A MORGAN-GRAMPIAN PUBLICATION

DIGITAL DESIGN

SYSTEMS ARCHITECTURE, INTEGRATION AND APPLICATIONS

SEPTEMBER 1984

- STD BUS: DIVERSITY AT THE LOW END
- SILICON SUPPORT FOR SOLID MODELING
- PROCESS CONTROL BRINGS INTELLIGENCE TO THE FACTORY FLOOR

- DEVELOPMENT TOOLS
- EEPROMs
- ERGONOMICS
- DISK CONTROLLERS

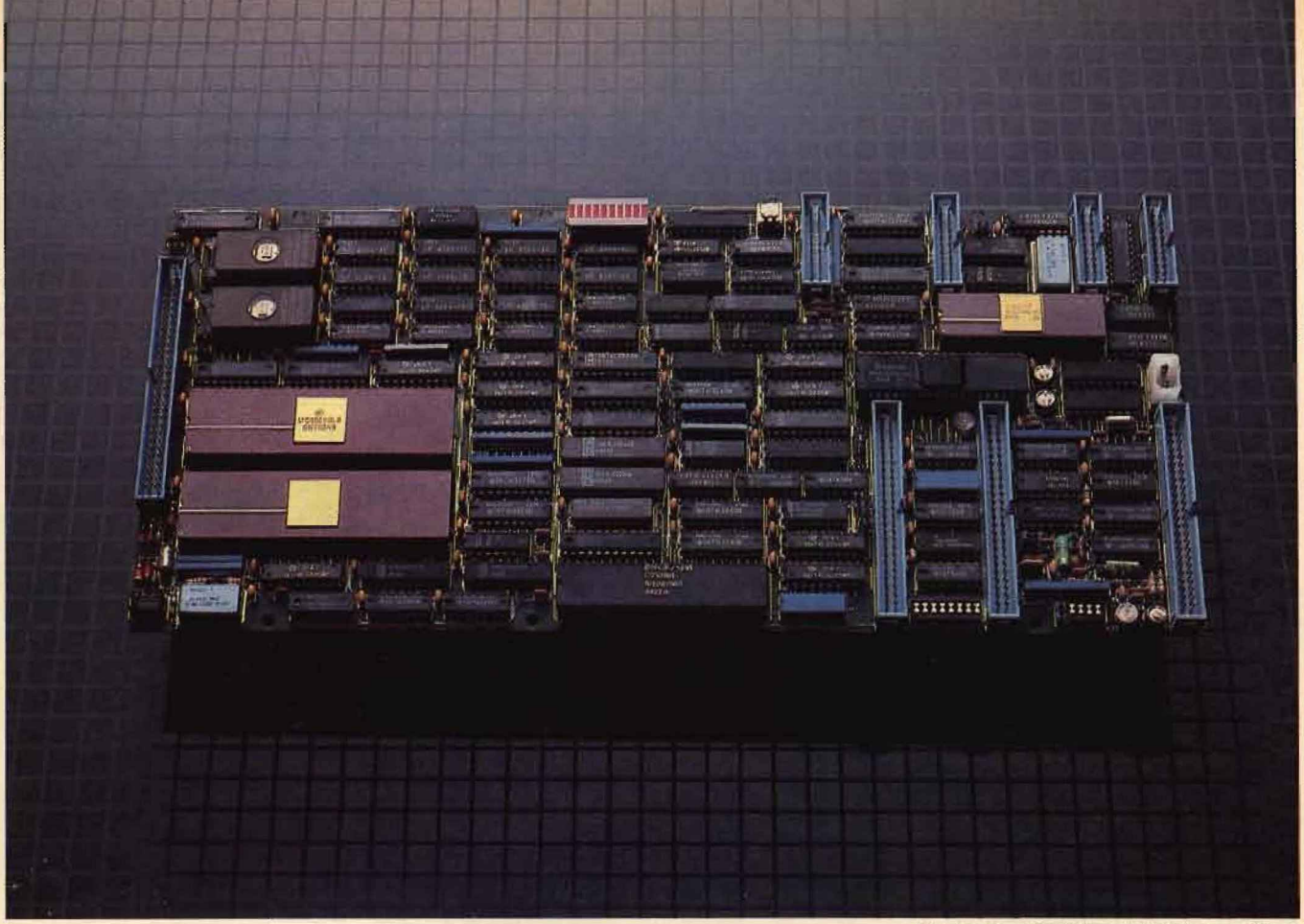

PHOTO COURTESY DISTRIBUTED PROCESSING TECHNOLOGY

Caching Disk Controller Relieves System Bottlenecks

by Stephen Goldman

Processors have become faster, disks have become larger, and data rates have increased, but one bottleneck remains in many small computer systems. Disk rotational speed and thus data access time have not been significantly enhanced and are not likely to be in the near future. This dependence upon a moving part for what is, in most systems, the key memory device, can cause severe performance degradation.

