(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
20 August 2009 (20.08.2009)

PCT

(10) International Publication Number
WO 2009/101050 A1

(51) International Patent Classification:
G06F 13/14 (2006.01)

(21) International Application Number:
PCT/EP2009/051445

(22) International Filing Date:
9 February 2009 (09.02.2009)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
12/030,961 14 February 2008 (14.02.2008) US

(71) Applicant *(for all designated States except US)*: **INTERNATIONAL BUSINESS MACHINES CORPORATION** [US/US]; New Orchard Road, Armonk, New York 10504 (US).

(71) Applicant *(for MG only)*: **IBM UNITED KINGDOM LIMITED** [GB/GB]; PO Box 41, North Harbour, Portsmouth Hampshire PO6 3AU (GB).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **CASPER, Daniel** [US/US]; 13 Brett Place, Poughkeepsie, New York 12603 (US). **BENDYK, Mark** [US/US]; 29 Hudson Drive, Hyde Park, New York 12538 (US). **HATHORN, Roger** [US/US]; 5820 East Placita De La Zurencia, Tucson, Arizona 85750 (US). **FLANAGAN, John** [US/US]; 15 Slate Hill Drive, Poughkeepsie, New York 12603 (US). **HARDY, Clint** [US/US]; 8230 South Camino Serpe, Tucson, Arizona 85747 (US). **HUANG, Catherine** [US/US]; 6 Thornberry Way, Poughkeepsie, New York 12603 (US). **KALOS, Matthew** [US/US]; 5435 East Heatherwood Way, Tucson, Arizona 85718 (US). **RICCI, Louis** [US/US]; 5 Spruce Road, Hyde Park, New York 12538 (US). **SITTMANN III, Gustav** [US/US]; 422 South Park Avenue, Webster Groves, Missouri 63119 (US). **YUDENFRIEND, Harry** [US/US]; 1 Nob Hill Road, Poughkeepsie, New York 12603-5545 (US).

(74) Agent: **WILLIAMS, Julian, David**; IBM United Kingdom Limited, Intellectual Property Law, Hursley Park, Winchester Hampshire SO21 2JN (GB).

[Continued on next page]

(54) Title: RESERVED DEVICE ACCESS CONTENTION REDUCTION

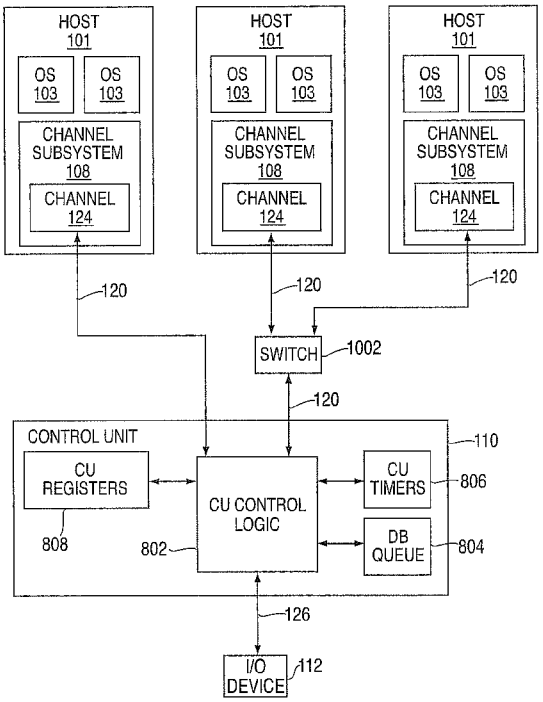

FIG. 10

(57) Abstract: A computer program product, an apparatus, and a method for reducing reserved device access contention at a control unit in communication with a plurality of operating systems via one or more channels are provided. The computer program product includes a tangible storage medium readable by a processing circuit and storing instructions for execution by the processing circuit for performing a method that includes receiving a command message at the control unit from a first operating system, including an I/O operation command for a device. A device busy indicator is received, indicating that a second operating system has reserved the device. The command message is queued on a device busy queue in response to the device busy indicator. The control unit monitors for a device end indicator. The device busy queue is serviced to perform the I/O operation command in response to the device end indicator.

WO 2009/101050 A1

(81) **Designated States** *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) **Designated States** *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:

— *with international search report (Art. 21(3))*

— *with amended claims (Art. 19(1))*

RESERVED DEVICE ACCESS CONTENTION REDUCTION

FIELD OF THE INVENTION

The present disclosure relates generally to input/output processing, and in particular, to reducing device contention issues associated with multiple requests to access a reserved device.

BACKGROUND OF THE INVENTION

Input/output (I/O) operations are used to transfer data between memory and I/O devices of an I/O processing system. Specifically, data is written from memory to one or more I/O devices, and data is read from one or more I/O devices to memory by executing I/O operations.

To facilitate processing of I/O operations, an I/O subsystem of the I/O processing system is employed. The I/O subsystem is coupled to main memory and the I/O devices of the I/O processing system and directs the flow of information between memory and the I/O devices. One example of an I/O subsystem is a channel subsystem. The channel subsystem uses channel paths as communications media. Each channel path includes a channel coupled to a control unit, the control unit being further coupled to one or more I/O devices.

The channel subsystem may employ channel command words (CCWs) to transfer data between the I/O devices and memory. A CCW specifies the command to be executed. For commands initiating certain I/O operations, the CCW designates the memory area associated with the operation, the action to be taken whenever a transfer to or from the area is completed, and other options.

During I/O processing, a list of CCWs is fetched from memory by a channel. The channel parses each command from the list of CCWs and forwards a number of the commands, each command in its own entity, to a control unit coupled to the channel. The control unit then processes the commands. The channel tracks the state of each command and controls when

the next set of commands are to be sent to the control unit for processing. The channel ensures that each command is sent to the control unit in its own entity. Further, the channel infers certain information associated with processing the response from the control unit for each command.

Performing I/O processing on a per CCW basis may involve a large amount of processing overhead for the channel subsystem, as the channels parse CCWs, track state information, and react to responses from the control units. Therefore, it may be beneficial to shift much of the processing burden associated with interpreting and managing CCW and state information from the channel subsystem to the control units. Simplifying the role of channels in communicating between the control units and an operating system in the I/O processing system may increase communication throughput as less handshaking is performed.

Additional problems can arise in managing requests from channels controlled by multiple operating systems to command a common I/O device via a control unit. The multiple operating systems can exist upon a common host system or across multiple host systems, with each host system including a channel subsystem and processing elements. When multiple operating systems attempt to access a common I/O device that has been reserved, the control unit typically receives a device busy indicator from the I/O device and reports the device busy indicator to the channels controlled by the operating systems requesting access. The access request may be a command to perform an I/O operation with or without reservation. Once the I/O device becomes non-busy, the control unit sends a device end indicator to the operating systems via their respective channels to notify them that the I/O device is available. The channel subsystems can then make the previously attempted request again, with the first-in-time channel winning the race condition relative to the other channels contending to access the I/O device. A faster responding host system can effectively block out slower responding host systems, as reservation requests are granted to the first-in-time requester. For example, an operating system that is a running on a host system that is further in distance from the I/O device may be prevented from accessing the I/O device for long periods of time, as an operating system running on a host system that is closer to the I/O device experiences a shorter communication transport delay. Thus, as contention for reserving the I/O device and subsequent access requests increases, the disparity between

operating systems in accessing the I/O device also increases. Accordingly, there is a need in the art for reserved I/O device contention reduction at a control unit in communication with a plurality of operating systems via one or more channels.

BRIEF SUMMARY OF THE INVENTION

Embodiments of the invention include a computer program product for reducing reserved device access contention at a control unit in communication with a plurality of operating systems via one or more channels. The computer program product includes a tangible storage medium readable by a processing circuit and storing instructions for execution by the processing circuit for performing a method. The method includes receiving a command message at the control unit from a first operating system of the plurality of operating systems via the one or more channels, where the command message includes an I/O operation command for a device in communication with the control unit. The method additionally includes receiving a device busy indicator from the device, where the device busy indicator notifies the control unit that the device is reserved by a second operating system of the plurality of operating systems. The method also includes queuing the command message on a device busy queue in response to the device busy indicator. The method further includes monitoring the device for a device end indicator, where the device end indicator notifies the control unit that the device is ready to receive a new I/O operation command. The method additionally includes servicing the device busy queue to perform the I/O operation command in response to the device end indicator.

Additional embodiments include an apparatus for reducing reserved device access contention. The apparatus includes a control unit in communication with a plurality of operating systems via one or more channels. The control unit performs a method including receiving a command message at the control unit from a first operating system of the plurality of operating systems via the one or more channels. The command message includes an I/O operation command for a device in communication with the control unit. The control unit receives a device busy indicator from the device. The device busy indicator notifies the control unit that the device is reserved by a second operating system of the plurality of operating systems. The control unit further queues the command message on a

device busy queue in response to the device busy indicator and monitors the device for a device end indicator. The device end indicator notifies the control unit that the device is ready to receive a new I/O operation command. The control unit additionally services the device busy queue to perform the I/O operation command in response to the device end indicator.

Further embodiments include a method for reducing reserved device access contention at a control unit in communication with a plurality of operating systems via one or more channels. The method includes receiving a command message at the control unit from a first operating system of the plurality of operating systems via the one or more channels, where the command message includes an I/O operation command for a device in communication with the control unit. The method also includes receiving a device busy indicator from the device, where the device busy indicator notifies the control unit that the device is reserved by a second operating system of the plurality of operating systems. The method additionally includes queuing the command message on a device busy queue in response to the device busy indicator and monitoring the device for a device end indicator. The device end indicator notifies the control unit that the device is ready to receive a new I/O operation command. The method further includes servicing the device busy queue to perform the I/O operation command in response to the device end indicator.

Other computer program products, apparatuses, and/or methods according to embodiments will be or become apparent to one with skill in the art upon review of the following drawings and detailed description. It is intended that all such additional computer program products, apparatuses, and/or methods be included within this description, be within the scope of the present invention, and be protected by the accompanying claims.

BRIEF DESCRIPTION OF THE DRAWINGS

The subject matter which is regarded as the invention is particularly pointed out and distinctly claimed in the claims at the conclusion of the specification. The foregoing and other objects, features, and advantages of the invention are apparent from the following detailed description taken in conjunction with the accompanying drawings in which:

FIG. 1 depicts one embodiment of an I/O processing system incorporating and using one or more aspects of the present invention;

FIG. 2A depicts one example of a prior art channel command word;

FIG. 2B depicts one example of a prior art channel command word channel program;

FIG. 3 depicts one embodiment of a prior art link protocol used in communicating between a channel and control unit to execute the channel command word channel program of FIG. 2B;

FIG. 4 depicts one embodiment of a transport control word channel program, in accordance with an aspect of the present invention;

FIG. 5 depicts one embodiment of a link protocol used to communicate between a channel and control unit to execute the transport control word channel program of FIG. 4, in accordance with an aspect of the present invention;

FIG. 6 depicts one embodiment of a prior art link protocol used to communicate between a channel and control unit in order to execute four read commands of a channel command word channel program;

FIG. 7 depicts one embodiment of a link protocol used to communicate between a channel and control unit to process the four read commands of a transport control word channel program, in accordance with an aspect of the present invention;

FIG. 8 depicts one embodiment of a control unit and a channel, in accordance with an aspect of the present invention;

FIG. 9 depicts one embodiment of a response message communicated from a control unit to a channel, in accordance with an aspect of the present invention;

FIG. 10 depicts one embodiment of a control unit in communication with a plurality of host systems, in accordance with an aspect of the present invention;

FIG. 11 depicts one embodiment of a process for reserved device access contention reduction; and

FIG. 12 depicts one embodiment of a computer program product incorporating one or more aspects of the present invention.

The detailed description explains the preferred embodiments of the invention, together with advantages and features, by way of example with reference to the drawings.

DETAILED DESCRIPTION OF THE INVENTION

In accordance with an aspect of the present invention, input/output (I/O) processing is facilitated with reduced contention for accessing a reserved I/O device. For instance, I/O processing is facilitated by readily enabling access to information, such as status and measurement data, associated with I/O processing. Further, I/O processing is facilitated, in one example, by reducing communications between components of an I/O processing system used to perform the I/O processing. For instance, the number of exchanges and sequences between an I/O communications adapter of a host system, such as a channel, and a control unit is reduced. This is accomplished by sending a plurality of commands from the I/O communications adapter to the control unit as a single entity for execution by the control unit, and by the control unit sending the data resulting from the commands, if any, as a single entity.

The plurality of commands are included in a block, referred to herein as a transport command control block (TCCB), an address of which is specified in a transport control word (TCW). The TCW is sent from an operating system or other application to the I/O communications adapter, which in turn forwards the TCCB in a command message to the control unit for processing. The control unit processes each of the commands absent a tracking of status relative to those individual commands by the I/O communications adapter. The plurality of commands is also referred to as a channel program, which is parsed and executed on the control unit rather than the I/O communications adapter.

In an exemplary embodiment, the control unit generates a response message including status and extended status information in response to executing the channel program. The control unit may also generate a response message without executing the channel program under a limited number of communication scenarios, e.g., to inform the I/O communications adapter

that the channel program will not be executed. The control unit may include a number of elements to support communication between the I/O communications adapter and I/O devices, as well as in support of channel program execution. For example, the control unit can include control logic to parse and process messages, in addition to one or more queues, timers, and registers to facilitate communication and status monitoring. The I/O communications adapter parses the response message, extracting the status and extended status information, and performs further calculations using the extracted information.

