



- (51) International Patent Classification:
G06F 1/32 (2006.01) *G06F 5/16* (2006.01)
- (21) International Application Number:
PCT/US2017/017023
- (22) International Filing Date:
8 February 2017 (08.02.2017)
- (25) Filing Language: English
- (26) Publication Language: English
- (30) Priority Data:
15/087,095 31 March 2016 (31.03.2016) US
- (71) Applicant: INTEL CORPORATION [US/US]; 2200 Mission College Boulevard, Santa Clara, California 95054 (US).
- (72) Inventors; and
- (71) Applicants : BODAS, Devadatta V. [US/US]; 35224 7th Avenue SW, Federal Way, Washington 98023 (US). RA-JAPPA, Muralidhar [US/US]; 3195 W. Tyson Place, Chandler, Arizona 85226 (US). SONG, Justin J. [CN/US]; 6923 47th Loop NE, Olympia, Washington

98516 (US). HOFFMAN, Andy [US/US]; 231 Clay Court SE, Olympia, Washington 98513 (US).

(74) Agents: MALLIE, Michael J. et al.; Blakely, Sokoloff, Taylor & Zafman LLP, 1279 Oakmead Parkway, Sunnyvale, California 94085 (US).

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

(84) Designated States (unless otherwise indicated, for every kind of regional protection available): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU,

[Continued on next page]

(54) Title: METHOD AND APPARATUS TO IMPROVE ENERGY EFFICIENCY OF PARALLEL TASKS

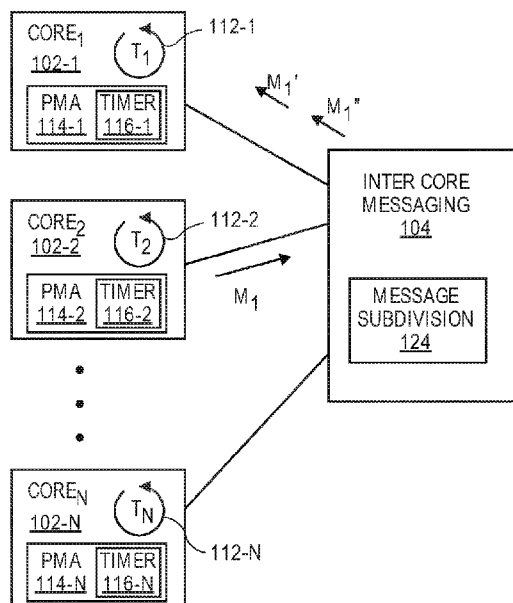


FIG. 1

(57) Abstract: A system with improved power performance for task executed in parallel. A plurality of processing cores each to execute tasks. An inter-core messaging unit to conveys messages between the cores. A power management agent transitions a first core into a lower power state responsive to the first core waiting for a second core to complete a second task. In some embodiments long messages are subdivided to allow a receiving core to resume useful work sooner.



TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:

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

METHOD AND APPARATUS TO IMPROVE ENERGY EFFICIENCY OF PARALLEL TASKS

FIELD

Embodiments of the invention relate to high-performance computing. More
5 specifically, embodiments of the invention relate to improving power consumption
characteristics in system executing parallel tasks.

BACKGROUND

Generally, power consumption has become an important issue in high-
performance computing (HPC). Typical HPC environments divide the processing task
10 between a number of different computing cores so these tasks can be performed in parallel.
At different points, data is required to be exchanged between the different tasks. Such times
are generally referred to as “synchronization points” because they require that the tasks be
synchronized, that is, have reached the same point in execution so that the exchanged data is
valid. Because all tasks do not require the same amount of time to reach the synchronization
15 point, early arrivers must wait for the other tasks to get to that synchronization point.
Generally, the task calls a wait routine and executes a spin loop until other tasks arrive at the
same synchronization point. Unfortunately, in the spin loop, the core continues to consume
significant energy.

BRIEF DESCRIPTION OF THE DRAWINGS

20 Embodiments of the invention are illustrated by way of example and not by
way of limitation in the figures of the accompanying drawings in which like references
indicate similar elements. It should be noted that different references to "an" or "one"
embodiment in this disclosure are not necessarily to the same embodiment, and such
references mean at least one.

25 **Figure 1** is a block diagram of a system according to one embodiment of
the invention.

Figures 2A-2D show timing diagrams of operation according to
embodiments of the invention.

