



(51) International Patent Classification:
G06F 1/32 (2006.01)

(21) International Application Number:
PCT/US2012/031751

(22) International Filing Date:
31 March 2012 (31.03.2012)

(25) Filing Language: English

(26) Publication Language: English

(71) Applicant (for all designated States except US): **INTEL CORPORATION** [US/US]; 2200 Mission College Boulevard, MS: RNB-4-150, Santa Clara, California 95052 (US).

(72) Inventor; and

(75) Inventor/Applicant (for US only): **KUMAR, Anil K.** [US/US]; 6312 W. Linda Lane, Chandler, Arizona 85226 (US).

(74) Agents: **JORDAN, B. Delando** et al.; JORDAN IP LAW, PC, c/o CPA Global, P.O. Box 52050, Minneapolis, Minnesota 55402 (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, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, 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, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, 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, ML, MR, NE, SN, TD, TG).

Published:

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

(54) Title: CONTROLLING POWER CONSUMPTION IN MULTI-CORE ENVIRONMENTS

(57) Abstract: Systems and methods of enabling modulation of a frequency of a first core in a multi-core environment include logic to determine a power limit assigned to the first core, logic to determine a stall count of the first core, and logic to modulate the frequency of the first core based at least on the power limit assigned to the first core and the stall count of the first core. The first core is included in a first tile of a socket in the multi-core computer environment.

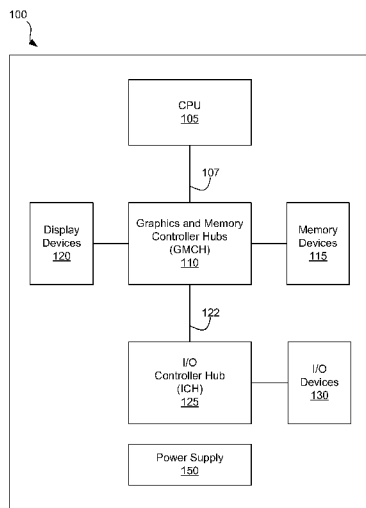


FIG. 1

CONTROLLING POWER CONSUMPTION IN MULTI-CORE ENVIRONMENTS**BACKGROUND**

High performance computing (HPC) and supercomputing environments may require
5 integration of multiple cores. However, power consumption in these environments may be
significant.

BRIEF DESCRIPTION OF THE DRAWINGS

The various advantages of the embodiments of the present invention will become apparent
to one skilled in the art by reading the following specification and appended claims, and by
10 referencing the following drawings, in which:

FIG. 1 is a block diagram that illustrates an example computer system, in accordance with
some embodiments;

FIG. 2 is a block diagram that illustrates an example of a multi-core processor, in
accordance with some embodiments;

15 FIG. 3 is a block diagram that illustrates example of a socket power control unit (PCU), in
accordance with some embodiments;

FIG. 4 is a block diagram that illustrates an example of a core local power unit (CLPU)
that may be used to control a frequency of a core, in accordance with some embodiments; and

FIG. 5 is a flowchart of an example method of modulating the frequency of a core in a tile,
20 in accordance with some embodiments.

DETAILED DESCRIPTION

Embodiments may involve an apparatus that enables modulation of a frequency of a first
core in a multi-core environment, wherein the apparatus may include logic to determine power
limit assigned to a first core, logic to determine a stall count of the first core, and logic to
25 modulate the frequency of the first core based at least on the power limit assigned to the first
core and the stall count of the first core. The first core may be included in a first tile of a socket
in the multi-core computer environment.

Embodiments may involve a system in which a phase locked loop (PLL) is configured to
be associated with a clock signal in a multi-core environment. The system may include a socket
30 coupled with the PLL and be configured to include multiple tiles. At least one of the tiles may
include a first core and a second core. The first core may be configured to include logic to
determine a power limit assigned to a first core, determine a stall count of the first core, and
modulate the frequency of the first core based at least on the power limit assigned to the first
core and the stall count of the first core. The modulation of the frequency of the first core may
35 be performed independently of a frequency of the tiles not associated with the first core.

Embodiments may involve a computer implemented method that provides for modulating a frequency of a core in a first tile of a multi-core environment at least independently of cores in other tiles based at least on an estimated power requirement of the core, a power limit assigned to the core and stall count of the core. The first tile and the other tiles may be associated with a phase locked loop (PLL) of a socket.