When multiple operating systems executing on one or more host systems attempt to access a reserved I/O device via a control unit in communication with one or more I/O adapters in the one or more host systems, access contention can arise. In order to perform certain an I/O operations, the I/O device may be reserved for exclusive access to a set of channels (path group) under the control of a requesting operating system. A path group can be established in order to allow a device reservation to exist from multiple channels back to the same host. Once the I/O device is reserved, subsequent attempts by other operating systems to reserve or use the I/O device are blocked, with the I/O device returning a device busy indicator. In an exemplary embodiment, multiple commands including reservation requests received at the control unit for the I/O device are queued while the device busy indicator is present. In response to the I/O device removing the device busy indicator, e.g., a device end indicator, the control unit services the queue and determines the next command to process. The queue can employ a number of queue management techniques, such as first in first out (FIFO) servicing, priority based servicing, and round robin servicing. FIFO servicing handles commands in the order that they are received. Priority based servicing allows higher priority commands to be serviced ahead of lower priority commands. Round robin servicing handles each communication connection from different operating systems in turn. Using a queue to manage requests to access a reserved I/O device simplifies communication between the control unit and I/O adapters controlled by operating systems, and reduces contention to prevent slower responding host systems from being blocked disproportionately by faster responding host systems.

One example of an I/O processing system incorporating and using one or more aspects of the present invention is described with reference to FIG. 1. I/O processing system 100 includes a host system 101, which further includes for instance, a main memory 102, one or more

central processing units (CPUs) 104, a storage control element 106, and a channel subsystem 108. The host system 101 may be a large scale computing system, such as a mainframe or server. The I/O processing system 100 also includes one or more control units 110 and one or more I/O devices 112, each of which is described below.

Main memory 102 stores data and programs, which can be input from I/O devices 112. For example, the main memory 102 may include one or more operating systems (OSs) 103 that are executed by one or more of the CPUs 104. For example, one CPU 104 can execute a Linux® operating system 103 and a z/OS® operating system 103 as different virtual machine instances. The main memory 102 is directly addressable and provides for high-speed processing of data by the CPUs 104 and the channel subsystem 108.

CPU 104 is the controlling center of the I/O processing system 100. It contains sequencing and processing facilities for instruction execution, interruption action, timing functions, initial program loading, and other machine-related functions. CPU 104 is coupled to the storage control element 106 via a connection 114, such as a bidirectional or unidirectional bus.

Storage control element 106 is coupled to the main memory 102 via a connection 116, such as a bus; to CPUs 104 via connection 114; and to channel subsystem 108 via a connection 118. Storage control element 106 controls, for example, queuing and execution of requests made by CPU 104 and channel subsystem 108.

In an exemplary embodiment, channel subsystem 108 provides a communication interface between host system 101 and control units 110. Channel subsystem 108 is coupled to storage control element 106, as described above, and to each of the control units 110 via a connection 120, such as a serial link. Connection 120 may be implemented as an optical link, employing single-mode or multi-mode waveguides in a Fibre Channel fabric. Channel subsystem 108 directs the flow of information between I/O devices 112 and main memory 102. It relieves the CPUs 104 of the task of communicating directly with the I/O devices 112 and permits data processing to proceed concurrently with I/O processing. The channel subsystem 108 uses one or more channel paths 122 as the communication links in managing the flow of information to or from I/O devices 112. As a part of the I/O processing, channel subsystem 108 also performs the path-management functions of testing for channel path

availability, selecting an available channel path 122 and initiating execution of the operation with the I/O devices 112.

Each channel path 122 includes a channel 124 (channels 124 are located within the channel subsystem 108, in one example, as shown in FIG. 1), one or more control units 110 and one or more connections 120. In another example, it is also possible to have one or more dynamic switches (not depicted) as part of the channel path 122. A dynamic switch is coupled to a channel 124 and a control unit 110 and provides the capability of physically interconnecting any two links that are attached to the switch. In another example, it is also possible to have multiple systems, and therefore multiple channel subsystems (not depicted) attached to control unit 110.

Also located within channel subsystem 108 are subchannels (not shown). One subchannel is provided for and dedicated to each I/O device 112 accessible to a program through the channel subsystem 108. A subchannel (e.g., a data structure, such as a table) provides the logical appearance of a device to the program. Each subchannel provides information concerning the associated I/O device 112 and its attachment to channel subsystem 108. The subchannel also provides information concerning I/O operations and other functions involving the associated I/O device 112. The subchannel is the means by which channel subsystem 108 provides information about associated I/O devices 112 to CPUs 104, which obtain this information by executing I/O instructions.

Channel subsystem 108 is coupled to one or more control units 110. Each control unit 110 provides logic to operate and control one or more I/O devices 112 and adapts, through the use of common facilities, the characteristics of each I/O device 112 to the link interface provided by the channel 124. The common facilities provide for the execution of I/O operations, indications concerning the status of the I/O device 112 and control unit 110, control of the timing of data transfers over the channel path 122 and certain levels of I/O device 112 control.

Each control unit 110 is attached via a connection 126 (e.g., a bus) to one or more I/O devices 112. I/O devices 112 receive information or store information in main memory 102 and/or other memory. Examples of I/O devices 112 include card readers and punches, magnetic tape units, direct access storage devices, displays, keyboards, printers, pointing

devices, teleprocessing devices, communication controllers and sensor based equipment, to name a few.

One or more of the above components of the I/O processing system 100 are further described in "IBM® z/Architecture Principles of Operation," Publication No. SA22-7832-05, 6th Edition, April 2007; U.S. Patent No. 5,461,721 entitled "System For Transferring Data Between I/O Devices And Main Or Expanded Storage Under Dynamic Control Of Independent Indirect Address Words (IDAWS)," Cormier et al., issued October 24, 1995; and U.S. Patent No. 5,526,484 entitled "Method And System For Pipelining The Processing Of Channel Command Words," Casper et al., issued June 11, 1996, each of which is hereby incorporated herein by reference in its entirety. IBM is a registered trademark of International Business Machines Corporation, Armonk, New York, USA. Other names used herein may be registered trademarks, trademarks or product names of International Business Machines Corporation or other companies.

In one embodiment, to transfer data between I/O devices 112 and memory 102, channel command words (CCWs) are used. A CCW specifies the command to be executed, and includes other fields to control processing. One example of a CCW is described with reference to FIG. 2A. A CCW 200 includes, for instance, a command code 202 specifying the command to be executed (e.g., read, read backward, control, sense and write); a plurality of flags 204 used to control the I/O operation; for commands that specify the transfer of data, a count field 206 that specifies the number of bytes in the storage area designated by the CCW to be transferred; and a data address 208 that points to a location in main memory that includes data, when direct addressing is employed, or to a list (e.g., contiguous list) of modified indirect data address words (MIDAWs) to be processed, when modified indirect data addressing is employed. Modified indirect addressing is further described in U.S. Application Serial Number 11/464,613, entitled "Flexibly Controlling The Transfer Of Data Between Input/Output Devices And Memory," Brice et al., filed August 15, 2006, which is hereby incorporated herein by reference in its entirety.

One or more CCWs arranged for sequential execution form a channel program, also referred to herein as a CCW channel program. The CCW channel program is set up by, for instance, an operating system, or other software. The software sets up the CCWs and obtains the addresses of memory assigned to the channel program. An example of a CCW channel

program is described with reference to FIG. 2B. A CCW channel program 210 includes, for instance, a define extent CCW 212 that has a pointer 214 to a location in memory of define extent data 216 to be used with the define extent command. In this example, a transfer in channel (TIC) 218 follows the define extent command that refers the channel program to another area in memory (e.g., an application area) that includes one or more other CCWs, such as a locate record 217 that has a pointer 219 to locate record data 220, and one or more read CCWs 221. Each read CCW 220 has a pointer 222 to a data area 224. The data area includes an address to directly access the data or a list of data address words (e.g., MIDAWs or IDAWs) to indirectly access the data. Further, CCW channel program 210 includes a predetermined area in the channel subsystem defined by the device address called the subchannel for status 226 resulting from execution of the CCW channel program.

The processing of a CCW channel program is described with reference to FIG. 3, as well as with reference to FIG. 2B. In particular, FIG. 3 shows an example of the various exchanges and sequences that occur between a channel and a control unit when a CCW channel program is executing. The link protocol used for the communications is FICON (Fibre Connectivity), in this example. Information regarding FICON is described in "Fibre Channel Single Byte Command Code Sets-3 Mapping Protocol (FC-SB-3), T11/Project 1357-D/Rev. 1.6, INCITS (March 2003), which is hereby incorporated herein by reference in its entirety.

Referring to FIG. 3, a channel 300 opens an exchange with a control unit 302 and sends a define extent command and data associated therewith 304 to control unit 302. The command is fetched from define extent CCW 212 (FIG. 2B) and the data is obtained from define extent data area 216. The channel 300 uses TIC 218 to locate the locate record CCW and the read CCW. It fetches the locate record command 305 (FIG. 3) from the locate record CCW 217 (FIG. 2B) and obtains the data from locate record data 220. The read command 306 (FIG. 3) is fetched from read CCW 221 (FIG. 2B). Each is sent to the control unit 302.

The control unit 302 opens an exchange 308 with the channel 300, in response to the open exchange of the channel 300. This can occur before or after locate command 305 and/or read command 306. Along with the open exchange, a response (CMR) is forwarded to the channel 300. The CMR provides an indication to the channel 300 that the control unit 302 is active and operating.

The control unit 302 sends the requested data 310 to the channel 300. Additionally, the control unit 302 provides the status to the channel 300 and closes the exchange 312. In response thereto, the channel 300 stores the data, examines the status and closes the exchange 314, which indicates to the control unit 302 that the status has been received.

The processing of the above CCW channel program to read 4k of data requires two exchanges to be opened and closed and seven sequences. The total number of exchanges and sequences between the channel and control unit is reduced through collapsing multiple commands of the channel program into a TCCB. The channel, e.g., channel 124 of FIG. 1, uses a TCW to identify the location of the TCCB, as well as locations for accessing and storing status and data associated with executing the channel program. The TCW is interpreted by the channel and is not sent or seen by the control unit.

One example of a channel program to read 4k of data, as in FIG. 2B, but includes a TCCB, instead of separate individual CCWs, is described with reference to FIG. 4. As shown, a channel program 400, referred to herein as a TCW channel program, includes a TCW 402 specifying a location in memory of a TCCB 404, as well as a location in memory of a data area 406 or a TIDAL 410 (i.e., a list of transfer mode indirect data address words (TIDAWs), similar to MIDAWs) that points to data area 406, and a status area 408. TCWs, TCCBs, and status are described in further detail below.

The processing of a TCW channel program is described with reference to FIG. 5. The link protocol used for these communications is, for instance, Fibre Channel Protocol (FCP). In particular, three phases of the FCP link protocol are used, allowing host bus adapters to be used that support FCP to perform data transfers controlled by CCWs. FCP and its phases are described further in "Information Technology – Fibre Channel Protocol for SCSI, Third Version (FCP-3)," T10 Project 1560-D, Revision 4, September 13, 2005, which is hereby incorporated herein by reference in its entirety.

Referring to FIG. 5, a channel 500 opens an exchange with a control unit 502 and sends TCCB 504 to the control unit 502. In one example, the TCCB 504 and sequence initiative are transferred to the control unit 502 in a FCP command, referred to as FCP_CMND information unit (IU) or a transport command IU. The control unit 502 executes the multiple commands of the TCCB 504 (e.g., define extent command, locate record command, read

command as device control words (DCWs)) and forwards data 506 to the channel 500 via, for instance, a FCP_Data IU. It also provides status and closes the exchange 508. As one example, final status is sent in a FCP status frame that has a bit active in, for instance, byte 10 or 11 of the payload of a FCP_RSP IU, also referred to as a transport response IU. The FCP_RSP IU payload may be used to transport FICON ending status along with additional status information, including parameters that support the calculation of extended measurement words and notify the channel 500 of the maximum number of open exchanges supported by the control unit 502.

In a further example, to write 4k of customer data, the channel 500 uses the FCP link protocol phases, as follows:

1. Transfer a TCCB in the FCP_CMND IU.

2. Transfer the IU of data, and sequence initiative to the control unit 502.

3. Final status is sent in a FCP status frame that has a bit active in, for instance, byte 10 or 11 of the FCP_RSP IU Payload. The FCP_RSP_INFO field or sense field is used to transport FICON ending status along with additional status information, including parameters that support the calculation of extended measurement words and notify the channel 500 of the maximum number of open exchanges supported by the control unit 502.

By executing the TCW channel program of FIG. 4, there is only one exchange opened and closed (see also FIG. 5), instead of two exchanges for the CCW channel program of FIG. 2B (see also FIG. 3). Further, for the TCW channel program, there are three communication sequences (see FIGs. 4-5), as compared to seven sequences for the CCW channel program (see FIGs. 2B-3).

The number of exchanges and sequences remain the same for a TCW channel program, even if additional commands are added to the program. Compare, for example, the communications of the CCW channel program of FIG. 6 with the communications of the TCW channel program of FIG. 7. In the CCW channel program of FIG. 6, each of the commands (e.g., define extent command 600, locate record command 601, read command 602, read command 604, read command 606, locate record command 607 and read

command 608) are sent in separate sequences from channel 610 to control unit 612. Further, each 4k block of data (e.g., data 614-620) is sent in separate sequences from the control unit 612 to the channel 610. This CCW channel program requires two exchanges to be opened and closed (e.g., open exchanges 622, 624 and close exchanges 626, 628), and fourteen communications sequences. This is compared to the three sequences and one exchange for the TCW channel program of FIG. 7, which accomplishes the same task as the CCW channel program of FIG. 6.

As depicted in FIG. 7, a channel 700 opens an exchange with a control unit 702 and sends a TCCB 704 to the control unit 702. The TCCB 704 includes the define extent command, the two locate record commands, and the four read commands in DCWs, as described above. In response to receiving the TCCB 704, the control unit 702 executes the commands and sends, in a single sequence, the 16k of data 706 to the channel 700. Additionally, the control unit 702 provides status to the channel 700 and closes the exchange 708. Thus, the TCW channel program requires much less communications to transfer the same amount of data as the CCW channel program of FIG. 6.

