

(19) World Intellectual Property Organization
International Bureau



(43) International Publication Date
18 May 2007 (18.05.2007)

PCT

(10) International Publication Number
WO 2007/056597 A1

(51) International Patent Classification:
G06F 9/52 (2006.01)

(21) International Application Number:
PCT/US2006/043889

(22) International Filing Date:
10 November 2006 (10.11.2006)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
11/272,197 10 November 2005 (10.11.2005) US
11/271,535 10 November 2005 (10.11.2005) US

(71) Applicant (for all designated States except US):
HEWLETT-PACKARD DEVELOPMENT COMPANY L.P. [US/US]; 20555 S.H. 249, Houston, TX 77070 (US).

(72) Inventors; and

(75) Inventors/Applicants (for US only): **COLLARD, Jean-Francois** [FR/US]; 1501 Page Mill Road, Palo Alto, CA

94304 (US). **JOUPPI, Norman, Paul** [US/US]; 1501 Page Mill Road, Palo Alto, CA 94304 (US). **SAMPSON, John, Morgan** [US/US]; 8540 Costa Verde Blvd. Apt #4327, San Diego, CA 95122 (US).

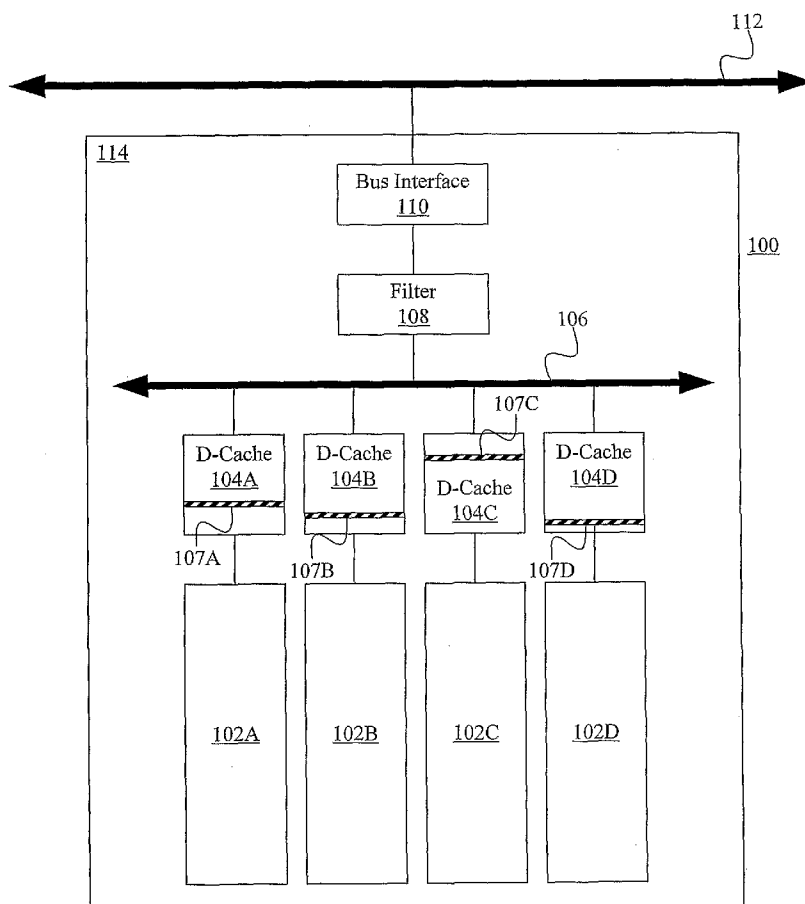
(74) Agent: **LEHMANN, Eileen, A.**; Hewlett-Packard Company, IP ADMINISTRATION, PO Box 272400, Ft. Collins, CO 80527-2400 (US).

(81) Designated States (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, 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, LV, LY, MA, MD, 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, 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,

[Continued on next page]

(54) Title: PROGRAM THREAD SYNCHRONIZATION



(57) Abstract: The present invention is a method of and system for program thread synchronization. In accordance with an embodiment of the invention, a method of synchronizing program threads for one or more processors is provided. An address (A) for data for each of a plurality of program threads to be synchronized is determined. For each processor (102A-D) executing one or more of the threads to be synchronized, execution of the thread is halted at a barrier by attempting a data operation to the determined address (A) and the address (A) being unavailable. Execution of the threads is resumed.



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, HU, IE, IS, IT, LT, LU, LV, MC, NL, 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

— before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

PROGRAM THREAD SYNCHRONIZATION

This application is related to U.S. Application Serial No. (Atty. Docket No. 200505971-1), filed (on the same day as this application), the entire contents of which are hereby incorporated by reference.

FIELD OF THE INVENTION

[0001] The present invention relates to synchronization of program threads.

BACKGROUND OF THE INVENTION

[0002] A computer system may include two or more interconnected processors that execute multiple program threads concurrently. Such a multiprocessor computing system allows a computing task to be completed more quickly by dividing the task into smaller tasks which are performed concurrently by the various processors.

[0003] A multiprocessor system may include a number of processors formed on separate computer chips which are then assembled into a multiprocessor system. Such systems are also referred to as parallel computing systems. Increasingly, computer chips are being manufactured that include multiple processors (or “cores”) on a single computer chip. Such single-chip multiprocessors tend to occupy less space and to better facilitate communication between the processors than traditional multiprocessor systems.

[0004] In either case, synchronization is typically required among the threads executed by a multiprocessor system. For example, interdependencies among the threads may require that one or more threads produce a result before one or more other threads make use of the result. Thus, it may be necessary to utilize a synchronization barrier in which each thread to be synchronized is allowed to execute until it reaches a barrier, the thread is stalled at the barrier until the other threads reach the barrier; and, then, the threads are allowed to resume execution.

[0005] Conventional synchronization techniques may require that a counter is shared by all of the threads to be synchronized. The counter is incremented by each thread that arrives at the barrier. Each thread repeatedly tests whether the counter has

reached a given value to determine whether the other threads have also reached the barrier. Because each thread contends for access to it, the counter may become a hot spot. To lessen contention for the counter, the threads may test less often; however, this may increase the time required for each thread to learn that the barrier has been reached.

[0006] Alternatively, a synchronization mechanism can be implemented by specialized hardware. However, this can complicate the design and manufacture of multiprocessor system that includes such specialized hardware and can limit flexibility of the system.

SUMMARY OF THE INVENTION

[0007] The present invention is a method of and system for program thread synchronization. In accordance with an embodiment of the invention, a method of synchronizing program threads for one or more processors is provided. An address for data for each of a plurality of program threads to be synchronized is determined. For each processor executing one or more of the threads to be synchronized, execution of the thread is halted at a barrier by attempting a data operation to the determined address and the address being unavailable. Execution of the threads is resumed.

[0008] In accordance with another embodiment of the invention, a system is provided for synchronizing program threads for one or more processors. One or more processors execute a plurality of threads to be synchronized. Execution of the threads is halted at a barrier by a corresponding determined data cache line being rendered unavailable. Execution of the threads resumes upon each of the plurality of threads by rendering the corresponding determined data cache line available. A filter is coupled to the plurality of processors for monitoring and selectively blocking fill requests for the data cache lines.

[0009] These and other embodiments are described in more detail herein.

BRIEF DESCRIPTION OF THE DRAWINGS

[0010] Figure 1 illustrates a multiprocessor system in accordance with an embodiment of the present invention;

[0011] Figure 2 illustrates a flow diagram for a barrier procedure in accordance with an embodiment of the present invention;

[0012] Figure 3 illustrates a table for a barrier filter in accordance with an embodiment of the present invention;

[0013] Figure 4 illustrates a state diagram for operation of a barrier filter in accordance with an embodiment of the present invention;

[0014] Figure 5 illustrates a flow diagram of an alternative barrier procedure in accordance with an embodiment of the present invention; and

[0015] Figure 6 illustrates an address structure in accordance with an embodiment of the present invention; and

[0016] Figure 7 shows an alternative embodiment of a multi-processor system in which functions of a barrier filter are distributed in accordance with an embodiment of the present invention.

