

DISPATCHING MULTIPLE THREADS IN A COMPUTER

BACKGROUND

[0001] The present invention relates generally to multithreading (MT), and more specifically, to a machine implementation for dispatching multiple threads in a computer.

[0002] Multithreading (MT) provides a means for increasing the number of processor threads that can operate in parallel within a single physical processor core without the need to add additional cores. Ideally, MT provides this increased capacity by having one or more threads use portions of the core hardware that are currently not being used by the other thread(s) running on the same core. For example, during the latency caused by a cache-miss or other delay in one thread, one or more other threads can utilize the core resources, thus increasing the utilization of the resources. Even though in practice, this sharing results in some interference between the threads and requires some additional hardware, MT still provides the ability to perform each thread's work using less hardware than would be required if each thread were to run on its own isolated core hardware. Often, additional benefit can be derived from MT when the sharing of hardware resources between threads also decreases the overall strain on the computer system to provide information, such as data from memory, to two unique cores.

[0003] Typically, although MT provides hardware savings, the addition of another work thread consumes the same coordination cost at the hypervisor level that would be required to provide increased capacity using an additional, separate core. In many instances, once a certain scaling ratio is achieved, the overhead to coordinate resources between work threads, whether run on a single or shared core, is substantial and can decrease or even outweigh the benefits seen by the ability to run an independent work thread. That is, in general there is more management overhead as the number of things to manage increases.

SUMMARY

[0004] Embodiments include a system, method, and computer program product for dispatching a multithreaded guest virtual machine (VM). According to one aspect, a computer system includes a configuration with a machine enabled to operate in a single thread (ST) mode and a multithreading (MT) mode. In addition, the machine includes physical threads. The machine is configured to perform a method that includes issuing a start-virtual-execution (start-VE) instruction to dispatch a guest entity comprising multiple logical threads on the core. The guest entity includes all or a part of a guest VM, and the issuing is performed by a host running on one of the physical threads on the core in the ST mode. The executing of the start-VE instruction by the machine includes mapping each of the logical threads to a corresponding one of the physical threads, initializing each of the mapped physical threads with a state of the corresponding logical thread, and starting execution of the guest entity on the core in MT mode.

[0005] According to another aspect, a computer implemented method for dispatching a multithreaded guest virtual machine (VM) in a configuration is provided. The configuration includes a machine having a single core enabled to operate in a single thread (ST) mode and a multithreading (MT) mode. The core includes physical threads. The method includes issuing a start-virtual-execution (start-VE) instruction to dispatch a guest entity comprising multiple logical threads on the core. The guest entity includes all or a part of a guest VM, and the issuing is performed by a host running on one of the physical threads on the core in the ST mode. The executing of the start-VE instruction by the machine includes mapping each of the logical threads to a corresponding one of the physical threads, initializing each of the mapped physical threads with a state of the corresponding logical thread, and starting execution of the guest entity on the core in MT mode.

[0006] A further aspect includes a computer program product for dispatching a multithreaded guest virtual machine (VM) in a configuration. The configuration includes a machine having a single core enabled to operate in a ST mode and a MT mode. In addition, the machine includes physical threads. The computer program product includes a computer readable storage medium having program instructions embodied therewith, where the

computer readable storage medium is not a signal. The program instructions are readable by a processing circuit to cause the processing circuit to perform a method. The method includes issuing a start-virtual-execution (start-VE) instruction to dispatch a guest entity comprising multiple logical threads on the core. The guest entity includes all or a part of a guest VM, and the issuing is performed by a host running on one of the physical threads on the core in the ST mode. The executing of the start-VE instruction by the machine includes mapping each of the logical threads to a corresponding one of the physical threads, initializing each of the mapped physical threads with a state of the corresponding logical thread, and starting execution of the guest entity on the core in MT mode.

BRIEF DESCRIPTION OF THE SEVERAL VIEWS OF THE DRAWINGS

[0007] The subject matter which is regarded as embodiments is particularly pointed out and distinctly claimed in the claims at the conclusion of the specification. The forgoing and other features, and advantages of the embodiments are apparent from the following detailed description taken in conjunction with the accompanying drawings in which:

FIG. 1 depicts a computing environment that may be implemented in accordance with an embodiment;

FIG. 2 depicts a physical processor that may be implemented in accordance with an embodiment;

FIG. 3 depicts a computing environment that may be implemented in accordance with an embodiment;

FIG. 4 depicts a state description of a multithreaded (MT) logical thread in accordance with an embodiment;

FIG. 5 depicts a block diagram of a thread-validity mask (TVM) in accordance with an embodiment;

FIG. 6 depicts a fixed offset state description group in accordance with an embodiment;

FIG. 7 depicts a state description group specified as an address list in accordance with an embodiment;

FIG. 8 depicts a state description group specified as a linked list in accordance with an embodiment;

FIG. 9 depicts a state description group specified as a circular list or ring in accordance with an embodiment;

FIG. 10 depicts a core dispatch process in accordance with an embodiment;

FIG. 11 depicts a coordinated exit from virtual-execution in accordance with an embodiment;

FIG. 12 depicts a block diagram of a system control area in accordance with an embodiment;

FIG. 13 depicts a process flow for coordinating between multithreaded cores in accordance with an embodiment;

FIG. 14, which includes FIGs. 14A and 14B, depicts a machine implementation of core dispatch in accordance with an embodiment;

FIG. 15, which includes FIGs. 15A and 15B, depicts a machine implementation of a coordinated exit from virtual-execution in accordance with an embodiment; and

FIG. 16 depicts a computer-readable medium according to an embodiment.

DETAILED DESCRIPTION

[0008] Embodiments described herein can be utilized to reduce hypervisor management overhead in a multithreading (MT) environment. As described herein, the management of multiple threads can be split between a hypervisor that manages the multiple threads as a single logical core, and a machine that manages interactions between the multiple threads as they access resources of the physical core. This can result in substantially reducing multithreading (MT) overhead costs by allowing the hypervisor to manage much of the hypervisor infrastructure resources on a logical core basis, and to allow the machine to manage other resources on a more granular, thread basis. An embodiment includes a core dispatch instruction that can be executed by a hypervisor that is running on a single thread (ST). Execution of the core dispatch instruction, referred to herein as a “start-VE instruction with MT specified”, can cause multiple guest logical threads that make up all or a portion of a guest virtual machine (VM) to be dispatched on a single physical core. In an embodiment, the instruction used by the hypervisor to dispatch the guest specifies whether the guest to be dispatched is single-threaded or multithreaded.

[0009] Embodiments described herein can include structures, such as a thread-validity mask for indicating which logical threads within a guest logical core are currently valid and a state description group that includes a state description ring, for managing the dispatch of a multithreaded logical core. In addition, primary and secondary state descriptions and field types (e.g., primary, core-common, thread specific) can be implemented to allow for the efficient management of the computer resources when a logical core with multiple threads is dispatched. Further, a coordinated exit where all threads within a logical core exit virtual-execution at the same time can be provided to simplify both hypervisor and logical core management functions.

[0010] Embodiments can include a control structure maintained by the hypervisor that is referred to herein as a core oriented system control area (COSCA). The COSCA is used by both the hypervisor and the machine to manage certain functions which may affect multiple logical processors in the guest configuration. An embodiment of the COSCA is implemented as a tree structure where the leaves represent logical cores and each leaf contains a list corresponding to the threads of that core. The COSCA structure can contain fields (e.g., state description addresses) that allow the hypervisor to easily access the state descriptions for all of the threads in a particular core.

[0011] Embodiments can also include a machine implementation of the core dispatch (e.g., the start-VE instruction with MT specified) where millicode located on the machine can be utilized to manage the core dispatch process.

[0012] As used herein, the term millicode is used to refer to Licensed Internal Code which is integral to the operation of processor hardware. Millicode is executed by the machine when it is running in an execution mode, referred to herein as millimode.

[0013] When a guest logical core includes multiple threads, one of the logical threads is designated as a primary thread and the rest are designated as secondary threads. The term “primary” applies to the logical thread. A physical thread is neither a primary nor a secondary thread from a hardware perspective; it becomes a primary thread once a start-VE instruction with MT specified is issued on it. The reason for this temporary distinction at the

physical level is because when control is passed back to the host, it is typically done on the primary thread, that is, control is passed back to the hypervisor on the same physical thread on which the start-VE was issued.

[0014] In an embodiment, millicode can be used to load almost the entire state of any thread (primary or secondary) from any other thread (primary or secondary). In embodiments, the millicode uses this state loading flexibility to load a very small portion of another thread in order to leverage the potential efficiency that can be gained by the parallel loading of its own state by each thread. Some commands (such as purge translation look-aside buffer or “PTLB”) and common resources can apply to all threads, and this allows them to be executed or loaded from only a single thread. Not only does this save on the time for the command or the load itself but, in some cases, it also saves on testing required to determine if the action is actually required. This flexibility included in the design can allow millicode to constantly adjust the algorithm used to support core dispatch as the design, development, and test cycle progress. Mechanisms can be provided to efficiently start and stop thread execution. In addition, millicode can also be used to account for the situation where internal firmware is running on a thread which is viewed as invalid at the software level.