Figure 3 is a flow diagram of operation of a system according to one embodiment of the invention.

DETAILED DESCRIPTION

Figure 1 is a block diagram of a system according to one embodiment of the invention. A plurality of processing cores 102-1, 102-2, 102-n (generically, core 102) are provided to process tasks in parallel. For example, core 102-1 processes task 112-1, core 102-2 processes task 112-2, and core 102-N processes task 112-N. “Task,” “thread” and “process” are use herein interchangeably to refer to and instance of executable software or hardware. The number of cores 102 can be arbitrarily large. Each core 120 includes a corresponding power management agent 114-1, 114-2-, 114-N (generically, power management agent 114). Power management agent 114 may be implemented as software, hardware, microcode etc. The power management agent 114 is used to place its core 120 in a lower power state when it reaches a synchronization point before other cores 120 processing other tasks 112. As used herein, “synchronization point” refers to any point in the processing where the further processing is dependent on receipt of data from another core in the system.

An inter-core messaging unit 104 provides messaging services between the different processing cores 120. In one embodiment, inter-core messaging unit 104 adheres to the message passing interface (MPI) protocol. When core 102-1 reaches a synchronization point, it calls a wait routine. For example, it may call MPI-wait from the inter-core messaging unit 104. In one embodiment, responsive to the call of the wait messaging routine, the power management agent 114-1 transitions core 102-1 into a lower power state. This may take the form of reducing core and/or its power domain power by employing whatever applicable power saving technology such as DVFS (dynamic voltage frequency scaling), gating, parking, offlining, throttling, non-active states, or standby states. In other embodiments, the power management agent 114 includes or has access to a timer 116-1, 116-2, 116-N, respectively (generically, timer 116) that delays entry into the low power state for a threshold period. Generally, there is a certain amount of overhead in entering and leaving the low power state. It has been found empirically that if all cores reach a synchronization point within a relatively short period of time, power consumption characteristics are not meaningfully improved, and in some cases, are diminished by immediate transition upon the call of the wait routine. However, since tasks 120 may execute for minutes, hours, or even longer beyond some relatively short threshold, the power savings of transitioning to a lower power state are quite

significant. In embodiments where the timer 116-1 is present, the core 102-1 will still enter the spin loop until transitioned into a lower power state. As used herein, “spin loop” refers to either a legacy spin-loop (checking one flag and immediately going to itself) or any other low latency state which can be immediately exited once a condition that caused a thread/core to wait has been met.

In one example, core 102-1 may be waiting for a message M_1 from core 102-2. In one embodiment, message M_1 is sent to inter-core messaging unit 104. If message M_1 exceeds a threshold length, inter-core messaging unit 104 subdivides the message into two submessages using a message subdivision unit 124. Submessages M_1' and M_1'' are sent sequentially to core 102-1. Each submessage includes its own message validation value, such as cyclic redundancy check (CRC) values, to allow the submessage to be validated individually. By subdividing the message, power savings can be achieved while improving processing performance. This is because power management agent 114-1 can transition core 102-1 into a higher power state once message M_1' is received and validated without waiting for the entire message (the remainder M_1'') to be received. Thus, core 102-1 exits the spin loop or enters the higher power state sooner so there is less power churn, and begins processing message M_1' while receiving message M_1'' . Of course, if M_1'' fails to validate, core 102-1 will need to invalidate message M_1' and request retransmission of the entire M_1 message, but as message failure transmissions are relatively infrequent, improved power savings and execution by virtue of the message subdivision generally results.

Figures 2A-2D show timing diagrams of operation according to embodiments of the invention. In **Figure 2A**, four tasks, task 1, task 2, task 3 and task 4 are shown as part of the execution environment. As shown, each of tasks 1, 2 and 3 finish (reach a synchronization point) before task 4. In this example, each thread executing the corresponding task transitions to a lower power state immediately when it reaches a respective synchronization point. **Figure 2B** is the same as Figure 2A, except that tasks 1, 2 and 3 each wait a hold off delay before entering the lower power state. During the delay, each task enters a spin loop during the delay.

Figure 2C and **2D** show behavior of the system with short and long messages respectively. Empirically message traffic tends to be bimodal characterized by either short or long messages. Where the messages are long message division can provide additional power savings. **Figure 2C** shows a receiving task R waiting for a sending task S to

send a message. When the message is receiving and validated, it exits the low power state, and the receiving task, task R, resumes. However, there is a finite delay to exit the low power state after the message has been received and validated. This is reflective of appropriate behavior when the message is relatively short.

