

Description

Invention relates to a multiple operating mode storage subsystems.

Peripheral memory systems which are attachable to a host (central processing unit and the like) serve diverse purposes with respect to the host. Some systems are for storing user data while others are for storing so-called paging and swapping data, such as can be used in connection with paging and swapping program data sets. An example of a paging store is the IBM 2305 Fixed Head Storage Module which is described in publication GA26-1589-3 "Reference Manual for IBM 2835 Storage Control and IBM 2305 Fixed Head Storage Module" available from International Business Machines Corporation, Armonk, New York. This peripheral memory system consists of a magnetic storage drum which provides rapid access to the stored data. Because of its limited storage capacity, larger hosts requiring larger capacities cannot always efficiently use the IBM 2305 paging store to its maximum efficiency; extra capacity is provided by disk type direct-access storage devices (DASD). Such is particularly true when so-called swapping data sets are used, i.e. large sequential sets of data are rapidly transferred between a host and a peripheral memory. To alleviate storage capacity limitations, while not sacrificing performance criteria, a hierarchical store can be substituted for the 2305 storage drum. An example of such a hierarchical store is shown in U.S. Patent 3,569,938 to Eden. This patent teaches the concept of an apparent store of high performance and high capacity through the use of a high speed cache memory operatively coupled to a backing store, such as a (DASD) or a magnetic tape recorder. Eden teaches that it is beneficial to page data from the backing store to the cache or front store upon demand, including paging the data that surrounds the requested data. While this arrangement is highly successful for general application of peripheral memories, when a series of backing stores share a common cache and the host is employing multi-tasking, the placement of large sets of serial data in the cache becomes troublesome. One solution is to try to make a larger cache. This solution unnecessarily adds to cost of the peripheral system and hence is undesirable. Accordingly, some better solution is needed.

U.S. patent 3,588,839 to Belady shows promoting a next word of data whenever a given word is requested. This arrangement works fine for a cache on a main memory. However, where large sets of data are being transferred, only promoting one additional set of data does not necessarily provide maximal utilization of the peripheral memory system by the host. This lack of maximal use is aggravated by the physical characteristics of the backing store. For example, in DASD backing stores there are several significant delay boundaries caused by the mechanical characteristics of the disk storage apparatus. For example, when selecting one disk storage apparatus or another

disk storage apparatus, substantial delays can be incurred. Additionally, most disk storage apparatus for controlling costs has but a one or two transducers per a recording surface. Access to all of the data areas on the disk storage apparatus is by radially moving the transducers for accessing various ones of the concentric record storage tracks. Such head movements are called cylinder seeks and require substantial delays when measured in terms of electronic speeds. Accordingly, the Belady solution, while eminently satisfactory for many applications, does not solve the problem of handling large sequential data sets in a multi-tasked multi-device environment.

U.S. patent 3,898,624 to Tobias shows a cache connected intermediate a host and main memory. According to this patent, an operator control panel is used to select when to promote one line of data from the main memory to the cache. The promotion always includes but one line of data. While this system may be employed to advantage in certain situations, such as in a main memory to host relationship, it does not in any manner address the large sequential data set problem mentioned above.

It is also desired that in any of these peripheral memory systems to minimize the amount of host intervention in the system to achieve desired goals. An example of such minimization of host intervention is shown in U.S. Patent 4,262,332 to Bass et al which uses high use characteristics and a define extent mechanism for minimizing host access to DASD peripheral systems. While this certainly provides for a minimization of host intervention in such devices, it does not indicate with any of the prior art how to handle large sequential data sets with a shared cache in a multi-tasking multi-device environment.

Thus the problem remains of providing an efficient facility in a store subsystem for handling large sequential data sets as well as random small data transfers and which can be readily accommodated to non-homogenous storage device aggregates of particular characteristics.

The present invention provides a storage subsystem capable of operating in various modes, for or in a data processing system which specifies such modes by the transmission of mode signals, the subsystem including a cache, backing storage and a controller, there being an internal interface between the cache and the backing storage and an external interface for or connecting the cache to the data processing system, across both of which interfaces records are transferred on demand, the controller retaining the current mode signal for the duration of a transfer operation and maintaining a directory of the records resident in the cache, one such mode signal indicating a "sequential" mode of transfer in which n sequential records are transferred to the cache per demand, characterized in that

a) the command that specifies the sequential mode transfer also specifies the maximum number of records to be transferred;

b) the controller is provided with registers, not forming part of the directory, for retaining, not only the current mode, but also the addresses of the individual sequential blocks specified to be transferred;

c) if, with the registers so set, a demand for a block is received whose address is retained in one of the registers, that block and all blocks addressed by subsequent such registers are transferred.

In other words, an arrangement is sought for reducing a number of delays in sequential data signal transfers while minimizing cache front store or buffer size while transferring such sequential data signals between a backing store through a cache to a host using chains of peripheral commands. As disclosed hereinafter by way of example a peripheral storage subsystem receives an indication from the host that sequential data will be requested by the host in an upcoming set of peripheral operations characterized by a sequence of peripheral commands. Along with the sequential indication is an electrical indication of a number of blocks of said data signals which will be sequentially accessed through to an end address of the first ones of said indicated sequential blocks. The peripheral system stores the electrical indications of said numbers and address along with said sequential indication such that subsequently received peripheral commands can be efficiently executed in a series of peripheral operations. The operation of the cache is such that each time a cache miss occurs, the peripheral storage system examines the cache contents and assigns a plurality of randomly selected areas of the cache to receive the sequential data. Then a predetermined number of blocks of sequential data are rapidly transferred from the backing store, such as a disk storage apparatus, to the cache. A directory is updated indicating which of the random locations within the cache contain the sequential data. Each time a block of such sequential data is requested, the directory points to the data in the cache as if it were being received sequentially from the backing store.

In theory all further data defined by the host as being sequential is promoted to the cache. For example, if eight blocks of data are indicated as being sequential, initially the eight blocks are promoted to cache. The host can read three blocks of such data. Assume that because of replacement algorithms, the other five blocks are replaced. Next time the host requests the fourth block there is a miss (fourth block was replaced) the peripheral system, when sequential data is indicated, automatically promotes the remainder of the sequential data, i.e. blocks 4 through 8, to the cache for use by the host.

Promotion of sequential data may incur a given error condition in a given block of data. The block in error is not promoted, but all preceding blocks are promoted to the cache; all other blocks could be promoted to cache. The host then reads the promoted blocks of data from the cache and any attempted access of the non-promoted block results in a miss. A retry may occur on the miss

data but once. If the retry is unsuccessful, a permanent error is indicated to the host and the other sequential data can be read from the cache. Recovery of the block in error can be provided by overriding action from the host for directly accessing the disk storage apparatus. This recovery may not enable recovery from errors in DASD.

If the backing storage is modular with an access delay inherent in crossing a module boundary, for example, a disk stack with an inbuilt delay in moving from cylinder to cylinder, the controller may be arranged to limit sequential record transfer to the current module, despite the number of records specified to be transferred.

The controller may be arranged to queue in substantially first in first out order read requests to backing storage resulting from access requests to cache resulting in a "miss", together with direct access requests to backing storage where a cache bypass mode facility is provided, the controller treating an access request in sequential mode resulting in a miss as a plural access request, making plural sequential queue entries corresponding thereto.

It is pointed out that, in a paper entitled "INTEGRATED DISK CACHE SYSTEM WITH FILE ADAPTIVE CONTROL" by Tokunaga, Hirai and Yamamoto, Comcon 1980, 21st IEEE Computer Society International Conference, 23rd-25th September 1980, pp 412-416, sequential file mode record transfer is discussed. However, the arrangement set forth is such that a fixed number n of records, absent error, is always transferred. No retention of record addresses, other than in the cache directory, nor any specification of a maximum number of block to be transferred is mentioned. Thus the arrangement set forth, given the assumption that some transferred block may be shipped out of the cache before they are accessed, will mean, on a subsequent demand therefor, that a different but overlapping n record will be promoted. No mention is made therein of any other claimed feature of the present invention.

IBM Technical Disclosure Bulletin, Vol. 24, No. 3, August 1981, published 29th July 1981, pages 1559-1562, discloses a similar environment to that of the present invention in which a form of sequentiality is used to prefetch the "next track" when a record near the end of the preceding track is accessed and the controller, for itself, decides that there is effective sequentiality. It does this by logging the actual record accesses for each device it manages in order of access and, if, by reference to the maintained log, it finds that more than a predetermined number of adjacent records have been accessed in order, it switches itself to a "sequential mode" but does nothing about it until it detects that a record in the sequence "near the end of the current track" has been accessed, when it prefetches the next track. The effect of this is much like that of the Compcon paper in that, if there is prefetching, a fixed amount is prefetched and not a predetermined "balance".