Turning now to FIG. 8, one embodiment of the control unit 110 and the channel 124 of FIG. 1 that support TCW channel program execution are depicted in greater detail. The control unit 110 includes CU control logic 802 to parse and process command messages containing a TCCB, such as the TCCB 704 of FIG. 7, received from the channel 124 via the connection 120. The CU control logic 802 can extract DCWs and control data from the TCCB received at the control unit 110 to control a device, for instance, I/O device 112 via connection 126 to perform one or more I/O operation commands. The CU control logic 802 sends device commands and data to the I/O device 112, as well as receives status information and other feedback from the I/O device 112. For example, the I/O device 112 may be busy because of a previous reservation request targeting I/O device 112. To manage potential device reservation contention issues that can arise when the control unit 110 receives multiple requests to access the same I/O device 112, the CU control logic 802 keeps track of and stores device busy messages and associated data in a device busy queue 804. In an exemplary embodiment, an OS 103 of FIG. 1 reserves I/O device 112 to keep other OSs 103 from accessing the I/O device 112 while the reservation is active. Although device

reservation is not required for all I/O operations, device reservation can be used to support operations that necessitate exclusive access for a fixed duration of time, e.g., disk formatting.

The CU control logic 802 can access and control other elements within the control unit 110, such as CU timers 806 and CU registers 808. The CU timers 806 may include multiple timer functions to track how much time a sequence of I/O operations takes to complete. The CU timers 806 may further include one or more countdown timers to monitor and abort I/O operations and commands that do not complete within a predetermined period. The CU registers 808 can include fixed values that provide configuration and status information, as well as dynamic status information that is updated as commands are executed by the CU control logic 802. The control unit 110 may further include other buffer or memory elements (not depicted) to store multiple messages or status information associated with communications between the channel 124 and the I/O device 112. The CU registers 808 may include a maximum control unit exchange parameter that defines the maximum number of open control unit exchanges that the control unit 110 supports.

The channel 124 in the channel subsystem 108 includes multiple elements to support communication with the control unit 110. For example, the channel 124 may include CHN control logic 810 that interfaces with CHN subsystem timers 812 and CHN subsystem registers 814. In an exemplary embodiment, the CHN control logic 810 controls communication between the channel subsystem 108 and the control unit 110. The CHN control logic 810 may directly interface to the CU control logic 802 via the connection 120 to send commands and receive responses, such as transport command and response IUs. Alternatively, messaging interfaces and/or buffers (not depicted) can be placed between the CHN control logic 810 and the CU control logic 802. The CHN subsystem timers 812 may include multiple timer functions to track how much time a sequence of I/O operations takes to complete, in addition to the time tracked by the control unit 110. The CHN subsystem timers 812 may further include one or more countdown timers to monitor and abort command sequences that do not complete within a predetermined period. The CHN subsystem registers 814 can include fixed values that provide configuration and status information, as well as dynamic status information, updated as commands are transported and responses are received.

One example of a response message 900, e.g., a transport response IU, communicated from the control unit 110 to the channel 124 upon completion of a TCW channel program is depicted in FIG. 9. The response message 900 provides status information to the channel 124 and may indicate that an open exchange between the channel 124 and the control unit 110 should be closed. The status information provided when a TCW channel program (e.g., as depicted in FIGs. 5 and 7) is executed includes additional information beyond the status information sent upon completion of a CCW channel program (e.g., as depicted in FIGs. 3 and 6). The response message 900 includes a status section 902 and an extended status section 904. When the channel 124 receives the response message 900, it stores parts of status section 902 in the subchannel for the device the TCW was operating with and the extended status section 904 in a memory location defined by the TCW associated with the TCW channel program that triggered the response message 900. For example, a TCW can designate a section of main memory 102 of FIG. 1 for storage of the extended status section 904.

The status section 902 of the response message 900 can include multiple fields, such as an address header 906, status flags one 908, maximum control unit exchange parameter 910, response flags 912, response code 914, residual count 916, response length 918, reserved location 920, SPC-4 sense type 922, status flags two 924, status flags three 926, device status 928, and a longitudinal redundancy check (LRC) word 930. Each field in the status section 902 is assigned to a particular byte address to support parsing of the response message 900. Although one arrangement of fields within the status section 902 is depicted in FIG. 9, it will be understood that the order of fields can be rearranged to alternate ordering within the scope of the disclosure. Moreover, fields in the response message 900 can be omitted or combined within the scope of the invention, e.g., combining status flags two 924 and three 926 into a single field. SPC-4 is further described in "SCSI Primary Commands – 4 (SPC-4)", Project T10/1731-D, Rev 11, INCITS (May 2007), which is hereby incorporated herein by reference in its entirety.

In an exemplary embodiment, the address header 906 is set to the same value as the value received by the control unit 110 in the TCCB that initiated the TCW channel program. Although the address header 906 is not required, including the address header 906 may

support testing to trace command and response messages on an I/O device 112 while multiple I/O devices 112 are being accessed.

Status flags one 908 may indicate information such as the success status of an I/O operation. Multiple bits within the status flags one 908 can provide additional status information.

The maximum control unit exchange parameter 910 identifies the maximum number of exchanges that the control unit 110 allows the channel 124 to open to it. A value of zero may inform the channel 124 that the control unit 110 is not altering the current value that the channel 124 is using. In an exemplary embodiment, the channel 124 establishes a default value for the maximum number of open exchanges, e.g. 64, which the control unit 110 can modify via the maximum control unit exchange parameter 910. The value of the maximum control unit exchange parameter 910 sent in the response message 900 may be the actual value desired or a seed value for an equation. For example, the value in the maximum control unit exchange parameter 910 can be incremented and/or multiplied by the channel 124 to determine the actual maximum number of open exchanges, e.g. a value of "1" interpreted as "32" by the channel 124.

Using a default value for the maximum number of open exchanges gives each control unit 110 and channel 124 a common starting point that can be modified as determined by the control unit 110. In one embodiment, the channel 124 checks the maximum control unit exchange parameter 910 received in the response message 900 from the control unit 110 to determine if the maximum control unit exchange parameter 910 is lower than the default value or a previously received value. If the new number is smaller than the current number of open exchanges, the channel 124 does not drive new I/O commands to the control unit 110 until the current number of exchanges used is less than the new limit.

In an exemplary embodiment, the response flags field 912 uses the standard definition as defined in FCP (previously referenced) and can be set to default value, e.g., two. The response code 914 may be equivalent to a Small Computer System Interface (SCSI) status field and can be set to a default value, such as zero. The residual count 916 for read or write commands indicates the difference between how many bytes were commanded to be read or written versus the number of bytes that actually were read or written. The response length 918 is an additional count of bytes of information in the response message 900 after the

reserved location 920. The response length 918 supports variable sized response messages 900. The SPC-4 sense type 922 can be assigned to a particular value based upon message type, e.g., a transport response IU = 7F hexadecimal. In one embodiment, the status flags two 924 is set to a value of 80 hexadecimal to indicate that the I/O operation completed, with a valid value of the residual count 916. Status flags three 926 is set to a value of one when the I/O operation completed, indicating that extended status 904 is included as part of the response message 900. The device status 928 relays status information generated by the I/O device 112. The LRC word 930 is a check word that covers the other fields in the status section 902 of the response message 900 to verify the integrity of the status section 902. The LRC word 930 can be generated through applying an exclusive-or operation to an initial seed value with each field included in the LRC calculation in succession.

The extended status section 904 provides information to the channel subsystem 108 and OS 103 associated with operating the control unit 110 in a transport mode capable of running a TCW channel program. The extended status section 904 may support configurable definitions with different type status definitions for each type. In an exemplary embodiment, the extended status section 904 includes a transport status header (TSH) 932, a transport status area (TSA) 934, and an LRC word 936 of the TSH 932 and the TSA 934. The TSH 932 may include extended status length 940, extended status flags 942, a DCW offset 944, a DCW residual count 946, and a reserved location 948. The TSH 932 is common for the different formats, with the each format defined by a type code in the extended status flags 942. The TSA 934 may include a total device time parameter 950, defer time parameter 952, queue time parameter 954, device busy time parameter 956, device active only time parameter 958, and appended device sense data 960. Each of these fields is described in greater detail in turn.

The extended status length 940 is the size of the extended status section 904. In an exemplary embodiment, the extended status flags 942 has the following definition:

Bit 0 - The DCW offset 944 is valid.

Bit 1 - The DCW residual count 946 is valid.

Bit 2 - This bit set to a one informs the OS 103 of FIG. 1 in a definitive manner when the control unit 110 had to access slow media for data, e.g., a cache miss.

Bit 3 - Time parameters 950 – 958 are valid. The type code set to a one and this bit set to a one indicates that all or the time parameters 950 – 958 are valid.

Bit 4 - Reserved.

Bits 5 to 7 - These three bits are the type code that defines the format of the TSA 934 of the extended status section 904. The names of the encodes are:

0. Reserved.

1. I/O Status. The extended status section 904 contains valid ending status for the transport-mode I/O operation.

2. I/O Exception. The extended status section 904 contains information regarding termination of the transport-mode I/O operation due to an exception condition.

3. Interrogate Status. The extended status section 904 contains status for an interrogate operation.

4. to 7. Reserved.

The DCW offset 944 indicates an offset in the TCCB of a failed DCW. Similarly, the DCW residual count 946 indicates the residual byte count of a failed DCW (i.e., where execution of the DCWs was interrupted).

In an exemplary embodiment, the TSA 934 definition when the type code of ES flags 942 indicates a type of I/O Status includes time parameters 950 – 958, as well as optionally appended device sense data 960. The time parameters 950 – 958 represent time values and can be scaled to any time units, such as microseconds. The CU timers 806 of FIG. 8 are used to calculate the time parameters 950 – 958, and the CU registers 808 can also be employed to capture values of the CU timers 806 on a triggering event.

The total device time parameter 950 is the elapsed time from when the control unit 110 received the transport command IU until it sent the transport response IU (i.e., response message 900) for the I/O operation. The defer time parameter 952 indicates control unit defer time. This is the time accumulated by the control unit 110 working with the I/O device

112 when no communication with the channel 124 is performed. On CCW channel programs, such as that depicted in FIG. 3, the control unit 302 disconnects from the channel 300 during this time.

The queue time parameter 954 is the time that an I/O operation is queued at the control unit 110, but does not include queue time for device busy time where the I/O device 112 is reserved by a different OS 103 on the same or another host system 101. The device busy time parameter 956 is the time that a transport command IU is queued at the control unit 110 waiting on a device busy caused by the I/O device 112 being reserved by a different OS 103 on the same or another host system 101.

The device active only time parameter 958 is the elapsed time between a channel end (CE) and a device end (DE) at the control unit 110, when the control unit 110 holds the CE until DE is available. The CE may indicate that the portion of an I/O operation involving a transfer of data or control information between the channel 124 and the control unit 110 has been completed. The DE may indicate that the device portion of an I/O operation is completed. The appended device sense data 960 is supplemental status that the control unit 110 provides conditionally in response to an active unit check (UC) bit in the device status 928.

The LRC word 936 is a longitudinal redundancy check word of the TSH 932 and the TSA 934, calculated in a similar fashion as the LRC word 930 in the status 902 section of the response message 900. The LRC word 936 can be calculated on a variable number of words, depending upon the number of words included in the appended device sense data 960.

Turning now to FIG. 10, multiple host systems 101 are depicted in communication with control unit 110 via connections 120. Each host system 101 includes channel subsystem 108 with one or more channels 124. Although only one channel 124 is depicted in each host system 101 of FIG. 10, it will be understood that each host system 101 can include multiple channels 124 controlled by multiple OSs 103. The host systems 101 also include other processing system elements, as previously depicted and described in reference to FIG. 1, i.e., one or more CPUs 104 coupled to storage element 106 and main memory 102. Each host system 101 may execute one or more OSs 103, with each OS 103 capable of making a

reservation request for exclusive access to I/O device 112. The OSs 103 in each host system 101 can individually control one or more channels 124 to initiate I/O operations. The OSs 103 can use different sub-channels (not depicted) on one or more channels 124 to communicate with the control unit 110.

Each connection 120 between a channel 124 and the control unit 110 can be a direct connection. Alternatively, connections 120 may pass through one or more dynamic switches 1002 as part of a Fibre Channel fabric to reduce the number of physical connections at the control unit 110.

As previously described in reference to FIG. 8, CU control logic 802 parses and processes command messages containing TCCBs, such as the TCCB 704 of FIG. 7, received from the channels 124 via the connections 120. Some commands received at the CU control logic 802 may include a device reservation request. While device reservation is not required for all I/O operations, device reservation may be requested on a per OS 103 basis for I/O operations that require exclusive access to I/O device 112. For example, different OSs 103 may both request to read blocks of data from I/O device 112. The control unit 110 can service each read request without reserving the I/O device 112. However, when one of the OSs 103 reserves I/O device 112 over one of the channels 124, other OSs 103 attempting to access I/O device 112 are blocked while the I/O device 112 is reserved.

In an exemplary embodiment, the CU control logic 802 receives a device busy indicator from the I/O device 112 when the I/O device 112 is reserved for an OS 103. As the CU control logic 802 receives command messages while the device busy indicator is present, the CU control logic 802 may place the command messages on the DB queue 804. The command messages can contain identification information establishing a particular OS 103 and/or channel 124 associated with each command message. In an alternate exemplary embodiment, the CU control logic 802 keeps track of the particular OS 103 and/or channel 124 associated with each command message in the CU registers 808. When I/O device 112 is no longer reserved, it notifies the CU control logic 802 via a device end indicator. In response to the device end indicator, the CU control logic 802 services the DB queue 804 to extract a command message for the I/O device 112 to perform. The extracted command message may again result in reserving the I/O device 112, causing further delays in servicing pending command messages. Alternatively, the extracted command message may not

require reservation of the I/O device 112 (e.g., exclusive access to the I/O device 112 is not needed), allowing additional servicing of the DB queue 804 to perform additional command messages in succession.

Servicing the DB queue 804 can be performed using a variety of techniques to manage the DB queue 804. For example, the DB queue 804 can be managed as a FIFO to extract each command message in the order that it was placed in the DB queue 804. Alternatively, the DB queue 804 can be serviced as a priority queue. The priority of the command messages written to the DB queue 804 may be included in a field within each command message that is transported to the control unit 110. Any number of priorities can be established for a range of scheduling options. When the DB queue 804 is serviced as a priority queue, the highest priority command message in the DB queue 804 is extracted before lower priority command messages upon servicing. The amount of time that command messages remain in the DB queue 804 can be monitored to increase the priority of command messages over a period of time to ensure that they are serviced. Another queue servicing approach for the DB queue 804 is round robin servicing. Using round robin servicing, the OS 103 (or channel 124) associated with each command message in the DB queue 804 is analyzed to service the DB queue 804 on a per OS 103 basis. Round robin servicing provides access for each communication link to prevent potential disparity that can arise if one OS 103 sends a burst of multiple access requests to the control unit 110.