DETAILED DESCRIPTION OF THE INVENTION

[0017] The present invention provides a method of and system for synchronizing program threads executing on one or more computer processors. Execution of multiple threads is synchronized by stalling execution of each of the program threads at a barrier until all of the threads to be synchronized reach the barrier. Each thread is stalled by invalidating the cache line for requested data in data cache (D-cache) for the processor executing the thread. A memory fence operation prevents speculative execution by the processor which might otherwise compromise the barrier.

[0018] To continue stalling each thread after the cache lines are invalidated, a barrier filter monitors D-cache fill requests and blocks the provision of selected cache lines to the processors. Once all of the threads to be synchronized reach the barrier,

the invalidated cache lines are validated and fill requests for the cache lines are allowed. As a result, the threads resume execution. Execution can be resumed quickly because reading the validated data cache line can be accomplished quickly, typically in one clock cycle.

[0019] Figure 1 illustrates a multiprocessor system 100 in accordance with an embodiment of the present invention. The system 100 includes multiple processors 102A-D. The processors 102A-D are coupled to corresponding data caches (D-caches) 104A-D and to an interconnect 106 via the D-caches 104A-D. A barrier filter 108 is also coupled to the interconnect 106. The interconnect 106 may comprise a local bus.

[0020] An interface 110 interfaces the processors 102A-D, D-caches 104A-D and filter 108 to a system bus 112. The system bus 112 allows the processors 102A-D to communicate with additional portions of the system 100, such as program memory, one or more outer cache levels and I/O devices (not shown). In an embodiment, the processors 102A-D, D-caches 104A-D, filter 108 and interface 110 are included in a multi-core integrated circuit chip 114. Alternatively, processors of the system 100 may be implemented by multiple integrated circuit chips. For example, the processors 102A-D may each be implemented as a separate integrated circuit chip.

[0021] Each of the processors 102A-D receives data values to be operated upon by the processor from its respective D-cache 104A-D. For example, the processor 102A receives its data values from the D-cache 104A; the processor 102B receives its data values from the D-cache 104B; and so forth. While four processors 102 are shown, it will be apparent that a multiprocessor system could have more or fewer processors.

[0022] Each D-cache 104A-D may have an associated D-cache controller (not shown). The D-cache controllers load data values into the D-caches 104A-D by retrieving the data values from program memory or an outer cache level (e.g., level-2 cache). As shown in Figure 1, the D-caches 104A-D receive the data values via the interconnect 106.

[0023] When threads are to be synchronized, each is stalled at a barrier while it awaits the arrival of the other threads to the barrier. As mentioned, to stall a thread executing on a processor, the data cache line for data to be operated upon by the processor is invalidated. The striped lines 107A-D in the D-caches 104 of Figure 1 represent invalidated cache lines. Because invalidating a cache line will generally result in a cache miss when the processor 102 attempts to retrieve a data value from the invalidated cache line, a fill request will be issued in an attempt to retrieve the data. The D-cache controller for that processor may issue such a fill request. To continue stalling the thread, such fill requests are blocked.

[0024] The filter 108 monitors, and selectively blocks, D-cache fill requests. This may be accomplished by the filter 108 monitoring the interconnect 106 for specified addresses contained in the fill requests. When a fill request is identified by the filter 108 as one that is to be blocked, the filter 108 blocks that request. This may be accomplished by the filter 108 preventing the D-cache from loading the identified cache lines or by failing to pass the identified fill request to level-2 cache.

[0025] In an embodiment, the filter 108 is informed of the presence of a barrier in advance of any of the threads reaching the barrier. Upon reaching the barrier, each of the threads to be synchronized may invalidate its own cache line. The filter 108 blocks fill requests for the invalidated cache lines until all of the threads reach the barrier. In another embodiment, the first thread that arrives at a barrier initiates synchronization by invalidating the appropriate cache lines for all of the threads to be synchronized. This thread may also notify the filter 108 to commence blocking fill requests for the invalidated cache lines.

[0026] The locations of barriers in each thread to be synchronized are determined in order to synchronize the threads. The locations will depend upon tasks to be performed by the various processors and interdependencies of the tasks. Accordingly, the locations of barriers may be determined when the threads are programmed or compiled. In an embodiment in which threads are capable of invalidating their own cache line, a barrier procedure call is placed in each thread to be synchronized at one or more appropriate synchronization points. The barrier filter 108 is informed of the presence of these barriers.

[0027] Figure 2 illustrates a flow diagram for a barrier procedure 200 in accordance with an embodiment of the present invention. As mentioned, a call to the barrier procedure 200 is placed at an appropriate synchronization location within each thread to be synchronized. The barrier procedure 200 may then be executed by each thread executing corresponding instructions. A thread may contain several calls to the barrier procedure 200, but need have access to only a single copy of the procedure

[0028] In a step 202 of the barrier procedure 200, a determined data cache line is invalidated. The determined data cache line is identified by its address, which may be given as address A. This invalidate instruction may be, for example, a data cache block invalidate instruction (DCBI), which is shown in Figure 2, a flush-cache (fc) instruction or another appropriate instruction. In a step 204, an attempt is made to load the data value at the address A into a specified register R. Because the data cache line at address A has been invalidated, execution of an instruction corresponding to the step 204 can be expected to result in a data cache miss.

[0029] The invalidate instruction issued in step 202 is preferably propagated throughout the cache hierarchy and eventually to the filter 108. This purges any copies of the invalidated cache line from cache levels between the processor 102 and the filter 108 in order to help ensure that the thread will stall on the attempt to load data from the invalidated cache line at address A. In addition, the filter 108 observes invalidation instructions issued by the threads, which prompts the filter 108 to block cache fill requests for the invalidated data cache line. Thus, upon observing the invalidation instruction for the address A, the filter commences to block fill requests specifying the address A. This helps to ensure that the processor will continue to stall at the load instruction attempted in step 204.

[0030] In step 206, a memory fence is implemented which may be accomplished by the processor executing a memory fence instruction. The memory fence prevents the processor from speculatively performing any subsequent memory accesses for the thread until all prior memory accesses have been completed. A processor may speculate instruction execution, that is, it may guess what action should be taken in advance of knowing what action is to be taken. For example, a processor may take a

conditional branch before the condition is known to the processor. As another example, a processor may guess the value of a register before the value is known. Such a processor has hardware to detect actions that were incorrectly performed and to undo those actions. However, if the speculation was correct, this execution will generally not be undone. If a processor speculatively executes an instruction that occurs after the barrier, this may compromise the barrier. For example, if a processor executing a thread speculatively loads a variable in accordance with an instruction that occurs after the barrier, but the correct value has not yet been written to the variable by another thread, the processor hardware will not detect this. Thus, the memory fence prevents the processor which executes the thread from speculatively performing any memory accesses until the load instruction attempted in step 204 has completed. This helps to ensure that memory instructions which occur after the barrier are not prematurely performed through speculation while the thread is stalled at the barrier.

[0031] The memory fence will generally not prevent the processor from speculatively performing operations that appear in the thread after the barrier and that do not require a memory access. For example, instructions that manipulate variables that are already present in processor registers may be speculatively performed. Because no memory access is performed, these operations can be performed without compromising the barrier.