GB-A-2 024 476, which deals with an "optional cache" discloses, in the arrangements for

handling it, means for detecting errors during record transfer and, quite independently, the use of a first-in-first-out directory for handling cache replacements.

None of these documents foreshadow the enhancements of the present invention.

The invention will be described further by way of example with reference to embodiments of the invention, as illustrated in the accompanying drawings.

In the drawings:-

Figure 1 is a logic diagram of one form of peripheral subsystem (connected to a host) according to the present invention. Also shown are channel command words and internal command words used therein;

Figure 2 is a logic block diagram of one form of controller for the subsystem;

Figure 3 illustrates various data structures used in the subsystem;

Figure 4 diagrammatically illustrates a mode of operation of the subsystem;

Figure 5 is a logic flow diagram showing execution of a set paging parameter command by the subsystem;

Figure 6 is a logic flow diagram for a read command execution together with a cache hit logic flow for the subsystem;

Figures 7 and 8 jointly show a logic flow diagram related to the promotion of sequential data from a backing store to a front or cache store including preparatory portions and internal command word portions; and

Figure 9 is a logic flow chart showing post data-promotion processing for large sequential data sets.

A hierarchical peripheral storage system 10 is attached as a subsystem to a host 11 for receiving and supplying data signals for host and storage utilization. In a typical application of storage system 10, host 11 consists of a central processing unit (CPU). In other variations, host 11 can be a virtual machine, a set of virtual machines running on a hardware CPU. Host 11 may also be a multi-processor, a uni-processor with attached processors and the like. While the invention can be applied to a great variety of storage systems 10, the preferred and illustrated embodiment shows a paging peripheral storage system for handling paging and swapping data sets. Generally, such paging and swapping data sets relate to storage of program data sets for host 11. As such, storage system 10 is attached to a single host while a general application peripheral storage system can be attached to a plurality of hosts. The invention can be applied to either type of peripheral storage system.

Communications between the paging storage system 10 and host 11 is via a plurality of input/output connections 12-15 which are constructed in accordance with the input/output peripheral channels of the IBM 370 series of computers available from International Business Machines Corporation, Armonk, New York. Such input/output connections, commonly referred to as

channels and subchannels, are so well known, their description is not necessary. Storage system 10 has a lower or backing storage portion consisting of a plurality of direct access storage devices (DASD) 16-18 and separately enumerated DO, D1,.... All accessing of data and storage of data by host 11 with respect to peripheral storage system 10 is by addressing the DASDs 16-18. This addressing is achieved by using the architecture of the input/output connections 12-15 which are summarized in a set of logic blocks 19. Logic blocks 19 represent a channel command word (CCW) as used in the channels for the IBM input/output connections. Typically each channel command word 19 includes an address byte 20. Each address byte 20 includes a plurality of bits for designating the control unit (CU) which is to receive the command. A second plurality of bits DEV which uniquely identify the devices 16-18 to be accessed. In a paging and swapping peripheral storage system 10, each of the devices 16-18 is provided with a plurality of logical device addresses, i.e. device DO for example can be addressed by any one of four addresses. Such multiple addressing has been practiced in the IBM 2305 paging storage system to a limited degree. The logical addresses for each device are indicated in the bits AC of address byte 20. Accordingly, AC has two bits for indicating which of the four logical addresses are being used by host 11 to address a device DO. In the presently constructed embodiment, one of the logical addresses 00 designates a direct access to devices 16-18. That is host 11 operates with devices 16-18 as if peripheral storage system 10 were not a hierarchical system; all the hierarchy is by-passed for direct access. For the AC bits being equal to 01, 10 or 11, the hierarchy, later described, is accessed for obtaining data from devices 16-18 or supplying data to those devices such that the apparent performance of those devices is enhanced on those three logical device addresses. The abbreviation AC is intended to indicate accessed path (logical) to the device indicated by bits DEV.

A second byte of CCW 19 is command byte 21 which contains a code permutation signifying to peripheral memory system 10 what function is to be performed. A third byte 22 is a command modifier byte having a plurality of control fields which electrically indicate to peripheral memory system 10 various modes of operation for executing the command indicated in byte 21. Of interest to the present invention is the bit pattern SEQ which host 11 uses to indicate to peripheral memory system 10 that the data to be transferred in an upcoming set of transfers will be sequential data. When SEQ portion of byte 22 indicates sequential data, then the additional command modifier byte 23 is included in the CCW 19 for indicating the number of blocks or segments of data which will be transferred from devices 16-18 to host 11, or in the reverse direction, as a sequential set of data. Such sequential sets of data in a paging environment are often referred to

as swapping data sets. Additionally, byte 22 can indicate read and discard in section RD which means that once host 11 obtains data from the hierarchy, that data in the hierarchy cache can be discarded; the data in the devices 16-18 is retained. Further controls are provided by so-called "guest operating system" GO. In a virtual computer environment for host 11, one of the operating systems can have cognizance of the paging peripheral memory system 10. Access to peripheral memory system 10 can be handed over to another operating system for accessing or storing data. Such other operating system is a guest of the first operating system and hence is not allowed to modify certain control aspects of the peripheral memory system. Other control fields are also used within byte 22 which are beyond the present description of the illustrated embodiment.

The hierarchy includes a system storage 30 of the semiconductor random access type which has a portion 40 designated as a cache for devices 16-18. Caching principles are sufficiently well known that the purposes and intent of cache 40, with respect to devices 16-18, need not be detailed. A control 31 receives the peripheral commands from host 11 for accessing devices 16-18 through one of the logical device addresses AC as well as providing access to cache 40 based upon the other three logical device addresses of AC. Data is transferred automatically by peripheral memory system 10 between cache 40 and devices 16-18. This transfer is achieved using the same principles of transfer as between host 11 and devices 16. That is, host 11 accesses devices 16-18 in a direct mode (AC = 00) via channel adaptors 32, individually denominated CAA, CAB, CAC and CAD, then over bus 70 through direct access control 56, data circuit 33, device adaptor 34 and device control attachment DCA 35. Received CCWs 19 are interpreted by control 31 for determining the direction of data flow between host 11 and devices 16-18 as well as other functions as is well known for controlling this type of storage apparatus. The relationships of cache 40 to devices 16-18 is substantially identical to the relationships between host 11 and devices 16-18. That is, while host 11 provides control via a series of CCWs 19, the control 31 provides access between cache 40 and devices 16-18 by using a plurality of internal control words (ICW) which are structured in a similar manner to the CCWs as will become apparent. Certain efficiencies and transfer of operation can be provided by altering the ICWs 24 with respect to the CCWs 19. Instead of going through the channel adaptors 32, control 31 has a cache access control CAC 61 which operates system storage 30 and provides access to devices 16-18 through direct access control DAC 56 using the ICWs 24. Instead of channel adaptors 32, a linkage port LKP 25 provides for transfers between CAC 61 and DAC 56. LKP 25 is described later with respect to Figure 3.

Each ICW 24 includes a command byte 26 corresponding to command byte 21. It should be

appreciated that the code permutations for identical commands are the same. Some additional commands are provided while some of the commands for byte 21 are dispensed with. A command modifier byte 27 includes a chain control bit "CHAIN" which replaces the chaining indication normally provided by host 11 to control 31 via channel adaptors 32. (The chaining indication by host 11 is the supplying of a SUPPRESS OUT tag signal.) When final status is due to be reported by peripheral memory system 10 to host 11; SUPPRESS OUT indication of chaining, i.e. an indication of a series of closely related peripheral commands as is fully described and used in connection with the input/output connections 12-15. Since CAC 61 does not use tag signals, command modifier byte 27 is used to replace that tag control signal. The bytes 28 of each ICW 24 points to the stored location of the address of the devices 16-18. No logical addresses are used in the ICWs. In fact, control 31 converts all of the logical addresses directed to the hierarchy into bits DEV. Address bytes 28 not only points to the stored location of DEV but also points to the cylinder address (C), the head or track address (H) and the record address (R). The record address corresponds to a sector address used in addressing most disk storage apparatus. In a preferred embodiment, four records were provided on a single track (H address); hence the record address is 1, 2, 3 or 4 corresponding to an effective orientation of 0°, 90°, 180° and 270° of the disk with respect to a reference rotational point. Design parameters may dictate actual rotational orientations that may differ from the orthogonal orientations.