Turning to FIG. 1, a block diagram that illustrates an example computer system 100 is shown, in accordance with some embodiments. The computer system 100 may include a central processing unit (CPU) 105, a graphics and memory controller hub (GMCH) 110, and an input/output controller hub (ICH) 125. The GMCH 110 may be coupled to the CPU 105 via a bus 107. The ICH 125 may be coupled to the GMCH 110 via a bus 122. The GMCH 110 may also be coupled to memory devices 115 and display devices 120. The ICH 125 may be coupled to I/O devices 130. The GMCH 110 may include a graphics system 200 (not shown). Although the CPU 105, the GMCH 110 and the ICH 125 may be illustrated as separate components, the functions of two or more of these components may be combined. A power supply 150 may be used to provide power to the computer system 100. The power supply 150 may be a battery or an external power source.

For some embodiments, the CPU 105 may be a multi-core processor. For example, the multi-core processor may be based on the Many Integrated Core (MIC) architecture of Intel Corporation of Santa Clara, California and may be implemented as a PCI Express (Peripheral Component Interconnect Express) card. The computer system 100 may also include many other components; however, for simplicity, they are not shown. For some embodiments, the computer system 100 may be a server computer system.

Turning to FIG. 2, a block diagram that illustrates an example multi-core processor is shown, in accordance with some embodiments. The multi-core processor 200 may include multiple cores 240, 242, 250, 252, 260, 262, 270 and 272 and multiple tiles 205, 215, 225 and 235. Each tile may include two cores. For example, the tile 205 may include the cores 240 and 242. It should be noted that the number of tiles and cores in the multi-core processor 200 may be many more. For example, there may be fifty (50) cores included in twenty five (25) tiles. The cores 240-272 and the tiles 205-235 may be associated with a socket.

Generally, the multi-core processor 200 may be implemented with a single phase locked loop (PLL) 280 providing a common reference signal and therefore the same frequency for all of the tiles 205-235 and cores 240-272. This may limit all of the cores 240-272 to a single frequency and therefore a single performance (P) state. One possible solution to overcome this limitation is to implement one PLL per core or tile. This may enable placing the core 240 of the

tile 205 into one P state (e.g., P0) and the core 250 of the tile 215 into a different P state (e.g., P1). This solution, however, may not be practical when there are design or power constraints.

Turning to FIG. 3, a block diagram that illustrates an example socket power control unit (PCU) is shown, in accordance with some embodiments. The PCU 305 may be configured to assign a power limit 310 to which each of the tiles in the socket may be assigned. The PCU 305 may also assign a thermal limit 315 to the tile. The PCU 305 may include a socket meter 325, which may be configured to receive a power estimate 320 from each of the tiles. For some embodiments, the power limit 310 assigned to a tile may be proportional to a power estimate 320 of the tile. The PCU 305 may be associated with a socket power limit (also referred to as a running average power limit (RAPL)) 330.

The PCU 305 may periodically (e.g., every few milliseconds) re-evaluate the tile power limit 310 based on the RAPL 330. The PCU 305 may be configured to compare the power estimate 320 received from the tile with the assigned power limit 310. For some embodiments, when the power estimate 320 is less than the power limit 310, the PCU 305 may reduce the power limit 310. For some embodiments, when the power estimate 320 is close to the power limit 310 within a predetermined range, the PCU 305 may increase the power limit 310.

Turning to FIG. 4, a diagram of a core local power unit that may be used to control the frequency of a core is shown, in accordance with some embodiments. Since the workloads running on the individual cores may be different, it may be advantageous to be able to control the P state for each core or tile independently of the other cores or tiles in the same socket while using the same PLL for the socket. In this example, core local power unit (CLPU) 400 may be associated with the core 240 (FIG. 3), and the CLPU 401 may be associated with the core 242 (FIG. 3).

The CLPU 400 may include a core energy monitor 405. For some embodiments, the power estimate 320 (shown in FIG. 3) may be determined by the core energy monitor 405 of the CLPU 400 and the core energy monitor (not shown) of the CLPU 401. The CLPU 400 may include Performance/Throttle (P/T) selection logic 410. The P/T selection logic 410 may be configured to control the power consumption of the core 240. This may include the placing the core 240 into different performance (P) states or throttling (T) states.

The P/T selection logic 410 may be configured to modify the clock rate or frequency at which the core 240 may be operating. The P/T selection logic 410 may also control a voltage operating point for the core 240. For example, when placed in a P0 state, the core 240 may operate at a relatively high frequency high performance level and may have more power consumption; when placed in a P1 state, the frequency and performance of the core 240 may be lower and the power consumption may be less; when placed in the T or throttled state, the core

240 may be throttled by modulating the frequency and the power consumption may be at its lowest. Having the core 240 operating at a low frequency level may also reduce the thermal load and cooling requirement associated with the core 240.