5 **Figure 2D** shows an embodiment for messages subdivided into two submessages, MSG 1 and MSG 2. This allows task R to initiate the transition from the low power state upon validation of MSG 1. This allows task R to resume processing and begin processing of MSG 1 while receiving MSG 2, thereby improving performance. Even in a system where task R is merely residing in a spin loop, this message subdivision can improve power because the time spent in the spin loop (not performing any useful work) is reduced over systems in which task R waits in a spin loop for the receipt of the entire lengthy message (here, the composition of MSG 1 and MSG2). This behavior is suitable where the message is long.

15 **Figure 3** is a flow diagram of operation of a system according to one embodiment of the invention. At block 302, a core completes its task (arrives at a synchronization point) and enters a wait condition. At block 304, the core notifies a messaging unit that it is waiting. At block 306, a delay timer is triggered to hold off entry into a lower power state. Some embodiments may omit the delay timer or have the delay set to zero. At decision block 308, a determination is made if the delay threshold has been achieved. As noted previously, system designers may select the delay threshold based on the overhead of entry and leaving the low power state. If the delay threshold has been achieved, the core is transitioned into the lower power state (LPS) at block 310.

25 At decision block 312, a determination is made whether there is a long message directed at the waiting core. Empirically, as noted above, it has been found that most messages fall into a bimodal length distribution, that is, most messages are either very short, or quite long. If the message is not long, it is processed normally at block 314, that is, the message is not subdivided and is merely sent as a unit. Then, at block 316, the core transitions to the higher power state or active state at block 316 once the entire message has been validated.

30 Conversely, at block 318, if the message is a “long message,” the message is subdividing into a first and second submessage, each with its validation values. At block 320, the core receives and validates the first submessage. A determination is made at block 322 if

the first submessage is valid. If the first submessage is not valid, the core remains in a lower power state and waits for subsequent valid message receipt. If the first message is valid, upon validating that first message, the core transitions into a higher power/active state, that is, it goes to a higher power, possibly C0 state, or exits a spin loop, for example. Then, at block 5 326, the core processes the submessage while receiving the second submessage. At block 328, a determination is made if the second submessage is valid. If the second submessage is not valid, the core invalidates both submessages and requests they be resent at block 330. If, however, the second submessage is valid (the usual case), it continues processing in the normal manner at block 332.

10 The following examples pertain to further embodiments. The various features of the different embodiments may be variously combined with some features included and others excluded to suit a variety of different applications. Some embodiments pertain to a multi-processing-core system in which a plurality of processing cores are each to execute a respective task. An inter-core messaging unit is to convey messages between the 15 cores. A power management agent, in use, transitions a first core into a lower power state responsive to the first core waiting for a second core to complete a second task.

In further embodiments, the system uses a delay timer to hold off a transitioning into the lower power state for a defined time after the first core begins waiting.

20 In further embodiments, the the messaging unit segments a message into a first submessage and a second submessage for transmission to the first core and the power management agent transitions the first core to a higher power state responsive to receipt and validation of the first submessage.

In further embodiments, the first core begins processing the first submessage while the second submessage is being received.

25 Some embodiments pertain to a method to reduce power consumption in a multi-core execution environment. A first core is transitioned into a reduced power state responsive to a task executing on the core reaching a synchronization point, a threshold before a second task executing on a second core reaches the synchronization point. The first core is returned to a higher power state responsive to the second task reaching the synchronization 30 point.

In further embodiments, a delay timer is initiated to hold off the transition until after the threshold.

In further embodiments, the delay timer is triggered based on a call of a messaging wait routine from the first task.

5 In further embodiments, when a message is to be sent to the first task in the reduced power state, the message is subdivided into a first submessage and a second submessage, each with a correction code value. The first submessage and the second submessage are sent to the first core. The return to the higher power state is initiated once the first submessage is validated.

10 In further embodiments, the first submessage begins processing while receiving the second submessage and the first submessage is invalidated responsive to a validation failure of the second submessage.

15 Some embodiments pertain to a method of reducing power consumption into a multi-core messaging. A first task is placed in a spin loop to wait for a message from a second task. The message is subdivided into a first submessage and a second submessage. The first task exits the spin loop responsive to a validation of the first submessage.

In further embodiments, the first submessage is processed in the first task while receiving the second submessage and invalidate responsive to a validation failure in the second submessage.