Cache 40 transfers data signals through channel adaptors 32 with host 11 via bus 41. In a similar manner, data signals are transferred between devices 16-18 through data circuits 33 to cache 40 via bus 42. When simultaneous transfers between cache 40 and host 11 or DASDs 16-18 are not desired, buses 41 and 42 are combined into a single bus time shared by the data transfers. Accessing cache 40, which can be a relatively large memory (several megabytes), requires CAC 61 to transfer the device address together with the cylinder and record addresses CHR over bus 64 to hash circuit 44. Hash circuit 44, which may be microcode implemented, converts the DASD address into a hash class indicator. Since the storage capacity of cache 40 is much less than devices 16-18, the address range of devices 16-18 are concentrated into classes called hash classes for ease of access. A scatter index table SIT 45 has one register for each of the classes defined by hash circuit 44. The contents of the registers in SIT 45 are address pointers to a directory DIR 43 which contains the address DCHR used to access devices 16-18. When data is stored in cache 40, the DASD 16-18 DCHR address together with the cache 40 address is stored in a so-called entry of DIR 43. Since a plurality of device 16-18 addresses corresponds to one hash class, a singly-linked hash class list is provided in the entries of DIR 43

such that scanning cache 40 using hashing only requires scanning the entries within a given hash class. Based upon the contents of directory 43, cache 40 is accessed using known techniques. If no related entries are found in directory 43, then a miss occurs requiring CAC 61 to either allocate space in cache 40 for receiving data from host 11 or to transfer data from devices 16-18 using ICWs 24 and linkage port LKP 25.

Control 31 includes the usual portion of control units that attach to hosts. For example, address and command evaluator ACE 50 communicates with channel adaptors 32 via buses 51, 52, 53 and 54 for receiving command signals from host 11 and supplying status signals to host 11. ACE 50 evaluates CCWs 19 and instructs the peripheral memory system 10 to perform the commanded function as well as indicating the chaining conditions and receiving status signals from the other portions of the peripheral system for relaying to host 11. In a direct mode, i.e. AC = 00, ACE 50 supplies command signals over bus 55 to DAC 56 such that data signals can be transferred between data circuits 33 and the appropriate channel adaptor 32 using known DASD peripheral storage device techniques. In executing its functions, DAC 56 exercises control over data circuit 33 in the usual manner.

Of importance to the present description is the operation of the hierarchy such that sequential data sets can be placed in cache 40 using a minimal size cache with minimal allocation controls while maintaining sequentiality in an efficient manner and maintaining a sufficient number of the data blocks in cache to satisfy the operating requirements of host 11. ACE 50, when receiving a logical device address in byte 20, indicating access to the hierarchy, supplies the received command signals over one of the three buses 60 to CAC 61. The three buses are logical buses indicating the respective cache 40 accesses. CAC 61 stores the received command and modifier data in a channel control block register 63, one register for each of the logical devices. Remember there are three logical device addresses for each of the devices. Therefore if there are eight devices 16-18 then there will be 24 registers in LDCB 62.

The identification and operational status of each logical device is kept in a respective one of logical device control block registers in LDCB 62. Access to the logical device, which is represented by allocation of registers in cache 40 to the address indicated in fields AC and DEV of byte 20, is via address bus 64 to hash circuit 44. In certain situations for sequential data, sequential addresses for devices 16-18 (CHR portion) successive registers in SIT 45 can be accessed. Accordingly, CAC 61 accesses SIT 45 via bus 65 to avoid the delay in hash circuit 44. This operation enhances the response of peripheral system 10 to host 11 when sequential data is being processed. When CAC 61 receives a miss indication from searching the hash class of DIR 43, a request for a data transfer from devices 16-18 to cache 40 is supplied over bus 66 to DAC 56 via LKP 25. The

bus 66 signal alerts DAC 56 to the request and indicates the ICWs are addressable via LKP 25. In the preferred microcode embodiment, LKP 25 is a microcode linkage port, as will become apparent. DAC 56 responds to the ICWs 24 in the same manner that it responds to the CCWs 19. Upon completion of the data transfer, as requested through LKP 25, DAC 56 supplies status signals over bus 67 to CAC 61. At that time, cache 40 has data available to host 11. Further communications between CAC 61 and DAC 56 are via bus 68, all such communications including storing message data in LKP 25. Because devices 16-18 are accessed through a plurality of logical device addresses, a set of queuing registers 69 queue device-related operations requested by CAC 61. In this manner, DAC 56 may not be concerned with the queuing requests through the logical devices but can operate in a direct-access DASD mode for either host 11 and for CAC 61. In this manner, DAC 56 cannot only be used in connection with the hierarchy, but can be used in those peripheral storage systems not employing a hierarchy.

CAC 61 also includes additional controls, for example, register ADEB 76 contains one entry of directory 43 with which CAC 61 is currently operating. That is, the address of device 16-18 resulted in a hit of cache 40 or a portion of cache 40 was allocated to data to be supplied by host 11; by placing the entry in register ADEB 76, operation of CAC 61 is enhanced. That is, directory 43 is a part of system storage 30; by placing the active entry in ADEB 76, system storage 30 is free to transfer data over buses 41 and 42 independent of control 31. Device buffer (DEV BUF) registers 77 contain control information relating to a device 16-18 and are used by CAC 61 in setting up accesses through DAC 56. Such registers are found in a writable control store in the microcoded implementation of the invention. Buffer 77 is merely an allocated portion of control store with no designated data structure. 8ST 78 is a buffer sequence table described later with respect to Figure 3 and used in connection with practicing the present invention within the illustrated peripheral system 10. It includes pointers to directory 43 for each of the data blocks to be transferred in a sequence of data blocks over bus 42 as well as a scanning control mechanism for determining which directory index is to be used for accessing cache 40 during the sequential transfer. In this manner, a sequential transfer can dispense with addressing setups such that a burst of blocks from a device 16-18 can be made without interruption, as will become apparent.

Figure 2 is a block diagram of a preferred embodiment of the Figure 1 illustrated system which employs a programmed microprocessor 31P corresponding to control 31. Bus 70 extends from channel adaptors 32 to data circuits 33 and operates in an identical manner as shown for Figure 1. Buses 41 and 42 extend respectively from channel adaptors 32 and data circuits 33 to system storage 30. Buses 41 and 42 may be combined into one bus with data transfers time

sharing the single bus. Processor 31P in controlling the transfer between data circuits 33 and system storage 30 provides control signals over bus 71 to circuits 33 and address and sequencing control signals over bus 72 to system storage 30. A plurality of system storage address registers SSAR 79 provide addresses to system storage 30. For example, 8 or 16 SSARs 79 may be provided.

Therefore when processor 31P accesses system storage 30, not only does it give the address of the system storage 30 to an SSAR 79 but indicates which of the SSARs is to be used in accessing the storage. Multiplex addressing registers to a memory are known and therefore not further described.

In practicing the present invention, the SSAR 79 for each of the burst of sequential data blocks, processor 31P primes system storage 30 by loading the addresses of cache 40 (a portion of subsystem storage 30) within an SSAR such that the address need not be loaded in SSAR 79 intermediate the successive sequential blocks.

Therefore during the sequential transfer, processor 31P merely refers to an SSAR for initiating the transfer of data signals between cache 40 and a device 16-18. It should be noted that cache 40 has a given address space within system storage 30 in a similar manner. Directory 43 has a different range of addresses. SSAR 79 are separate electronic registers outside the memory array of system storage 30. Processor 31P communicates with channel adaptors 32 over a single bus denominated as 51-54.

Operation of processor 31P is in accordance with microcode programs stored in a control store 73 which is preferably writable, although a portion can be writable while another portion containing certain programs can be read-only. Bus 74 couples the processor 31P to control store 73. Within control store 73 are programs ACE 50P which implement the function of address and command evaluator 50, DAC 56P which are programmed to implement the function of direct access control 56, CAC program 61P which implements the functions of cache access control 61 and OP 75 which are other programs necessary for operation of the storage system 10 but which are not necessary to an understanding of the present invention. The registers used by processor 31P to control the system 10 via the programs 50P, 56P and 61P include CCB 63, LDCB 62, queue registers 69, ADEB 76, SIT 45, buffer 77, LKP 25 and BST 78. For an extremely large cache 40, SIT 45 can be stored in system storage 30. To enhance performance, a set of registers for containing a page of SIT 45 can be reserved in control store 73.

Operation of the Figure 2 illustrated preferred embodiment is best understood by reference to Figures 3 through 9 which illustrates the data structures in detail as well as logic flow diagrams for the microcode portions necessary for an understanding of the operation of the present invention. Figure 3 illustrates the data structures used by a processor 31P to operate peripheral

system 10 in accordance with the invention. LDCB 62 is a series of registers containing data signals in control store 73 consisting of four sections. A first section 80 is a so-called foundation data structure which defines and supports the functions of peripheral system 10 in a general operational sense. Pparms 81 is that portion of LDCB 62 relating to the parameters defining a paging and swapping function established through the later described set paging parameters command. Cparms 82 contains the command parameters such as set sector, seek, search ID command issued by host 11. These commands are those used in connection with known disk storage apparatus peripheral storage systems. Rparms 83 contain the parameters for supporting read activity; i.e. transferring data signals from devices 16-18 to cache 40.