The core energy monitor 405 may be configured to receive an activity counter 407 from the core 240 to determine the core energy 420. The activity counter 407 may include information related to a number of times the core 240 is placed in the C0 state, the number of instructions retired, the number of core stalls, etc.

The P/T selection logic 410 may be configured to receive information regarding the core energy 420 from the core energy monitor 405, core stall count 409 from the core 240, thermal limit 315 from the PCU 305, and power limit 310 from the PCU 305. For some embodiments, when the power estimate 320 is determined to be greater than the assigned power limit 310, the CLPU 400 may cause the frequency of the core with the higher core stalls to be modulated. A threshold may be used to determine whether the core stall count 409 is at a level that may affect the modulation of the frequency of the core 240. For example, when the core stalls, it may not perform any instruction. As such, modulating the frequency of the core to a lower frequency may not affect its performance but may reduce its power consumption. The modulation of the frequency may be proportional to the core stall ratio (e.g., stall vs. not stall) and may be bounded by the power limit. The modulation of the frequency of the core may be performed by the core clock modulation module 420. The core clock modulation module 410 may be coupled with the core clock gating control 415. The core clock gating control 415 may be coupled with the PLL 280 (shown in FIG. 2). By modulating the frequency of the cores in the socket based on the estimated power 320 and the power limit 310, the performance and power consumption of the individual cores may be optimized. For some embodiments, the modulation of the frequency of the core may further be based on the assigned thermal limit 315. For example, when it is determined that a temperature of the core is near the assigned thermal limit 315, the P/T selection logic 410 may reduce the frequency of the core. The P/T selection logic 410 may also receive user requirement 490 and Operating System (OS) requirement 495 and use these requirements to determine how to modulate the frequency.

Turning to FIG. 5, an example flow diagram illustrating a process performed by a core local power unit (CLPU), in accordance with some embodiments. The process may correspond to the CLPU 400 managing the power consumption of the core 240. At block 505, the CLPU may determine a power limit assigned to a core. The power limit may be assigned by a PCU 305 (shown in FIG. 4). At block 510, the estimated power required by the core may be determined. The estimated power requirement may be determined by the core energy monitor 405 (shown in FIGS. 4 and 5). At illustrated block 515, the power limit is compared with the estimated power

requirement to determine whether the power assigned to the core by the PCU 305 is appropriate. For example, when the estimated power requirement is much less than the power limit assigned by the PCU 305, the power limit may be reduced. At block 520, a number of core stalls may be determined. At block 525, the frequency of the core may be modulated based on a result of the
5 comparison between the estimated power requirement and the power limit and based on the number of core stalls. For example, the frequency may be decreased when the number of core stalls is higher than a predetermined threshold and the estimated power requirement is less than the power limit. Although not shown in FIG. 5, the frequency of the core may also be modulated based on the thermal limit assigned to the core by the PCU 305.

10 Various embodiments may be implemented using hardware elements, software elements, or a combination of both. Examples of hardware elements may include processors, microprocessors, circuits, circuit elements (e.g., transistors, resistors, capacitors, inductors, and so forth), integrated circuits, application specific integrated circuits (ASIC), programmable logic devices (PLD), digital signal processors (DSP), field programmable gate array (FPGA), logic
15 gates, registers, semiconductor device, chips, microchips, chip sets, and so forth. Examples of software may include software components, programs, applications, computer programs, application programs, system programs, machine programs, operating system software, middleware, firmware, software modules, routines, subroutines, functions, methods, procedures, software interfaces, application program interfaces (API), instruction sets, computing code,
20 computer code, code segments, computer code segments, words, values, symbols, or any combination thereof. Determining whether an embodiment is implemented using hardware elements and/or software elements may vary in accordance with any number of factors, such as desired computational rate, power levels, heat tolerances, processing cycle budget, input data rates, output data rates, memory resources, data bus speeds and other design or performance
25 constraints.

One or more aspects of at least one embodiment may be implemented by representative instructions stored on a machine-readable medium which represents various logic within the processor, which when read by a machine causes the machine to fabricate logic to perform the techniques described herein. Such representations, known as “IP cores” may be stored on a
30 tangible, machine readable medium and supplied to various customers or manufacturing facilities to load into the fabrication machines that actually make the logic or processor.