[0032] Once all of the threads to be synchronized have reached the barrier, the filter 108 ceases blocking fill requests for the invalidated cache lines. Accordingly, the thread is allowed to resume execution and the instruction attempted in step 204 is allowed to complete. At this point, the value at the address A is loaded to register R. Thus, the load operation has served its purpose and the value loaded to register R may not be needed. Also, the condition of the memory fence that all memory accesses prior to the fence are complete is satisfied. Thus, memory accesses occurring after the memory fence are no longer prevented by the memory fence.

[0033] In a step 208, the thread informs the filter 108 that it has exited the barrier. This is so that the filter 108 can again block any future fill requests for address A coming from this thread. Blocking future fill requests avoids one thread running

ahead of the other threads such that it reaches and surpasses another barrier (or another instance of the same static call to the barrier procedure) before the other threads have exited the current barrier. This may be accomplished by the thread invalidating another, predetermined D-cache line that contains dead data (i.e. unaccessed data). The address of this line may be given as E. The identity of the processor 102 (i.e. the processor ID) making this invalidation is carried with the request, which enables the filter 108 to identify which thread it should stop servicing requests from. It will be apparent, however, that informing the filter 108 that a thread is exiting the barrier may be accomplished in other ways. For example, some other instruction may be issued in step 208 that is recognized by the filter 108. This instruction should be a type that is prevented by the memory fence from occurring prematurely (before the thread has exited the barrier).

[0034] In an embodiment, a step is added to the procedure 200 which causes the processor 102 to save its internal state for a sufficient period to allow the other threads time to reach the barrier. This allows the processor to resume execution of the thread at the next instruction following the barrier. Otherwise, if the processor stalls at the barrier for too long, it may trap and execute an error handling procedure, which could interfere with its execution of the thread. This step may be performed between the steps 206 and 208. Whether this additional step is included may depend upon the specific internal processes of the processor.

[0035] To cause the processor 102 to save its internal state, one or more instructions may be executed which cause the processor to check the value of the register R and to exit the thread conditioned on the value. For example, the value of R may be desired to be less than or equal to zero; the instruction exits the thread upon determining that the value is greater than zero. If the condition is met, the thread does not exit. If the condition is not met, this causes the processor to save its internal state so that the other threads have more time to reach the barrier.

[0036] Rather than the processor exiting the thread upon a condition that is not met as described above, a time-out may alternatively cause the processor to return to a prior step of the procedure 200. For example, one or more instructions may be inserted after the step 204 which cause the processor to return to step 204 in the event

a time-out occurs. For example, depending upon the internal processes of the processor 102, it may eventually encounter a time-out error upon stalling the thread at the barrier. When this occurs, the processor returns to step 204 and again attempts to load the value at address A. In this case, the processor may repeatedly attempt the load instruction for each occurrence of a time-out until all the other threads to be synchronized reach the barrier. This also has the effect of causing the processor to save its internal state so that execution reliably resumes at a next instruction after the barrier.

[0037] Figure 3 illustrates a table 300 for a barrier filter 108 in accordance with an embodiment of the present invention. The table 300 may be used by the filter 108 for keeping track of the status of various barriers. For example, the table 300 may be stored in level-2 cache, in hardware that implements cache control, or in hardware that implements control of the filter 108. An entry in the table 300 may be made for each barrier; as shown in Figure 3, information for each barrier is arranged in a row. The barriers may be identified by number, such as 1 through n as shown in Figure 3. For each barrier, the table 300 may include a field 302 for the address A, a field 304 for the address E, a counter C field 306, a pending vector field 308, an exited vector field 310 and a valid bit field 314. The pending vector and the exited vector may include one bit corresponding to each thread to be synchronized. Thus, where N threads are to be synchronized, these vectors each have N bits. Each application may only require one barrier which is used repeatedly by the application and, thus, the table 300 may have only a single entry.

[0038] The filter table 300 may be initialized at the beginning of an application. Thus, support of the operating system may be required to provide the physical addresses of the data cache lines used to implement the barrier, that is, the addresses A and E.

[0039] Figure 4 illustrates a state diagram for operation of a barrier filter 108 in accordance with an embodiment of the present invention. A finite state machine (or automaton) for a particular barrier has two main states: a blocking state 402 and a service state 404. When an entry is inserted into the table 300, its corresponding automaton is initially in the blocking state 402, its counter C (field 306), its pending

vector (field 308) and its exited vector (field 310) are set to zero. In addition, the valid bit (field 314) is set to indicate that the corresponding D-cache line is invalid.

[0040] In the blocking state 402, the filter 108 processes each incoming cache fill request by checking to see whether the target address is blocked. This may be accomplished by the filter 108 comparing the target address for fill requests to its address A entries in column 302 of the table 300. If the target address is blocked, the filter 108 does not pass the requested cache line to the corresponding D-cache. Instead, the filter 108 may obtain the requested cache line from memory and store it in the corresponding cache line content field 312 of the table 300. The filter 108 may then set the bit in the pending vector (field 308) that corresponds to the requesting processor 102 to indicate that the processor 102 has reached the barrier and has a pending request. Otherwise, if the target address is not blocked, the filter 108 passes the requested cache line to the corresponding processor 102.

[0041] In addition to monitoring fill requests in the blocking state 402, the filter 108 monitors invalidate instructions issued by the processors 102. For each invalidate that targets address A, the filter 108 increments the corresponding counter C (field 306) in the table 300. The filter 108 remains in the blocking state until the counter C reaches N, the number of threads to be synchronized. Once the counter C reaches this number, this indicates that all of the threads to be synchronized have reached the barrier. At this point, the automaton for the barrier moves to the service state 404 and resets the valid bit (field 312) to indicate that the cache line content is valid. Thus, the filter 108 monitors invalidate requests to determine when each thread has reached the barrier. It will be apparent that the filter 108 could monitor other messages from the processors to determine when each thread has reached the barrier. For example, the filter 108 could monitor stores to specific locations.

[0042] In the service state 404, the pending data cache fill requests are serviced. New fill requests arriving while in the state 404 are serviced as well. As in the blocking state 402, the filter 108 processes each incoming cache fill request by checking to see whether the target address is blocked. However, in the service state for a particular barrier, the corresponding address A is not blocked. Thus, those fill requests are serviced. In the service state 404 the filter 108 monitors invalidate

instructions targeting the address E for the barrier. The filter 108 identifies which processor issues each invalidate targeting the address E and sets corresponding bit in the exited vector (field 310) to indicate which threads have exited the barrier. Once the bit in the exited vector corresponding to a processor 102 is set, the filter 108 will not service any further fill requests from that processor 102 that target the address A for the barrier.

[0043] When setting the corresponding bit in the exited vector (field 310) would cause the vector 310 to have all of its bit equal to one, this indicates that all of the N threads have been synchronized and have exited the barrier. Accordingly, the automaton for the barrier returns to the blocking state 402. In addition, the counter C (field 306), the pending vector (field 308) and the exited vector (field 310) are cleared. The cache line content (field 312) need not be cleared since it may be reused. This process of selectively blocking fill requests (in the blocking state 402) and then servicing the requests (in the service state 404) is repeated for each barrier instance encountered as the threads are executed.

[0044] In another embodiment, the first thread that arrives at a barrier initiates synchronization by causing the appropriate data cache lines to be invalidated for all of the threads to be synchronized. This thread also notifies the filter 108 to commence blocking fill requests for the invalidated cache lines. Threads that reach the barrier subsequent to the first thread already have their cache lines invalidated so that when they attempt to read this cache line, they stall at the barrier. Preferably in this embodiment, threads that reach the barrier after the first thread do not attempt to invalidate the cache lines of the other threads.