When a command message is queued on the DB queue 804, a device busy timer in the CU registers 808 is initiated. Upon servicing the DB queue 804 to perform an I/O operation command, the value of the device busy timer in the CU registers 808 is read to determine how long the command message was pending in the DB queue 804. The value of the device busy timer is reported in the device busy time parameter 956 of the response message 900 of FIG. 9. There may be multiple device busy timers in the CU registers 808 to support multiple OSs 103. Alternatively, the device busy timer in the CU registers 808 can be a continuously running timer with device busy time values captured in the CU registers 808 for multiple OSs 103, and output to each respective OS 103 via channels 124.

As multiple command messages are queued to the DB queue 804, the depth of the DB queue 804 is monitored. If the DB queue 804 is full such that no more command messages can be queued, the CU control logic 802 may output a device busy message to all OSs 103 that send

new command messages while the DB queue 804 full. The CU control logic 802 can then send a device end message to indicate that the I/O device 112 is ready, which may trigger the OSs 103 to resend the command messages via one or more channels 124. Alternatively some of the command messages in the DB queue 804 can be returned to the OSs 103 with a device busy and flushed from the DB queue 804. Again, the CU control logic 802 can send a device end message to indicate that the I/O device 112 is ready, which may trigger the OSs 103 to resend the command messages via one or more channels 124.

In managing the DB queue 804, the CU control logic 802 may notify OSs 103 of a busy condition under a variety of scenarios beyond a full queue condition. For example, the CU control logic 802 can use CU timers 806 to monitor time that command messages remain on the DB queue 804. When a command message is queued for a period of time greater than a command timeout period while I/O device 112 is reserved (e.g., device end indicator has not been received), then the command message is removed from the DB queue 804, and a device busy message is sent in an FCP_RSP IU to the originator of the command message. The command timeout period may be set to a fixed value, such as thirty seconds, or configurable. In an exemplary embodiment, when a new command message is received while I/O device 112 has been reserved for greater than a device busy timeout period, the CU control logic 802 does not place the new command message on the DB queue 804. The CU control logic 802 sends a device busy message in an FCP_RSP IU to the originator of the new command message.

OSs 103 may also monitor elapsed time for a requested command message to complete. In response to determining that an operating system timeout period has elapsed, OS 103 may send a reset allegiance message to the control unit 110 to attempt to free up I/O device 112. In response thereto, the control unit 110 enters an operating system timeout recovery period. If a new command message is received during the operating system timeout recovery period, the CU control logic 802 does not place the new command message on the DB queue 804. The CU control logic 802 responds with a device busy message in an FCP_RSP IU to the originator of the new command message.

Turning now to FIG. 11, a process 1100 for reducing reserved device access contention at a control unit in communication with a plurality of OS via one or more channels will now be described in accordance with exemplary embodiments, and in reference to the I/O processing

system 100 of FIG. 1 and the detailed view of control unit 110 of FIG. 10. At block 1102, the control unit 110 receives a command message from a first OS of a plurality of OSs 103 via one or more channels 124, where the command message includes an I/O operation command for I/O device 112 in communication with the control unit 110. The command message may be a transport command IU, including a TCCB with multiple DCWs as part of a TCW channel program. At block 1104, the control unit 110 receives a device busy indicator from I/O device 112. The device busy indicator notifies the control unit 110 that a second OS of the plurality of OSs 103 has reserved I/O device 112.

At block 1106, the control unit 110 queues the command message on DB queue 804 in response to the device busy indicator. As additional command messages are received at the control unit 110, command messages are queued on the DB queue 804.

At block 1108, the control unit 110 monitors I/O device 112 for a device end indicator, where the device end indicator notifies the control unit 110 that the I/O device 112 is ready to receive a new I/O operation command. At block 1110, the control unit 110 services the DB queue 804 to extract a command message and perform I/O operation commands in response to the device end indicator. The DB queue 804 can be serviced using FIFO servicing, priority based servicing, or round robin servicing, as previously described.

Technical effects of exemplary embodiments include reducing reserved device access contention in an I/O processing system. Using a device busy queue to temporarily store command messages received while a device is reserved allows a control unit to manage the servicing order of command messages from different OSs without burdening channel subsystems of host systems that originated the command messages. Advantages include handling multiple command messages without interrupting the execution of a TCW channel program on a control unit. Thus, device busy queuing handles access contention to a reserved device, while also gaining advantages of higher communication throughput due in part to exchanging fewer messages per channel program. A variety of device busy queue servicing techniques can be used depending upon the preferences of the system designer or customer. In exemplary embodiments, command messages originating from slower responding host systems are serviced equitably with respect to faster responding host systems.

As described above, embodiments can be embodied in the form of computer-implemented processes and apparatuses for practicing those processes. In exemplary embodiments, the invention is embodied in computer program code executed by one or more network elements. Embodiments include a computer program product 1200 as depicted in FIG. 12 on a computer usable medium 1202 with computer program code logic 1204 containing instructions embodied in tangible media as an article of manufacture. Exemplary articles of manufacture for computer usable medium 1202 may include floppy diskettes, CD-ROMs, hard drives, universal serial bus (USB) flash drives, or any other computer-readable storage medium, wherein, when the computer program code logic 1204 is loaded into and executed by a computer, the computer becomes an apparatus for practicing the invention. Embodiments include computer program code logic 1204, for example, whether stored in a storage medium, loaded into and/or executed by a computer, or transmitted over some transmission medium, such as over electrical wiring or cabling, through fiber optics, or via electromagnetic radiation, wherein, when the computer program code logic 1204 is loaded into and executed by a computer, the computer becomes an apparatus for practicing the invention. When implemented on a general-purpose microprocessor, the computer program code logic 1204 segments configure the microprocessor to create specific logic circuits.

While the invention has been described with reference to exemplary embodiments, it will be understood by those skilled in the art that various changes may be made and equivalents may be substituted for elements thereof without departing from the scope of the invention. In addition, many modifications may be made to adapt a particular situation or material to the teachings of the invention without departing from the essential scope thereof. Therefore, it is intended that the invention not be limited to the particular embodiment disclosed as the best mode contemplated for carrying out this invention, but that the invention will include all embodiments falling within the scope of the appended claims. Moreover, the use of the terms first, second, etc. do not denote any order or importance, but rather the terms first, second, etc. are used to distinguish one element from another. Furthermore, the use of the terms a, an, etc. do not denote a limitation of quantity, but rather denote the presence of at least one of the referenced item.

CLAIMS

1. A computer program product for reducing reserved device access contention at a control unit in communication with a plurality of operating systems via one or more channels, the computer program product comprising:

a tangible storage medium readable by a processing circuit and storing instructions for execution by the processing circuit for performing a method comprising:

receiving a command message at the control unit from a first operating system of the plurality of operating systems via the one or more channels, wherein the command message includes an I/O operation command for a device in communication with the control unit;

receiving a device busy indicator from the device, wherein the device busy indicator notifies the control unit that the device is reserved by a second operating system of the plurality of operating systems;

queuing the command message on a device busy queue in response to the device busy indicator;

monitoring the device for a device end indicator, wherein the device end indicator notifies the control unit that the device is ready to receive a new I/O operation command; and

servicing the device busy queue to perform the I/O operation command in response to the device end indicator.

2. The computer program product of claim 1 wherein the method further comprises:

receiving additional command messages including additional I/O operation commands at the control unit from the plurality of operating systems via the one or more channels;

queuing the additional command messages on the device busy queue in response to the device busy indicator; and

servicing the device busy queue to perform the additional I/O operation commands.

3. The computer program product of claim 2 wherein the servicing of the device busy queue is performed as a first in first out (FIFO) servicing, the FIFO servicing outputting the additional command messages in order of queuing on the device busy queue.

4. The computer program product of claim 2 wherein the servicing of the device busy queue is performed as a priority based servicing, the priority based servicing outputting the additional command messages from the device busy queue in order of a higher priority command message before a lower priority command message.

5. The computer program product of claim 2 wherein the servicing of the device busy queue is performed as a round robin servicing, the round robin servicing outputting the additional command messages from the device busy queue on a per operating system basis.

6. The computer program product of claim 1 wherein the method further comprises:

outputting a device busy message in response to one or more of:

receiving an additional command message at the control unit while the device busy queue is full;

receiving the additional command message at the control unit while the device end indicator has not been received within a device busy timeout period;

receiving the additional command message at the control unit within an operating system timeout recovery period; and

determining that the device end indicator has not been received within a command timeout period.

7. The computer program product of claim 1 wherein the method further comprises:

initiating a device busy timer in response to queuing the command message on the device busy queue;

reading a value of the device busy timer in response to servicing the device busy queue to perform the I/O operation command; and

outputting the value of the device busy timer in a transport response information unit message to the first operating system via the one or more channels.

8. The computer program product of claim 1 wherein the command message is a transport command information unit message, including a transport command control block (TCCB) holding the I/O operation command as part of a transport control word (TCW) channel program.

9. An apparatus for reducing reserved device access contention, the apparatus comprising:

a control unit in communication with a plurality of operating systems via one or more channels, the control unit performing a method comprising:

receiving a command message at the control unit from a first operating system of the plurality of operating systems via the one or more channels, wherein the command message includes an I/O operation command for a device in communication with the control unit;

receiving a device busy indicator from the device, wherein the device busy indicator notifies the control unit that the device is reserved by a second operating system of the plurality of operating systems;

queuing the command message on a device busy queue in response to the device busy indicator;

monitoring the device for a device end indicator, wherein the device end indicator notifies the control unit that the device is ready to receive a new I/O operation command; and

servicing the device busy queue to perform the I/O operation command in response to the device end indicator.

10. The apparatus of claim 9 wherein the control unit further performs:

receiving additional command messages including additional I/O operation commands at the control unit from the plurality of operating systems via the one or more channels;

queuing the additional command messages on the device busy queue in response to the device busy indicator; and

servicing the device busy queue to perform the additional I/O operation commands.

11. The apparatus of claim 10 wherein the servicing of the device busy queue is performed as a first in first out (FIFO) servicing, the FIFO servicing outputting the additional command messages in order of queuing on the device busy queue.

12. The apparatus of claim 10 wherein the servicing of the device busy queue is performed as a priority based servicing, the priority based servicing outputting the additional command messages from the device busy queue in order of a higher priority I/O command message before a lower priority I/O command message.

13. The apparatus of claim 10 wherein the servicing of the device busy queue is performed as a round robin servicing, the round robin servicing outputting the additional command messages to perform from the device busy queue on a per operating system basis.

14. The apparatus of claim 9 wherein the control unit further performs:

outputting a device busy message in response to one or more of:

receiving an additional command message at the control unit while the device busy queue is full;

receiving the additional command message at the control unit while the device end indicator has not been received within a device busy timeout period;

receiving the additional command message at the control unit within an operating system timeout recovery period; and

determining that the device end indicator has not been received within a command timeout period.

15. The apparatus of claim 9 wherein the control unit further performs:

initiating a device busy timer in response to queuing the command message on the device busy queue;

reading a value of the device busy timer in response to servicing the device busy queue to perform the I/O operation command; and

outputting the value of the device busy timer in a transport response information unit message to the first operating system via the one or more channels.

16. The apparatus of claim 9 wherein the command message is a transport command information unit message, including a transport command control block (TCCB) holding the I/O operation command as part of a transport control word (TCW) channel program.

17. A method for reducing reserved device access contention at a control unit in communication with a plurality of operating systems via one or more channels, the method comprising:

receiving a command message at the control unit from a first operating system of the plurality of operating systems via the one or more channels, wherein the command message includes an I/O operation command for a device in communication with the control unit;

receiving a device busy indicator from the device, wherein the device busy indicator notifies the control unit that the device is reserved by a second operating system of the plurality of operating systems;

queuing the command message on a device busy queue in response to the device busy indicator;

monitoring the device for a device end indicator, wherein the device end indicator notifies the control unit that the device is ready to receive a new I/O operation command; and

servicing the device busy queue to perform the I/O operation command in response to the device end indicator.

18. The method of claim 17 further comprising:

receiving additional command messages including additional I/O operation commands at the control unit from the plurality of operating systems via the one or more channels;

queuing the additional command messages on the device busy queue in response to the device busy indicator; and

servicing the device busy queue to perform the additional I/O operation commands, wherein the servicing of the device busy queue is performed as one of a first in first out (FIFO) servicing, a priority based servicing, and a round robin servicing.

19. The method of claim 17 further comprising:

outputting a device busy message in response to one or more of:

receiving an additional command message at the control unit while the device busy queue is full;

receiving the additional command message at the control unit while the device end indicator has not been received within a device busy timeout period;

receiving the additional command message at the control unit within an operating system timeout recovery period; and

determining that the device end indicator has not been received within a command timeout period.

20. The method of claim 17 wherein the command message is a transport command information unit message, including a transport command control block (TCCB) holding the I/O operation command as part of a transport control word (TCW) channel program, and further comprising:

initiating a device busy timer in response to queuing the command message on the device busy queue;

reading a value of the device busy timer in response to servicing the device busy queue to perform the I/O operation command; and

outputting the value of the device busy timer in a transport response information unit message to the first operating system via the one or more channels.