20 Some embodiments pertain to a non-transitory computer-readable medium having instructions stored therein, which when executed by a processor, cause the processor to perform a set of operations to reduce power consumption in a multi-core execution environment. A first core is transitioned into a reduced power state responsive to a task executing on the core reaching a synchronization point, a threshold before a second task
25 executing on a second core reaches the synchronization point. The first core is returned to a higher power state responsive to the second task reaching the synchronization point.

In further embodiments, a delay timer is initiated to hold off the transition until after the threshold.

In further embodiments, the delay timer is triggered based on a call of a messaging wait routine from the first task.

5 In further embodiments, when a message is to be sent to the first task in the reduced power state, the message is subdivided into a first submessage and a second submessage, each with a correction code value. The first submessage and the second submessage are sent to the first core. The return to the higher power state is initiated once the first submessage is validated.

10 In further embodiments, the first submessage begins processing while receiving the second submessage and the first submessage is invalidated responsive to a validation failure of the second submessage.

Some embodiments pertain to high-performance computing system having a plurality of processing cores. The system includes means for inter-core messaging and means for reducing power consumption on a first core when the first core is waiting for a second core.

15 In further embodiments, the means for inter-core messaging has means for subdividing a message into a first submessage and a second submessage, and wherein a receiving core begins processing of the first submessage before the second submessage is fully received.

20 In further embodiments, the means for reducing power consumption includes means for reducing a clock frequency in the processing core.

In further embodiments, the means for reducing power consumption includes means for transitioning the waiting core into a lower power state.

25 While embodiments of the invention are discussed above in the context of flow diagrams reflecting a particular linear order, this is for convenience only. In some cases, various operations may be performed in a different order than shown or various operations may occur in parallel. It should also be recognized that some operations described with respect to one embodiment may be advantageously incorporated into another embodiment. Such incorporation is expressly contemplated.

Elements of embodiments of the present invention may also be provided as

a machine-readable medium for storing the machine-executable instructions. The machine-readable medium may include, but is not limited to, flash memory, optical disks, compact disks read only memory (CD-ROM), digital versatile/video disks (DVD) ROM, random access memory (RAM), erasable programmable read-only memory (EPROM), electrically erasable programmable read-only memory (EEPROM), magnetic or optical cards.

In the foregoing specification, the invention has been described with reference to the specific embodiments thereof. It will, however, be evident that various modifications and changes can be made thereto without departing from the broader spirit and scope of the invention as set forth in the appended claims. The specification and drawings are, accordingly, to be regarded in an illustrative rather than a restrictive sense.

CLAIMS

What is claimed is:

1. A multi-processing-core system:
a plurality of processing cores, each core to execute a task;
5 an inter-core messaging unit to convey messages between the cores; and
a power management agent to transition a first core into a lower power state responsive to
the first core waiting for a second core to complete a second task.
2. The system of claim 1, further comprising:
10 a delay timer to hold off a transitioning into the lower power state for a defined time after
the first core begins waiting.
3. The system of claim 2, wherein the messaging unit segments a message into a first
submessage and a second submessage for transmission to the first core, and wherein the power
15 management agent transitions the first core to a higher power state responsive to receipt and
validation of the first submessage.
4. The system of claim 3, wherein the first core begins processing the first submessage
while the second submessage is being received.
20
5. A method to reduce power consumption in a multi-core execution environment, the
method comprising:
transitioning a first core into a reduced power state responsive to a task executing on the
core reaching a synchronization point, a threshold before a second task executing on a second
25 core reaches the synchronization point; and
returning the first core to a higher power state responsive to the second task reaching the
synchronization point.
6. The method of claim 5, further comprising:
30 initiating a delay timer to hold off the transition until after the threshold.
7. The method of claim 6, further comprising:
triggering the delay timer based on a call of a messaging wait routine from the first task.

8. The method of any of claims 5-7, wherein a message is to be sent to the first task in the reduced power state, the method further comprising:

subdividing the message into a first submessage and a second submessage, each with a
5 correction code value;
sending the first submessage and the second submessage to the first core; and
initiating the return to the higher power state once the first submessage is validated.

9. The method of claim 8, further comprising:

10 processing the first submessage while receiving the second submessage; and
invalidating the first submessage responsive to a validation failure of the second
submessage.