[0045] So that a thread can determine whether it is the first thread to reach the barrier, each thread may request exclusive access to a memory location in the system 100 (Figure 1) when it reaches the barrier. For example, each thread may acquire a lock from the OS of the system 100 so that the thread has exclusive access to a value stored at the memory location. The thread then checks the value and, from the value, the thread infers whether it is the first thread to reach the barrier. For example, a value of zero may indicate that the thread is the first thread. If the thread is the first thread, it also changes the value so that when the other threads check the value, they

will infer from the value that they are not the first thread. For example, a value of one may indicate that the first thread has already arrived at the barrier.

[0046] For this embodiment, the barrier procedure may be modified from that shown in Figure 2 so that each thread causes the designated memory location to be checked to determine whether a thread is the first thread to reach the barrier and, if so, to invalidate the cache lines of the other threads. These invalidates must also invalidate any prefetch buffers between the other threads' processor data fetch mechanism and the filter as well as their L1 caches. The barrier procedure may also include code to cause the first thread to insert the current barrier into the table 300 of the filter 108, if there is not an entry corresponding to the thread already in the table 300. In this embodiment, once the filter table 300 has an entry for the barrier and the filter 108 has been informed that the first thread has reached the barrier, operation of the filter 108 may commence in the blocking state 402 as explained above in connection with the blocking state 402. Once all of the threads have reached the barrier, operation of the filter 108 may commence in the service state 404, as explained above, and may then return to the blocking state 402.

[0047] In an embodiment of the invention, the OS of the system 100 provides support for implementing the barriers. In this case, the OS accesses the filter 108 so as to save and to restore the state of the filter 108, possibly including the table 300, in the event of context changes, such as when an application is de-scheduled or a page holding program text is swapped out of memory. In case the OS does not provide such support, compensating actions may be performed elsewhere, such as by the filter 108 itself. For example, the filter 108 may monitor the OS to identify context changes by associating process identifications of the threads to contexts. The filter 108 may then associate the process identifications with the barriers in its table 300. In this way, the filter 108 may invalidate cache lines and block fill requests for each processor according to the process identification for the thread currently executing on the processor. So that the filter 108 is informed of the process identification for each thread, the invalidate instructions of the barrier procedure (e.g., procedure 300 of Figure 3) preferably carry the corresponding process identification.

[0048] In accordance with embodiments explained above, each thread issues an invalidate instruction to invalidate the cache line at address A in step 202 of the barrier procedure. In an alternative embodiment, this step is omitted and, instead, the address A is determined for an uncachable address. An uncachable address A designates an address whose value cannot be loaded into cache and is, therefore, unavailable. For example, the address A may be selected to be within a range of addresses in main memory which cannot be loaded into cache memory. In this case, in step 204, an attempt to load this uncachable address into the data cache is made which causes the thread to stall. The filter monitors load requests for any directed to the address A. When the filter observes such a load request, this notifies the filter 108 that the thread has reached the barrier. Upon all of the threads reaching the barrier, the filter 108, responds to the load requests by providing a value, which causes the threads to resume execution. The provided value may be used only for the purpose of implementing the barrier and, thus, could be any value. The barrier procedure may otherwise be implemented as explained above.

[0049] In accordance with embodiments explained above, a load operation is attempted in step 204. Alternatively, this instruction may instead be replaced with a store operation to the invalidated data cache address A when used with a level 1 data cache that allocates lines before processing write operations (i.e., a write-allocate cache). A write to a write-allocate cache will stall until the corresponding cache line is present in the cache. This behavior is similar to the embodiment above using an invalidated cache line and a load instruction. In step 202, in addition to invalidating the data cache line for address A (including flushing any dirty data it contains to lower levels of the memory hierarchy), an appropriate line or line segment in a write buffer for the processor may also be flushed to lower levels in the memory hierarchy. Alternatively, the filter 108 may flush the write buffer in response to observing the invalidate instruction to address A. The store operation attempted in step 204 will be unsuccessful so long as the appropriate lines in data cache and in the write buffer are not present. When all of the threads have reached the barrier, execution of the threads is allowed to resume by allowing the store operation to be completed. This may be accomplished by the filter 108 ceasing to block fill requests for the invalidated data cache line or by allowing the line or a segment thereof to become valid in the write buffer. The barrier procedure may otherwise be implemented as explained above.

[0050] In accordance with embodiments explained above, in order to prevent a thread from passing through a next barrier before the other threads to be synchronized have reached the current barrier, the filter 108 determines that a thread has exited the current barrier and resumes blocking fill requests for the determined cache line for the thread after it has exited the barrier. In this case, the address A may be remain fixed for each instance of the barrier. Alternatively, two or more barrier instances may be distinguished by using a different address A for each barrier. So that the same barrier procedure can be called by the threads, the address A is preferably computed within the barrier procedure. In other words, rather than being fixed, the address A is changed dynamically.

[0051] Figure 5 illustrates a flow diagram of an alternative barrier procedure 500 in accordance with an embodiment of the present invention. In step 502, a toggle variable is loaded from a specified location in memory. For example, the toggle variable may take on two allowed values, such as zero and D. The toggle variable may be initialized to one of these allowed values prior to the first time the barrier procedure is called. In step 504, the toggle variable is changed to other allowed value. For example, if the toggle variable was initially zero, it is changed to the value D in step 504. This may be accomplished in a number of ways, such as by performing an exclusive-OR of the current value of the toggle value with the value D or by performing a conditional load depending upon the current value of the toggle variable.

[0052] In step 506, the address A is determined from the value of the toggle variable. The address A may take on one of two values, which may be given as A_a or A_b , depending upon the value of the toggle variable. For example, the address A may be computed as a sum of a base address B and the toggle variable, such that:

$$A_a = B + \text{zero}; \text{ and}$$

$$A_b = B + D.$$

[0053] In step 508, the new value of the toggle variable is stored. This value will be different from the value loaded in step 502, such that each successive time the barrier procedure 500 is called the toggle variable is toggled between its two values,

zero and D. Also in this embodiment, the address A for a thread is toggled back and forth between the two values, A_a and A_b , for each barrier instance.

[0054] In a step 510 of the barrier procedure 500, the determined data cache line is invalidated. The determined data cache line is identified by its address, which may be given as address A_a or A_b . In a step 512, an attempt is made to load the data value at the address A_a or A_b into a specified register R. Because the data cache line at address A_a or A_b has been invalidated, execution of an instruction corresponding to the step 510 can be expected to result in a data cache miss. As before, the invalidate instruction issued in step 510 is preferably propagated throughout the cache hierarchy and eventually to the filter 108. The filter 108 observes invalidation instructions issued by the threads, which prompts the filter 108 to block cache fill requests for the invalidated data cache line, causing the thread to stall.

[0055] In step 512, a memory fence is implemented which prevents the processor from speculatively performing any subsequent memory accesses for the thread while it is stalled.

[0056] Once all of the threads to be synchronized have reached the barrier, the filter 108 ceases blocking fill requests for the invalidated cache lines. Accordingly, the thread is allowed to resume execution and the instruction attempted in step 512 is allowed to complete. Also, memory accesses occurring after the memory fence are no longer prevented by the memory fence.

[0057] Because the filter 108 is able to distinguish between the current barrier and a next barrier depending upon whether the thread issues an invalidate instruction to the address A_a or to the address A_b , there is no need for each thread to inform the filter 108 that it is exiting the current barrier (as in step 208 of Figure 2). Accordingly, in this embodiment, the thread preferably does not issue an invalidate request specifying address E. As described above, the address A is toggled between two values; it will be apparent that it may be toggled among more than two values.

[0058] In the embodiment of Figure 5, a step may be added to the procedure 500 which causes the processor 102 to save its internal state for a sufficient period to

allow the other threads time to reach the barrier. This allows the processor to resume execution of the thread at the next instruction following the barrier.

Otherwise, if the processor stalls at the barrier for too long, it may execute an error handling procedure, which could interfere with its execution of the thread. This step may be performed after the step 514. Whether this additional step is included may depend upon the specific internal processes of the processor.