[0015] Additional embodiments are directed to machine implementation of a coordinated exit from a dispatched MT guest logical core back to a ST host (e.g., a hypervisor). In an embodiment, millicode is used to synchronize the system, and this includes the coordination of all the different threads by taking into account each of their current states. An embodiment can also include the handling of high-priority interruptions while holding lower-priority interruptions which may delay the exit. The shutting down of secondary threads can be done in a way that allows for the most efficient use of the core resources after the exit is complete. For example, the millicode can disable certain interruptions to prevent core resources from being used to dispatch the millicode interruption handler when it is not necessary. Millicode can also be used to indicate that certain physical registers are no longer being used so that they are free for use by the ST that is running.

[0016] As used herein, the term “thread” refers to a single instruction stream and its associated state. That is, at an architecture level, each logical thread represents an independent CPU or processor. At a hardware level, a physical thread is the execution of an instruction stream associated with a logical thread, combined with the maintaining of that guest state, when that thread is dispatched. It is the maintaining of that thread state by the machine that reduces the management required at the hypervisor level. The total number of logical threads available for use by logical cores is limited by the total number of physical threads available to the physical cores.

[0017] As used herein, the term “physical core” refers to a hardware processing unit that executes one or more independent instruction streams or threads but shares many base resources such as execution units and low-level caches. This sharing can be done a number of ways including by having each thread use the same hardware resources at independent times or by having the resources be logically shared with each physical entry tagged with a thread identifier. Proper synergy between the threads, for example one thread that needs resource A often but resource B only rarely and another thread that typically uses resource B but not resource A, can improve the efficiency of this sharing. As used herein, the term “machine” refers to hardware included in the physical core as well as millicode and other hardware used in support of the physical core.

[0018] As used herein, the terms “guest VM” and “guest” are used interchangeably to refer to a single guest configuration which can include a single CPU or multiple CPUs. As used herein, the term “logical core” refers to the group of logical guest threads or CPUs that are defined to be dispatched together as part of a start-VE instruction where MT is specified. A guest VM can be made up of a single logical core (either ST or MT) or multiple logical cores (also each of which can be ST or MT).

[0019] As used herein, the term “software” refers to either the hypervisor program (e.g. PR/SM or zVM) or the guest operating system or application program which is dispatched as a result of the start-VE instruction.

[0020] As used herein, the terms “hypervisor” and “host” refer to the program which manages the system resources and dispatches the guest logical processor(s) to run on the physical hardware.

[0021] The operand of the start-VE instruction used to dispatch a guest points to a state description or group of state descriptions which defines the state of that guest processor or core. The state description itself has pointers to “satellite blocks” which can be viewed as an extension to the state description and include additional information which further defines the state of that guest core or processor. As used herein, the term “state description” refers not only to the state description itself but also to those satellite blocks. The core-oriented system control area (COSCA), one of these satellite blocks, is depicted in FIG. 12.

[0022] Turning now to FIG. 1, a computing environment 100 that may be implemented by an exemplary embodiment is generally shown. Computing environment 100 can be based, for example, on the z/Architecture offered by International Business Machines Corporation, Armonk, New York. The z/Architecture is described in an IBM® publication entitled, “z/Architecture Principles of Operation,” IBM Publication No. SA22-7832-09, August 2012, which is hereby incorporated herein by reference in its entirety. In one example, a computing environment based on the z/Architecture includes an eServer zSeries, offered by International Business Machines Corporation, Armonk, New York.

[0023] As one example, computing environment 100 can include a processor complex 102 coupled to a system controller 120. Processor complex 102 can include, for example, one or more partitions 104 (e.g., logical partitions LP1-LPn), one or more physical cores 106 (e.g., Core 1-Core m), and a level-0 hypervisor 108 (e.g., a logical partition manager), each of which is described below.

[0024] Each logical partition 104 can be capable of functioning as a separate system. That is, each logical partition 104 can be independently reset, initially loaded with an operating system 110, if desired, and operate with different programs. An operating system 110 or application program running in a logical partition 104 can appear to have access to a full and complete system, but in reality, only a portion of it is available. A combination of

hardware and Licensed Internal Code (commonly referred to as microcode or millicode or firmware) keeps a program in one logical partition 104 from interfering with a program in a different logical partition 104. This allows several different logical partitions 104 to operate on a single or multiple physical cores 106 in a time sliced manner. In an embodiment, each physical core includes one or more central processors (also referred to herein as “physical threads”). In the example shown in FIG. 1, each logical partition 104 has a resident operating system 110, which may differ for one or more logical partitions 104. The operation system 110 running in each logical partition 104 is an example of a virtual machine or guest configuration. In one embodiment, operating system 110 is the z/OS® operating system, offered by International Business Machines Corporation, Armonk, New York.

[0025] Physical cores 106 include physical processor resources that are allocated to the logical partitions 104. A logical partition 104 can include one or more logical processors, each of which represents all or a share of the physical processor resources allocated to the partition 104. The physical cores 106 may be either dedicated to the logical cores of a particular partition 104, so that physical processor resources of the underlying core(s) 106 are reserved for that partition 104; or shared with the logical cores of another partition 104, so that physical processor resources of the underlying core(s) resources are potentially available to another partition 104.

[0026] In the embodiment shown in FIG. 1, the logical partitions 104 are managed by level-0 hypervisor 108 which is implemented by firmware running on physical cores 106. Logical partitions 104 and hypervisor 108 each comprise one or more programs residing in respective portions of central storage (memory) associated with the physical cores 106. One example of hypervisor 108 is the Processor Resource/Systems Manager (PR/SM™), offered by International Business Machines Corporation, Armonk, New York.

[0027] System controller 120, which in FIG. 1 is coupled to the central processor complex 102, can include centralized logic responsible for arbitrating between different processors issuing requests. For instance, when system controller 120 receives a memory access request, it determines if access to that memory location is allowed and, if so, provides

the contents of that memory location to the central processor complex 102 while maintaining memory consistency between processors within that complex.

[0028] Turning now to FIG. 2, a block diagram of a processing circuit 200 for implementing a machine or physical core, such as physical core 106 in FIG. 1, is generally shown in accordance with an embodiment. The processing circuit 200 can include one physical core of a plurality of physical cores in a multi-processing environment. The processing circuit 200 shown in FIG. 2 includes a system controller interface unit 202 that can couple the processing circuit 200 to other cores and peripheral devices. The system controller interface unit 202 can also connect Dcache 204, which reads and stores data values, Icache 208, which reads program instructions, and a cache interface unit 206 to external memory, processors, and other peripheral devices.

[0029] The Icache 208 can provide loading of instruction streams in conjunction with an instruction fetch unit (IFU) 210, which pre-fetches instructions and may include speculative loading and branch prediction capabilities. The fetched instructions can be provided to an instruction decode unit (IDU) 212 for decoding into instruction processing data.

[0030] The IDU 212 can provide the instructions to an issue unit 214 which can control the issuing of the instructions to various execution units, such as one or more fixed point units (FXU) 216 for executing general operations and one or more floating point units (FPU) 218 for executing floating point operations. The FPUs 218 can include a binary floating point unit (BFU) 220, a decimal floating point unit (DFU) 222, or any other floating point unit. The issue unit 214 can also be coupled to one or more load/store units (LSU) 228 via one or more LSU pipelines. The multiple LSU pipelines are treated as execution units for performing loads and stores and address generation for branches. Both the LSU 228 and the IFU 210 can utilize a translation-lookaside-buffer (TLB) 230 to provide buffered translations for the operand and instruction addresses.

[0031] The FXU 216 and FPU 218 are coupled to various resources such as general-purpose registers (GPR) 224 and floating point registers (FPR)

226. The GPR 224 and FPR 226 provide data value storage for data values loaded and stored from the Dcache 204 by a LSU 228.

[0032] Turning now to FIG. 3, a computing environment 300 that may be implemented by an embodiment is generally shown. The computing environment 300 shown in FIG. 3 is similar to the computing environment 100 shown in FIG. 1 with the addition of a level-1 hypervisor 302 executing in the logical partition 104 labeled LP2. As shown in FIG. 3, the level-1 hypervisor 302 can provide the same hypervisor functions described previously with respect to hypervisor 108 (also referred to herein as a “level-0 hypervisor”), such as transparent time-slicing of resources between multiple operating systems (e.g. OS1 314, OS2 312, and OS3 310 running in virtual machines VM1 304, VM2 306, and VM3 308) and isolation of those operating systems from one another, within the logical partition 104 labeled LP2. The embodiment shown in FIG. 3 includes three virtual machines as an example and other embodiments may include more or fewer virtual machines based on application requirements.

[0033] As shown in FIG. 3, the logical partition 104 labeled LP1 has a resident operating system 110, and the logical partition 104 labeled LP2 runs a level-1 hypervisor 302 which in turns creates virtual machines 304 306 308 each of which runs its own resident operating systems 314 312 310. Any number of the logical partitions 104 may run a level-1 hypervisor 302. In an embodiment, the level-1 hypervisor 302 is the z/VM hypervisor, offered by International Business Machines Corporation, Armonk, New York. The resident operating systems running in the various logical partitions may differ and, when running under a level-1 hypervisor 302, the resident operating systems (e.g., operating systems 314 312 310) within a single partition 104 (e.g., LP2) may also differ. In an embodiment, operating system 110 in the logical partition 104 labeled LP1 is the z/OS operating system, which is offered by International Business Machines Corporation, Armonk, New York. In an embodiment, operating systems 310 and 312 are Linux and operating system 314 is z/OS.