The foundation portion 80 includes a bit ODE 90 which signifies whether or not a device end (DE) is owed by peripheral storage system 10 to host 12. CNL mask 91 contains a bit pattern indicating which channel adaptor 32 received the current command, i.e. which channel the logical device has an affinity to. LDADDR 92 contains a code permutation indicating a logical address received with the command, i.e. the bit patterns of AC and DEV of byte 20 in Figure 1. CMD 23 contains the code permutation from byte 21 of Figure 1. SEQ 94 contains the contents of SEQ section of byte 22 of Figure 1.

CCR 95 indicates whether a channel command retry has been sent to host 11 by system 10. In this regard, when a cache miss is indicated in section 96, a channel command retry was sent to host 11. Therefore LDCB 62 signifies when a miss has occurred for cache 40 and whether or not system 10 has supplied the appropriate CCR signal. Channel command retries merely signifies to host 11 that a delay in executing the peripheral command is required. System 10 upon reaching a state in which the command can be executed will send a device end (DE) signal to the host. The host then sends the peripheral command for the second time such that the command can then be executed by system 10.

Pparms 81 include a sequential bit 100 corresponding to the sequential bit SEQ in byte 22 as well as the RD indicator 101 from RD section of byte 22. B COUNT 102 contains the number of blocks from byte 23. As each block of the sequential data is transferred to host 11, B COUNT 102 is decremented by one. Therefore, it indicates the number of blocks yet to be transmitted to host 11 through cache 40. BASE CYL 103 contains the cylinder address C from which the sequential data will be transmitted from devices 16-18, i.e. in a multicylinder request BASE CYL 103 contains the value C of a virtual machine (VM) minidisk.

Cparms 82 contains the DASD seek address in SEEK ADDR 104, the last or current search ID argument in SID 105 and the last or current set sector value in SECT 106.

Rparms 83 includes REQD 110 indicating that a data transfer from a device 16-18 to cache 40 is

required, RIP 111 indicates that a read is in progress from a device 16-18 to cache 40. RA 112 indicates that a read has been completed from a device 16-18 and that certain post-processing functions are being performed. DADDR 113 contains the bit pattern of DEV from byte 20 (Fig. 1) for indicating the actual device 16-18 being addressed. DIR INDEX 114 contains a directory 43 index value for indicating which directory entry register contains the entry corresponding to the logical device identified in the particular LDCB 62 register. SSAR 115 identifies which SSAR 79 will be used in accessing cache 40 in a data transfer between a device 16-18 and cache 40. SAVE 119 indicates an area of the LDCB 62 registers which processor 31P uses to save control data signals during various operations, including interruption operations.

ADEB 76 is structured in the same way that each entry of directory 43 is structured. Accordingly, description of ADEB 76 amounts to a description of directory 43. In each entry of directory 43 as well as ADEB 76, INDEX 107 is the logical address of the directory entry. This field contains self-identifying data for each entry. Section 108 contains the address of devices 16-18 corresponding to the data stored in cache or allocated for storage. CCP is the physical cylinder address, i.e. the actual physical address of the cylinder for a device 16-18, H is the head address, R is the record, address, P is the device address bit pattern corresponding to DEV section of byte 20, sector is the actual sector value, i.e. rotational position of the disk from which reading will begin. The R value for tracks having four records can vary from one to four while the sector value is the actual sector address. In addressing the DASD, the R value is translated into a rotational position indicator at the byte level as in usual DASD addressing techniques. The R value in some host operating systems can range from 1-120 or other numbers; in such cases the larger R values are reduced to a valued modulo the number of records N in a track. Then the R value, modulo N, is converted to a rotational address of the disk. Such sector value is suitable for initiating access to a record with a minimal latency delay. CCL is the logical cylinder address such as provided for logical devices which are defined on physical devices. Link 109 contains the data signal code permutation of the singly-linked list for linking all entries of a hash class together. The last entry of a given hash class will have a particular ccde pattern (zeroes) indicating end of chain or end of class. M bit 269 indicates whether or not the data in cache 40 has been modified since it was received from a device 16-18. Other code permutations can be added to each directory 43 entry and which are not pertinent to an understanding of the present invention. For example, an MRU-LRU list may be included in each entry.

LKP 25 is an area in control store 73 accessible by programs ACE 50P, DAC 56P and CAC 61P which make up a linkage port or message area for controlling the interaction of the execution of

these microcode units. In one embodiment, ACE 50P and DAC 56P were treated as one code segment such that LKP 25 was accessed by those two microcode sections as a single unit. In any event, the structure of the port includes a code point CP 110 which identifies the portion of the code which lodged the control data in the port. That is when CAC 61P lodges an entry in LKP 25, DAC 56P will fetch the control data and execute the function. Then when DAC 56P enters new data in LKP 25 responding to the request by CAC 61P, CP 110 indicates the CAC 61P which point in code execution the DAC 56P relates to therefore can continue processing based upon DAC 56P response. Priority section 111 contains code permutations indicating whether the request lodged in LKP 25 is high priority, low priority or a continued processing indication. V bit 112 indicates whether or not LKP 25 entry is valid, i.e. is it a recent entry requiring action. DADDR section 113 contains the DEV code permutations from byte 20 for identifying which device 16-18 is associated with the current LKP 25 control data signals. PARMS 114 contains various parameters associated with the message, i.e. what function is to be performed, status and the like.

BST 78 has a set of registers for each of the devices 16-18. A first register includes section DELEP 120 which contains an index value of 1 to 8 pointing to the directory indices 122-123. These indices identify the directory entries to be deleted. EK 121 contains a count of the number of valid entries in the table. It is also used as an address, for example the first directory pointer index is always stored in 122 while the 8th one is always stored at 123. For a value of 3 in EK 121, a third directory index is accessed. Directory index, remember, is a logical address of a directory 43 entry, hence is a rapid access into directory 43.

Figure 4 illustrates a sequence of CCWs and ICWs in a read or write data transfer. A read transfer transfers signals from a device 16-18 to host 11, while the write transfer is a data transfer in the reverse direction. A chain of CCWs 130 begins with set paging parameters (SPP) CCW 132. Figure 5 illustrates the execution of such a command by storage system 10. Fundamentally, SPP 132 sets whether or not sequential data is to be transferred from the peripheral storage system 10 to host 11 as well as other parameters identified in byte 22 of CCW 19 (Fig. 1). Once SPP has indicated parameters of operation to system 10, a seek CCW 133 results in a seek command being transferred to the peripheral storage system; in one embodiment the seek parameters were embedded in the SPP command. Using normal DASD architecture, seek is followed by set sector CCW 134 which in turn is followed by a search ID equal (SIDE) 135. Now the storage system is ready to read data from an addressed device 16-18 by read CCW 136. Upon receipt of the read command, peripheral storage system 10 provides the action indicated in column 131. First of all, the seek set sector and ID commands are stacked as at 140. At 137 a directory 43 search, as explained

with respect to Figure 1, is conducted. For a hit, i.e. the requested data is in cache 40, the data is immediately transferred as indicated by arrow 138 from cache 40 to host 11 via the channel adaptor 32 which received the command. On the other hand, if directory 43 indicated the data was not in the cache, then a miss has occurred as indicated at arrow 141. A channel command retry (CCR) is supplied by system 10 as indicated by arrow 142. This tells host 11 that when a DEVICE END signal is received from system 10, that the read CCW 136 must be reexecuted by the channel by sending the same read command to system 10. While this is occurring, system 10 constructs a chain of ICWs 143-148 beginning with a seek ICW 143 which is derived from the stacked seek commands received from host 11. For a multitrack operation, the ICWs are derived from search ID parameters. The seek ICW 143 is followed by a set sector ICW 144 which has the sector calculated from the record number. At 145, the local input results in a set cache ICW 145. This ICW causes DAC 56P to insert into the appropriate SSAR 79 the address of system storage 30 at which the data to be read will be stored. If a plurality of blocks of data are to be transferred, then a plurality of set cache ICWs occur as indicated by numeral 146. Then a search ID equal ICW 147 corresponding to the SIDE CCW 135 occurs. The search ID equal ICW 147 corresponds to the first set cache ICW 145. This means a plurality of blocks of data are read in sequence using but one SIDE ICW 147. Then a number of read ICWs 148 commands equal to the number of data blocks to be transferred are given to DAC 56P for reading a predetermined number of blocks of data indicated by the number of set cache ICWs. Upon completion of the read, which transfers data from the address device 16-18 to cache 40 at the addresses set in SSARs 97, system 10 supplies a device end (DE), as indicated by arrow 150, to host 11. Host 11 immediately responds by reissuing a peripheral command at 151 corresponding to the CCW 136. Of course, system 10 searches directory 43 at 152 resulting in a hit because of the just executed ICW chain. Data is then transferred from cache 40 to host 11 as indicated by arrow 153. In the event that the data was not transferred for the requested data block at 136, another miss will occur and an error status will be reported to host 11.

