applications this is unacceptable as it is imperative that the RAID subsystem be able to service any request, at any time, regardless if there has been a drive failure.

RAID-3 fits for two primary reasons. First, because all user data is striped across all drives, transfer rate is very high. This is true for either reading or writing. In general, a RAID-3 disk array will have a sustained transfer rate equal to the product of sustained transfer rate of the disks used in the array and the number of data drives in the array. Second, RAID-3 doesn't suffer any performance degradation after a drive fails. Because all of the drives are accessed for each data request, there always is sufficient information being transferred from the array that can be combined with the error correction code (which is also always transferred on every data request) to generate the failed drive's data. Special hardware on a RAID-3 controller is able to perform the failed drive's data reconstruction on-the-fly, with no performance loss.

Other Factors to Consider

In addition to the media redundancy inherent in RAID, other subsystem components should be protected against failure. For instance, most RAID subsystems include AC to DC power supplies. These units have failure rates similar to disk drives. Power supply redundancy should also be considered. One good approach is to incorporate dual, load-sharing power supplies in the RAID subsystem. Each power supply has sufficient power to operate the entire subsystem in case the other should fail.

Another subsystem component worth considering for redundancy are the cooling fans. Fans, being a mechanical device, are also prone to failures. A RAID subsystem can incorporate redundant fans to protect against overheating in case of a fan failure.

All redundant components, drives, power supplies, and cooling fans, can support 'hot swapping'. Hot swapping is the ability to replace a failed component without shutting the subsystem down or taking it offline. Most hot swap components will be housed in canisters or carriers which slide into the RAID subsystem.

Another factor to consider is the host interface. The host interface directly affects the performance a RAID disk array will be able to deliver. The most common interface found is SCSI-2. It is a 16-bit wide parallel interface which clocks data at 10 MHZ for a burst rate of 20 MB/sec. Sustained rates of over 19 MB/sec are possible with SCSI-2 RAID-3 disk array.

The successor to SCSI-2, SCSI-3, includes a performance improvement to 40 MB/sec. This capability, sometimes referred to as UltraSCSI, is backward compatible with SCSI-2. SCSI-3 uses the same 16-bit wide parallel interface as SCSI-2, but data is clocked at 20 MHz, instead of SCSI-2's 10 MHz. UltraSCSI RAID-3 disk arrays are capable of sustained data rates in excess of 38 MB/sec.

Another interface which offers excellent high bandwidth performance is Fibre Channel. This is a serial interface which is clocked at 1 Gbit/sec with a sustained interface capability of 100 MB/sec. Fibre Channel is not physically compatible with SCSI-2 or -3 but is software compatible. Fibre Channel supports a number of software protocols which are encapsulated in 'frames' which are the data packets that are transferred between Fibre Channel nodes. SCSI is one of the software protocols supported. The first Fibre Channel compatible RAID-3 disk arrays are becoming available in 1996 with sustained data rates of nearly 90 MB/sec.

A good example of high bandwidth RAID-3 disk arrays are those available from Ciprico, Inc. (Minneapolis, MN). Ciprico offers a full line of high bandwidth disk arrays which are well suited to real-time, high bandwidth applications. Ciprico's arrays all offer high data transfer rate, no performance degradation after drive failures, and media redundancy. There are a number of interface, redundancy, and capacity options, designed to support a variety of applications. Table 1 summarizes the capabilities of Ciprico's disk array.

Model	Interface	Burst Transfer Rate	Sustained Transfer Rate	Redundancy	Hot Swap
6500	UltraSCSI	40 MB/sec	38 MB/sec	Drives	No
6700	SCSI-2	20 MB/sec	19 MB/sec	Drives Power	YES
6900	UltraSCSI	40 MB/sec	38 MB/sec	Drives Power	YES
7000	Fibre Channel	100 MB/sec	80+ MB/sec	Drives Power Fans	YES

Table 1 - Ciprico's RAID-3 disk arrays offer a variety of performance and redundancy options. Users can select an array which best fits their application.

Summary

High bandwidth applications require high sustained data transfer rates under all operating conditions. RAID storage technology, while offering differing methodologies for a variety of applications, supports the performance and redundancy required in real-time

applications. Of the various RAID levels, RAID-3 is the only one which provides high data transfer rate under all operating conditions, including after a drive failure.