10. A method of reducing power consumption into a multi-core messaging system:

15 placing a first task in a spin loop to wait for a message from a second task;
subdividing the message into a first submessage and a second submessage; and
exiting the spin loop in the first task responsive to a validation of the first submessage.

11. The method of claim 10, further comprising:

20 processing the first submessage in the first task while receiving the second submessage;
and
invalidating the first submessage responsive to a validation failure in the second
submessage.

25 12. A non-transitory computer-readable medium having instructions stored therein, which
when executed by a processor, cause the processor to perform a set of operations to reduce
power consumption in a multi-core execution environment, comprising:

transitioning a first core into a reduced power state responsive to a task executing on the
core reaching a synchronization point, a threshold before a second task executing on a second
30 core reaches the synchronization point; and

returning the first core to a higher power state responsive to the second task reaching the
synchronization point.

13. The non-transitory computer-readable medium of claim 12, wherein the instructions cause the processor to perform a set of operations further comprising:

initiating a delay timer to hold off the transition until after the threshold.

5 14. The non-transitory computer-readable medium of claim 13, wherein the instructions cause the processor to perform a set of operations further comprising:

triggering the delay timer based on a call of a messaging wait routine from the first task.

15. The non-transitory computer-readable medium of any of claim 12-14, wherein the
10 instructions cause the processor to perform a set of operations further comprising:

sending a message to the first task in the reduced power state;

subdividing the message to a first submessage and a second submessage, each with a correction code value;

sending the first submessage and the second submessage to the first core; and

15 initiating the return to the higher power state once the first submessage is validated.

16. The non-transitory computer-readable medium of claim 15, wherein the instructions cause the processor to perform a set of operations further comprising:

processing the first submessage while receiving the second submessage; and

20 invalidating the first submessage responsive to a validation failure of the second submessage.

17. A high-performance computing system comprising:

a plurality of processing cores;

25 means for inter-core messaging; and

means for reducing power consumption on a first core when the first core is waiting for a second core.

18. The system of claim 17, wherein the means for inter-core messaging comprises:

30 means for subdividing a message into a first submessage and a second submessage, and wherein a receiving core begins processing of the first submessage before the second submessage is fully received.

19. The system of either of claim 17 or 18, wherein the means for reducing power consumption comprises:

means for reducing a clock frequency in the processing core.

5 20. The system of either of claim 17 or 18, wherein the means for reducing power consumption comprises:

means for transitioning the waiting core into a lower power state.