Multi-user or multi-tasking operating systems such as UNIX can aggravate the problem by requiring data to be simultaneously accessed on widely separated areas of a disk. With average cylinder seek times of around 85msec for a typical rigid disk drive, system throughput can be excessively slow.

A family of caching disk controllers, the PM-3010 series, has been designed to relieve this bottleneck. The combination of fast access time (400 μsec. worst case) and high sustained data rate (1.0 Mbyte/sec) along with up to 16 Mbytes of cache, allow disk throughput to be substantially increased. Up to eight rigid and flexible disk drives can be controlled by a single 5¼" extended form factor board, using one 5VDC power source. The disk controller family controls a range of drives including Winchesters and flexible disks. ST506, SA1000 and SMD compatible Winchester drives with up to 16 heads and 1024 cylinders are supported along with SA460 and SA860 compatible flexible disk drives.

The controllers use the Small Computer System Interface (SCSI) for host communications. The SCSI specification, proposed by the American National Standards Institute (ANSI), defines both hardware and software protocol for device controllers. The primary objective of the SCSI interface is to provide host computers with device independence. Disk drives, tape drives, printers, and communications devices, of different types, can be added to the host computer(s) without modifying the interface hardware and with little or no change in generic system software.

Standard SCSI Command Set

The PM-3010 supports the standard and

Stephen Goldman received a BS in Physics from the University of Florida and was a member of the technical staff of Georgia Tech before founding Distributed Processing Technology in 1977.

The primary technical problem to be overcome during the design of a caching disk controller is the limitation on cache access time.

extended SCSI command set. An important objective during its design was to assure that caching operations were transparent to the host computer and did not require modification of existing software. The same basic SCSI commands are used by both the controller and other SCSI compatible disk controllers. The primary difference is in the execution speed of those commands.

The worst case access time for cache resident data with the PM-3010 is 400μsec. Access time is measured starting with receipt of the last byte of the host command and ending with the first byte of the data transfer between controller and host. Without cache, the access time would be limited by the rotational latency and cylinder seek time of the disk drive. For ST506 compatible Winchester drives, the average rotational latency is 8.3msec and the average cylinder seek time is typically 85msec. This means that if the data can be accessed in cache, the PM-3010 will begin its data transfer in 1/200th the time it takes a conventional disk controller to do a random disk access. Even if the disk read/write heads are already positioned at the correct cylinder so that no seek is required, the disk access will still take 20 times longer than the worst case PM-3010 cache access.

Once the data transfer is initiated, it proceeds asynchronously at a rate determined by the host and the controller. The PM-3010 is capable of handling data transfers at up to 1.0 Mbyte/sec. Although some non-caching disk controllers can also transfer data across the host interface at that rate, the data must additionally be transferred between the controller and the disk. Since data transfers to ST506 compatible drives are slower than 1.0 Mbyte/sec, a non-caching disk controller will not be able to maintain a sustained data rate of 1.0 Mbyte/sec.

In theory, the best possible sustained data transfer rate that a disk controller without cache could maintain is limited by the average data rate for sequential sectors on a disk. Using the example of a controller capable of handling a sector interleave of zero, at 256 bytes/sector, the best average data rate for an ST506 compatible drive is less than 500 Kbytes/sec. Even if no extra time was taken up by the subsequent transfer of data to the host, the PM-3010 would still transfer data at twice the rate of a conventional "high-performance" disk controller.

Cache Expandable To 16 Mbytes

Since cache accesses are so much more rapid than disk accesses, the key to optimizing performance with a caching disk controller is to maintain the highest possible "hit ratio." Hit ratio is defined as the number of times that data requested by the host is found in cache, divided by the total number of data requests.