AMENDED CLAIMS
received by the International Bureau on 06 July 2009 (06.07.2009)

CLAIMS

1. A computer program product for reducing reserved device access contention at a control unit in communication with a plurality of operating systems via one or more channels, the computer program product comprising:

a tangible storage medium readable by a processing circuit and storing instructions for execution by the processing circuit for performing a method comprising:

receiving a command message at the control unit from a first operating system of the plurality of operating systems via the one or more channels, wherein the command message includes an I/O operation command for a device in communication with the control unit;

receiving a device busy indicator from the device, wherein the device busy indicator notifies the control unit that the device is reserved by a second operating system of the plurality of operating systems;

queuing the command message on a device busy queue in response to the device busy indicator;

monitoring the device for a device end indicator, wherein the device end indicator notifies the control unit that the device is ready to receive a new I/O operation command; and

servicing the device busy queue to perform the I/O operation command in response to the device end indicator;

wherein the method further comprises:

initiating a device busy timer in response to queuing the command message on the device busy queue;

reading a value of the device busy timer in response to servicing the device busy queue to perform the I/O operation command; and

outputting the value of the device busy timer in a transport response information unit message to the first operating system via the one or more channels.

2. The computer program product of claim 1 wherein the method further comprises:

receiving additional command messages including additional I/O operation commands at the control unit from the plurality of operating systems via the one or more channels;

queuing the additional command messages on the device busy queue in response to the device busy indicator; and

servicing the device busy queue to perform the additional I/O operation commands.

3. The computer program product of claim 2 wherein the servicing of the device busy queue is performed as a first in first out (FIFO) servicing, the FIFO servicing outputting the additional command messages in order of queuing on the device busy queue.

4. The computer program product of claim 2 wherein the servicing of the device busy queue is performed as a priority based servicing, the priority based servicing outputting the additional command messages from the device busy queue in order of a higher priority command message before a lower priority command message.

5. The computer program product of claim 2 wherein the servicing of the device busy queue is performed as a round robin servicing, the round robin servicing outputting the additional command messages from the device busy queue on a per operating system basis.

6. The computer program product of claim 1 wherein the method further comprises:

outputting a device busy message in response to one or more of:

receiving an additional command message at the control unit while the device busy queue is full;

receiving the additional command message at the control unit while the device end indicator has not been received within a device busy timeout period;

receiving the additional command message at the control unit within an operating system timeout recovery period; and

determining that the device end indicator has not been received within a command timeout period.

7. The computer program product of claim 1 wherein the command message is a transport command information unit message, including a transport command control block (TCCB) holding the I/O operation command as part of a transport control word (TCW) channel program.

8. An apparatus for reducing reserved device access contention, the apparatus comprising:

a control unit in communication with a plurality of operating systems via one or more channels, the control unit performing a method comprising:

receiving a command message at the control unit from a first operating system of the plurality of operating systems via the one or more channels, wherein the command message includes an I/O operation command for a device in communication with the control unit;

receiving a device busy indicator from the device, wherein the device busy indicator notifies the control unit that the device is reserved by a second operating system of the plurality of operating systems;

queuing the command message on a device busy queue in response to the device busy indicator;

monitoring the device for a device end indicator, wherein the device end indicator notifies the control unit that the device is ready to receive a new I/O operation command; and

servicing the device busy queue to perform the I/O operation command in response to the device end indicator;

wherein the control unit further performs:

initiating a device busy timer in response to queuing the command message on the device busy queue;

reading a value of the device busy timer in response to servicing the device busy queue to perform the I/O operation command; and

outputting the value of the device busy timer in a transport response information unit message to the first operating system via the one or more channels.

9. The apparatus of claim 8 wherein the control unit further performs:

receiving additional command messages including additional I/O operation commands at the control unit from the plurality of operating systems via the one or more channels;

queuing the additional command messages on the device busy queue in response to the device busy indicator; and

servicing the device busy queue to perform the additional I/O operation commands.

10. The apparatus of claim 9 wherein the servicing of the device busy queue is performed as a first in first out (FIFO) servicing, the FIFO servicing outputting the additional command messages in order of queuing on the device busy queue.

11. The apparatus of claim 9 wherein the servicing of the device busy queue is performed as a priority based servicing, the priority based servicing outputting the additional command messages from the device busy queue in order of a higher priority I/O command message before a lower priority I/O command message.

12. The apparatus of claim 9 wherein the servicing of the device busy queue is performed as a round robin servicing, the round robin servicing outputting the additional command messages to perform from the device busy queue on a per operating system basis.

13. The apparatus of claim 8 wherein the control unit further performs:

outputting a device busy message in response to one or more of:

receiving an additional command message at the control unit while the device busy queue is full;

receiving the additional command message at the control unit while the device end indicator has not been received within a device busy timeout period;

receiving the additional command message at the control unit within an operating system timeout recovery period; and

determining that the device end indicator has not been received within a command timeout period.

14. The apparatus of claim 8 wherein the command message is a transport command information unit message, including a transport command control block (TCCB) holding the I/O operation command as part of a transport control word (TCW) channel program.

15. A method for reducing reserved device access contention at a control unit in communication with a plurality of operating systems via one or more channels, the method comprising:

receiving a command message at the control unit from a first operating system of the plurality of operating systems via the one or more channels, wherein the command message includes an I/O operation command for a device in communication with the control unit;

receiving a device busy indicator from the device, wherein the device busy indicator notifies the control unit that the device is reserved by a second operating system of the plurality of operating systems;

queuing the command message on a device busy queue in response to the device busy indicator;

monitoring the device for a device end indicator, wherein the device end indicator notifies the control unit that the device is ready to receive a new I/O operation command; and

servicing the device busy queue to perform the I/O operation command in response to the device end indicator;

wherein the command message is a transport command information unit message, including a transport command control block (TCCB) holding the I/O operation command as part of a transport control word (TCW) channel program, and further comprising:

initiating a device busy timer in response to queuing the command message on the device busy queue;

reading a value of the device busy timer in response to servicing the device busy queue to perform the I/O operation command; and

outputting the value of the device busy timer in a transport response information unit message to the first operating system via the one or more channels.

16. The method of claim 15 further comprising:

receiving additional command messages including additional I/O operation commands at the control unit from the plurality of operating systems via the one or more channels;

queuing the additional command messages on the device busy queue in response to the device busy indicator; and

servicing the device busy queue to perform the additional I/O operation commands, wherein the servicing of the device busy queue is performed as one of a first in first out (FIFO) servicing, a priority based servicing, and a round robin servicing.

17. The method of claim 15 further comprising:

outputting a device busy message in response to one or more of:

receiving an additional command message at the control unit while the device busy queue is full;

receiving the additional command message at the control unit while the device end indicator has not been received within a device busy timeout period;

receiving the additional command message at the control unit within an operating system timeout recovery period; and

determining that the device end indicator has not been received within a command timeout period.

5

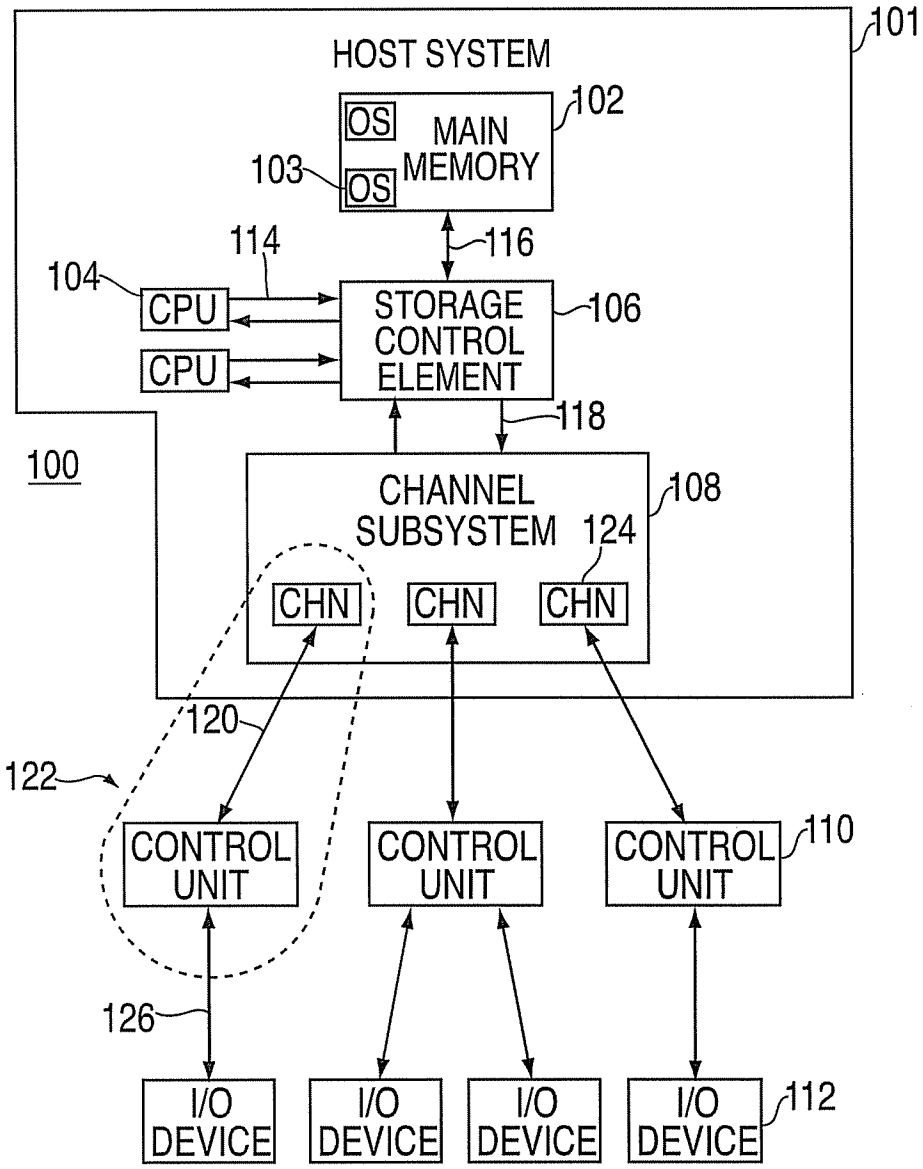
HOST SYSTEM
101
OS
MAIN MEMORY
102
103
OS
114
116
104
CPU
CPU
STORAGE CONTROL ELEMENT
106
118
100
CHANNEL SUBSYSTEM
108
124
CHN
CHN
CHN
120
122
CONTROL UNIT
CONTROL UNIT
CONTROL UNIT
110
126
I/O DEVICE
I/O DEVICE
I/O DEVICE
I/O DEVICE
112