1/4

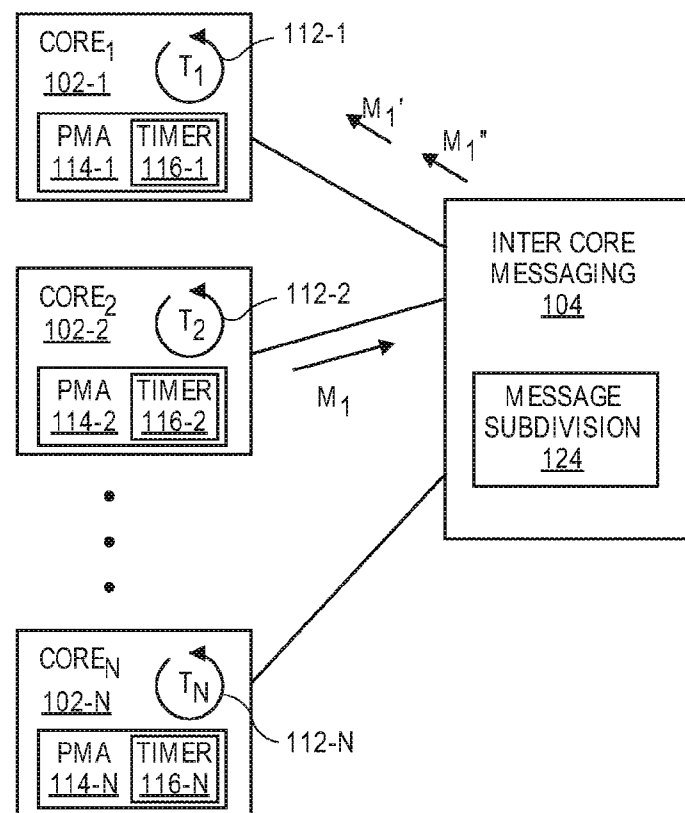
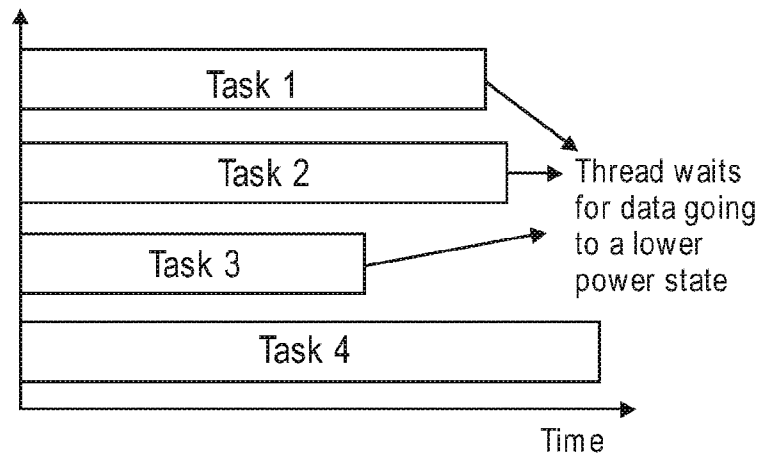
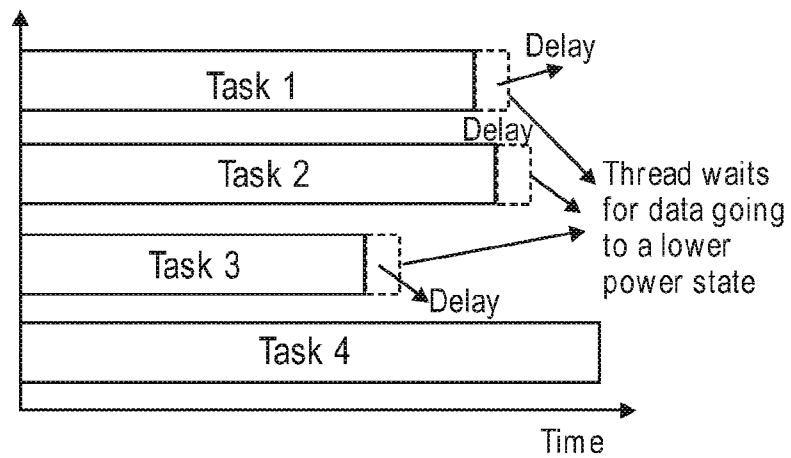