Example sizes/models/values/ranges may have been given, although embodiments of the present invention are not limited to the same. As manufacturing techniques (e.g., photolithography) mature over time, it is expected that devices of smaller size could be
35 manufactured. In addition, well known power/ground connections to integrated circuit (IC)

chips and other components may or may not be shown within the figures, for simplicity of illustration and discussion, and so as not to obscure certain aspects of the embodiments of the invention. Further, arrangements may be shown in block diagram form in order to avoid obscuring embodiments of the invention, and also in view of the fact that specifics with respect to implementation of such block diagram arrangements are highly dependent upon the platform within which the embodiment is to be implemented, i.e., such specifics should be well within purview of one skilled in the art. Where specific details (e.g., circuits) are set forth in order to describe example embodiments of the invention, it should be apparent to one skilled in the art that embodiments of the invention can be practiced without, or with variation of, these specific details. The description is thus to be regarded as illustrative instead of limiting.

The term “coupled” may be used herein to refer to any type of relationship, direct or indirect, between the components in question, and may apply to electrical, mechanical, fluid, optical, electromagnetic, electromechanical or other connections. In addition, the terms “first”, “second”, etc. might be used herein only to facilitate discussion, and carry no particular temporal or chronological significance unless otherwise indicated.

Those skilled in the art will appreciate from the foregoing description that the broad techniques of the embodiments of the present invention can be implemented in a variety of forms. Therefore, while the embodiments of this invention have been described in connection with particular examples thereof, the true scope of the embodiments of the invention should not be so limited since other modifications will become apparent to the skilled practitioner upon a study of the drawings, specification, and following claims.

CLAIMS

We claim:

1. An apparatus comprising:

logic to determine a power limit assigned to a first core;

5 logic to determine a stall count of the first core; and

logic to modulate a frequency of the first core based at least on the power limit assigned to the first core and the stall count of the first core, wherein the first core is to be included in a first tile of a socket in a multi-core computer environment.

10 2. The apparatus of claim 1, further comprising:

logic to determine an estimated power requirement of the first core, wherein the frequency of the first core is to be modulated further based on the estimated power requirement of the first core.

15 3. The apparatus of claim 2, wherein the power limit is to be assigned to the first core by a power control unit (PCU) associated with the socket.

4. The apparatus of claim 1, wherein the logic to modulate the frequency of the first core is coupled with a phase locked loop (PLL) associated with the socket.

20 5. The apparatus of claim 4, wherein the socket is to be configured to include the first tile and a second tile, and wherein the first tile and the second tile are to be associated with the PLL.

25 6. The apparatus of claim 5, wherein the frequency of the first core is to be modulated independently of a frequency associated with the second tile.

7. The apparatus of any of the claims 1 to 6, wherein the first tile is to include the first core and a second core, and wherein the frequency of the first core is to be modulated
30 independently of a frequency associated with the second core.

8. The apparatus of claim 3, further comprising logic to determine a thermal limit assigned to the first core by the PCU.

9. The apparatus of claim 8, wherein the frequency of the first core is to be
35 modulated based on the thermal limit.

10. The apparatus of claim 2, wherein the frequency of the first core is decreased when the stall count is higher than a threshold and the estimated power requirement is less than the power limit.

5

11. The apparatus of claim 2, wherein the frequency of the first core is modulated proportionally to a core stall ratio.

12. The apparatus of any of the claims 1 to 11, wherein the frequency of the first core is modulated within a boundary of the power limit.

10

13. The apparatus of claim 2, wherein the estimated power requirement of the first core is determined by a core energy monitor of a core local power unit (CLPU) associated with the first core.

15

14. A computer-implemented method comprising:
modulating a frequency of a core in a first tile of a multi-core environment at least independently of cores in other tiles based at least on an estimated power requirement of the core, a power limit assigned to the core and a stall count of the core, wherein the first tile and the other tiles are associated with a phase locked loop (PLL) of a socket.

20

15. The method of claim 14, further comprising:
determining the estimated power requirement of the core;
determining the power limit assigned to the core; and
determining the stall count of the core.

25

16. The method of claim 15, wherein modulating the frequency of the core in the first tile includes decreasing the frequency of the core when the number of core stalls is higher than a threshold and the estimated power requirement is less than the power limit.

30

17. The method of claim 15, wherein the frequency of the core is modulated proportionally to a core stall ratio.

18. The method of claim 17, wherein the frequency of the core is modulated within a boundary of the power limit.

35

19. The method of claim 14, wherein the estimated power requirement of the core is determined by a core energy monitor of a core local power unit (CLPU) associated with the core.