Although highly dependent upon the application program, hit ratios as a rule, can be increased by raising the amount of cache memory. The PM-3010 comes with 128 Kbytes of cache memory integrated onto one 5.75" by 12.5" circuit board. Memory expansion boards may be added to increase the size of the cache. With a maximum supported cache size of 16 Mbytes, approximately 64,000 disk sectors of 256 bytes each can simultaneously reside in cache with sector sizes of 128, 512, and 1024 bytes also supported. The ability to expand cache as needed allows the system designer to configure the controller for maximum performance without requiring more cache than needed for his application.

Sector Vs. Track Caching

The primary technical problem to be overcome during the design of a caching disk controller is the limitation on cache access time. If the access time for cache resident data is too long, a cache system may, in certain cases, slow down an application rather than speed it up. In order to reduce cache access time, many caching disk controllers resort to track rather than sector caching.

Track caching schemes access all sectors on a track as a group when moving data between cache and disk. Since entire tracks are stored in cache instead of individual sectors, the search algorithms are much simpler. The primary problem with track caching schemes is the additional overhead needed to access and store entire tracks of disk data when only one sector may be needed by the host. Track caching works well when large blocks of physically contiguous data are accessed, but performance may be seriously degraded for random access file structures such as hierarchical or non-contiguous files.

To avoid this the PM-3010 uses sector rather than track caching. Each sector in the cache is handled separately and only the sectors needed by the host are read into cache. However, even with up to 64,000 separate sectors residing in the full 16 Mbytes of cache, worst case cache access time will never exceed 400μsec. When sequential sectors are accessed, throughput can be further increased by addressing multiple sectors with a single command. In this case, the rated access time applies only to the first sector in the group. All additional sectors addressed by the command will be transferred across the host interface without significant delay.

High Performance Internal Architecture

To search 16 Mbytes of cache in less than 400μsec required development of new caching algorithms executed in firmware by an on-board 68000 microprocessor. The proprietary algorithm organizes the

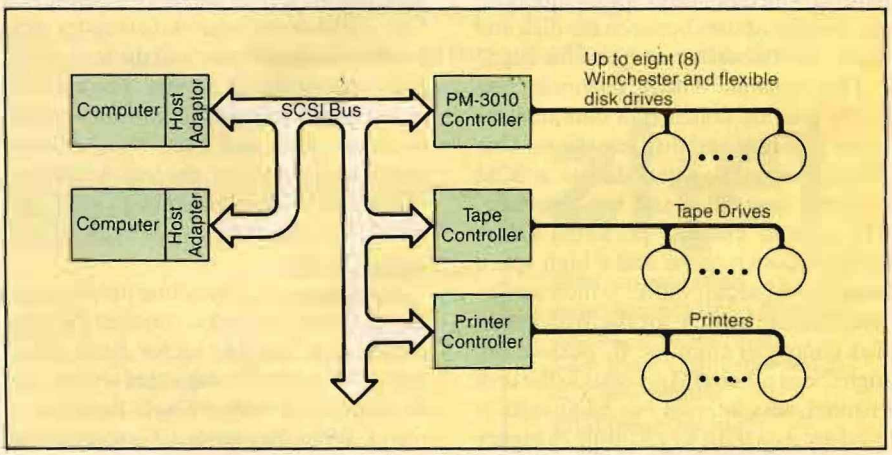

Figure 1: Up to eight devices such as disk or tape controllers, printers, communications devices, and computers can share a single SCSI bus.

COMMAND	DESCRIPTION
INQUIRY	Returns information regarding the Class of Device of the controller and attached drives.
MODE SENSE	Returns drive parameters and media format.
MODE SELECT	Sets drive parameters and media format.
READ CAPACITY	Returns size of usable media or locates areas of contiguous storage.
TEST UNIT READY	Confirms the drive is powered on and ready.
REQUEST SENSE	Returns information regarding command completion including errors and statistical reporting.
READ	Reads data from disk (and cache).
READ EXTENDED	Same as READ but supports larger address space on disk.
WRITE	Writes data to disk (and cache).
WRITE EXTENDED	Same as WRITE but supports larger address space on disk.
WRITE AND VERIFY	Writes data and then verifies CRC (or ECC) or performs byte-by-byte comparison of data.
VERIFY	Verifies CRC (or ECC) or performs byte-by-byte comparison of data with data from host.
*PRE-READ	Pre-fetches data into cache before needed by host.
REZERO UNIT	Recalibrates drive to cylinder zero.
SEEK	Moves head to desired cylinder (not mandatory).
START/STOP UNIT	Controls spindle motor and seeks head landing zone before power-down.
LOCK/UNLOCK MEDIUM	Clears cache and allows media removal.
*LOCK/UNLOCK CACHE	Allows or prevents data from being "paged out" of cache.
RESERVE	Prevents other SCSI initiators from accessing drive.
RELEASE	Allows other SCSI initiators to access drive.
FORMAT UNIT	Writes sector boundaries onto the disk. Optionally maps bad sectors and saves device parameters on media.
REASSIGN BLOCKS	Adds additional bad sectors to defect map after formatting.
SEND DIAGNOSTIC	Directs the controller to perform a self-test.
RECEIVE DIAGNOSTIC	Returns diagnostic results from self-test.