[0059] To cause the processor 102 to save its internal state, the processor may execute one or more instructions which would cause it to exit the thread upon a condition that is not met, as described above in connection with Figure 2.

Alternatively, a time-out may alternatively cause the processor to return to a prior step of the procedure 500, as was also explained above in connection with Figure 2. For example, one or more instructions may be inserted after the step 514 which cause the processor to return to step 510 in the event a time-out occurs.

[0060] In accordance with embodiments described above, it is assumed that a barrier is used to synchronize one thread per processor. Thus, there are N threads to be synchronized, one for each of N processors. In this case, the pending vector 308 and the exited vector 310 may each have N bits. Also, the counter C (field 306) indicates that all of the threads have reached the barrier when it reaches a count of N. Having one thread per processor also allows the barrier filter 108 to use the processor ID to identify threads that enter or exit a barrier. Thus, the processor ID may be carried with the invalidate request issued by threads exiting a barrier to enable the filter 108 to identify which thread it should stop servicing requests from. Accordingly, allowing one thread per processor simplifies the thread tracking performed by the filter 108. In other embodiments, there may be more or fewer threads to be synchronized than processors.

[0061] In an embodiment in which there are fewer threads to be synchronized than processors, a bit vector which indicates the threads to be synchronized for each barrier may be added to each barrier entry in the filter table 300. This vector may be referred to a participation vector. For example, where there are eight processors, but for a particular barrier, only the first five and the last one of the processors execute threads to be synchronized, the participation vector may be given as 11110001. Thus, a

column may be added to the table 300 in which the participation vector entered may be different for different barriers in the table 300 so as to indicate which threads and corresponding processors participate in each barrier.

[0062] A number of bits set in the participation vector indicates the number of threads that are to participate in the corresponding barrier. This number may be used by the filter 108 to determine the count required to be reached by the counter C (field 306) to indicate that the required number of threads have arrived at the barrier. Thus, when the counter C for a barrier reaches the particular number of threads participating in the barrier, this indicates that all of the threads to be synchronized have reached the barrier. Also, the participation vector indicates which bits in the exited vector (field 310) are needed to indicate that all of the threads have exited the barrier. Thus, the filter 108 may monitor the bits of the exited vector which have been set to one, and when a thread exiting the barrier would cause the exited vector to match the participation vector, this indicates to the filter 108 that all of the threads being synchronized have exited the barrier.

[0063] In another embodiment, one or more of the processors may execute multiple threads that to be synchronized. Thus, when the first of the threads of such a processor reaches the barrier, that thread is stalled; however, the processor continues to execute each additional thread until it also reaches the barrier. To accomplish this, invalidated D-cache lines are uniquely associated with each of the threads. Thus, each thread stalls on an invalidated D-cache line that is unique to the thread. Upon exiting the barrier, each thread may invalidate a unique cache line that contains dead code to identify itself to the filter 108 and to indicate that it is exiting the barrier.

[0064] Request coalescing might otherwise occur if two or more threads executing on the same processor attempted to invalidate the same D-cache line address. This is because the system 100 may perceive all but one of the requests as superfluous and may drop them or merge the requests. Thus, using distinct cache lines for the threads defeats request coalescing and enables the filter 108 to identify the individual threads even if executing on the same processor.

[0065] So that each thread executing on the same processor invalidates a unique D-cache line, the threads may each execute a unique version of the barrier procedure (e.g., procedure 200 of Figure 2). For example, the program code may include as many versions of the barrier procedure as there are threads to be synchronized. Each version of the barrier procedure specifies a unique address A (to be invalidated for stalling the thread) and a unique address E (to be invalidated to inform the filter 108 when the thread has exited the barrier). When a thread makes a call to the barrier procedure, it jumps to its private version of the barrier procedure. In this case, an appropriate barrier procedure call may be inserted into each thread. Alternatively, another procedure (e.g., a trampoline code) may be called that determines the appropriate barrier procedure to call based on the identification of the thread (e.g., the process ID). In this case, the process ID is passed to this procedure.

[0066] Alternatively, rather than each thread executing a unique version of the barrier procedure, the barrier procedure may be modified to compute unique addresses A and E for each thread. For example, the addresses A and E may be computed as a function of each thread's process ID. In this case, the process ID is passed to the barrier procedure.

[0067] However the addresses A and E are determined for each thread, the filter 108 monitors the invalidate requests issued by the threads. The filter 108 selectively blocks fill requests for invalidated D-cache lines based on the invalidate requests issued by the threads that target each unique address A. Because fill requests are blocked selectively for each thread based on its unique address A, a thread may be stalled from executing on a processor while another thread may continue to execute on the same processor until it reaches the barrier. Also, because each thread issues an invalidate targeting a unique address E when exiting a barrier, the filter 108 is able to determine which threads have exited the barrier even when multiple threads are executed by the same processor.

[0068] In an embodiment, the filter 108 identifies invalidate requests as belonging to a particular barrier based on the addresses A and E targeted by the request (in addition to identifying the particular thread, as explained above). This allows the filter 108 to determine when all of the threads to be synchronized have arrived at a

particular barrier and when they have exited the particular barrier. This may be accomplished by the filter 108 storing a bit pattern or mask for the addresses A and E for each barrier in the fields 302 and 304 of the table 300 rather than the addresses themselves. The pattern for the field 302 may be given as AP, while the pattern for field 304 may be given as EP. Thus, given a predetermined mask M, an invalidate request targeting address X is identified as belonging to the barrier if $M \text{ AND } X$ equals AP, where "AND" is a logical AND operation.

[0069] The mask M may be selected such that higher order bits of X specify the particular barrier and the lower order bits specify the particular thread. Assume that addresses can be encoded on 8 bits, and that the size of cache lines is four bytes. Therefore, the first four addresses 00000000 through 00000011 are contained in the first cache line, and 00000100 is the address of the beginning of the second cache line. In general, the beginning address of a cache line has its two lowest-order (rightmost) bits equal to 0. We can also decide that cache lines participating in the same barrier have the same high-order bits. For example, an 8-bit address can have a structure as shown in Figure 6, where x, y and z represent a bit. Thus, the first four bits are represented as xxxx identify a specific barrier, the next two bits are represented as yy and identify a specific thread and the last two bits are represented as zz and specify offset within a cache line. In the example, cache lines are four bytes, so zz is 00.

[0070] A valid mask M may be 11110000. Now assume the addresses A targeted by two different threads are: $A_1=11010000$ and $A_2=11010100$. In this case, $M \text{ AND } A_1$ yields 11010000, as does $M \text{ AND } A_2$. Thus, the first four bits of the addresses A_1 and A_2 , which are 1101, are the same and, thus, specify the same barrier. In this case, the value AP of field 302 for this particular barrier is 1101, with the lower order bits being "don't cares." The next two bits of the targeted address specify the particular thread. Thus, the thread that issued the invalidate request for address A_1 or A_2 can be distinguished by the filter 108 examining bits five and six (bits "yy"). For address A_1 these bits are 00 and for address A_2 , these bits are 01. Bits 7 and 8 can be ignored by the filter 108.

[0071] Invalidate requests for address E may be handled similarly. For example, the value EP for field 304 for this same barrier may be given as 1100. This value differs from the value AP so that the filter 108 is able to distinguish between invalidates that target the address A and those that target the address E. However, bits five and six may still be used to identify the particular thread. Thus, in the example, an invalidate request targeting address 11000100 can be identified as an E invalidate issued by thread 01 for the same barrier as in the example above. It will be apparent that different bits of the targeted addresses A and E may be used to identify the particular barrier and the particular thread.