[0034] When a level-1 hypervisor 302 is running in a logical partition 104 it can provide the same virtualization of resources provided by a level-0 hypervisor, such as hypervisor 108, to logical partitions 104 to the operating systems 310 312 314 running in virtual

machines 308 306 304. As at the first level, each virtual machine may include multiple virtual processors.

[0035] Physical cores 106 include physical processor resources that can be dedicated or shared, as described for FIG 1, between the logical partitions 104 LP1, LP2, LP3 and LP4. When logical partition LP2 is dispatched on one or more physical cores, the level-1 hypervisor 302 can then transparently share those resources between its virtual machines VM1 304, VM2 306 and VM3 308. In one embodiment, level-0 hypervisor 108 uses a start-VE instruction with MT specified to dispatch a multithreaded level-1 hypervisor 302 which then uses a start-VE instruction with ST specified to dispatch single-threaded virtual machines VM1 304, VM2 306 and VM3 308. In a different embodiment, level-0 hypervisor 108 uses a start-VE instruction with ST specified to dispatch a single threaded level-1 hypervisor 302 which then uses a start-VE instruction with MT specified to dispatch multithreaded virtual machines VM1 304, VM2 306 and VM3 308. In another embodiment both level-1 hypervisor 302 and its guest VMs 304 306 308 are all single-threaded.

[0036] In a guest multiprocessing (MP) environment, the hypervisor can maintain a control structure, known as the system control area (SCA), which is used by both the hypervisor and the machine to manage certain functions which may affect multiple logical processors in the guest configuration. The same SCA origin (SCAO) is specified in the state description for all the guest processors in the configuration or virtual machine. In an embodiment, this area can include a common area (used, in general, to coordinate guest-configuration-wide functions) and separate, processor-specific entries. The common area, for example, holds information about which virtual processors within the guest configuration are valid. The separate, processor-specific area within the SCA can, for example, be used to interpret or emulate inter-processor guest functions such as an inter-processor interruption or to provide easily accessible pointers to each logical processor's respective state description. In an embodiment the SCA used for ST is extended for MT use by adding additional thread-specific entries for each potential guest thread.

[0037] An embodiment of a core dispatch can allow a hypervisor that is running on a single thread to dispatch a multithreaded guest on its core using a variation of the start-VE

instruction, sometimes referred to as start multithreaded virtual-execution (start-MVE). Each thread in the multithreaded guest can represent a guest logical central processing unit (CPU), or guest thread. The start-VE instruction can enable multithreading (MT) guest execution on the physical core, via a control field in the state description. The operand of the start-VE instruction when used for core dispatch can specify either a single state description which contains the state of all the guest threads or a group of state descriptions each of which, for example, represents the state of a single guest thread. In an embodiment, the logical core includes this group of state descriptions. Core dispatch requires virtual-execution entry to load the state of the logical core and each of these guest logical threads into a physical core thread and its threads. These threads can be instruction streams which operate independently from one another. In various embodiments, a group of state descriptions can be specified a number of ways including as fixed offsets from one another, as a list of state description addresses or state descriptions, or as a circular list (ring) of state descriptions that applies to the core with each state description in that group representing a separate guest thread. Such techniques allow for easy access by the hypervisor and the machine to other threads within the logical core and allow for fields which apply to the entire logical core to be maintained in a single place.

[0038] The guest OS can make use of multithreading simply by issuing an MT-setting instruction which enables multithreading in the guest. This allows the guest OS to treat these new threads as additional, independent CPUs and manage them as it would in the absence of multithreading. In addition, the guest OS may use these threads in a way that leverages the fact that they share a core or it can have them operate in a more interdependent manner. This is all transparent to the hypervisor and the machine. The hypervisor then provides these additional threads to the guest OS while the hypervisor itself continues to run on a single thread per core and to manage much of the guest MT environment on a core basis. The OS enablement of multithreading is described in more detail in U.S. Patent Application 14/226895, (our reference POU920140064US1) entitled “Thread Context Preservation in a Multithreading Computer System,” which is filed concurrently herewith, the contents of which are incorporated by reference herein in its entirety. A copy is included on the file.

[0039] In an embodiment of core dispatch, the state description which is specified as the operand of the start-VE instruction with MT specified is a “primary” state description and the associated guest logical thread is the “primary” thread. The other state descriptions in the group are referred to herein as “secondary” state descriptions and, if applicable, apply to secondary logical threads. When the state description group is implemented as either a list or a ring, there can be a next-state-description (NSD) field in the primary state description that points to the first secondary state description which in turn, either 1) points to the next secondary state description in the group or 2) contains a value to indicate the end of a group. The NSD value in the state description for the last in the list may be the address of the primary state description in which case the list forms a ring of state descriptions.

[0040] In a non-MT implementation, the hypervisor dispatches one guest logical processor (also referred to herein as a “logical thread”) on a given physical core at a time. If a particular logical processor is in an invalid state, for example, in the stopped state or in a disabled wait, the hypervisor will not dispatch that guest. In an MT environment, the core dispatch allows the hypervisor to dispatch multiple guest threads on the core at the same time. In order to accommodate the possibility that one or more of the threads in that logical core’s state-description group is invalid, an embodiment utilizes a thread-validity mask (TVM) in the primary state description, each bit of which indicates the validity, from a software perspective, of the logical thread in the corresponding state description in the group.

[0041] In another embodiment, only valid threads are included in the state description group and no validity indication is necessary. An embodiment which includes invalid logical threads in the state description group allows the hypervisor to maintain the state associated with these invalid threads and these threads may become valid again in the future. The machine will only initialize and run those threads which have a valid state. The hypervisor will only dispatch a guest logical core if at least one thread in the group is valid.

[0042] Turning now to FIG. 4, a state description of a logical thread that includes most of the architected state of the guest is generally shown in accordance with an embodiment.

In this context, the term “state description” includes not only the state description itself but also the satellite blocks, whose pointers reside in the state description, that act as an extension. As shown in FIG. 4, the state description 400 can include the guest general registers (GRs) 402, access registers (ARs) 404, control registers (CRs) 406, guest timers 408 (including clock comparator and CPU timer), guest prefix register 410, virtual CPU number (VCN) 412, program-status word (PSW) and instruction address (IA) 414. In addition, it can include control information such as interception control (IC) bits 420 to indicate if certain instructions (e.g., Load Program Status Word (LPSW) and Invalidate Page Table Entry (IPTE)) require interception to the host or if purging of the guest translation lookaside buffer (TLB) is required before guest instruction execution can begin. The state description also contains the next-state-description (NSD) 422 which is used to define state description lists and rings as described in FIGs 6-9. The primary state description also includes the TVM 430 as described in FIG 5 and the logical-partition number (LPN) 432. The virtual CPU number (VCN) 412 is equivalent to the CPU number, potentially adjusted to include the thread number in MT mode, as described in U.S. Patent Application 14/226947, (our reference POU920140072US1), entitled “Address Expansion and Contraction in a Multithreading Computer System” which is filed concurrently herewith, the contents of which are incorporated by reference herein in its entirety. A copy is included on the file.

[0043] Threads within a core may be identified by a binary thread identification (TID). For brevity in the below figures, thread x is often referred to by the term TID x , in which case the meaning is “the thread having TID x ”.

[0044] Referring now to FIG. 5, a block diagram of a thread-validity mask (TVM) 520 is generally shown in accordance with an embodiment. As shown in FIG. 5, bit 0 530 of the TVM 520 represents the validity of logical thread 0 in the state description group, bit 1 531 represents the validity of thread 1, bit 2 532 represents the validity of thread 2, bit 3 533 represents the validity of thread 3 and so forth until bit n 537 which represents the validity of thread n , the last possible logical thread in the state description group associated with this core. The TVM can reside in the primary state description for the group.

[0045] Turning now to FIG. 6, a fixed offset state description group structure is generally shown in accordance with an embodiment. As shown in FIG. 6, the state description group is specified at fixed offsets (N) from one another. In this case, the operand of the start-VE instruction 602 points to the primary state description 603 for logical thread 0. The secondary state description for logical thread x 605 is located at a fixed offset of N bytes after the primary state description and the secondary state description for logical thread y 607 is located N bytes after the secondary state description for thread x. This continues for all the threads in the group. The number of threads in the group can be determined a number of ways including by a count in the primary state description or by an end marker following the last state description address in the list.

[0046] FIG. 6 can represent two cases, the first case being where the group includes state descriptions for all logical threads in the group, whether they are valid or not, and the second case being where only valid state descriptions are included in the group. In the first case, the state description for thread x 605 represents the state of thread 1 and that for thread y 607 represents the state of thread 2. The TVM 620, which is only needed in this first case, represents the validity of each of these logical threads. In the second case, the state description for thread x 605 represents the state of the first valid logical secondary thread and the state description for logical thread y 607 represents the state of the second valid secondary thread. For example, if thread 1 is not valid and threads 2 and 3 are both valid, then thread x 605 would represent thread 2 and thread y 607 would represent thread 3. There would be no state description included in the group for thread 1 since it is invalid. These same two cases can also apply to the embodiments shown in FIGS. 7-9 below, however only case 1 is described and pictured.

[0047] Turning now to FIG. 7, a state description group structure specified as a list is generally shown in accordance with an embodiment. In this case, the operand of the start-VE instruction 702 represents a list of state description addresses with the first entry in the list 704 pointing to the primary state description 705 for thread 0, the second entry in the list 706 pointing to the secondary state description for thread 1 707, the third entry in the list 708 pointing to the secondary state description for thread 2 709, and so on, continuing for all threads in the group. The TVM 720 represents the validity of each of these threads.