*vendor unique command

Figure 2: The primary objective of the SCSI interface is to provide host computers with device independence. Since the PM-3010 supports the standard and extended SCSI command set, little or no change is required in generic SCSI software drivers in order to be compatible with the PM-3010.

sectors in cache into a tree structure and constantly rearranges the structure to minimize search time. The cache search and cache maintenance tasks run in parallel with other tasks which supervise the transfer of data between the disk and cache and between cache and the host.

Two separate DMA channels into cache provide concurrent data transfer across the host and disk interfaces. One DMA channel transfers data to a SCSI protocol controller and bus interface. The second channel performs DMA bursts between cache and a high speed static RAM sector buffer which acts as intermediate storage for the Winchester disk controller circuitry. By performing high speed block DMA bursts to the disk channel, less internal bus bandwidth is used for data transfer resulting in higher overall execution speed for the on-board 68000 microprocessor.

Two queues are used by the firmware to store requests for the two tasks which transfer sectors in and out of cache. These requests are generated by the task which searches for sectors asked for by the host. One queue stores requests to transfer sectors between the cache and the host computer across the SCSI bus. The second queue stores requests to transfer sectors between cache and disk. By queueing data transfer requests, the search task can "get ahead" of the transfer tasks and keep both DMA channels busy a large percentage of the time.

As an example illustrating the processing of concurrent tasks, consider the execution of a multiple sector Read command. Up to 65,535 sequential sectors can be transferred with a single Read command. When the command is received by the PM-3010, a cache search for the first sector in the request is immediately performed. This will typically take from 100 to 250μsec. If the sector is found in cache, an entry is made in the SCSI Channel Request Queue, requesting the DMA channel to begin the sector transfer across the SCSI bus. As soon as the request has been added to the queue, the 68000 is free to begin the search for the next sequential sector. Since this sector sequentially follows the first sector, the search proceeds faster than the first totally random search.

Typical search times range from 10 to 50μsec. At a SCSI bus transfer rate of 1.0 Mbyte/sec the first sector will take a minimum of 256μsec to complete its transfer to the host, assuming 256 bytes per sector. By the time the transfer is complete, the cache search task will have completed several sector look-ups and will have added more entries to the queue. The next sector transfer can begin without delay.

During execution of the Read command, if one or more sectors are not found in cache, then the missing sectors must be read into cache from disk. The controller uses the second Channel Request Queue to queue up requests to the disk controller circuitry. By using queues to store internally generated requests for data transfers between disk, cache, and the host, the cache search and maintenance routines can execute concurrently with the loading of missing data into cache from disk and the transfer of cache resident data to the host. This results in zero access time for all but the first sector in multiple sector transfers. Since the search routines are faster than the data transfer, the PM-3010 can look ahead to read missing sectors into cache from the disk while cache-resident sectors are still being transferred to the host.

Vendor Unique Commands Enhance Performance

The SCSI proposed by the ANSI Task Group X3T9.2 defines not only hardware interface standards but in addition a standardized software command set. Within this standard a command may be issued to a device controller such as the PM-3010 by sending a Command Description Block over the SCSI bus. These blocks consist of a one byte command code, followed by the parameters for that command. Although commands are standardized, certain bits have been left available for vendor unique options.