[0072] Figure 7 shows an alternative embodiment of a multi-processor system 700 in which functions of a barrier filter 108 are distributed in accordance with an embodiment of the present invention. Similarly to the system 100 of Figure 1, the system 700 of Figure 7 includes multiple processors 102A-D which are coupled to corresponding data caches (D-caches) 104A-D and to an interconnect 106 via the D-caches 104A-D.

[0073] The system 700 of Figure 7 differs from that of Figure 1 in that the filter 108 comprises distributed filter elements 108A-D. Each distributed filter element 108A-D is associated with a corresponding one of the processors 102A-D and a corresponding one of the D-caches 104A-D. The filter elements 108A-D are coupled to the interconnect 106 and to level-2 cache memories 116A-D. In a preferred embodiment, each filter element 108A-D is incorporated into a level-2 cache controller which controls the corresponding level-2 cache 116A-D. In addition, each filter element 108A-D may be identical to the others, except that only one filter element 108A-D performs all of functions for implementing a particular barrier. Thus, each element 108A-D may have a different table 300 which includes only those barriers assigned to it.

[0074] As shown in Figure 7, the filter elements 108A-D may be outside the communication paths between the interconnect 106 and the level-2 caches 116A-D. Accordingly, the filter elements 108A-D are outside the critical communication path for incoming requests and outbound responses between the processors and memory; this avoids the filter 108 causing delay in such communications. However, the filter

elements 108A-D monitor these communications so that they can invalidate operations and fill requests and can selectively block fill requests for any of the processors 102A-D.

[0075] The system 700 may also include memory controllers 118A-D which control accesses to memories 120A-D. Similarly to the system 100 of Figure 1, the processors 102A-D may be implemented as a multi-core integrated circuit chip 122. Alternatively, processors of the system 700 may be implemented by multiple integrated circuit chips.

[0076] In systems with multiple multi-core chips, several techniques for large-scale barrier synchronization are presented. In one embodiment, filter hardware (which implements filter 108) is present within each chip, as explained above, and conventional software barrier techniques are used for synchronization between chips. In this case, the barrier is broken down into two stages. Once the participating cores on one chip have reached the first stage, one core on the chip executes a software barrier with other threads each on different chips by inter-chip communications among the participating chips. Once this software barrier is complete, the threads on each chip are signaled to proceed.

[0077] In an alternative embodiment, barriers are processed hierarchically by a combination of intra- and inter-chip communication. One chip's filter is designated as a home filter for a given barrier address, in the way that memory addresses in a NUMA multiprocessor have home locations. Non-home filters participating in a barrier aggregate their arrivals at the barrier before communicating with the home filter. For example, if a chip has eight cores participating in a barrier, but is not the home node, its filter can wait until each of its 8 cores has reached the barrier (using techniques described above) before communicating to the home filter. The home filter requires state bits corresponding to the summarized status of each other processor chip in the shared memory system. This can be faster than the software method described above, but may be less scalable for systems with very large numbers of chips.

[0078] In yet another embodiment, inter-chip synchronization may be performed with conventional hardware synchronization networks, such a global AND/OR tree networks. This results in a hybrid hardware mechanism.

[0079] It will be apparent that computer-readable media may be provided having stored thereon computer code for performing any of the functions described herein. For example, the computer code may be loaded from the computer-readable media to the system 100 or 700 for causing the system 100 or 700 to operate in accordance with the teachings herein.

[0080] While the foregoing has been with reference to particular embodiments of the invention, it will be appreciated by those skilled in the art that changes in these embodiments may be made without departing from the principles and spirit of the invention, the scope of which is defined by the appended claims.

CLAIMS

What is claimed is:

1. A method of synchronizing program threads for one or more processors comprising:
 - determining an address (A) for data for each of a plurality of program threads to be synchronized;
 - for each processor (102A-D) executing one or more of the threads to be synchronized, halting execution of the thread at a barrier for a determined address (A) that is unavailable by attempting a data operation to the determined address (A); and
 - resuming execution of the threads.
2. The method according to claim 1, wherein the determined address (A) is rendered unavailable for each thread by invalidating a data cache line (107A-D) for each of the threads.
3. The method according to claim 2, wherein said rendering the data cache line (107A-D) unavailable further comprises inhibiting a fill request for the cache line.
4. The method according to claim 3, wherein said inhibiting is performed by a filter (108) that monitors cache fill requests and selectively blocks the fill request for the cache line (107A-D).
5. The method according to claim 4, further comprising implementing a memory fence for each thread before resuming execution.
6. The method according to claim 4, wherein the filter monitors invalidate requests issued by the threads to determine when each thread has reached the barrier.

7. The method according to claim 2, wherein each of the plurality of the threads invalidates its corresponding cache line (107A-D).

8. The method according to claim 7, wherein each of the plurality of threads invalidates its corresponding cache line (107A-D) by executing a called procedure, the procedure comprising invalidating the corresponding cache line (202), attempting the operation on the invalidated cache line (204) and a memory fence (206).

9. The method according to claim 2, wherein a first thread of the plurality to arrive at the barrier invalidates cache lines (107A-D) for other threads of the plurality.

10. A system for synchronizing program threads for one or more processors comprising:

one or more processors (102A-D) executing a plurality of threads to be synchronized, execution of the threads being halted at a barrier by a corresponding determined data cache line (107A-D) being rendered unavailable and execution of the threads resuming by rendering the corresponding determined cache line (107A-D) available; and

a filter (108) coupled to the plurality of processors for monitoring and selectively blocking fill requests for the data cache lines.