[0048] Turning now to FIG. 8, a state description group structure specified as a linked list is generally shown in accordance with an embodiment. In this case, as in the case depicted in FIG 6, the operand of the start-VE instruction 802 points to the primary state description for thread 0 803 but, instead, the pointer 804 for the secondary state description for thread 1 805 is provided as a next state description (NSD) field 804 in the primary state description. In turn, the pointer 806 for the secondary state description for thread 2 807 is provided as an NSD 806 in the secondary state description for thread 1. This would continue for all threads in the group with the NSD 810 in the state description for the last thread n 809 being specified as zeroes or some other unique value indicating the end of the list. The TVM 820 provided in the primary state description 803 represents the validity of each of these threads.

[0049] Turning now to FIG. 9, a state description group structure specified as a circular list or ring is generally shown in accordance with an embodiment. This case is identical to the case shown in FIG. 8 in that the operand of the start-VE instruction 902 points to the primary state description 903 for thread 0, which contains an NSD 904 for the secondary state description for thread 1 905, which contains an NSD 906 for the secondary state description for thread 2 907 and this continues for all threads up to the last thread n. In the embodiment shown in FIG. 9, however, the NSD 910 in the state description for thread n 909 forms a circular list and points back to the primary state description 903. The TVM 920 provided in the primary state description 903 represents the validity of each of these threads.

[0050] The core dispatch allows the hypervisor to manage many aspects of the logical threads at the core level. Core dispatch not only often simplifies the hypervisor code required for thread management by pushing the coordination of virtual-execution of multiple threads of a core into the machine, but it can also reduce the overhead required for managing more processors in the configuration. Priority management for logical partitions (or guests) can continue to be done at the logical core level, reducing scaling pressure on this type of management. The hypervisor itself still needs to manage the collection of threads associated with a logical core to make sure its needs (such as instruction interceptions) are all met before reissuing the start-VE instruction.

[0051] Referring now to FIG. 10, a core dispatch process in accordance with an embodiment is generally shown. As shown in FIG. 10, a hypervisor is running single-threaded on physical core N 1010 and physical thread A 1020. In block 1022, the hypervisor issues the start-VE instruction with MT specified to dispatch the multithreaded guest core. The machine determines that the guest is multithreaded and, in block 1024, makes physical threads B and C available to run software. The machine loads the guest state from the state description for each of the threads into a corresponding physical thread. In the embodiment depicted in FIG. 10, the machine uses multiple physical threads to perform this function, that is, millicode running on physical thread A 1020 loads the state of guest logical thread X into physical thread A, as shown in block 1026. Likewise, millicode running on physical threads B 1040 and C 1060 load the state of guest logical threads Y and Z into physical threads B and C as shown in blocks 1046 and 1066. Once the guest state is loaded, the software running on guest logical threads X, Y, and Z executes on physical threads A, B, and C as shown in blocks 1028, 1048 and 1068.

[0052] Referring now to FIG. 11, a coordinated exit from virtual-execution is generally shown in accordance with an embodiment. As shown in FIG. 11, guest logical threads X, Y and Z are executing guest software on physical threads A 1120, B 1140, and C 1160 as indicated in blocks 1128, 1148 and 1168. One or multiple guest threads determine that an exit from virtual-execution is required. Referring to FIG. 11, guest logical thread Y running on physical thread B 1140 determines that it must exit virtual-execution, as shown in block 1150, causing the machine to signal physical threads A 1120 and C 1160 to exit virtual-execution, as shown in block 1152. In blocks 1136, 1154 and 1174, millicode running on each of the physical threads coordinates the exit from virtual-execution and then makes physical threads B 1140 and C 1160 unavailable for use by software as indicated in blocks 1156 and 1176. Millicode on physical thread A 1120 reloads the host state into the hardware as shown in block 1138 which results in the execution of the hypervisor software on physical thread A as shown in block 1140. The hypervisor will then process any pending guest interceptions and host interruptions as needed.

[0053] FIG. 12 depicts a block diagram of a core-oriented system control area (COSCA) for a single guest configuration that includes multiple logical cores in accordance with an

embodiment. The COSCA shown in FIG. 12 can be used to provide coordination both between logical threads within a core and between logical threads on different cores. The COSCA can include a common area representing the entire guest configuration with pointers, one for each logical core, to separate core-description areas. Each core-description includes a common area representing that core and a series of contiguous, separate, thread-specific areas or thread descriptions for that core. In another embodiment the core description provides the locations of the thread descriptions. The provided locations may be implied (e.g., they are a list contained in the core description, or they may be in memory blocks that are consecutive to the core description). In other embodiments, pointers to other memory locations containing the thread descriptions may be provided. As used herein, the term “indicates a location” is used to refer to any of these, or any additional manner, of determining a location of an item (e.g., the thread descriptions or other elements in the COSCA). This structure maintains the tree-like representation of the MT guest configuration which facilitates in some situations, particularly at the hypervisor level, things to be managed on a core basis but in other situations things to be managed on a thread or processor basis.

[0054] The same COSCA origin (COSCAO) can be provided in the SCA origin (SCAO) field in the state descriptions for all the guest threads within the guest configuration and the same core-description area address (CDAA) can be provided for all threads within a given core. An advantage of this embodiment is that it does not require as much contiguous real storage which can be difficult for some hypervisors to provide. Another embodiment could add an additional level of indirection and have each core description include a list of pointers for each thread-specific area removing the need for the control blocks containing these areas to be contiguous.

[0055] Referring now to FIG. 12, an example embodiment of a COSCA for a single guest configuration which includes two logical cores with three logical threads in each core is generally shown. In an embodiment the COSCA includes the contents of the COSCA common area 1260 (shown in FIG. 12 as “COSCACA 1260”), core description area 1270, and core description area 1280. The primary state description for the state description group associated with logical core 0 1203 is specified as the operand of the start-VE instruction

used by the hypervisor to dispatch guest core 0 1202. In addition, the primary state description for the state description group associated with logical core 1 1233 is specified as the operand of the start-VE instruction used to dispatch core 1 1232. This primary state description for “core 0 thread 0” 1203 contains NSD01 1205, which points to the secondary state description for core 0 thread 1 1213 which, in turn, contains NSD02 1215, which points to the final secondary state description for core 0 thread 2 1223 in the group. Similarly, the state description group for logical core 1 begins with the primary state description for core 1 thread 0 1233 containing NSD11 1235, which points to the secondary state description for core 1 thread 1 1243 containing NSD12 1245, which points to the final secondary state description for core 1 thread 2 1253. The state descriptions for all six threads in this guest configuration 1203 1213 1223 1233 1243 1253 contain the same value in SCAO 1204 1214 1224 1234 1244 1254, pointing to COSCA common area 1260.

[0056] The COSCA common area 1260, as shown in FIG. 12, contains core-level information used to coordinate guest configuration-wide functions. The COSCA common area 1260 includes an SCA core validity mask (SCVM) 1261 which indicates the validity of each logical core within the guest configuration and also includes a core-description area address (CDAA) for each core 1262 1264. Both the bits in the SCVM and the array of core-description addresses may be indexed by the core number. CDAA0 1262, which points to the core description area (CDA) for core 0 1270, is included in the COSCA common area 1260. Additionally, the CDAA field in the state descriptions for all threads in core 0 1206 1216 1226 also point to the CDA for core 0 1270. CDAA1 1264, which points to the CDA for core 1 1280, is also included in the COSCA common area 1260, and likewise the CDAA field in the state descriptions for all the threads in core 1 1236 1246 1256 also point to the CDA for core 1 1280. The core description area (CDA) for core 0 1270 contains an SCA thread validity mask (STVM0) 1271 which indicates the validity of each logical thread within core 0. It also contains the thread description areas for core 0 thread 0 1272, thread 1 1274, and thread 2 1276. The CDA for core 1 1280, similarly contains STVM1 1281 and the thread description areas for core 1 thread 0 1282, thread 1 1284, and thread 2 1286). Each of these thread description areas 1272 1274 1276 1282 1284 1286 contain a state description address (SDA) 1273 1275 1277 1283 1285 1287 for the thread corresponding to that thread description area, core 0 thread 0, core 0 thread 1, core 0 thread 2, core 1 thread 0, core 1

thread 1, and core 1 thread 2, respectively. Both the bits in the STVM and the array of thread description areas may be indexed by the thread identification. These SDAs make it easier for the hypervisor to manage threads within a core and for the machine to present guest inter-processor interruptions.

[0057] FIG. 13 depicts a process flow for managing multithreaded cores in accordance with an embodiment that uses the COSCA shown in FIG. 12. In the example shown in FIG. 13, at block 1302, a guest operating system (OS) running on a first physical thread (e.g., core 0 thread 1 defined by state description 1213) has determined that it will signal a second logical thread, or target thread, (e.g., core 1 thread 2 defined by state description 1253). At block 1304, the guest OS does this, for example, by issuing an inter-processor-interrupt instruction. The machine, as part of executing the inter-processor-interrupt instruction, uses the COSCA to emulate the guest inter-processor-interrupt instruction. The inter-processor-interrupt instruction is emulated by the machine since the logical core that includes the target logical thread may or may not be dispatched at the time the signaling is being done. At block 1306, the machine locates (e.g., via SCA0 1214 since the inter-processor-interrupt instruction was executed by logical core 0 thread 1) a common area (e.g., COSCA common area 1260) for the guest configuration in order to access an SCVM (e.g., SCVM 1261) to verify the validity of the target core and to obtain the appropriate CDAA (e.g. CDAA1 1264 since the target thread is on core 1).