Several of the standard SCSI commands in the PM-3010 command set utilize the defined vendor unique bits to allow the selection of optional command execution modes. These options may be

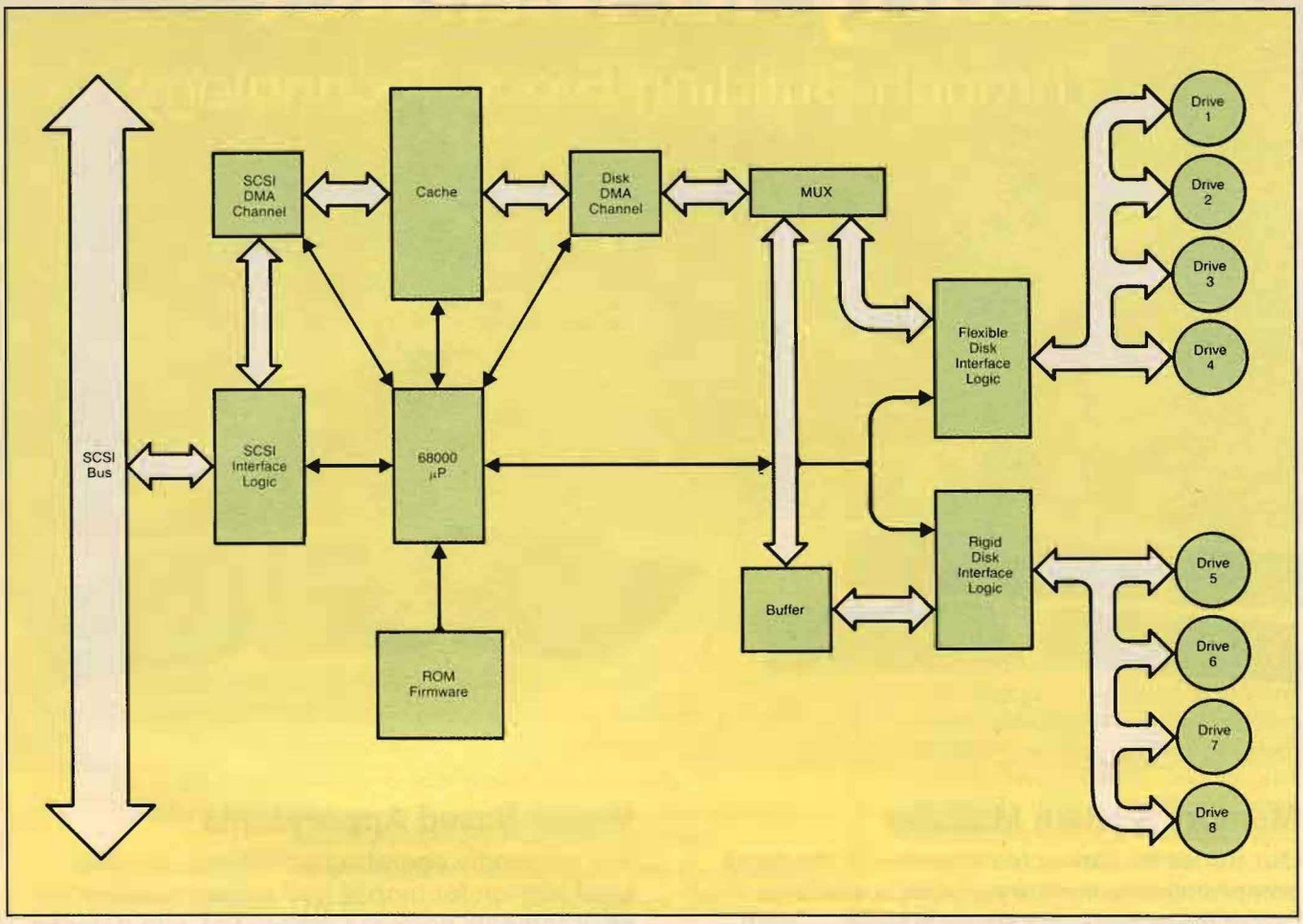

Figure 3: Two separate DMA channels into cache provide concurrent data transfers across host and disk interfaces.

used to fine-tune system performance for specific applications. Since these bits may be set to zero without degrading performance, the PM-3010 may be used with generic software with few or no software modifications required.

A vendor unique bit in the Read and Write commands allows the cache memory to be bypassed for any new sectors accessed which do not already reside in cache. This prevents the "paging out" and removal of sectors which are already cache-resident in order to make room for the new data. This option may be selected for data files which will be accessed infrequently, thereby avoiding unnecessary allocation of cache to store this data.

Another vendor unique bit in the Write command causes the controller to write all data to disk before the Write command is terminated. Normally disk writes are postponed until after command completion status has been returned and the controller has remained idle for one second. The necessary sectors are then copied back to disk starting with the least recently used sector. If another command is received during the writes to disk, the new command will be immediately executed. When the controller again becomes idle, the writing is resumed. This allows the cache to be used to "spool" data to disk, thereby freeing up the host computer as soon as the data has been transferred into cache.