1/7

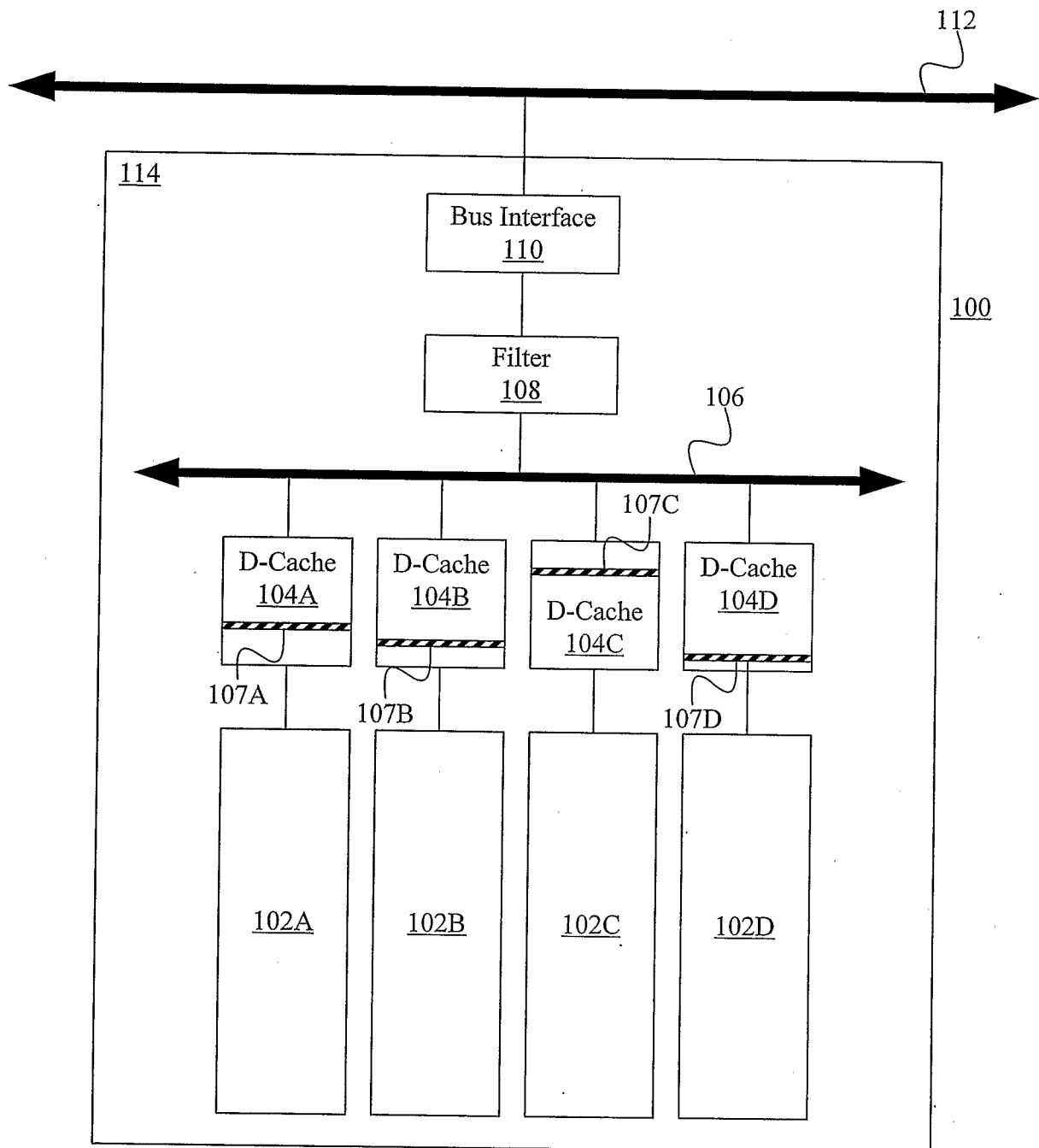


FIG. 1

2/7

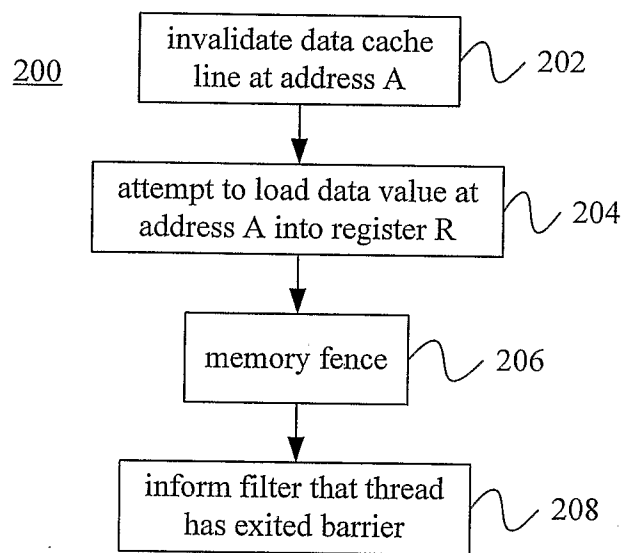


FIG. 2

Barrier #	Address A 302	Address E 304	Counter C 306	Pending Vector 308	Exited Vector 310	Cache Line Content 312	Valid Bit
1	A ₁	E ₁	C ₁	PV ₁	EV ₁	Content ₁	VB ₁
2	A ₂	E ₂	C ₂	PV ₂	EV ₂	Content ₂	VB ₂
⋮							
n	A _n	E _n	C _n	PV _n	EV _n	Content _n	VB _n