[0058] Next, at block 1308, the machine locates (e.g., via CDA1 1264) the core description area for the target core (e.g., CDA 1280). The machine verifies that the target thread is valid by accessing an STVM in the core description area (e.g., STVM1 1281 in CDA 1280). At block 1310, the machine locates the thread description area (e.g. thread description area 1286 corresponding to thread 2 since the target thread is thread 2). At block 1312, information about the interruption is recorded in the thread description area for the target thread (e.g., it places the identity of the sending thread into thread description area 1286). At block 1314, the machine locates (e.g. via SDA12 1287 in thread description area 1286) the state description for the target thread (e.g. secondary state description for core 1 TID2 1253). At block 1316, the interruption is made pending in the target state description (e.g., the IP bit 1257 is set in state description for core 1 TID2 1253). As a result, when the

target logical processor (e.g. core 1 thread 2) is dispatched on a physical thread and is enabled for the interruption, the machine will present the interruption, if enabled, to the guest operating system. If the target logical processor is already dispatched at the time that the interruption becomes pending, it will take the interruption once it is enabled.

[0059] There are instances where the machine can also make use of the fact that threads within a logical core have attributes in common. For example, core dispatch naturally lends itself to having all guest threads on a logical core reside in the same LPAR zone or partition. The design can minimize hardware by only having to implement things associated with the zone once per core instead of once per thread. In addition, complicated control logic (handling of system-wide interruptions, for example) can also be simplified since it must only deal with a single core value.

[0060] In one embodiment, each field (or bit within a field) in the group of state descriptions representing a multithreaded guest is classified as primary, core-common, or thread-specific. A primary field resides only in the primary state description and applies to all processors in the logical core; any access made to a primary field on behalf of any thread of a core must use the value from the associated primary state description. This classification is used for fields which define the overall state of the core, such as the thread-validity mask. A core-common field is common among all processors within a logical core and this field has the same value in every state description in the group; any access made to one of these fields on behalf of a processor may use the value from any state description in the group. This classification is used for fields that apply across the entire core, such as the LP number. The hypervisor is required to maintain the core-common fields in all the state descriptions but the machine is allowed to access this field in the state description of any thread, whichever provides the best performance. Since these fields are not often changed by the hypervisor but are often accessed by the machine on every entry into virtual-execution, defining a field as core-common rather than thread-specific allows virtual-execution entry, for example, to load a secondary thread facility from the primary thread using the value in the primary state description. A thread-specific field is specific to each logical thread; any access made to one of these fields on behalf of any given thread

must use the value from that thread's state description. This classification is used for fields that are typically unique between threads, such as the guest prefix.

[0061] An embodiment includes a machine implementation of the core dispatch instruction. When a hypervisor issues a core dispatch, or start-VE instruction with MT specified, the logical core that is described by the associated state description group is loaded into a physical core by the virtual-execution-entry (VE-entry) millicode. As part of this process, the state of each valid logical thread is loaded into a physical thread. The mapping of logical threads to physical threads can be a direct one-to-one mapping or can be virtualized. Before VE-entry begins, the contents of each physical thread contain the state of whichever virtual thread was last run on it. Therefore, VE-entry millicode replaces the entire state with the state of the newly dispatched guest thread.

[0062] When core dispatch is invoked by a single-threaded hypervisor, it is the responsibility of millicode to load the individual guest thread (logical processor) state into the hardware and to set the hardware up to begin multithreaded execution. In order to improve efficiency by allowing each physical thread to load the majority of its own state in parallel, the millicode can load a small number of hardware registers for each of the secondary threads (either by the primary thread or another, already initiated, secondary thread). This can require that a secondary thread which is currently inactive from a hardware perspective be "woken up" to begin execution of a millicode routine which will complete the initialization of its own guest state and eventually begin execution of the guest program. There are cases that although no hypervisor or guest program code is running on a secondary thread, internal firmware may be running in order to handle, for example, some internal system management function. If this is the case, the machine must coordinate this with the dispatch of the guest threads.

[0063] There are some operations, such as purging of the TLB, which can be specified to apply to the entire core. This eliminates the need for each thread to determine if the purge is necessary and, when necessary, to perform that purge. In addition, there are some core resources which are shared or common between the physical threads within a core. Millicode can take advantage of the fact that shared resources need only be loaded from a

single thread and that a single thread can load all copies of common thread resources if savings may be recognized to do so. VE-entry millicode can also use the guest-multithreading-enabled and thread-validity bits to bypass initialization of invalid logical threads in an effort to speed-up the execution of initialization millicode on the valid threads. The flexible hardware design allows embodiments of the millicode to optimize their implementations as the design develops.

[0064] Turning now to FIG. 14, which includes FIGs. 14A and 14B, a machine implementation of core dispatch is generally shown in accordance with an embodiment. As shown in FIG. 14A, at block 1400, a core dispatch is initiated by the hypervisor running on a single thread (referred to as the primary thread) using a start-VE instruction where MT is specified. Based on the hypervisor issuing the start-VE instruction, VE-entry millicode is invoked on the primary thread, that is the thread that issued the instruction, and it starts by initiating multithreading on the hardware for the entire core. Much of this initialization can include testing hardware state bits from all applicable secondary threads and/or setting hardware control bits or fields in all applicable secondary threads. The control and state bits 1450 shown in FIG. 14A may logically reside in the hardware of the secondary thread itself or they may logically reside in common hardware resources shared between the threads but which represent and control the secondary thread. At block 1402, the VE-entry millicode determines, using the thread-validity mask (TVM) in the state description, which guest logical thread(s) are valid, from a software perspective, and therefore need to be loaded into the hardware. It also determines the mapping of the valid logical threads to physical threads.

[0065] Blocks 1404 through 1414 of FIG. 14A are performed by the millicode to verify that the physical threads requested are available. At block 1404 the millicode prevents the applicable secondary hardware threads from taking new interruptions which may invoke internal code or from beginning to run any new program code. This can be accomplished by setting appropriate hardware controls in the secondary thread(s). At block 1406, it is determined whether software is running on the secondary threads. In an embodiment, the hypervisor that is issuing the start-VE instruction on the primary thread guarantees that other hardware threads on this core are not running any program code (e.g.

hypervisor or guest code). In another embodiment, the hypervisor issuing the start-VE instruction may not be able to easily determine if any other hardware threads required for this new start-VE are running program code, for example, code is associated with an independent start-VE. In this case, the VE-entry millicode checks the appropriate hardware state bit(s) on the secondary thread(s) at block 1406.

[0066] If it is determined, at block 1406, that program code is running on hardware needed to run a guest secondary thread, the new start-VE instruction is completed and at block 1408 the hypervisor is informed by the machine that the start-VE instruction was unsuccessful and, potentially, informed of the number of currently available hardware threads. In response, the hypervisor may take appropriate action such as reducing the number of valid guest threads to be dispatched or waiting for some predetermined time, as indicated by block 1410, before issuing the start-VE instruction again at block 1400. If it is determined, at block 1406, that the hardware is available, then processing continues to block 1412. At block 1412, the millicode determines, for example by checking the appropriate state bit(s) in the hardware, if any of the applicable secondary hardware thread(s) are running internal firmware code. If so, in one embodiment, the primary thread waits for the secondary thread(s) to finish running the internal code and, while waiting, in order to avoid hangs, the primary thread may honor certain interruptions. It may, however, block the secondary thread(s) from taking other interruptions so they can reach the idle state more quickly. Processing then continues at block 1416. In another embodiment, if a hardware thread is running internal firmware code, as determined at block 1412, the machine may nullify the start-VE instruction and return control back to the hypervisor at block 1414. This gives the primary thread the opportunity to take internal firmware interruptions and avoid potential hangs and, once no interruption is pending on the primary thread, the start-VE instruction is executed again. Both of these embodiments have benefits when compared to pausing the internal firmware and restarting it again on the hardware when multithreading work is completed since firmware code operation is often essential to system operation (concurrent upgrade, for example) and, in addition, threads run internal firmware infrequently enough that waiting for it to finish is a viable option.

[0067] Processing then continues at block 1416 to start the loading of the logical thread(s) into the physical threads on the physical core. At block 1416, the millicode checks for and takes exceptions related to certain exception conditions. Some of the exception conditions may apply to the start-VE instruction itself (e.g., invalid state description ring structure) and others are related to the conditions applicable to a secondary logical thread (e.g., an access exception on an NSD). At block 1418, entries from guest hardware lookaside buffers (including TLBs) can be purged. This may include purging lookaside buffers for the secondary thread(s), when applicable. At block 1420, the minimal states from the primary and secondary state descriptions are loaded and the required hardware facilities are initialized including each of the valid threads. In an embodiment, the minimal states include the state description addresses for the secondary threads. At block 1422, the hardware controls for the secondary thread(s) are set to stop fetching any internal instruction stream(s). This can simplify switching from a single-thread execution to multithread execution. At block 1424, the millicode instruction address (milli-IA) for each of the other valid secondary thread(s) is loaded. The milli-IA is the location where the secondary threads start executing once they begin fetching an internal instruction stream and it typically points to a location in millicode that will complete initialization of each valid logical thread. The primary thread continues its execution of VE-entry millicode and, therefore, no new milli-IA need be loaded there. At block 1426, in one embodiment, the primary thread wakes up the secondary thread(s) by setting hardware controls which change their execution mode to millimode which causes them to begin execution at the previously loaded milli-IA. In another embodiment, the primary thread (Ta) can wake up the first secondary thread (Tb); Tb can wake up the next secondary thread (Tc); and so on until all valid threads are active and running in the hardware. In an embodiment, secondary threads will not begin executing millicode instructions until another thread sets its execution mode to millicode execution mode, or millimode.