20. The method of claim 14, wherein the frequency of the core in the first tile is modulated independently of another core in the first tile.

21. The method of claim 14, wherein the frequency of the core in the first tile is decreased when the stall count is higher than a threshold and the estimated power requirement is less than the power limit.

22. The method of any of the claims 14 to 21, wherein the frequency of the core is modulated proportionally to a core stall ratio.

23. The method of any of the claims 14 to 21, wherein the frequency of the core is modulated within a boundary of the power limit.

24. A system comprising:
a phase locked loop (PLL) configured to be associated with a clock signal in a multi-core environment;
a socket coupled with the PLL and configured to include multiple tiles, at least one of the tiles including a first core and a second core, wherein the first core is configured to include logic to:

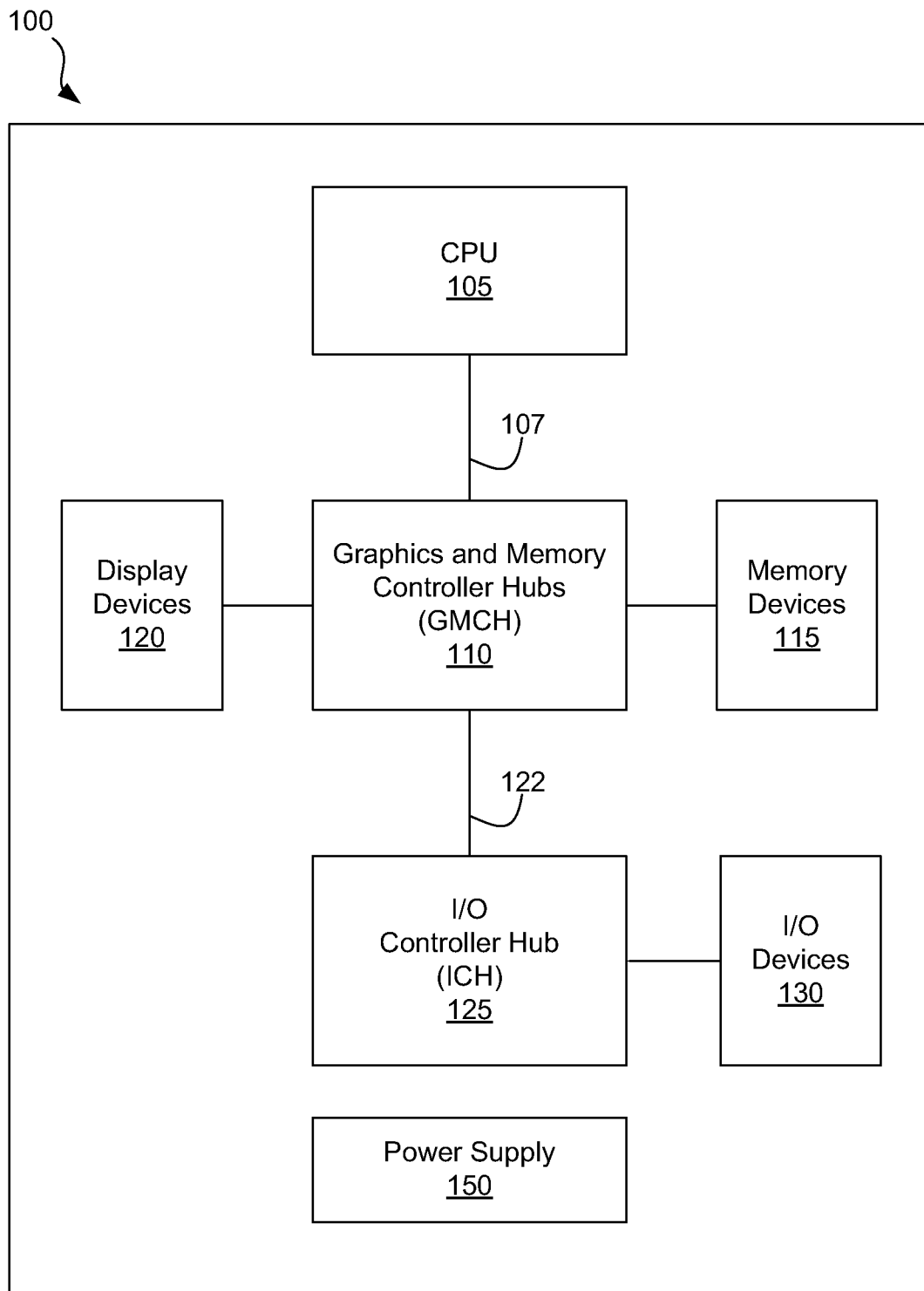
determine a power limit assigned to a first core;
determine a stall count of the first core; and
modulate a frequency of the first core based at least on the power limit assigned to the first core and the stall count of the first core independently of a frequency of tiles not associated with the first core.