This error status will reflect the fact that system 10 was unable to transfer data from the addressed device 16-18 at the cylinder and head address. Host 11 then can use the direct access (AC = 00) for attempting recovery using standard disk storage apparatus recovery techniques beyond the scope of the present description. Ellipsis 154 indicates that the above-described operation is highly repetitive as well as indicating that various CCW chains for various devices 16-18 can be interleaved. The ICW chains do not necessarily follow the sequence of chains of CCWs. Depending upon the circumstances, an ICW chain may be constructed and used by a later occurring CCW

chain. Such possibility indicates the asynchronous aspect of the ICW chains with respect to the CCW chains. Usually, the first CCW chain will result in a first occurring ICW chain. At any instant, a separate ICW chain can be active for each DASD 16-18.

Figure 5 illustrates the execution by system 10 of the SPP command. ACE 50P receives and decodes the SPP command. As a result of that decoding, processor 31P activates CAC 61P. Upon activation, processor 31P via CAC 61P performs certain nonpertinent logic functions at 155. Then at 156, LDCB 62 (Fig. 3) is accessed for setting SIO 105 in CPARMS 82 to unity, setting ODE 90 in foundation section 80 to 0, setting CCR bit 95 to 0, setting SEQ 100 to the value received in byte 22 (indicated by X), RD section 101 is set to the value in RD section of byte 22 and B COUNT 102 of PPARMS 81 is set to the value indicated in byte 23. Following setting LDCB 62, processor 31P performs some nonpertinent logic functions at 155. Then processor 31P at 157 examines LDCB 62 section SEQ 100 to see if sequential data is involved. If not, processor 31P returns to ACE 50P via LKP 25. If the sequential data is indicated, processor 31P at 158 transfers the number of blocks indicated in B COUNT 102 to an internal register IR (not shown) of processor 31P. In the illustrated embodiment a maximum of eight blocks were to be transferred in a given burst of data blocks. In that embodiment when B COUNT was greater than or equal to eight, eight blocks were transferred and B COUNT reduced by eight. For B COUNT less than eight, a number of blocks equal to B COUNT were transferred. Then at 159 the value of the block count is examined. If it is nonzero, then an appropriate SPP command execution has occurred. If the block count is 0, then the sequential indicator or the block count must be in error. Accordingly, processor 31P leaves step 159 to go to an error status reporting procedure beyond the scope of the present description.

The seek, set sector, and SIDE CCWs 133-135 are not described since they are well known. A change in system 10 operation from the prior art occurs upon the receipt of a read command based on read CCW 136 for sequential data. Figure 6 illustrates the machine operations for transferring sequential data. At 160 the received command is processed by ACE 50P. Then through LKP 25, CAC 61P is activated by processor 31P. The command is again decoded at 161. Since it is a read command, directory 43 is searched at 162 as described with respect to Figure 1. At 163 processor 31P determines whether or not the directory search resulted in a hit or a miss. For a miss, the received command is enqueued at 164 by placing the command and its control information in queue registers 69. A CCR is sent to host 11. Since queue registers 69 can use any format they are not further described except to say that the queue is a first-in first-out queue for each of the addressable devices, i.e. for 8 devices 16-18 there are 8 queues. The importance of having a FIFO

queue is to ensure that the sequence of responses to the host for a given device corresponds to the sequence of commands sent by the host. From queue 69, CAC 61P will initiate a read from the addressed device 16-18, as explained with respect to Figures 7-9.

A hit condition in the directory search at 163 results in cache 40 automatically transferring data to host 11 via the appropriate channel adaptor 32 at 170. Such automatic cache to host transfers are well known and not described for that reason. During the automatic data transfer an error can occur; accordingly, upon an error detection, processor 31P goes to an error reporting and analyzing routine at 171. Generally the data transfers will be error free. At 172 following the successful completion of a data transfer, processor 31P accesses LDCB 62 to examine RD section 101. If discard after read is indicated, processor 31P sets the just read block of data cached for destage if modified; and free, if not modified. Destaging is performed by processor 31P when no commands are being executed. Destaging the data prior to requirement of a replacement algorithm being invoked reduces the control required for efficiently managing cache 40, i.e. free spaces are made available before they are needed. Then through logic path 174, from either steps 172 or 173, processor 31P at 175 determines from directory 43 in a field (not shown) whether or not the data is pinned to cache 40. Pinning data to cache 40 means that it cannot be transferred to devices 16-18 until a pinning flag (not shown) of directory 43 has been erased. If the data is not pinned to cache, then the block that was just read is made the most recently used (MRU) block at 176, in the LRU list (not shown) for the replacement algorithm. This is achieved by accessing directory 43 and updating the least recently used list of known design in that directory. At 177, nonpertinent logic steps are performed by processor 31P. Then at 180, LDCB 62 is again accessed for examination of SEQ 100. If sequential data has been indicated, then processor 31P at 182 examines LDCB B COUNT 102 to see if the block count is equal to 0, i.e. is the just-transferred block the last block in the sequence of data. If it is not the last block transferred, then at 183 the block count (BK) is decremented by 1. Following steps 180, 182 or 183 logic path 181 leads processor 31P back to ACE 50P for performing final status reporting to host 11 in the usual manner.

Figures 7 and 8 illustrate scanning the read queues in queue 69 and generating an ICW chain of internal system 10 commands. After the requested read has been enqueued, processor 31P causes system 10 to perform diverse functions, such as responding to commands received over various ones of the channel adaptors, additional commands received from the channel adaptor which had transferred the read command (which was CCR'd). When there is a lull in receipt of peripheral commands from the host, seek and set sector device commands are sent to devices 16-18. When there is a lull in control activity which

may occur while cache 40 is transferring data to host 11, receiving data from host 11, transferring or receiving data from an addressed device 16-18, processor 31P through its dispatcher microcode, which is a portion of OP 75 (Fig. 2) scans its work tables, including queue registers 69. If the queues are empty, i.e. no reading is to occur, processor 31P follows logic path 192 returning to dispatcher 190. If a read has been enqueued as detected at 191 by scanning the queue flags, the queue entry is transferred from a queue register 69 at 193 to an internal register (not shown) of processor 31P. If an error occurs in this transfer, an error reporting and recovery technique is instituted at 194. Upon successful reading the queue entry from queue register 69, LDCB 62 is accessed at 195 to set ODE section 90 to unity to indicate that a device end is owed upon completion of a successful read (such as indicated in Figure 5 by arrow 150). At 196 some nonpertinent functions are performed. Then at 200, in the device buffer area 77 corresponding to the addressed device, a bit is set to indicate that logical chaining will occur, i.e. more than one ICW will be used in the upcoming access to the addressed device 16-18. At 201, LDCB 62 is again accessed to examine the value of SEQ 100. For sequential data being indicated, processor 31P proceeds to 202 to set up the block count for the upcoming ICW chain equal to the paging parameters (PA) in the illustrated embodiment.

A maximum number of blocks which can be transferred through a given ICW chain is equal to the number of SSARs 97. For example, for 8 SSARs the number of blocks transferred will be a maximum of 8. Further, delay boundaries are a consideration, for example, if the 8 blocks to be transferred require accessing 2 cylinders; then only those blocks in the first cylinder will be transferred. For example, if the 8 blocks have 4 blocks in a first cylinder and 4 blocks in a second cylinder, then the number of blocks would be set to 4. This action minimizes the time required to transfer a series of blocks and enables all transfers to proceed to completion at electronic speeds. In the event of a miss on the first block of a given cylinder, then up to 8 blocks could be automatically transferred. Also the maximum number of blocks is never greater than the remaining value in B COUNT 102. The ICW chains are instructed such that cylinder boundaries are never crossed by any given ICW chain. These calculations follow usual computer programming techniques and are not described further for that reason. If sequential data is not indicated at 201, then the number of blocks to be transferred is set to 1 at 203. These numbers are supplied to the device buffer 77 along with the chaining flag, device addresses and other device control data. At 204, the SSAR 79 identification is set to 0. This means that processor 31P will access the SSAR having identification 0.

At 205, the logical address LDADDR including AC and DEV from CCW 19 of Figure 1 is converted to a physical device address. In the illustrated

embodiment, this action is achieved by merely masking the AC portion from the logical address. Certain nonpertinent functions are performed at 206. point 207 is reentry point B from a continuation of the logic flow diagram described with respect to Figure 8, i.e. all of the logic steps from 190 through 206 are preparatory steps with the following described steps being repeatable as a loop for setting up a succession of block transfers.