FIG. 1

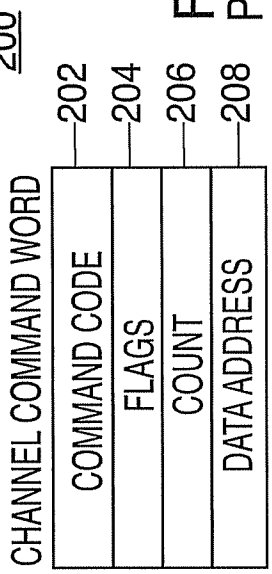

FIG. 2A
PRIOR ART

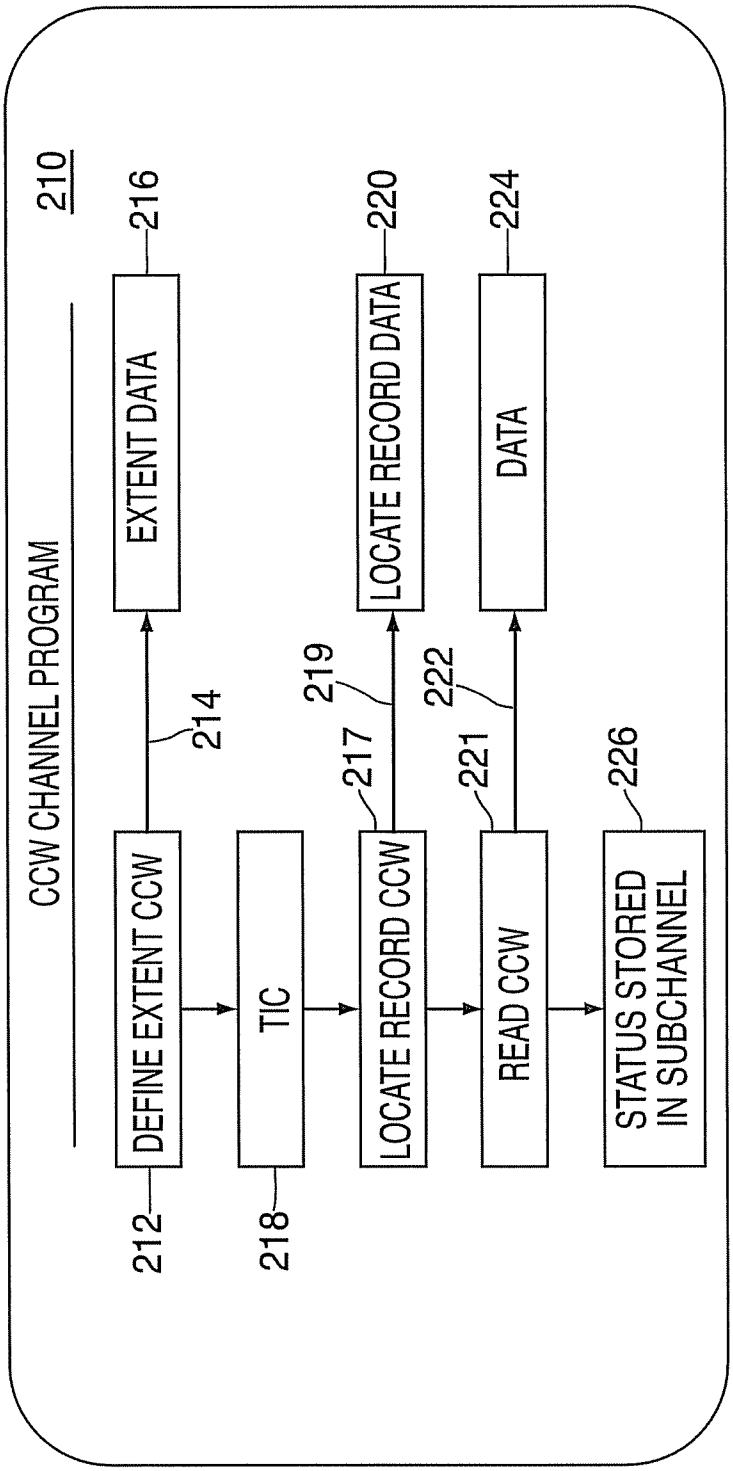

FIG. 2B
PRIOR ART

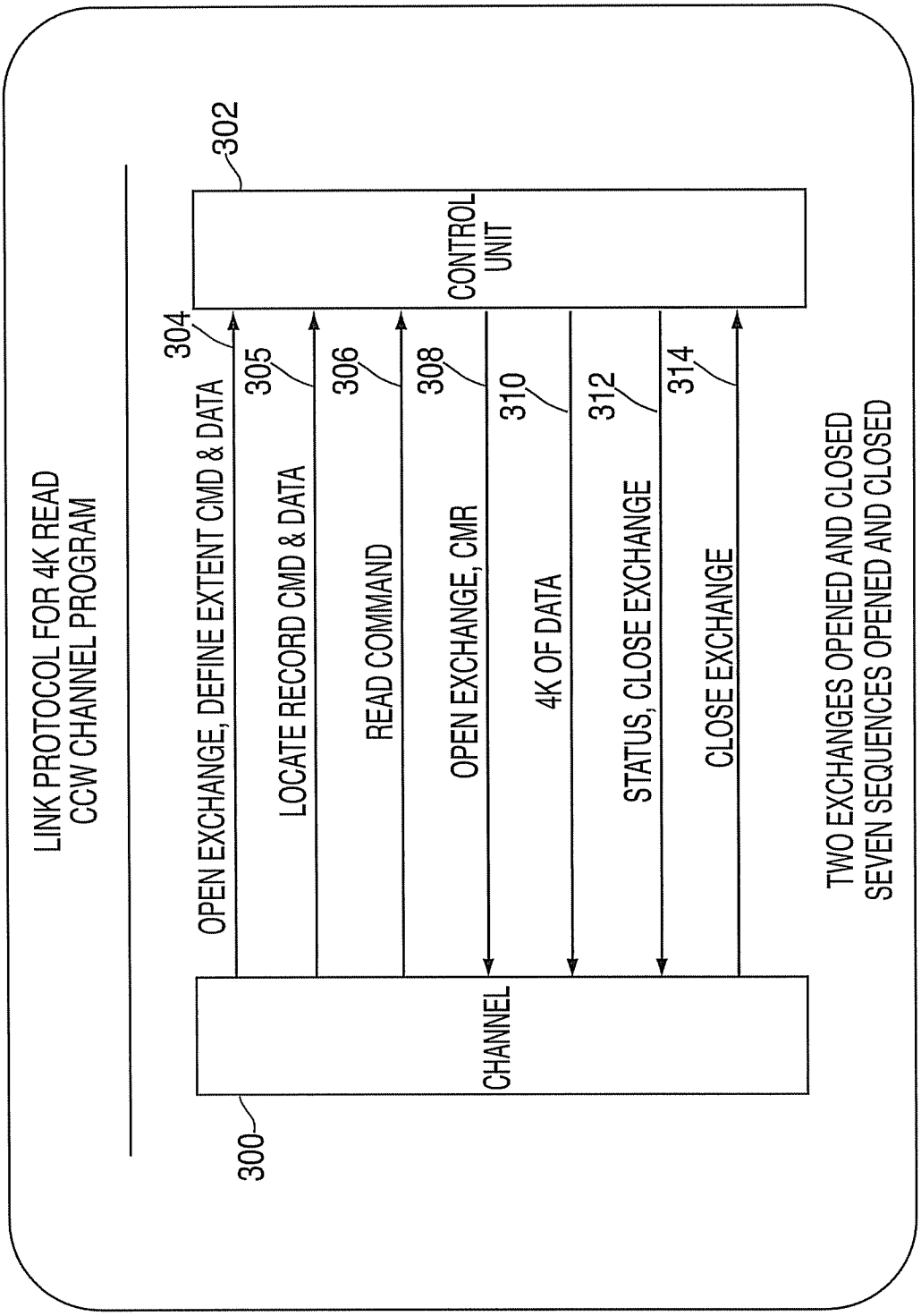

FIG. 3
PRIOR ART

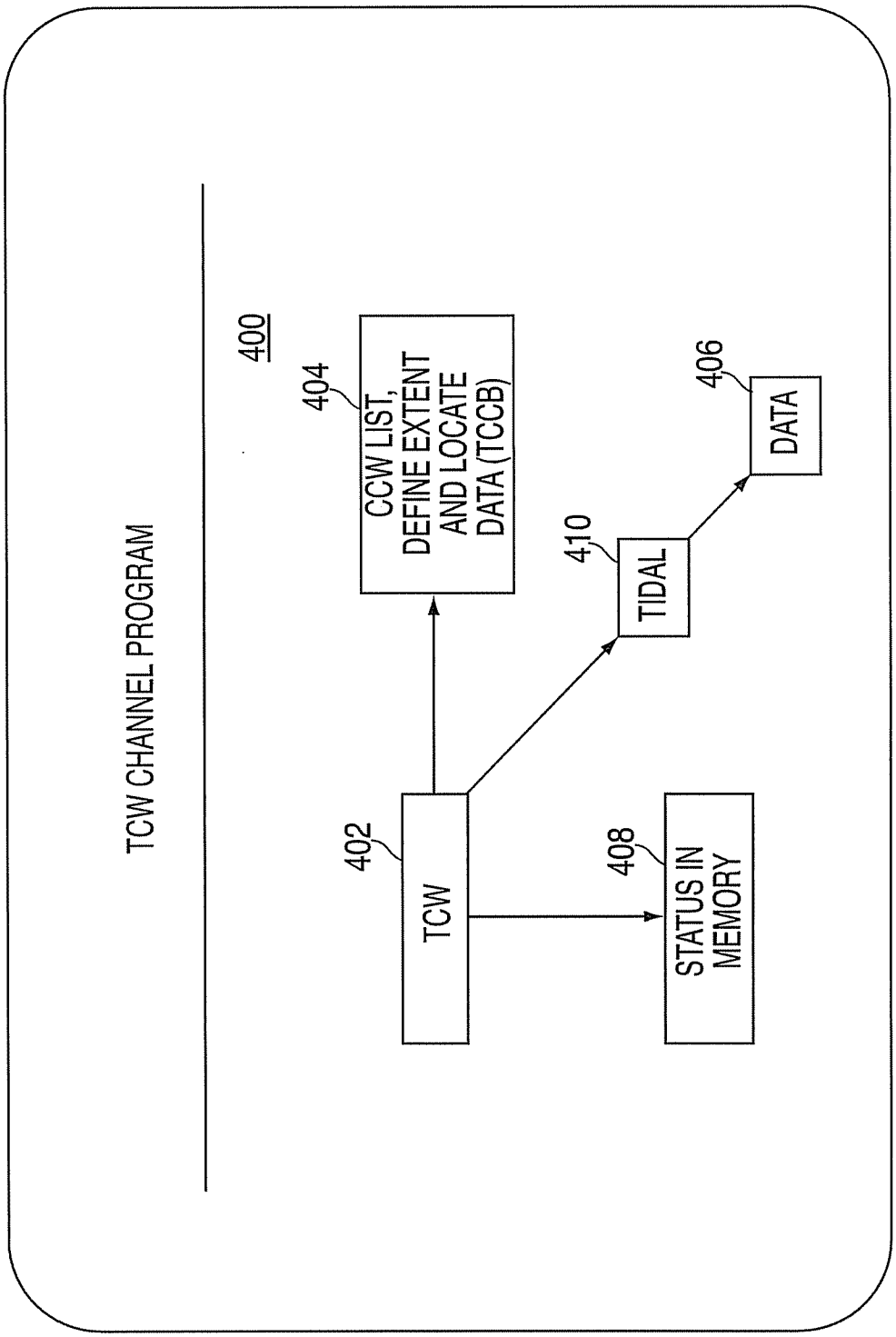
TCW CHANNEL PROGRAM
400
404
CCW LIST, DEFINE EXTENT AND LOCATE DATA (TCCB)
406
DATA
410
TIDAL
402
TCW
408
STATUS IN MEMORY

FIG. 4

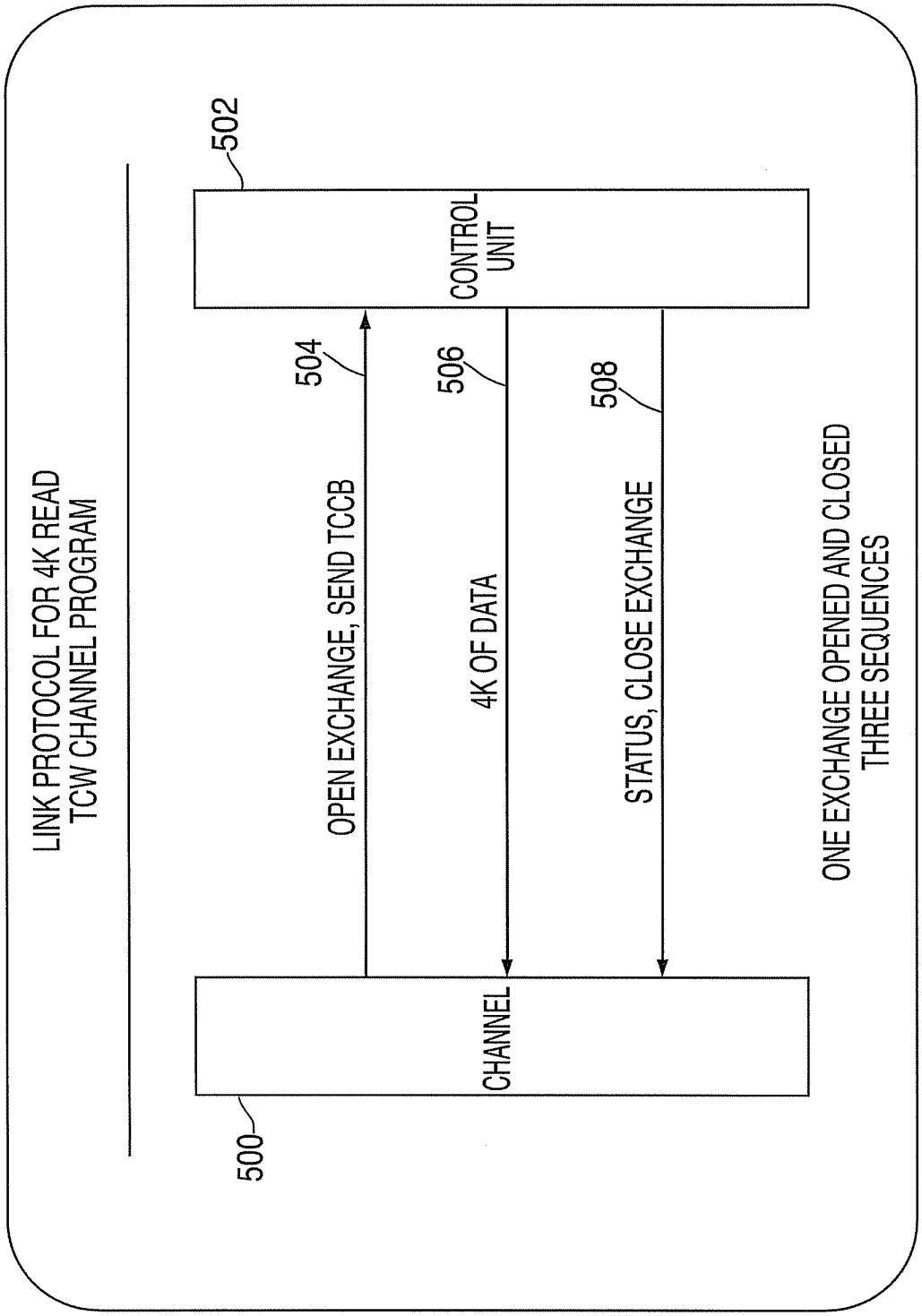
LINK PROTOCOL FOR 4K READ
TCW CHANNEL PROGRAM
502
CONTROL
UNIT
504
OPEN EXCHANGE, SEND TCCB
506
4K OF DATA
508
STATUS, CLOSE EXCHANGE
ONE EXCHANGE OPENED AND CLOSED
THREE SEQUENCES
CHANNEL
500