[0068] Referring now to FIG. 14B, at blocks 1428 and 1478, the primary thread and each valid secondary thread execute the millicode needed to complete loading of the associated guest logical thread state into hardware. In an embodiment, the state includes the guest general registers (GRs) 402, access registers (ARs) 404, control registers (CRs) 406, prefix register 408, guest timers 410, virtual CPU number (VCN) 412, program-status word (PSW)

and instruction address (IA) 414, interception control (IC) bits 420, and logical-partition number (LPN) 432 as shown in FIG. 4. At blocks 1430 and 1480, the primary thread and secondary threads complete execution of VE-entry and they exit millicode mode (e.g., millimode). At blocks 1432 and 1482, they begin execution of the guest thread program instruction stream. In an embodiment, each thread's completion of VE-entry is independent from that of the other threads. In another embodiment, the threads sync-up before completing VE-entry.

[0069] In an embodiment, in order to support the use of core dispatch and the hypervisor running single threaded, a coordinated exit from virtual-execution (VE-exit) in which all the guest threads in a given core exit back to the ST host at the same time can be provided. In the context of coordinated VE-exit, types of VE-exit can be divided into three categories: (1) host interruptions which pertain to host operation; (2) host interruptions which pertain to guest operation; and (3) guest interceptions. Host external, I/O, and some machine check interruptions fall into VE-exit category (1). For this case, all guest threads are required to exit virtual-execution mode in order to allow the host to handle the interruption. This interruption will likely cause the host to dispatch a different guest. If the interruption occurs while running in virtual-execution mode, the host interruption can either be detected on all threads so that they can exit virtual-execution mode or be detected on a single thread which then signals the other threads if they should exit.

[0070] VE-exit category (2), host interruptions which pertain to the guest, can include some machine check interruptions (such as an uncorrectable storage error). In a non-multithreaded situation, these conditions are presented as host interruptions. With core dispatch there is only one host thread but since these exceptions pertain to guest operation it is possible for multiple guest threads to detect distinct and different reasons for the same host interruption. To accommodate this, for core dispatch, when applicable, these host interruptions are instead presented in the corresponding guest state description as a new type of guest interception and are handled the same as category (3), described below. In an embodiment, host address-translation fault interruptions arising due to guest memory references also fall into category (2), and may be presented as another new type of guest interception.

[0071] Guest interceptions, even in a guest multithreaded environment, for both VE-exit categories (2) and (3) (above) pertain to a single guest thread and are independent from guest execution of another thread. It is further possible that multiple guest threads recognize such conditions concurrently, requiring the host to handle all of them. Typically, when presented with an interception, the host will simulate some behavior on the guest's behalf and then re-dispatch that same guest. For these cases, since the host is running single-threaded, all guest threads must exit virtual-execution mode before the host can handle the interception(s). This can be accomplished by either waiting for all threads to exit naturally or by signaling the other threads to exit when one thread has determined that it must intercept back to the host. This is referred to as "coordinated VE-exit".

[0072] As each thread determines it must exit virtual-execution mode, it enters VE-exit, and waits in the initial VE-exit sync-up loop until all the other valid threads are also ready to exit. If the implementation requires, then it signals the other threads to exit before entering this sync-up loop. While in the VE-exit sync-up loop, only a minimum of interruptions are handled. In order to allow for the situation where a guest thread is required to exit virtual-execution mode when no host interruption and no guest interception apply, a "no-action" interception is defined to indicate to the host that no interception action is required on behalf of this guest.

[0073] Once all the threads have entered the initial VE-exit sync-up loop, the storing of guest data in all of the valid state descriptions can complete. That is, the current guest state which resides in the hardware is saved in the corresponding state description so this logical guest thread can be re-dispatched at a later time. A final VE-exit sync-up point is required after this storing is complete to guarantee all updates to the secondary thread state descriptions are complete before control is passed back to the hypervisor (which is typically running on the primary thread). Once VE-exit completes, the hypervisor can process each thread in the ring to determine if an interception was presented and, if so, handle it appropriately. After doing so it can then either re-dispatch this same guest logical core or a different one on the physical processor.

[0074] An embodiment includes a machine implementation of a coordinated exit from virtual-execution that includes the suspension of multithreaded guest execution and the return of control back to a single-threaded host. In a MT environment, the virtual-execution-exit (VE-exit) millicode is responsible, under most conditions, for signaling the other valid guest threads to exit virtual-execution as soon as it is feasible. It is possible for the delay seen by the thread requesting exit to be substantial, particularly if a thread is currently running internal code which cannot be interrupted to exit virtual-execution. While a thread is waiting for other threads to exit virtual-execution, millicode can be used to handle certain interruptions such as those which may result in hang conditions or those which may affect overall system performance. It may not be optimal for millicode to handle other interruptions if they may just delay the eventual exit from virtual-execution now that other threads (whose work may have been interrupted) have been signaled. This must all be done while still maintaining proper interruption priority.

[0075] In addition, much of the coordination required to exit virtual-execution which is done by millicode must take into account the state of each thread. For example, it must account for whether each logical thread is valid and if the physical thread is running internal code. Regardless of whether the primary thread is invalid, it must still be signaled since this is typically the thread on which the host program will be resumed. This is all accomplished using an internal millicode-requested interruption system.

[0076] The millicode may also perform some actions to free up physical registers that had been used by the secondary guest thread(s) but are no longer needed since the guest state has been saved and the host does not use the secondary thread(s). For example, millicode may zero out mapped registers as a means to free resources since hardware can, in certain implementations, map all of the logical registers with zero contents to a single physical register leaving the other physical registers for use by the primary thread. There may be other resources that are shared between the threads and can be freed up for use solely by the primary thread. Millicode can also mask off certain interruptions so that core resources are not used to invoke millicode unnecessarily.

[0077] Once any thread determines an exit from virtual-execution is required, it will signal the other threads as described below, for example, in reference to FIG. 15 which includes FIGs. 15A and 15B. The process used can vary based on whether the thread that is exiting virtual-execution is the primary thread or a secondary thread.

[0078] In an embodiment, a primary thread can signal valid secondary threads to exit virtual-execution if all of the following conditions are met: (1) it wants to exit virtual-execution mode and return control back to the hypervisor; (2) all other valid threads are not already in the process of exiting (i.e. at least one thread is still running guest instructions); (3) all the other threads have not already been signaled to exit; and (4) there is at least one other valid thread. If these conditions are not met, each secondary thread is given time to clean up and exit virtual-execution independently. In an embodiment, a secondary thread signals the primary thread (even if it is invalid) and all other valid secondary threads to exit virtual-execution mode if all of conditions (1)-(3) above are met. In an embodiment, a secondary thread must send a signal to the primary thread even if it is the only valid thread since it must signal the primary thread to complete the coordinated exit from virtual-execution.

[0079] Turning now to FIG. 15, which includes FIGs. 15A and 15B, a machine implementation of a coordinated exit from virtual-execution is generally shown in accordance with an embodiment. In the example shown in FIG. 15, primary thread P is executing guest instruction stream P at 1502, secondary thread A is executing guest instruction stream A at 1542, and secondary thread B is executing guest instruction stream B at 1572. At block 1544, secondary thread A determines that it must exit virtual-execution and, using the criteria described above, determines at block 1546 that it must signal the other threads to also exit. Secondary thread A signals the other threads by making an internal exit-virtual-execution request interruption pending in the hardware state control of primary thread P 1508 and the hardware state control of secondary thread B 1578. Also shown in FIG. 15A is the hardware state control of secondary thread A 1548. When responding to the internal interruption at block 1510 for primary thread P and at block 1580 for secondary thread B, each thread checks the reason for the internal interruption and determines, at block 1512 for primary thread P and at block 1582 for secondary thread B, if the internal interruption is an

exit-virtual-execution request. If it is determined not to be an exit-virtual-execution request, then then primary thread P and secondary thread B perform blocks 1514 and 1584 respectively to handle the other interruption request.

[0080] If the interruption is an exit-virtual-execution request, then in an embodiment, each thread can independently perform the following process. At block 1516 for primary thread P and at block 1586 for secondary thread B, it is determined if the thread is valid. Secondary thread A which requested the exit from virtual-execution can be assumed to be valid. For a thread determined to be valid, the majority of the associated guest state from the hardware is saved into its own state description. This includes, as shown in FIG. 4, the guest general registers (GRs) 402, access registers (ARs) 404, control registers (CRs) 406, program-status word (PSW) and instruction address (IA) 414. Guest state which did not change during virtual-execution, such as the virtual CPU number (VCN) 412 and the logical-partition number (LPN) 432, do not need to be saved. This is shown in FIG. 15A at block 1518 for primary thread P, at block 1558 for secondary thread A, and at block 1588 for secondary thread B. Each thread updates its internal state to indicate that it is at the initial VE-exit sync-up point. This is shown in FIG. 15A at block 1520 for primary thread P, at block 1560 for secondary thread A, and at block 1590 for secondary thread B.