Since all sectors are immediately cached, the hit ratio for Write commands without the Write Immediate option selected is always 100%. The Write Immediate option may be selected for cases where a data recovery routine must be executed by the host computer if the disk write fails.

In addition to providing vendor unique

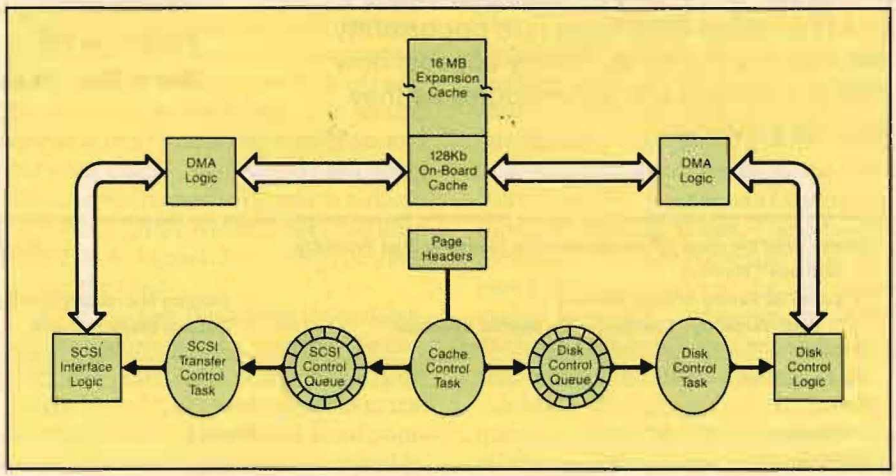

Figure 4: Concurrent execution of cache search and data transfer tasks allow these processes to be overlapped, resulting in zero access time for all but the first sector in multiple sector transfers.

bits within the standard SCSI command definitions, the SCSI specification also allows certain command codes to be used for vendor unique implementations. The vendor unique commands in the PM-3010 command set may be used to optimize controller performance for specific applications. A Lock/Unlock Cache command prevents critical sectors from being "paged out" and removed from cache. A Pre-Read command is also provided to allow data to be "pre-fetched" from the disk and loaded into cache before it is actually needed by the host computer. These commands, although not mandatory, can be used in certain applications to fine-tune system performance.

Automatic Features Simplify Programming

To simplify software driver design, the SCSI command set uses "logical" rather than "physical" addressing. Physical drive parameters are initially specified by the host computer or in some cases, automatically sensed by the controller. When a disk is formatted, these parameters can be optionally stored in a reserved area of the disk, along with the defective sector map. Drive parameters and defect map are then automatically loaded by the controller at power-up. Since all physical characteristics of the drive and media are known by the controller, it can assume the burden of translating logical block addresses into cylinder, head, and sector number, plus mapping defective sectors without host intervention.

The controller handles media defect mapping on a sector rather than a track level. Any reference by the host computer to a block which is known to be defective, is automatically mapped to an alternate location on the disk. The mapping is transparent to the host computer and does not require the entire track to be declared defective. The defective sector map may be augmented as additional media defects show up during normal disk utilization.

Extensive Performance Reporting

The controller utilizes the Return Sense command defined by the SCSI specification to return optional performance statistics to the host upon request. The host has the option of limiting the Return Sense data to error reports only, or to request full performance data at any time. Information provided includes Cache Pages Utilized, Cache Pages Dirty, Cache Hits, Cache Misses, Background Disk Accesses, Foreground Disk Accesses, Post Completion Disk Accesses, Soft Disk Error Retries, and SCSI Bus Disconnects. This data may be used to assess the results of adding more cache expansion boards to the controller or more devices to the SCSI bus. In addition, the data can be used to evaluate the need to pre-read or lock files into cache.

Diagnostics are embedded in the on-board firmware and are accessible through software command (Send Diagnostic) or by pressing a self-test pushbutton on the controller. Results are returned by software to the host computer or may be read by viewing a bank of on-board LEDs.

Caching operations will become increasingly important as disk controller manufacturers attempt to resolve some of the performance issues of computer systems. □

How useful did you find this article? Please write in the appropriate number on the Reader Inquiry Card.

Very Useful	609
Useful	610
Somewhat Useful	611