The first step 210 in the loop allocates a slot or space in cache 40. Usual allocation procedures are followed, i.e. an addressable unit (slot) on a so called free list (not shown) is identified as the one to receive the first block of signals from the addressed device 16-18. That slot is then removed from the free list and identified within an internal register (not shown) within processor 31P for identifying which directory 43 entry is to be used for identifying the slot in cache 40. Note that there is one entry register in directory 43 for each addressable slot in cache 40. Accordingly, the actual address in cache 40 of the data can be derived directly from which register of directory 43 contains the entry. Upon the attempted allocation of the number of slots equal to the number of blocks set in steps 202 or 203, processor 31P at 211 determines whether or not any error occurred in the allocation process. If an error has occurred, the total number of blocks may not be successfully transferred from the address device 16-18 to cache 40. Accordingly, for an error condition, at 212 processor 31P examines LDCB 62 SEQ 100 to determine if the data transfer is a sequential transfer. If it is not a sequential transfer, processor 31P follows logic path 213 returning to ACE 50P to wait for a replacement algorithm control to make space available for one block. For a sequential transfer, processor 31P at 214 determines whether or not the error occurred on the first block to be transferred. If so, the entire operation will be aborted. Accordingly, at 215 LDCB 62 is accessed setting a unit check (UC) flag (not shown) to unity, a disconnect owed status flag (DS) to unity and resets the other flags to 0. Then processor 31P returns via logic path 216 to ACE 50P. If the allocation error is not for the first block, then data transfers of the remaining blocks occur. Processor 31P follows path 217 to 220 for truncating the number of blocks to be transferred in the unallocated area from the ICW list.

Returning to step 211, if there were no allocation errors, then at 218 some nonpertinent functions are performed. These functions include analyzing microcode logic errors not related to allocation.

If a slot was not allocated due to such microcode errors, then the truncate step 220 is also performed for reducing the number of blocks transferred from the addressed device 16-18 to cache 40. Without an error or after truncation, processor 31P performs some nonpertinent logic steps at 221. At 222, LDCB 62 SEQ 100 is examined. If SEQ is 0, i.e. nonsequential data, then at 223 the index of the directory 43 entry corresponding to the slot in cache 40 to receive

the data is entered into LDCB 62 section 114 of RPARMS 83. For sequential data or after the index is entered into LDCB 62, at 224 the cache address to be inserted later into an SSAR 79 is generated from the directory index just inserted into LDCB 62. This generation is merely adding an offset to each of the directory indices. Then at 225, LDCB 62 SEQ 100 indicating sequential mode causes processor 31P to examine B COUNT 102 to see if the count is greater than one. If the count is greater than 1, then at 232 processor 31P examines to see if the first block in the sequence of blocks being transferred is being handled. If not, at 233 a new cache address for the second block is provided. Then at 234 in the device buffer area 77, the SSAR 79 corresponding to the second or other blocks is set to the cache address, flags are set, pointer to the directory 43 is set and the SSAR 79 to receive the cache address is identified. Other functions to be performed may also be defined in the device buffer 77.

Returning to steps 225, 231 and 232, the logic path 226 leads to nonpertinent steps 227 followed by processor 31P accessing LDCB 62 to store the generated cache address in section 118 of RPARMS 83. Then following nonpertinent steps 229, processor 31P proceeds through connector 235 to the logic steps shown in Figure 8.

The connection between Figures 7 and 8 is through connector A denominated by numerals 235 and 240, respectively, in Figures 7 and 8. At 241 processor 31P updates the pointer to SSAR 97 by incrementing EK 121 of Figure 3. At 242 processor 31P determines whether or not all of the blocks to be transferred to cache 40 have received allocations in cache 40. If not, through connector B 243 processor 31P returns to Figure 7 flow chart at B 207 to allocate cache 40 space another block. This loop is repeated until EK 121 contains a count equal to the number of blocks to be transferred (not more than eight).

After completing the loop, some nonpertinent logic steps are performed at 244. At 245, the read code command is set into the ICW representing a read data command for DASD 16. At 250, LDCB 62 is accessed to determine whether or not the sequential flag SEQ 100 in PPARMS 81 is set or reset. When set, processor 31P at 251 determines whether or not the received block count is greater than 1. If it is greater than 1, then a chaining indication is set in command modifier byte 27 of ICW 24 (Fig. 1); otherwise from steps 250 or 251 the end of chain indication EOC is indicated in byte 27 by resetting the chain indicator. At 254 the device buffer 77 in control store 73 receives the ICW, i.e. the code permutation flags and other storage operation (STOROP) indications. At 255, processor 31P again examines SEQ 100 of LDCB 62 for nonsequential, i.e. SEQ = 0. If only one block will be transferred, processor 31P follows logic path 256 to execute logic step 257 for transmitting the just constructed ICW to DAC 56P via LKP 25.

For a sequential data transfer, processor 31P leaves step 255 to execute logic step 260 for

adjusting EK 121 to the next entry (set next). Then at 261, if the remaining block count is not greater than 1 then the ICWs are transmitted to DAC 56 in step 257. For a number of blocks remaining greater than 1, loop 270 is executed for setting up the remaining ICWs for a chain of such ICWs. At 271 the command is set for read count, key, data and multi-track commands. At 272 processor 31P determines whether or not the last block in a sequential group of blocks is to be processed. If not, the chaining flag in byte 27 of the ICW being built is set to unity. Otherwise at 274 the end of chaining condition is indicated by resetting the chaining flag. At 275 the ICW is transferred to the device buffer 77. At 276 the cache address is stored in the device buffer such that it can be transferred immediately to SSAR 97 for the burst transfer. At 277 processor 31P determines if the block is the last block; if not, the loop is indexed at 278 adjusting a count in an internal register (not shown) using usual control techniques. Otherwise, step 257 is performed. When the loop is indexed at 278 the steps 271 through 277 are again performed.

DAC 56P upon receiving the ICW chain executes the chain in the same manner that it executes received commands through channel adaptors 32. Since this latter operation is well known, the execution of the ICW chains is not further described. It should be noted that in transferring signals from DASD 16 to cache 40, DAC 56P not only provides the addressing to DASD 16 but also transfers the cache address contents of device buffer 77 into SSAR 79 such that several blocks of data can be transferred in a single data stream, i.e. can be a multi-track transfer. Upon completion of that transfer, DAC 56P loads the resulting status, including error indications, into LKP 25. Processor 31P operation then switches from DAC 56P to CAC 61P.

Figure 9 illustrates the portion of CAC 61P called post processing, i.e. logic steps performed following the transfer of data signals from DASD 16 to cache 40. First the contents of LKP 25 are transferred to work registers at step 280. This includes the device address, the pointer to LDCB 62 and any flags that may be generated by DAC 56P. At 281, processor 31P accesses LDCB 62 for resetting RIP 111 to 0 for indicating that no read from a DASD is in progress. At 282 some nonpertinent logic steps are performed. At 283 processor 31P examines the DAC 56P return code (RC) for an error free condition; for a successful transfer of all requested blocks of data to cache 40 a return code of 0 is provided. For such a successful operation, the contents of an internal work register of processor 31P is set to unity at 284. This initializes the count. At 285, LDCB 62 is accessed for examining SEQ 100. If the transfer is not sequential, i.e. only one data block is to be transferred, then the number of blocks being transferred is set to unity at 287. Otherwise at 286 the number of blocks allocated as stored in device buffer 77 is transferred to EK 121 to indicate the number of entries in BST 78.

For an error condition at step 283, i.e. RC does not equal 0, an error analysis set of logic steps are performed at 290. If a permanent error is indicated by such analysis, processor 31P follows path 291 to a permanent error recovery and reporting, procedure beyond the scope of the present description. Otherwise at 292, processor 31P determines whether or not the command was a multi-track command, i.e. more than one block of data was to be automatically transferred. If not, an error condition would effect a single block therefore a permanent error has to be handled causing processor 31P to follow logic path 291. If a plurality of blocks were transferred, then further action can occur, i.e. one block may be in error while all of the preceding blocks were transferred error free. DAC 56P has identified via LKP 25 which block caused the error. CAC 61P from the DAC 56P information can identify which ICW is associated with the error. Processor 31P then fetches the immediately preceding ICW at 293 from device buffer 77. At 294, if the command was for reading data from DASD 16 to cache 40, processor 31P at 295 adjusts the block count by subtracting the number of blocks in error from the slots allocated.

At 294 for a command in error other than a read data command, the error occurred before any data transfer; processor 31P goes to a permanent error routine (not described). From steps 286, 287 or 295, some nonpertinent logic steps are performed at 300. Such logic steps pertain to internal addressing, not pertinent to an understanding of the present invention. At 301, processor 31P accesses LDCB 62 SEQ 100 to determine whether or not a sequential transfer is indicated. For a sequential transfer logic path 302 is followed to step 304 to determine whether or not the block transferred was the first block. For the first block being transferred in a sequence of blocks or in a nonsequential mode, wherein only one block is transferred, processor 31P at 305 transfers the contents of LDCB 62 to a work register (not shown) for the ensuing logic steps. This includes transferring the logical cylinder indication CCL (see 108), record number R, the logical device address D and other control data not pertinent to an understanding of the invention. For blocks which are not the first block, processor 31P at 306 accesses BST 78 (Fig. 3) to obtain the directory index for that block as indicated by DELEP 121. For the first block or nonsequential data, directory 43 index 114 of LDCB provides the same information.