[0081] The threads each wait for the primary thread (e.g., primary thread P) and all valid secondary threads (e.g., secondary threads A and B) to reach the initial sync-up point before continuing. This is shown in FIG. 15A at block 1521 for primary thread P, at block 1561 for secondary thread A, and at block 1591 for secondary thread B. Once all of the threads reach the initial VE-sync-up point in the exit-virtual-execution millicode, each of the threads finish updating the appropriate information in the state description. Examples of this information can include the guest CPU timer which is part of the guest timers 408 in FIG. 4. This is shown in FIG. 15A at block 1522 for primary thread P, at block 1562 for secondary thread A, and at block 1592 for secondary thread B. Delaying this state saving until the sync-up point improves consistency in guest timing between the threads.

[0082] Referring now to FIG. 15B, each thread updates its own internal state to indicate that it is at the final sync-up point in the VE-exit millicode. This is shown in FIG. 15B at

blocks 1524 for primary thread P, at blocks 1564 for secondary thread A, and at blocks 1594 for secondary thread B. In addition, primary thread A waits for all of the threads to reach the final sync-up point 1599 at block 1526 before continuing to block 1528. Each secondary thread sets an internal hardware bit to indicate that it is no longer executing a valid program instruction stream which causes the machine to exit millimode and stop executing code. This is shown in FIG. 15B at block 1566 for secondary thread A, and at block 1596 for secondary thread B. In an embodiment, each secondary thread will not execute any program code until the next Start-VE however it may execute internal code, as needed, to handle internal interruptions.

[0083] Once all threads have reached the final sync-up point in VE-exit millicode, shown as point 1599 in FIG. 15B, single-threaded execution is started on primary thread P. Primary thread P completes core cleanup at block 1528, loads the hypervisor state including the host program IA at block 1530 into the hardware, and exits millimode at block 1532. As a result, primary thread P is running in host mode and will handle any appropriate host interruptions, and the hypervisor will handle any guest interceptions.

[0084] Technical effects and benefits include the ability to dispatch a multithreaded guest virtual machine.

[0085] Embodiments include a system, method, and computer program product for dispatching a multithreaded guest virtual machine (VM). According to one aspect, a computer system includes a configuration with a machine enabled to operate in a single thread (ST) mode and a multithreading (MT) mode. In addition, the machine includes physical threads. The machine is configured to perform a method that includes issuing a start-virtual-execution (start-VE) instruction to dispatch a guest entity comprising multiple logical threads on the core. The guest entity includes all or a part of a guest VM, and the issuing is performed by a host running on one of the physical threads on the core in the ST mode. The executing of the start-VE instruction by the machine includes mapping each of the logical threads to a corresponding one of the physical threads, initializing each of the mapped physical threads with a state of the corresponding logical thread, and starting execution of the guest entity on the core in MT mode.

[0086] According to another aspect, a computer implemented method for dispatching a multithreaded guest virtual machine (VM) in a configuration is provided. The configuration includes a machine having a single core enabled to operate in a single thread (ST) mode and a multithreading (MT) mode. The core includes physical threads. The method includes issuing a start-virtual-execution (start-VE) instruction to dispatch a guest entity comprising multiple logical threads on the core. The guest entity includes all or a part of a guest VM, and the issuing is performed by a host running on one of the physical threads on the core in the ST mode. The executing of the start-VE instruction by the machine includes mapping each of the logical threads to a corresponding one of the physical threads, initializing each of the mapped physical threads with a state of the corresponding logical thread, and starting execution of the guest entity on the core in MT mode.

[0087] A further aspect includes a computer program product for dispatching a multithreaded guest virtual machine (VM) in a configuration. The configuration includes a machine having a single core enabled to operate in a ST mode and a MT mode. In addition, the machine includes physical threads. The computer program product includes a computer readable storage medium having program instructions embodied therewith, where the computer readable storage medium is not a signal. The program instructions are readable by a processing circuit to cause the processing circuit to perform a method. The method includes issuing a start-virtual-execution (start-VE) instruction to dispatch a guest entity comprising multiple logical threads on the core. The guest entity includes all or a part of a guest VM, and the issuing is performed by a host running on one of the physical threads on the core in the ST mode. The executing of the start-VE instruction by the machine includes mapping each of the logical threads to a corresponding one of the physical threads, initializing each of the mapped physical threads with a state of the corresponding logical thread, and starting execution of the guest entity on the core in MT mode.

[0088] In addition to one or more of the features described above, or as an alternative, further embodiments can include where only valid logical threads are mapped to a corresponding physical thread.

[0089] In addition to one or more of the features described above, or as an alternative, further embodiments can include where the initializing includes loading, for each of the logical threads, a portion of the state of the logical thread into the corresponding physical thread; and completing, by each mapped physical thread, the loading of the rest of the state of the corresponding logical thread, the rest of the state determined based on the portion of the state.

[0090] In addition to one or more of the features described above, or as an alternative, further embodiments can include where the portion of the state includes an address of a state description of the rest of the state.

[0091] In addition to one or more of the features described above, or as an alternative, further embodiments can include where the one of the physical threads performs the loading, for each of the logical threads, a portion of the state of the logical thread.

[0092] In addition to one or more of the features described above, or as an alternative, further embodiments can include where the executing includes verifying that the mapped physical threads are not currently running any program code or internal code prior to the initializing.

[0093] In addition to one or more of the features described above, or as an alternative, further embodiments can include where the starting execution is performed for each of the physical threads based on the initializing completing for the physical thread.

[0094] In addition to one or more of the features described above, or as an alternative, further embodiments can include where the starting execution is performed for each of the physical threads based on the initializing completing for all the physical threads.

[0095] In addition to one or more of the features described above, or as an alternative, further embodiments can include where the executing includes determining that software is running on one of the mapped physical threads; and based on the determining, completing the start-VE instruction and notifying the host that the start-VE instruction failed.

[0096] In addition to one or more of the features described above, or as an alternative, further embodiments can include where the executing includes determining that internal firmware is running on one of the mapped physical threads; and based on the determining, performing at least one of: nullifying the start-VE instruction and returning control to the host; and waiting for the mapped physical thread to become available.

[0097] The terminology used herein is for the purpose of describing particular embodiments only and is not intended to be limiting of the invention. As used herein, the singular forms “a”, “an” and “the” are intended to include the plural forms as well, unless the context clearly indicates otherwise. It will be further understood that the terms “comprises” and/or “comprising,” when used in this specification, specify the presence of stated features, integers, steps, operations, elements, and/or components, but do not preclude the presence or addition of one or more other features, integers, steps, operations, element components, and/or groups thereof.

[0098] The corresponding structures, materials, acts, and equivalents of all means or step plus function elements in the claims below are intended to include any structure, material, or act for performing the function in combination with other claimed elements as specifically claimed. The description of the present invention has been presented for purposes of illustration and description, but is not intended to be exhaustive or limited to the invention in the form disclosed. Many modifications and variations will be apparent to those of ordinary skill in the art without departing from the scope and spirit of the invention. The embodiment was chosen and described in order to best explain the principles of the invention and the practical application, and to enable others of ordinary skill in the art to understand the invention for various embodiments with various modifications as are suited to the particular use contemplated.

[0099] The descriptions of the various embodiments of the present invention have been presented for purposes of illustration, but are not intended to be exhaustive or limited to the embodiments disclosed. Many modifications and variations will be apparent to those of ordinary skill in the art without departing from the scope and spirit of the described embodiments. The terminology used herein was chosen to best explain the principles of the

embodiments, the practical application or technical improvement over technologies found in the marketplace, or to enable others of ordinary skill in the art to understand the embodiments disclosed herein.

[0100] Referring now to FIG. 16, in one example, a computer program product 1600 includes, for instance, one or more storage media 1602, wherein the media may be tangible and/or non-transitory, to store computer readable program code means or logic 1604 thereon to provide and facilitate one or more aspects of embodiments described herein.

[0101] The present invention may be a system, a method, and/or a computer program product. The computer program product may include a computer readable storage medium (or media) having computer readable program instructions thereon for causing a processor to carry out aspects of the present invention.

[0102] The computer readable storage medium can be a tangible device that can retain and store instructions for use by an instruction execution device. The computer readable storage medium may be, for example, but is not limited to, an electronic storage device, a magnetic storage device, an optical storage device, an electromagnetic storage device, a semiconductor storage device, or any suitable combination of the foregoing. A non-exhaustive list of more specific examples of the computer readable storage medium includes the following: a portable computer diskette, a hard disk, a random access memory (RAM), a read-only memory (ROM), an erasable programmable read-only memory (EPROM or Flash memory), a static random access memory (SRAM), a portable compact disc read-only memory (CD-ROM), a digital versatile disk (DVD), a memory stick, a floppy disk, a mechanically encoded device such as punch-cards or raised structures in a groove having instructions recorded thereon, and any suitable combination of the foregoing. A computer readable storage medium, as used herein, is not to be construed as being transitory signals *per se*, such as radio waves or other freely propagating electromagnetic waves, electromagnetic waves propagating through a waveguide or other transmission media (e.g., light pulses passing through a fiber-optic cable), or electrical signals transmitted through a wire.

[0103] Computer readable program instructions described herein can be downloaded to respective computing/processing devices from a computer readable storage medium or to an external computer or external storage device via a network, for example, the Internet, a local area network, a wide area network and/or a wireless network. The network may comprise copper transmission cables, optical transmission fibers, wireless transmission, routers, firewalls, switches, gateway computers and/or edge servers. A network adapter card or network interface in each computing/processing device receives computer readable program instructions from the network and forwards the computer readable program instructions for storage in a computer readable storage medium within the respective computing/processing device.