FIG. 5

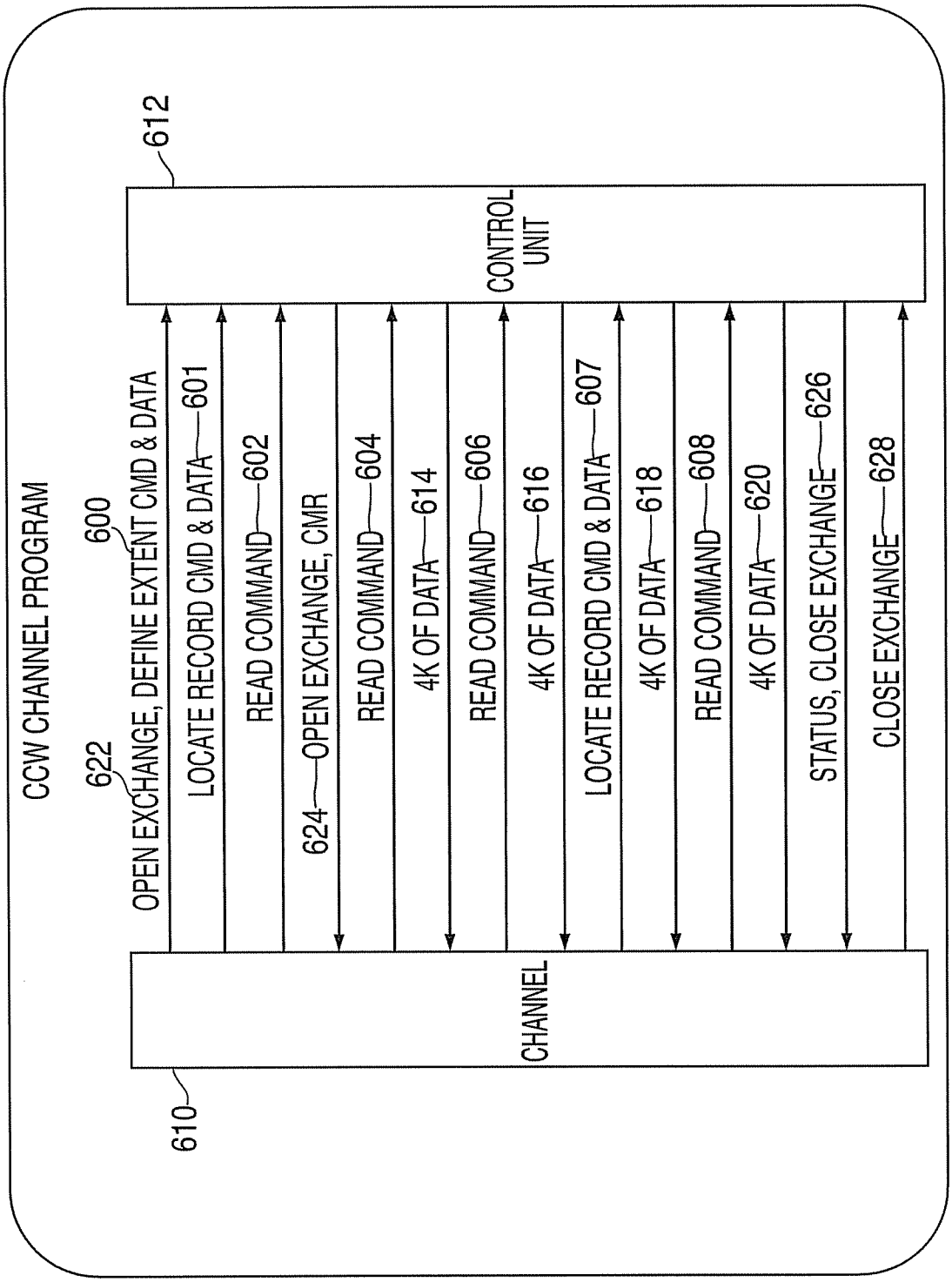

FIG. 6
PRIOR ART

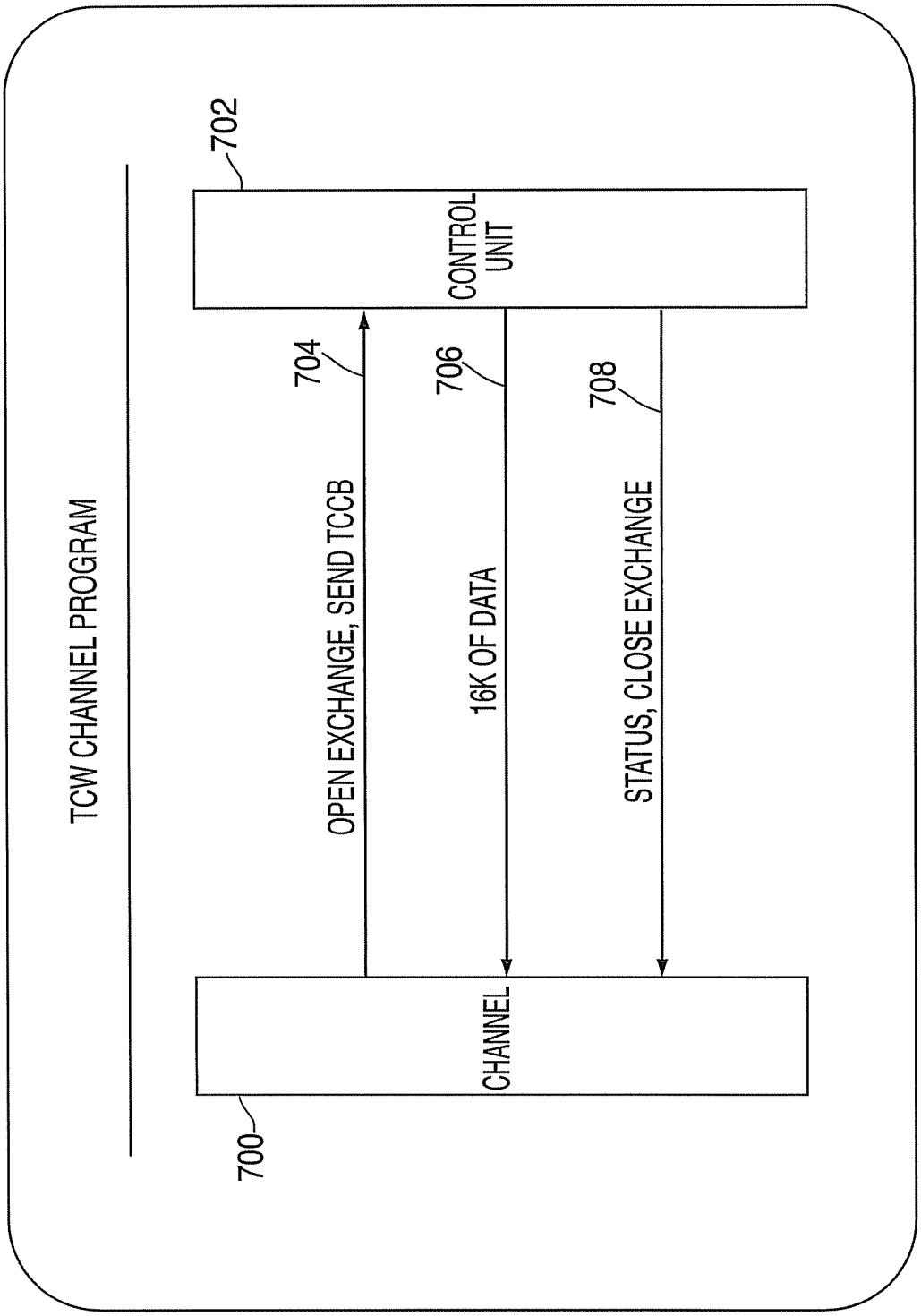
TCW CHANNEL PROGRAM
702
CONTROL UNIT
704
706
708
OPEN EXCHANGE, SEND TCCB
16K OF DATA
STATUS, CLOSE EXCHANGE
CHANNEL
700

FIG. 7

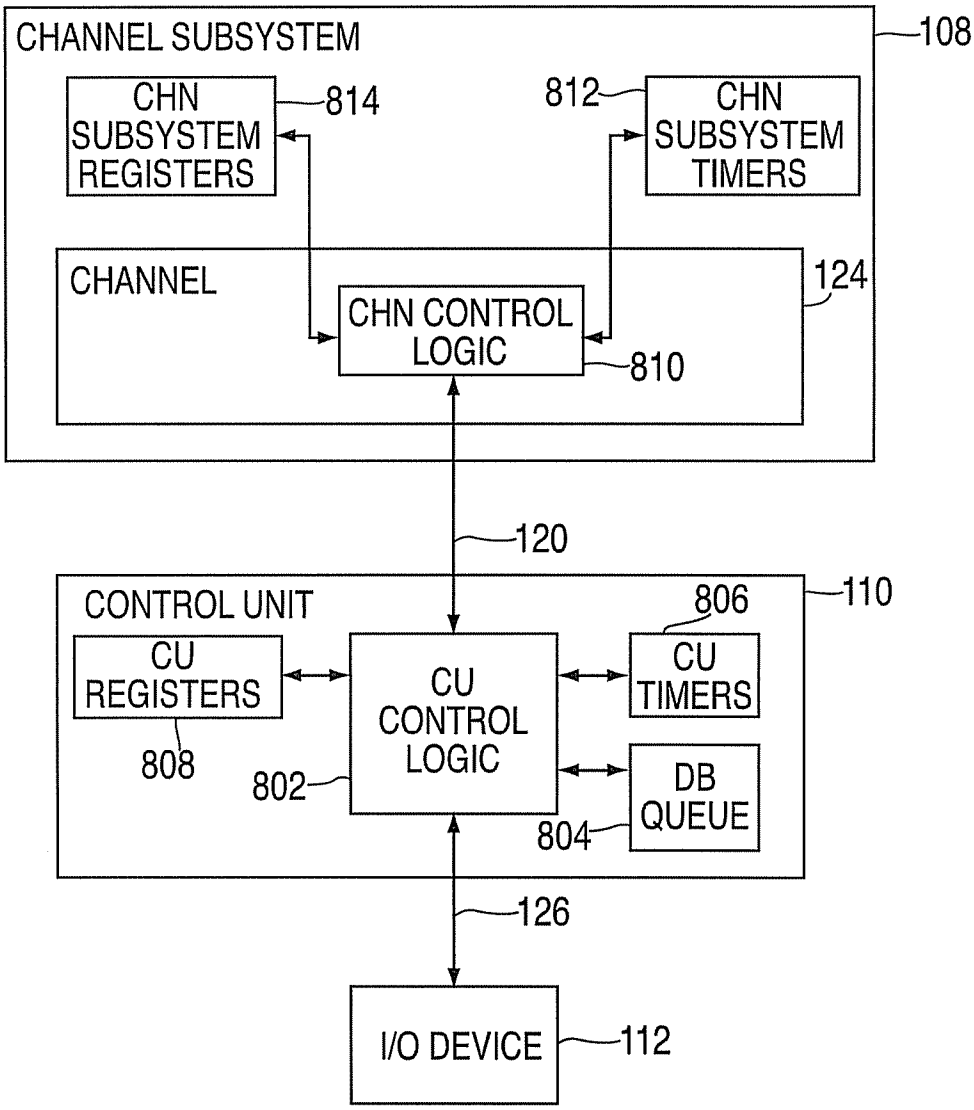
CHANNEL SUBSYSTEM
108
CHN SUBSYSTEM REGISTERS
814
812
CHN SUBSYSTEM TIMERS
CHANNEL
124
CHN CONTROL LOGIC
810
120
CONTROL UNIT
110
806
CU REGISTERS
808
CU CONTROL LOGIC
802
CU TIMERS
DB QUEUE
804
126
I/O DEVICE
112

FIG. 8

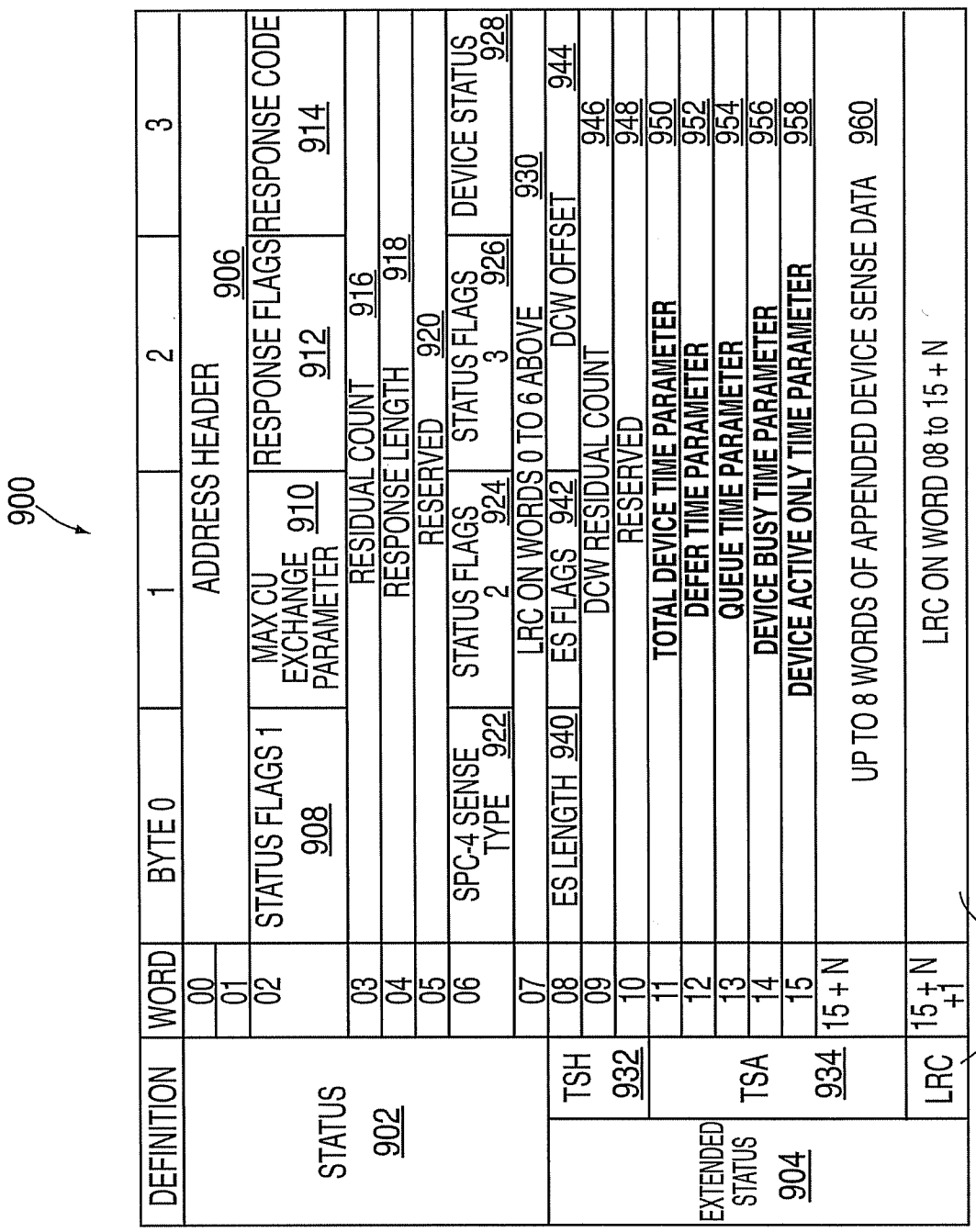
900
DEFINITION
WORD
BYTE 0
1
2
3
STATUS 902
00
01
02
ADDRESS HEADER 906
STATUS FLAGS 1 908
MAX CU EXCHANGE PARAMETER 910
RESPONSE FLAGS 912
RESPONSE CODE 914
03
RESIDUAL COUNT 916
04
RESPONSE LENGTH 918
05
RESERVED 920
06
SPC-4 SENSE TYPE 922
STATUS FLAGS 2 924
STATUS FLAGS 3 926
DEVICE STATUS 928
07
LRC ON WORDS 0 TO 6 ABOVE 930
EXTENDED STATUS 904
TSH 932
08
ES LENGTH 940
ES FLAGS 942
DCW OFFSET 944
09
DCW RESIDUAL COUNT 946
10
RESERVED 948
TSA 934
11
TOTAL DEVICE TIME PARAMETER 950
12
DEFER TIME PARAMETER 952
13
QUEUE TIME PARAMETER 954
14
DEVICE BUSY TIME PARAMETER 956
15
DEVICE ACTIVE ONLY TIME PARAMETER 958
15 + N
UP TO 8 WORDS OF APPENDED DEVICE SENSE DATA 960
LRC
15 + N +1
LRC ON WORD 08 to 15 + N
936