For the directory 43 search, processor 31P at 307 looks for the cache address corresponding to the device address. This search includes a hashing operation to determine whether or not the directory 43 has an entry corresponding to the block of data just transferred from DASD 16 to cache 40. Remember that several parallel accesses to the same DASD 16 are possible in an asynchronous manner; therefore it is important that one and only one replication of DASD 16 data be in cache 40. This requirement provides data

integrity, i.e. if duplicate copies were in cache 40, one copy could be updated while a second copy could be erroneous. Then the updated copy could be stored in DASD 16, that entry being erased; later access by a host to the system 10 could result in the erroneous data residing in cache being sent to the host via channel adaptor 32.

Following a search, error indications are checked at 308. For no errors, at 310 processor 31P determines whether or not a duplicate copy was found. For no duplicate at 311, the BST 78 entry is calculated, i.e. the DELEP 120 value, BST 78 is accessed for pointing to the directory index (numerals 122-123 of Figure 3) such that the directory 43 entry corresponding to the DASD 16 address is transferred from directory 43 to ADEB 76 for convenient access by processor 31P, all action occurring at 312. At 313 processor 31P resets M 269 in ADEB 76. M indicates modified data in cache 40. Resetting M to O indicates that the copy in cache 40 is identical to the copy on DASD 16. At 314, BST 78 is again accessed for incrementing DELEP 120 and decrementing EK 121. At 315 a directory 43 entry is added corresponding to the data just transferred to cache 40; that is, when DAC 56P caused the block of data to be transferred to cache 40, directory 43 had not yet been updated, i.e. the data in cache 40 is not yet addressable. By creating a directory entry in the usual manner at 315, the just transferred data from DASD 16-18 to cache 40 becomes addressable.

On the other hand, if a duplicate is found at 310, it is assumed that the data already in cache 40 is the correct copy, i.e. it may have been modified. Therefore, it is desired not to make the just-transferred data block addressable and to proceed to the next data block. At 316 BST 78 is accessed for decrementing EK and incrementing DELEP. The block just transferred is freed at 317 leaving only a single copy of the data in cache 40. Freeing the block makes it nonaddressable.

The final housekeeping logic functions occur beginning with step 320 determining whether or not all of the data blocks transferred from DASD 16 to cache 40 have been post processed. If not, at 321 the number of blocks to be post processed is decremented by 1. Processor 31P then follows logic path 302 to execute a loop including steps 304 through 315.

At 323 BST 78 is accessed for resetting EK and DELEP to O. Step 323 is also entered from PE logic path 291 via logic path 325. At 324, queue registers 69 are accessed for removing the queue entry from the queue such that a duplicate read will not occur. You will recall that in Figure 6 the read was enqueued at 164. Then at 322 LDCB 62 is accessed for resetting control flags, such as CCR 95, miss 96, RA 112 and the like and for pinning for a device end, such as by setting ODE 90. At 190 status is presented to the host via ACE 50P. In all of the above description it should be noted that the transfer of data signals from DASD 16 to cache 40 and in the reverse direction is on an asynchronous basis with respect to the operation

of the channel adaptors 32. Under certain circumstances, a request from the host through channel adaptor 32 can take priority over the transfer, therefore some of the just-described operations may be interleaved with higher priority operation. Since multiprocessing is well known this detail is not further discussed.

Claims

1. A storage subsystem capable of operating in various modes, for or in a data processing system which specifies such modes by the transmission of mode signals, the subsystem including a cache (40), backing storage (16-18) and a controller (31), there being an internal interface (33) between the cache and the backing storage and an external interface (32) for or connecting the cache to the data processing system (11), across both of which interfaces records are transferred on demand, the controller retaining the current mode signal for the duration of a transfer operation and maintaining a directory (43) of the records resident in the cache, one such mode signal indicating a "sequential" mode of transfer in which n sequential records are transferred to the cache per demand, characterised in that

a) the command that specifies the sequential mode transfer also specifies the maximum number of records to be transferred;

b) the controller is provided with registers (LDCB; SSAR 79), not forming part of the directory, for retaining, not only the current mode, but also the addresses of the individual sequential blocks specified to be transferred;

c) if, with the registers so set, a demand for a block is received whose address is retained in one of the registers, that block and all blocks addressed by subsequent such registers are transferred.

2. A subsystem as claimed in claim 1 wherein the controller is arranged to limit the number of sequential records transferred from the backing storage to the cache to the record capacity of the cache.

3. A subsystem as claimed in either claim 1 or claim 2 wherein the backing storage is modular with access delay inherent in crossing a module boundary, and the controller is arranged to limit sequential record transfer from the backing storage to the cache to the current module despite the number of records specified to be transferred.

4. A subsystem as claimed in any preceding claim wherein the controller is arranged to decrement the retained mode record number each time a record is transferred under the mode across the external interface.

5. A subsystem as claimed in any preceding claim wherein means for detecting errors in records read from backing storage are provided, the controller being responsive to such error detecting means to terminate sequential record transfer to cache at the record next before the record detected to be in error.

integrity, i.e. if duplicate copies were in cache 40, one copy could be updated while a second copy could be erroneous. Then the updated copy could be stored in DASD 16, that entry being erased; later access by a host to the system 10 could result in the erroneous data residing in cache being sent to the host via channel adaptor 32.

Following a search, error indications are checked at 308. For no errors, at 310 processor 31P determines whether or not a duplicate copy was found. For no duplicate at 311, the BST 78 entry is calculated, i.e. the DELEP 120 value, BST 78 is accessed for pointing to the directory index (numerals 122-123 of Figure 3) such that the directory 43 entry corresponding to the DASD 16 address is transferred from directory 43 to ADEB 76 for convenient access by processor 31P, all action occurring at 312. At 313 processor 31P resets M 269 in ADEB 76. M indicates modified data in cache 40. Resetting M to 0 indicates that the copy in cache 40 is identical to the copy on DASD 16. At 314, BST 78 is again accessed for incrementing DELEP 120 and decrementing EK 121. At 315 a directory 43 entry is added corresponding to the data just transferred to cache 40; that is, when DAC 56P caused the block of data to be transferred to cache 40, directory 43 had not yet been updated, i.e. the data in cache 40 is not yet addressable. By creating a directory entry in the usual manner at 315, the just transferred data from DASD 16-18 to cache 40 becomes addressable.

On the other hand, if a duplicate is found at 310, it is assumed that the data already in cache 40 is the correct copy, i.e. it may have been modified. Therefore, it is desired not to make the just-transferred data block addressable and to proceed to the next data block. At 316 BST 78 is accessed for decrementing EK and incrementing DELEP. The block just transferred is freed at 317 leaving only a single copy of the data in cache 40. Freeing the block makes it nonaddressable.

The final housekeeping logic functions occur beginning with step 320 determining whether or not all of the data blocks transferred from DASD 16 to cache 40 have been post processed. If not, at 321 the number of blocks to be post processed is decremented by 1. Processor 31P then follows logic path 302 to execute a loop including steps 304 through 315.

At 323 BST 78 is accessed for resetting EK and DELEP to 0. Step 323 is also entered from PE logic path 291 via logic path 325. At 324, queue registers 69 are accessed for removing the queue entry from the queue such that a duplicate read will not occur. You will recall that in Figure 6 the read was enqueued at 164. Then at 322 LDCB 62 is accessed for resetting control flags, such as CCR 95, miss 96, RA 112 and the like and for pinning for a device end, such as by setting ODE 90. At 190 status is presented to the host via ACE 50P. In all of the above description it should be noted that the transfer of data signals from DASD 16 to cache 40 and in the reverse direction is on an asynchronous basis with respect to the operation

of the channel adaptors 32. Under certain circumstances, a request from the host through channel adaptor 32 can take priority over the transfer, therefore some of the just-described operations may be interleaved with higher priority operation. Since multiprocessing is well known this detail is not further discussed.

Claims

1. A storage subsystem capable of operating in various modes, for or in a data processing system which specifies such modes by the transmission of mode signals, the subsystem including a cache (40), backing storage (16-18) and a controller (31), there being an internal interface (33) between the cache and the backing storage and an external interface (32) for or connecting the cache to the data processing system (11), across both of which interfaces records are transferred on demand, the controller retaining the current mode signal for the duration of a transfer operation and maintaining a directory (43) of the records resident in the cache, one such mode signal indicating a "sequential" mode of transfer in which n sequential records are transferred to the cache per demand, characterised in that

a) the command that specifies the sequential mode transfer also specifies the maximum number of records to be transferred;

b) the controller is provided with registers (LDCB; SSAR 79), not forming part of the directory, for retaining, not only the current mode, but also the addresses of the individual sequential blocks specified to be transferred;

c) if, with the registers so set, a demand for a block is received whose address is retained in one of the registers, that block and all blocks addressed by subsequent such registers are transferred.