FIG. 1

2/4

**FIG. 2A****FIG. 2B**

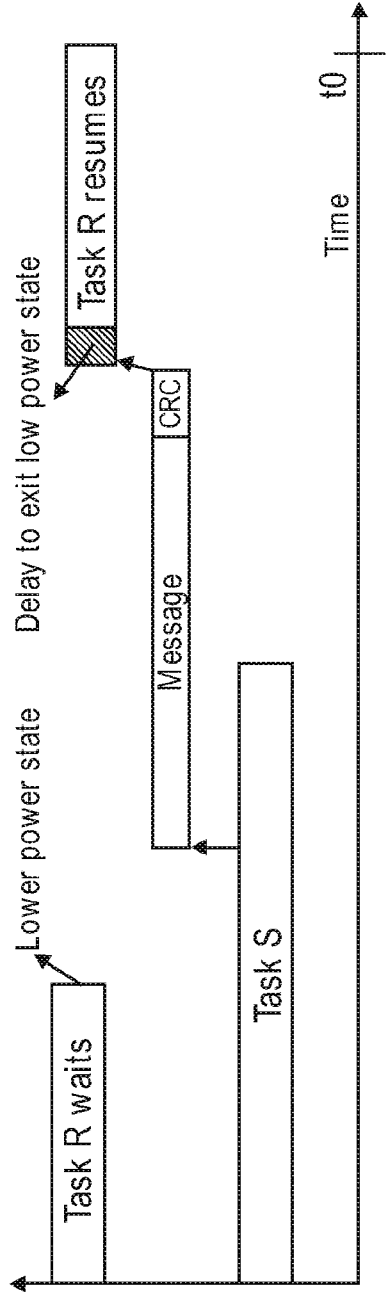


FIG. 2C

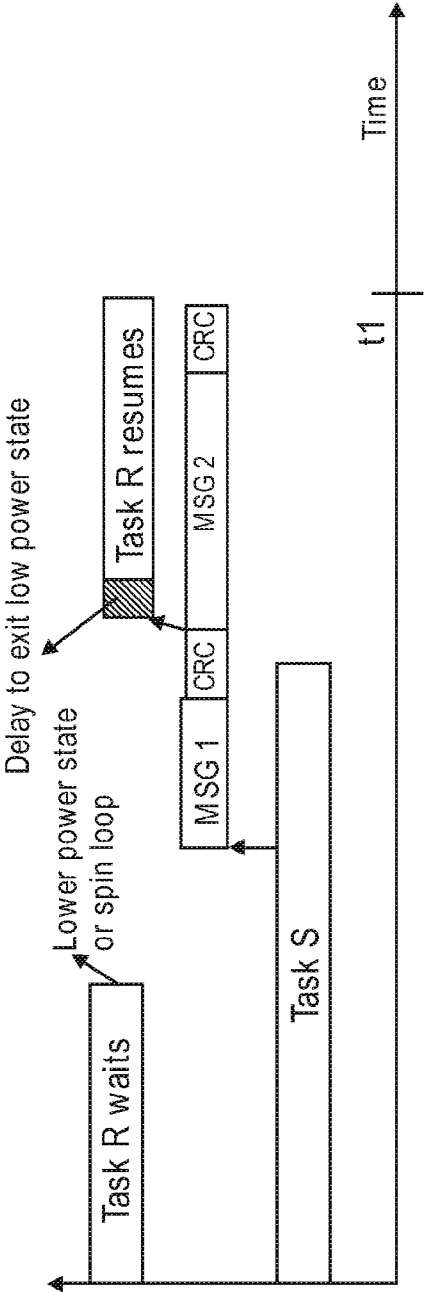
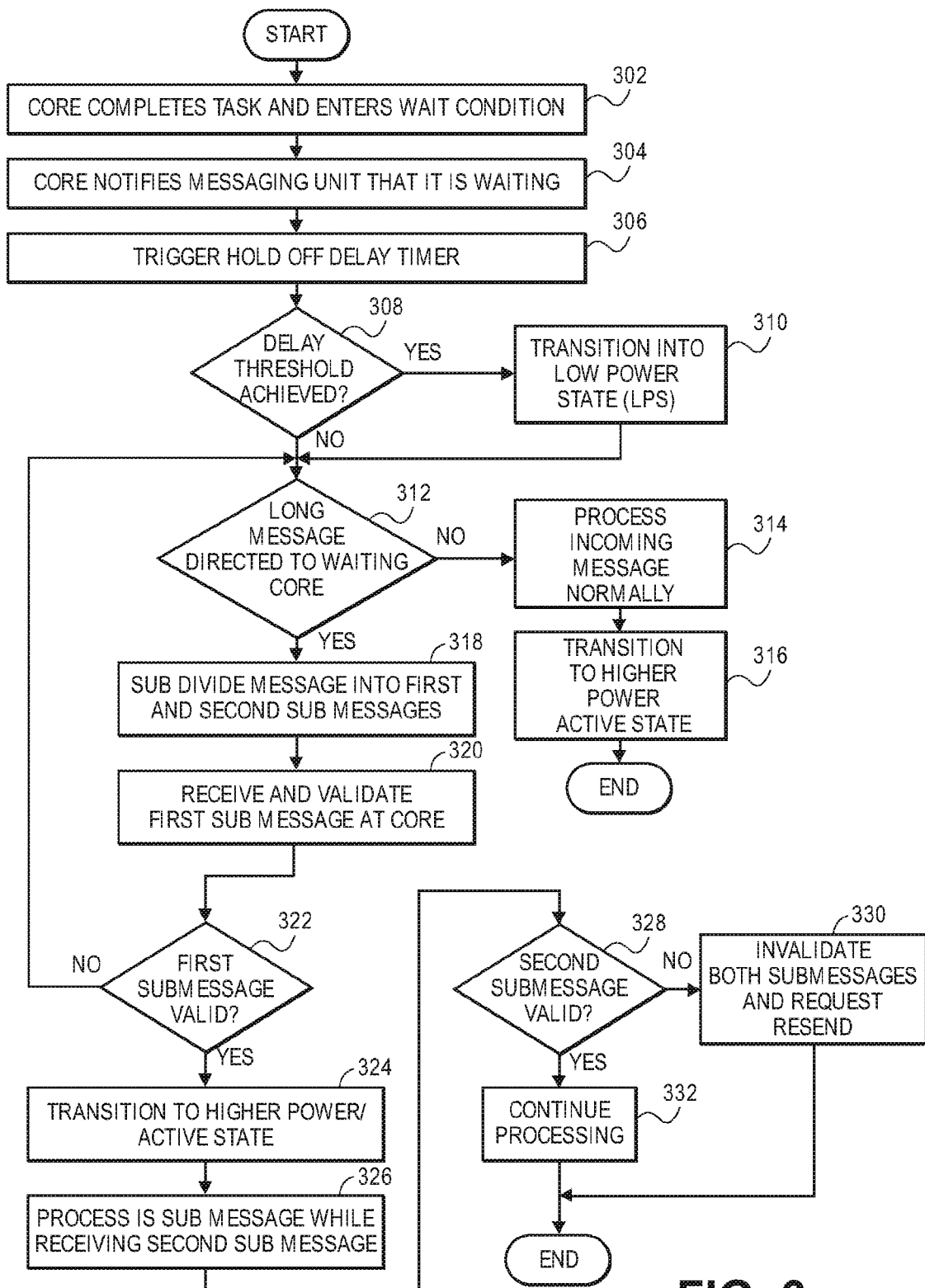