[0104] Computer readable program instructions for carrying out operations of the present invention may be assembler instructions, instruction-set-architecture (ISA) instructions, machine instructions, machine dependent instructions, microcode, firmware instructions, state-setting data, or either source code or object code written in any combination of one or more programming languages, including an object oriented programming language such as Smalltalk, C++ or the like, and conventional procedural programming languages, such as the "C" programming language or similar programming languages. The computer readable program instructions may execute entirely on the user's computer, partly on the user's computer, as a stand-alone software package, partly on the user's computer and partly on a remote computer or entirely on the remote computer or server. In the latter scenario, the remote computer may be connected to the user's computer through any type of network, including a local area network (LAN) or a wide area network (WAN), or the connection may be made to an external computer (for example, through the Internet using an Internet Service Provider). In some embodiments, electronic circuitry including, for example, programmable logic circuitry, field-programmable gate arrays (FPGA), or programmable logic arrays (PLA) may execute the computer readable program instructions by utilizing state information of the computer readable program instructions to personalize the electronic circuitry, in order to perform aspects of the present invention.

[0105] Aspects of the present invention are described herein with reference to flowchart illustrations and/or block diagrams of methods, apparatus (systems), and computer program

products according to embodiments of the invention. It will be understood that each block of the flowchart illustrations and/or block diagrams, and combinations of blocks in the flowchart illustrations and/or block diagrams, can be implemented by computer readable program instructions.

[0106] These computer readable program instructions may be provided to a processor of a general purpose computer, special purpose computer, or other programmable data processing apparatus to produce a machine, such that the instructions, which execute via the processor of the computer or other programmable data processing apparatus, create means for implementing the functions/acts specified in the flowchart and/or block diagram block or blocks. These computer readable program instructions may also be stored in a computer readable storage medium that can direct a computer, a programmable data processing apparatus, and/or other devices to function in a particular manner, such that the computer readable storage medium having instructions stored therein comprises an article of manufacture including instructions which implement aspects of the function/act specified in the flowchart and/or block diagram block or blocks.

[0107] The computer readable program instructions may also be loaded onto a computer, other programmable data processing apparatus, or other device to cause a series of operational steps to be performed on the computer, other programmable apparatus or other device to produce a computer implemented process, such that the instructions which execute on the computer, other programmable apparatus, or other device implement the functions/acts specified in the flowchart and/or block diagram block or blocks.

[0108] The flowchart and block diagrams in the Figures illustrate the architecture, functionality, and operation of possible implementations of systems, methods, and computer program products according to various embodiments of the present invention. In this regard, each block in the flowchart or block diagrams may represent a module, segment, or portion of instructions, which comprises one or more executable instructions for implementing the specified logical function(s). In some alternative implementations, the functions noted in the block may occur out of the order noted in the figures. For example, two blocks shown in succession may, in fact, be executed substantially concurrently, or the blocks may sometimes

be executed in the reverse order, depending upon the functionality involved. It will also be noted that each block of the block diagrams and/or flowchart illustration, and combinations of blocks in the block diagrams and/or flowchart illustration, can be implemented by special purpose hardware-based systems that perform the specified functions or acts or carry out combinations of special purpose hardware and computer instructions.

CLAIMS

1. A computer system for dispatching a multithreaded guest virtual machine (VM), the computer system comprising:

a configuration comprising a machine that includes a single physical core enabled to operate in a single thread (ST) mode and a multithreading (MT) mode, wherein the single physical core includes multiple physical threads, and

wherein the machine is configured to perform a method comprising:

issuing a start-virtual-execution (start-VE) instruction to dispatch a guest entity comprising multiple logical threads on the single physical core, wherein the guest entity includes all or a part of a guest VM, and wherein the issuing of the start-VE instruction is performed by a host running on one of the multiple physical threads on the single physical core in the ST mode; and

executing the start-VE instruction by the machine, wherein the executing of the start-VE instruction by the machine comprises:

mapping each of the multiple logical threads to a corresponding one of the multiple physical threads;

initializing each of the multiple mapped physical threads with a state of a corresponding logical thread; and

starting execution of the guest entity on the single physical core in the MT mode.

2. The computer system of claim 1, wherein only multiple valid logical threads are mapped to a corresponding physical thread.

3. The computer system of claim 1, wherein the initializing includes:

loading, for each of the multiple logical threads, a portion of the state of the corresponding logical thread into the corresponding physical thread; and

completing, by each mapped physical thread, the loading of the rest of the state of the corresponding logical thread, wherein the rest of the state of the corresponding logical thread is determined based on the portion of the state of the corresponding logical thread.

4. The computer system of claim 3, wherein the portion of the state of the corresponding logical thread includes an address of a state description of the rest of the state of the corresponding logical thread.
5. The computer system of claim 3, wherein one of the multiple physical threads performs the loading, for each of the multiple logical threads, a portion of the state of the corresponding logical thread.
6. The computer system of claim 1, wherein the executing further comprises verifying that the multiple mapped physical threads are not currently running any program code or internal code prior to the initializing.
7. The computer system of claim 1, wherein the starting execution is performed for each of the multiple physical threads based on the initializing completing for the physical thread.
8. The computer system of claim 1, wherein the starting execution is performed for each of the multiple physical threads based on the initializing completing for all the multiple physical threads.
9. The computer system of claim 1, wherein the executing further comprises:
determining that software is running on one of the multiple mapped physical threads; and
based on the determining, completing the start-VE instruction and notifying the host that the start-VE instruction failed.
10. The computer system of claim 1, wherein the executing further comprises:
determining that internal firmware is running on one of the multiple mapped physical threads; and
based on the determining, performing at least one of:
nullifying the start-VE instruction and returning control to the host; and

waiting for the mapped physical thread to become available.

11. A computer implemented method for dispatching a multithreaded guest virtual machine (VM) in a configuration comprising a machine that includes a single physical core enabled to operate in a single thread (ST) mode and a multithreading (MT) mode, wherein the single physical core includes multiple physical threads, and wherein the machine is configured to perform a method comprising:

issuing a start-virtual-execution (start-VE) instruction to dispatch a guest entity comprising multiple logical threads on the single physical core, wherein the guest entity includes all or a part of a guest VM, and wherein the issuing of the start-VE instruction is performed by a host running on one of the multiple physical threads on the single physical core in the ST mode; and

executing the start-VE instruction by the machine, wherein the executing of the start-VE instruction by the machine comprises:

mapping each of the multiple logical threads to a corresponding one of the multiple physical threads;

initializing each of the multiple mapped physical threads with a state of a corresponding logical thread; and

starting execution of the guest entity on the single physical core in the MT mode.

12. The computer implemented method of claim 11, wherein only multiple valid logical threads are mapped to a corresponding physical thread.

13. The computer implemented method of claim 11, wherein the initializing includes:

loading, for each of the multiple logical threads, a portion of the state of the corresponding logical thread into the corresponding physical thread; and

completing, by each mapped physical thread, the loading of the rest of the state of the corresponding logical thread, wherein the rest of the state of the corresponding logical thread is determined based on the portion of the state of the corresponding logical thread.

14. The computer implemented method of claim 11, wherein one of the multiple physical threads performs the loading, for each of the multiple logical threads, a portion of the state of the corresponding logical thread.

15. The computer implemented method of claim 11, wherein the executing further comprises verifying that the multiple mapped physical threads are not currently running any program code or internal code prior to the initializing.

16. The computer implemented method of claim 11, wherein the starting execution is performed for each of the multiple physical threads based on the initializing completing for the physical thread.

17. A computer program product for dispatching a multithreaded guest virtual machine (VM) in a configuration comprising a machine that includes a single physical core enabled to operate in a single thread (ST) mode and a multithreading (MT) mode, wherein the single physical core includes multiple physical threads, and wherein the computer program product comprises:

- a computer readable storage medium having program instructions embodied therewith, wherein the computer readable storage medium is not a signal, and wherein the program instructions are readable by a processing circuit to cause the processing circuit to perform a method comprising:

- issuing a start-virtual-execution (start-VE) instruction to dispatch a guest entity comprising multiple logical threads on the single physical core, wherein the guest entity includes all or a part of a guest VM, and wherein the issuing of the start-VE instruction is performed by a host running on one of the multiple physical threads on the single physical core in the ST mode; and

- executing the start-VE instruction by the machine, wherein the executing of the start-VE instruction by the machine comprises:

- mapping each of the multiple logical threads to a corresponding one of the multiple physical threads;

- initializing each of the multiple mapped physical threads with a state of a corresponding logical thread; and

- starting execution of the guest entity on the single physical core in the MT mode.

18. The computer program product of claim 17, wherein only multiple valid logical threads are mapped to a corresponding physical thread.

19. The computer program product of claim 17, wherein the initializing includes:

loading, for each of the multiple logical threads, a portion of the state of the corresponding logical thread into the corresponding physical thread; and

completing, by each mapped physical thread, the loading of the rest of the state of the corresponding logical thread, wherein the rest of the state of the corresponding logical thread is determined based on the portion of the state of the corresponding logical thread.

20. The computer program product of claim 17, wherein the starting execution is performed for each of the multiple physical threads based on the initializing completing for the physical thread.