2. A subsystem as claimed in claim 1 wherein the controller is arranged to limit the number of sequential records transferred from the backing storage to the cache to the record capacity of the cache.

3. A subsystem as claimed in either claim 1 or claim 2 wherein the backing storage is modular with access delay inherent in crossing a module boundary, and the controller is arranged to limit sequential record transfer from the backing storage to the cache to the current module despite the number of records specified to be transferred.

4. A subsystem as claimed in any preceding claim wherein the controller is arranged to decrement the retained mode record number each time a record is transferred under the mode across the external interface.

5. A subsystem as claimed in any preceding claim wherein means for detecting errors in records read from backing storage are provided, the controller being responsive to such error detecting means to terminate sequential record transfer to cache at the record next before the record detected to be in error.





6. A subsystem as claimed in any preceding claim wherein the controller is arranged to queue in substantially first in first out order read requests to backing storage resulting from access requests to cache resulting in a "miss", the controller treating an access request in sequential mode resulting in a miss as a plural access request, making plural sequential queue entries corresponding thereto.

Patentansprüche

1. Speicher-Subsystem für den Betrieb in unterschiedlichen Betriebsarten für bzw. in einem Datenverarbeitungssystem, welches die Betriebsarten durch die Übertragung von Betriebsartsignalen bestimmt, mit einem Cache (40), einem Hilfsspeicher (16—18) und einer Steuereinheit (31), wobei sich zwischen Cache und Hilfsspeicher ein internes Interface (33) befindet und ein externes Interface (32) den Cache mit dem Datenverarbeitungssystem (11) verbindet, wobei über beide Interfaces auf Anforderung Datensätze übertragen werden, wobei die Steuereinheit das aktuelle Betriebsartssignal für die Dauer einer Übertragung festhält und ein Inhaltsverzeichnis (43) der im Cache befindlichen Datensätze pflegt, wobei eines der Betriebsartsignale einen "sequentiellen" Übertragungsmodus anzeigt, in dem n sequentielle Datensätze auf Anforderung zum Cache übertragen werden,

dadurch gekennzeichnet, daß

a) das den sequentiellen Übertragungsmodus bezeichnende Kommando ebenfalls die Höchstzahl von zu übertragenden Datensätzen bezeichnet;

b) die Steuereinheit Register (LDCB; SSAR 79) aufweist, die nicht Teil des Inhaltsverzeichnisses sind und nicht nur die aktuelle Betriebsart, sondern auch die Adressen der einzelnen sequentiellen zu übertragenden Blöcke festhält;

c) und daß wenn mit den derart gesetzten Registern eine Anforderung für einen Block erhalten wird, dessen Adresse in einem der Register gehalten wird, dieser Block sowie sämtliche von derartigen nachfolgenden Registern adressierten Blöcke übertragen werden.

2. Subsystem nach Anspruch 1, dadurch gekennzeichnet, daß die Steuereinheit so angeordnet ist, daß sie die Anzahl der aus dem Hilfsspeicher zum Cache zu übertragenden sequentiellen Datensätze auf die Datensatz-Kapazität des Caches begrenzt.

3. Subsystem nach Anspruch 1 oder 2, dadurch gekennzeichnet, daß der Hilfsspeicher modular aufgebaut ist mit Zugriffsverzögerung aufgrund einer erforderlichen Überschreitung einer Modulgrenze und daß die Steuereinheit so angeordnet ist, daß sie sequentielle Datenübertragungen vom Hilfsspeicher zum Cache auf das aktuelle Modul ungeachtet der Anzahl der zu übertragenden Datensätze begrenzt.

4. Subsystem nach einem der vorhergehenden Ansprüche, dadurch gekennzeichnet, daß die Steuereinheit so angeordnet ist, daß sie die ver-

bleibende Anzahl von Modus-Datensätzen jedes mal dekrementiert, wenn ein zu dem Modus gehörender Datensatz über das externe Interface übertragen wird.

5. Subsystem nach einem der vorhergehenden Ansprüche, dadurch gekennzeichnet, daß eine Einrichtung zum Erkennen von Fehlern in den aus dem Hilfsspeicher gelesenen Datensätzen vorgesehen ist, wobei die Steuereinheit auf diese Erkennungseinrichtung so reagiert, daß die Übertragung der sequentiellen Datensätze zum Cache bei dem Datensatz beendet wird, welcher dem als fehlerhaft erkannten Datensatz vorausgeht.

6. Subsystem nach einem der vorhergehenden Ansprüche, dadurch gekennzeichnet, daß die Steuereinheit so angeordnet ist, daß sie Leseanforderungen im wesentlichen in FIFO-Folge an den Hilfsspeicher einreicht, die aus Zugriffsanforderungen an den Cache resultieren, die zu einer "Verfehlung" führen, wobei die Steuereinheit eine Zugriffsanforderung im Sequentiellen Modus, welche zu einer "Verfehlung" führen, wie eine Mehrfach-Zugriffsanforderung behandelt, indem mehrere sequentielle Warteschlangeneinträge entsprechend dazu gemacht werden.

Revendications

1. Sous-ensemble de mémoires pouvant fonctionner en divers modes, pour ou dans un ensemble de traitement de l'information qui détermine ces modes par la transmission de signaux de mode, ce sous-ensemble comprenant une antémémoire (40), une mémoire auxiliaire (16 à 18) et un contrôleur (31), une jonction intérieure (33) existant entre l'antémémoire et la mémoire auxiliaire, et une jonction extérieure (32) étant prévue pour ou connectant l'antémémoire à l'ensemble (11) de traitement de l'information, des enregistrements étant transférés à la demande à travers ces deux jonctions, le contrôleur retenant le signal du mode en cours pendant la durée d'une opération de transfert et maintenant un répertoire (43) des enregistrements résidant dans l'antémémoire, ce signal de mode indiquant un mode "séquentiel" de transfert dans lequel "n" enregistrements séquentiels par demande sont transférés dans l'antémémoire, caractérisé en ce que:

a) l'ordre qui spécifie le transfert en mode séquentiel spécifie également le nombre maximum d'enregistrements à transférer;

b) le contrôleur est doté de registres (LDCB; SSAR 79) ne faisant pas partie du répertoire, pour retenir non seulement le mode en cours mais également les adresses des blocs séquentiels individuels à transférer suivant la spécification;

c) si, les registres étant ainsi réglés, une demande de bloc est reçue dont l'adresse est retenue dans l'un des registres, ce bloc et tous les blocs adressés par les registres qui suivent sont transférés.

2. Sous-ensemble suivant la revendication 1, dans lequel le contrôleur est agencé de manière à limiter à la capacité d'enregistrement de l'antémémoire le nombre d'enregistrements séquen-

tiels à transférer de la mémoire auxiliaire à l'antémémoire.

3. Sous-ensemble suivant l'une des revendications 1 ou 2, dans lequel la mémoire auxiliaire est modulaire et présente un retard inhérent d'accès à la traversée d'une limite de module, et le contrôleur est agencé de manière à limiter au module en cours le transfert d'enregistrements séquentiels de la mémoire auxiliaire à l'antémémoire, malgré le nombre d'enregistrements à transférer suivant la spécification.

4. Sous-ensemble suivant l'une quelconque des revendications précédentes, dans lequel le contrôleur est agencé de manière à décrémenter le nombre retenu d'enregistrements de mode chaque fois qu'un enregistrement est transféré selon le mode à travers la jonction extérieure.

5. Sous-ensemble suivant l'une quelconque des revendications précédentes, dans lequel des moyens sont prévus pour détecter des erreurs

dans les enregistrements extraits de la mémoire auxiliaire, le contrôleur étant sensible à ces moyens de détection d'erreurs afin de mettre fin au transfert d'enregistrements séquentiels vers l'antémémoire à l'enregistrement suivant avant que l'enregistrement détecté ne soit erroné.

6. Sous-ensemble suivant l'une quelconque des revendications précédentes, dans laquelle le contrôleur est agencé pour mettre en file d'attente, sensiblement suivant le système premier entré premier sorti, les demandes d'extraction envoyées à la mémoire auxiliaire et résultant de demandes d'accès envoyées à l'antémémoire et ayant pour résultat une "lacune," le contrôleur traitant une demande d'accès dans le mode séquentiel résultant dans une lacune comme une demande déaccès multiples, leur faisant correspondre plusieurs entrées de files d'attente séquentielles.

25

30

35

40

45

50

55

60

65

15



* C L M S *