FIG. 2D

**FIG. 3**

A. CLASSIFICATION OF SUBJECT MATTER**G06F 1/32(2006.01)i, G06F 5/16(2006.01)i**

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

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F 1/32; G06F 1/26; G06F 9/46; G06F 13/00; G06F 13/12; G06F 5/16

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

Korean utility models and applications for utility models

Japanese utility models and applications for utility models

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

eKOMPASS(KIPO internal) & keywords: processing cores, power management, lower power state, waiting

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2015-0067356 A1 (ADVANCED MICRO DEVICES, INC.) 05 March 2015 See paragraphs [0003]-[0004], [0021]-[0030]; claims 1-19; and figures 3-4.	5, 12
Y		1-4, 6-11, 13-20
Y	US 8578079 B2 (APPLE INC.) 05 November 2013 See column 1, line 61 - column 2, line 3; column 3, line 34 - column 6, line 43; and figure 1.	1-4, 6-11, 13-16
Y	US 2010-0169528 A1 (AMIT KUMAR et al.) 01 July 2010 See paragraphs [0041]-[0066], [0125]-[0128], [0141]; and figures 2, 13A, 17.	3-4, 8-11, 15-20
A	US 2015-0113304 A1 (WISCONSIN ALUMNI RESEARCH FOUNDATION) 23 April 2015 See paragraphs [0056]-[0063]; and figures 2-3.	1-20
A	US 2013-0047011 A1 (DAVID DICE et al.) 21 February 2013 See paragraphs [0028]-[0032]; and figure 1.	1-20



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents:

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

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

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

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

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

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

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

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

"&" document member of the same patent family

Date of the actual completion of the international search

22 May 2017 (22.05.2017)

Date of mailing of the international search report

24 May 2017 (24.05.2017)

Name and mailing address of the ISA/KR

International Application Division

Korean Intellectual Property Office

189 Cheongsa-ro, Seo-gu, Daejeon, 35208, Republic of Korea

Facsimile No. +82-42-481-8578

Authorized officer

Lee, Eun Kyu

Telephone No. +82-42-481-3580



INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No.

PCT/US2017/017023

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2015-0067356 A1	05/03/2015	None	
US 8578079 B2	05/11/2013	CN 102918474 A CN 102918474 B EP 2261771 A2 EP 2261771 A3 JP 2012-527042 A JP 5491622 B2 KR 10-1165151 B1 KR 10-2010-0122875 A US 2010-0293401 A1 US 2012-0167107 A1 US 2013-0067257 A1 US 8156275 B2 US 8332559 B2 WO 2010-132298 A2 WO 2010-132298 A3	06/02/2013 12/10/2016 15/12/2010 22/04/2015 01/11/2012 14/05/2014 12/07/2012 23/11/2010 18/11/2010 28/06/2012 14/03/2013 10/04/2012 11/12/2012 18/11/2010 20/03/2014
US 2010-0169528 A1	01/07/2010	US 8645596 B2 WO 2010-078017 A2 WO 2010-078017 A3	04/02/2014 08/07/2010 26/08/2010
US 2015-0113304 A1	23/04/2015	US 9519330 B2	13/12/2016
US 2013-0047011 A1	21/02/2013	US 8775837 B2	08/07/2014