FIG. 3

300

314

4/7

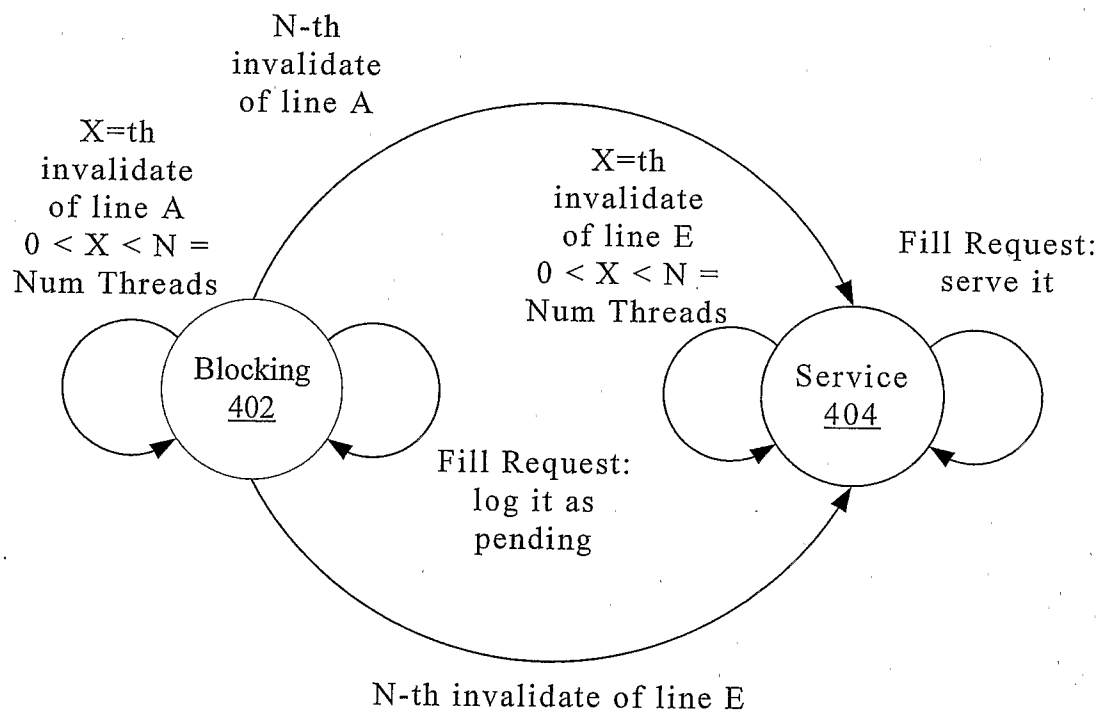


FIG. 4

5/7

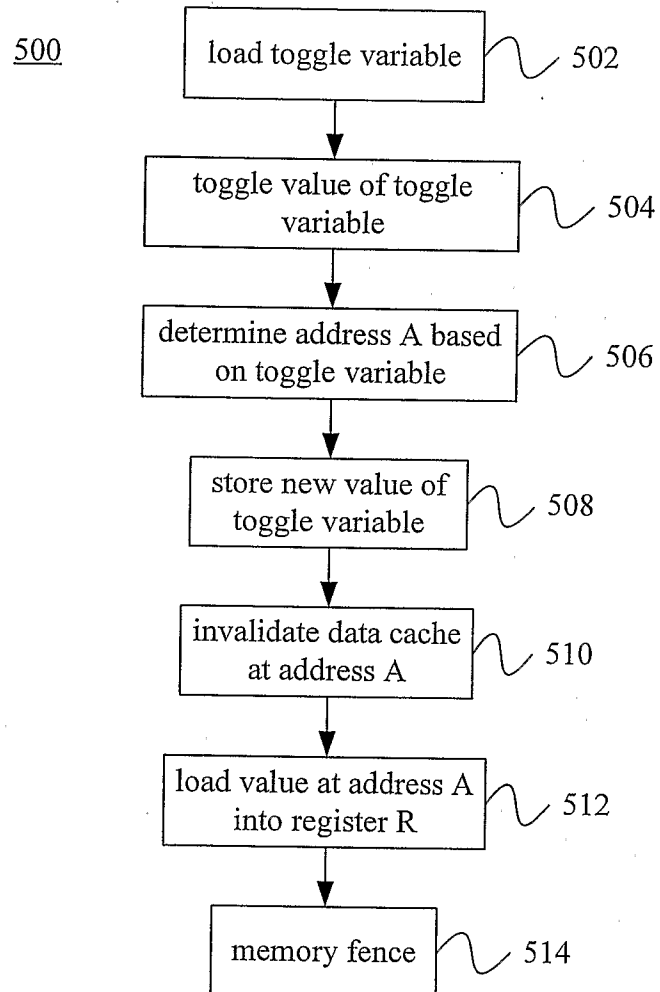


FIG. 5

6/7

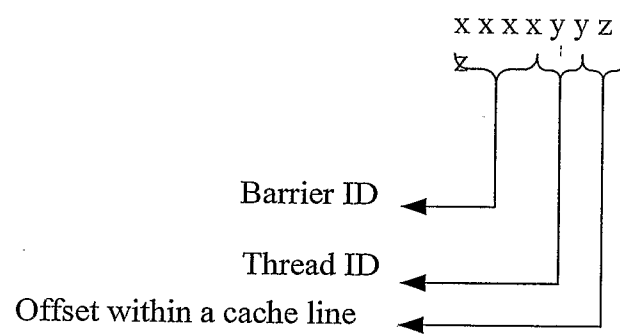


FIG. 6

7/7

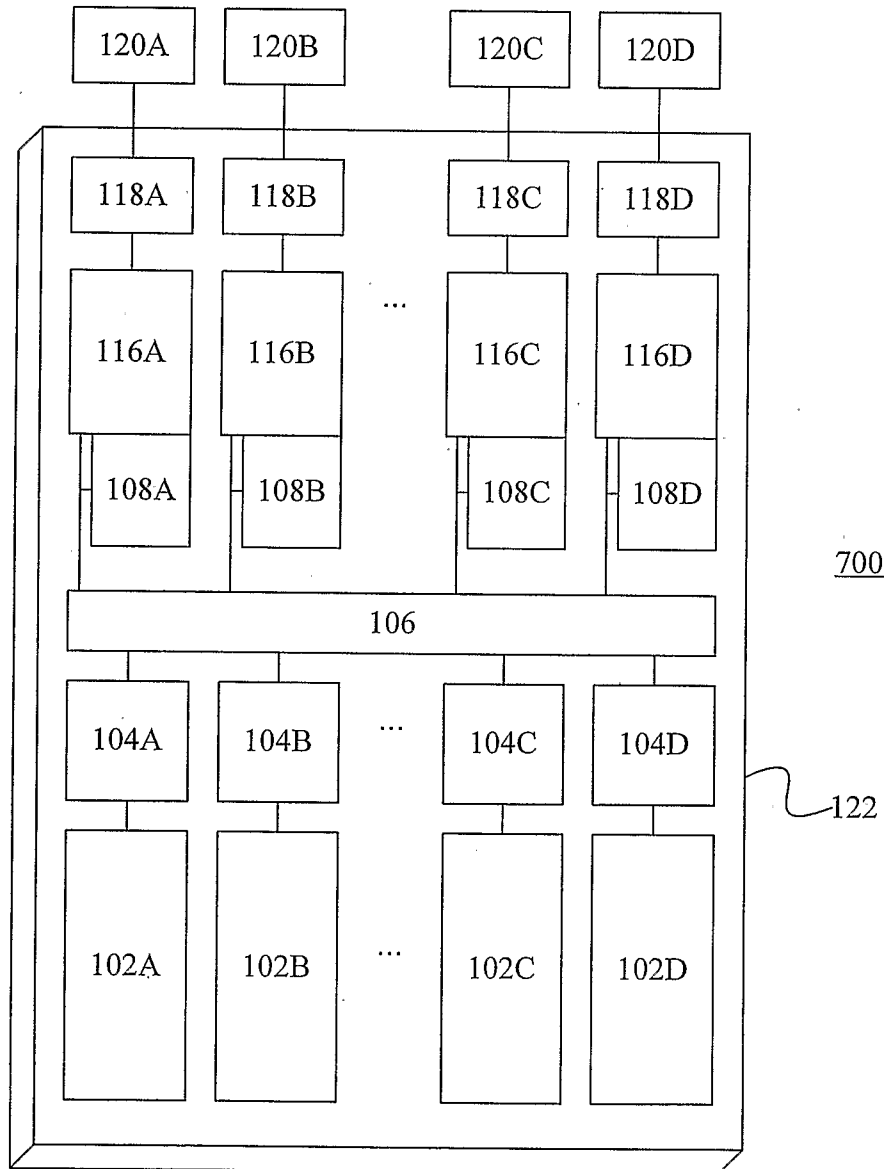


FIG. 7

INTERNATIONAL SEARCH REPORT

International application No

PCT/US2006/043889

A. CLASSIFICATION OF SUBJECT MATTER
INV. G06F9/52

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, INSPEC

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	G. ZHANG ET AL: "Busy-Wait Barrier Synchronization Using Distributed Counters with Local Sensor" WOMPAT 2003, [Online] 26 June 2003 (2003-06-26), - 27 June 2003 (2003-06-27) pages 84-98, XP002417541 toronto, canada Retrieved from the Internet: URL: http://www.springerlink.com/content/at05ttvlfngbqgpp/fulltext.pdf [retrieved on 2007-01-29] the whole document	1,10
A	WO 99/26148 A2 (CRAY RESEARCH INC [US]) 27 May 1999 (1999-05-27) the whole document	1,10
-/--		

☒ 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

28 February 2007

Date of mailing of the international search report

06/03/2007

Name and mailing address of the ISA/

European Patent Office, P.B. 5818 Patentlaan 2
NL - 2280 HV Rijswijk
Tel. (+31-70) 340-2040, Tx. 31 651 epo nl,
Fax: (+31-70) 340-3016

Authorized officer

Michel, Thierry

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	MASAHICO IWANE ET AL: "TAGGED COMMUNICATION AND SYNCHRONIZATION MEMORY FOR MULTIPROCESSOR-ON-A-CHIP" SYSTEMS & COMPUTERS IN JAPAN, WILEY, HOBOKEN, NJ, US, vol. 32, no. 4, April 2001 (2001-04), pages 1-13, XP001116315 ISSN: 0882-1666 page 4, left-hand column, line 17 - page 4, right-hand column, line 4 -----	1,10
A	LEE J ET AL: "Synchronization with multiprocessor caches" PROCEEDINGS OF THE ANNUAL INTERNATIONAL SYMPOSIUM ON COMPUTER ARCHITECTURE. SEATTLE, MAY 28 - 31, 1990, LOS ALAMITOS, IEEE COMP. SOC. PRESS, US, vol. SYMP. 17, 28 May 1990 (1990-05-28), pages 27-37, XP010019262 ISBN: 0-8186-2047-1 page 33, left-hand column, line 6 - line 15; figure 5 -----	1,10
A	BECKMANN C J ET AL: "FAST BARRIER SYNCHRONIZATION HARDWARE" PROCEEDINGS OF THE SUPERCOMPUTING CONFERENCE. NEW YORK, NOV. 12 - 16, 1990, WASHINGTON, IEEE COMP. SOC. PRESS, US, vol. CONF. 3, 12 November 1990 (1990-11-12), pages 180-189, XP000267419 ISBN: 0-8186-2056-0 the whole document -----	1,10
P,X	J. SAMPSON ET AL: "Fast Synchronization for Chip Multiprocessors" ACM SIGARCH SPECIAL ISSUE DAS CMP'05, [Online] 13 November 2005 (2005-11-13), pages 64-69, XP002417542 Retrieved from the Internet: URL:http://www.cse.ucsd.edu/{rakumar/dasCM P05/paper07.pdf paper07.pdf (application/pdf Object)> [retrieved on 2007-01-23] the whole document -----	1-10

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No

PCT/US2006/043889

Patent document cited in search report	Publication date	Patent family member(s)	Publication date	
WO 9926148	A2	27-05-1999	US 6085303 A	04-07-2000