FIG. 9

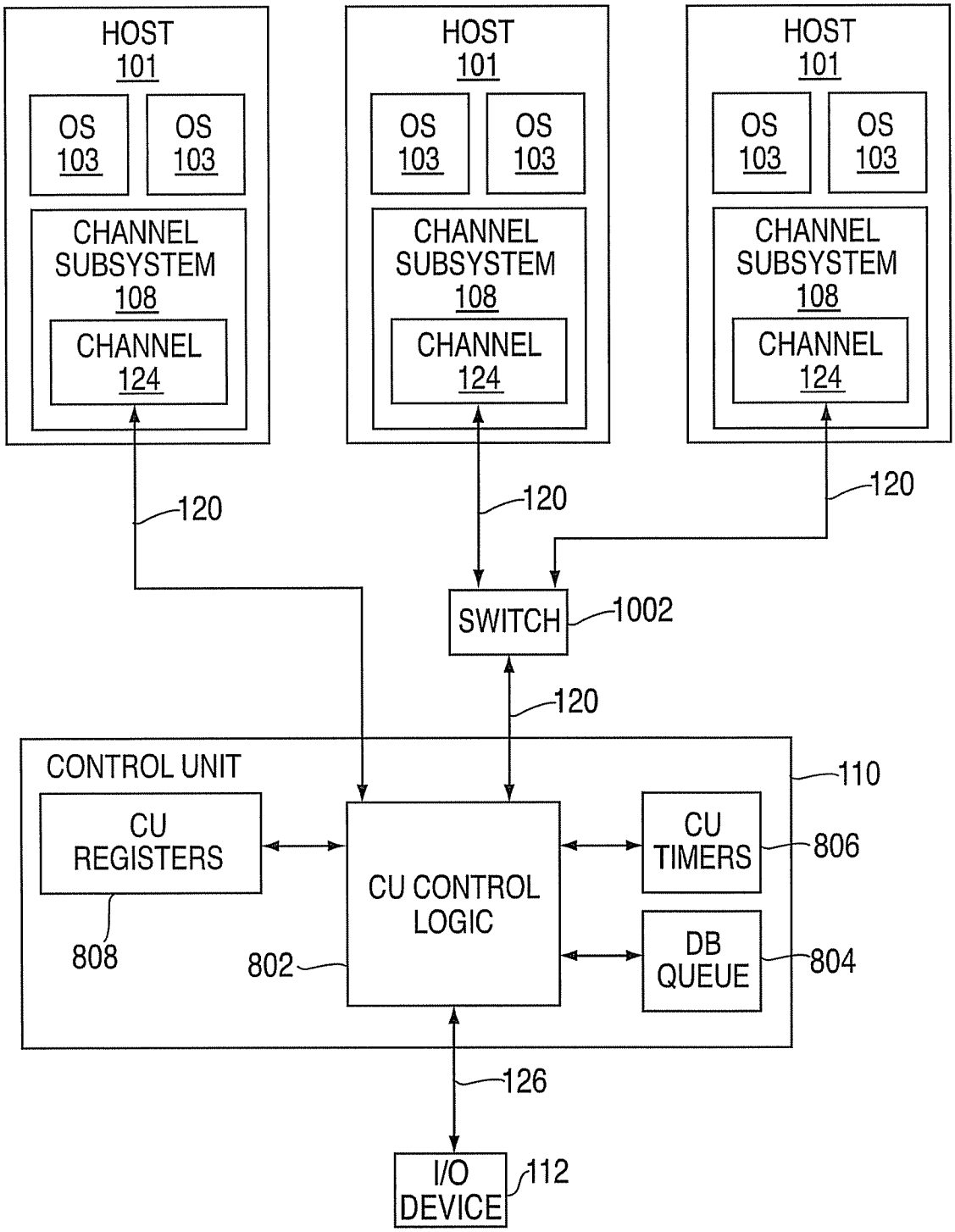
HOST
101
OS
103
OS
103
CHANNEL
SUBSYSTEM
108
CHANNEL
124
HOST
101
OS
103
OS
103
CHANNEL
SUBSYSTEM
108
CHANNEL
124
HOST
101
OS
103
OS
103
CHANNEL
SUBSYSTEM
108
CHANNEL
124
120
120
120
SWITCH
1002
120
CONTROL UNIT
110
CU
REGISTERS
808
CU CONTROL
LOGIC
802
CU
TIMERS
806
DB
QUEUE
804
126
I/O
DEVICE
112

FIG. 10

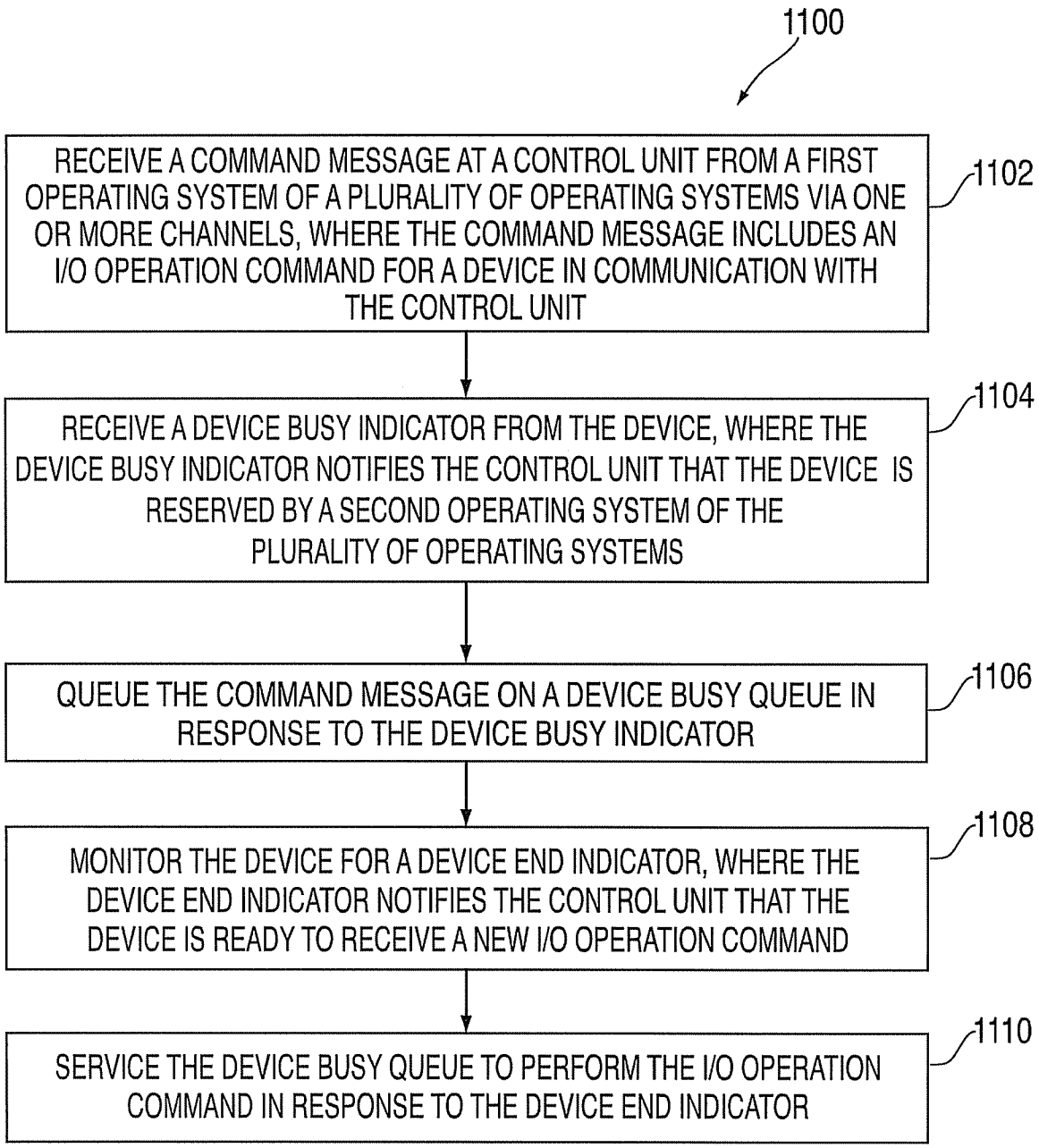
1100
RECEIVE A COMMAND MESSAGE AT A CONTROL UNIT FROM A FIRST OPERATING SYSTEM OF A PLURALITY OF OPERATING SYSTEMS VIA ONE OR MORE CHANNELS, WHERE THE COMMAND MESSAGE INCLUDES AN I/O OPERATION COMMAND FOR A DEVICE IN COMMUNICATION WITH THE CONTROL UNIT
1102
RECEIVE A DEVICE BUSY INDICATOR FROM THE DEVICE, WHERE THE DEVICE BUSY INDICATOR NOTIFIES THE CONTROL UNIT THAT THE DEVICE IS RESERVED BY A SECOND OPERATING SYSTEM OF THE PLURALITY OF OPERATING SYSTEMS
1104
QUEUE THE COMMAND MESSAGE ON A DEVICE BUSY QUEUE IN RESPONSE TO THE DEVICE BUSY INDICATOR
1106
MONITOR THE DEVICE FOR A DEVICE END INDICATOR, WHERE THE DEVICE END INDICATOR NOTIFIES THE CONTROL UNIT THAT THE DEVICE IS READY TO RECEIVE A NEW I/O OPERATION COMMAND
1108
SERVICE THE DEVICE BUSY QUEUE TO PERFORM THE I/O OPERATION COMMAND IN RESPONSE TO THE DEVICE END INDICATOR
1110

FIG. 11

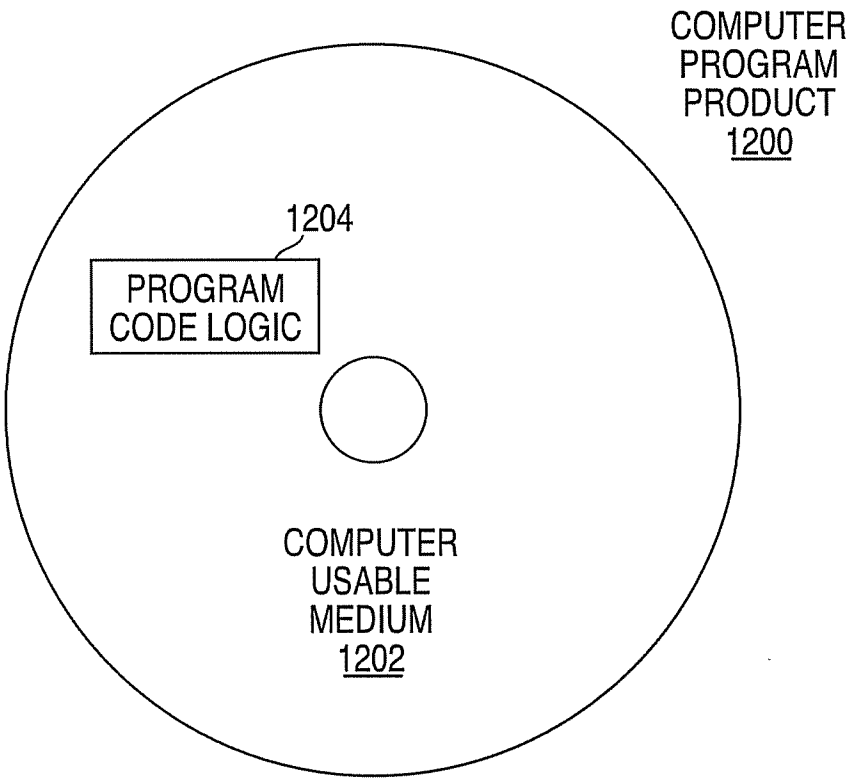
COMPUTER
PROGRAM
PRODUCT
1200
1204
PROGRAM
CODE LOGIC
COMPUTER
USABLE
MEDIUM
1202

FIG. 12

INTERNATIONAL SEARCH REPORT

International application No
PCT/EP2009/051445

A. CLASSIFICATION OF SUBJECT MATTER
INV. G06F13/14

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)
EPO-Internal, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2005/102456 A1 (KANG SHIN-WOOK [KR]) 12 May 2005 (2005-05-12) the whole document	1-6, 8-14, 16-19
A	US 6 694 390 B1 (BOGIN ZOHAR [US] ET AL) 17 February 2004 (2004-02-17) the whole document	1-20
A	JP 63 236152 A (FUJITSU LTD) 3 October 1988 (1988-10-03) abstract	1-20
A	DE 39 31 514 A1 (HITACHI LTD [JP]) 22 March 1990 (1990-03-22) the whole document	1-20

☐ Further documents are listed in the continuation of Box C.

☒ See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
5 June 2009	25/06/2009

Name and mailing address of the ISA/	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL – 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016	Rudolph, Stefan

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/EP2009/051445

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2005102456 A1	12-05-2005	KR 20050043426 A	11-05-2005
US 6694390 B1	17-02-2004	US 2004078507 A1	22-04-2004
		US 2004059839 A1	25-03-2004
JP 63236152 A	03-10-1988	NONE	
DE 3931514 A1	22-03-1990	JP 1831242 C	15-03-1994
		JP 2083757 A	23-03-1990
		JP 5044052 B	05-07-1993