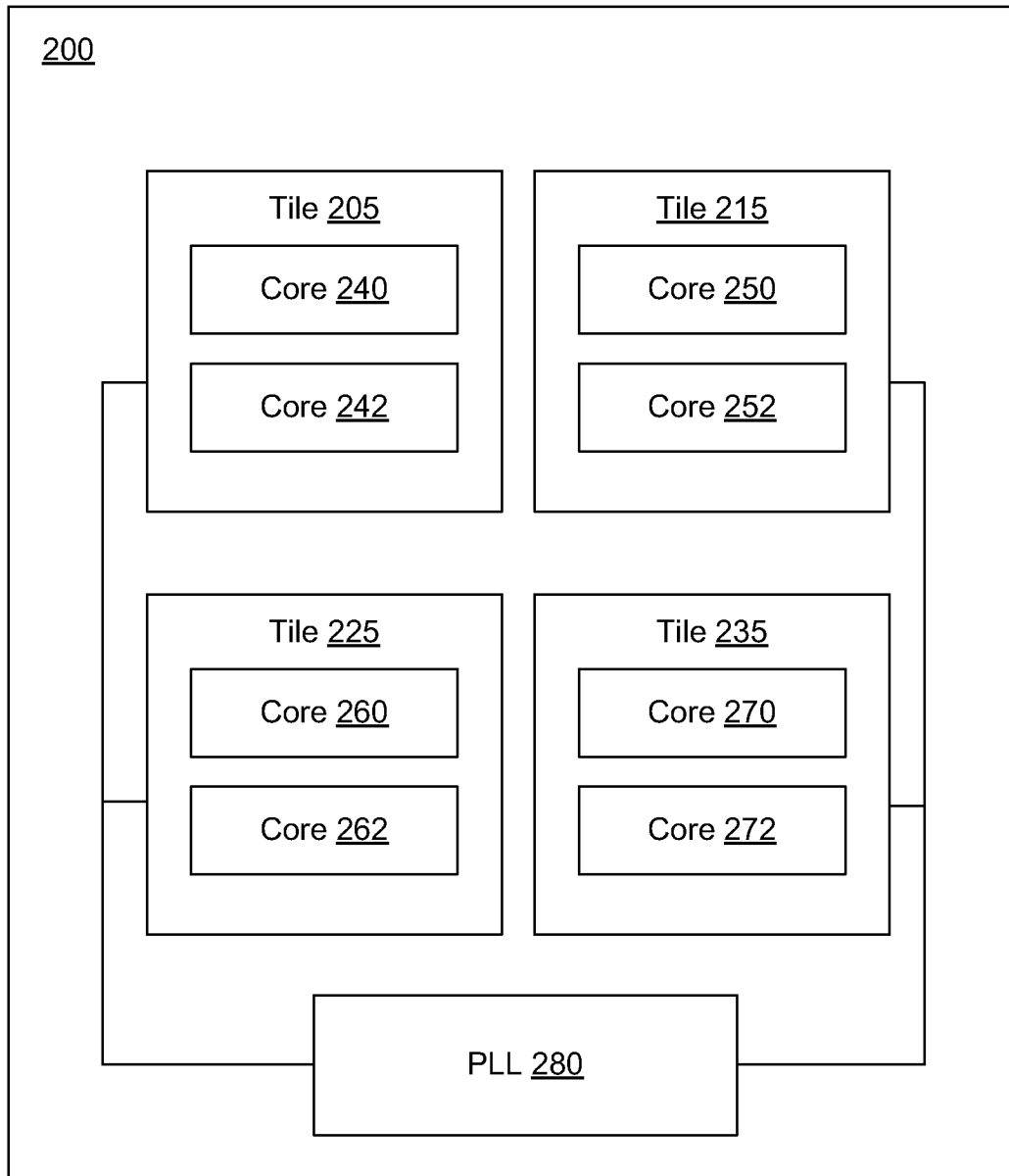
25. The system of claim 24, wherein the first core is further configured to include logic to determine an estimated power requirement of the first core.

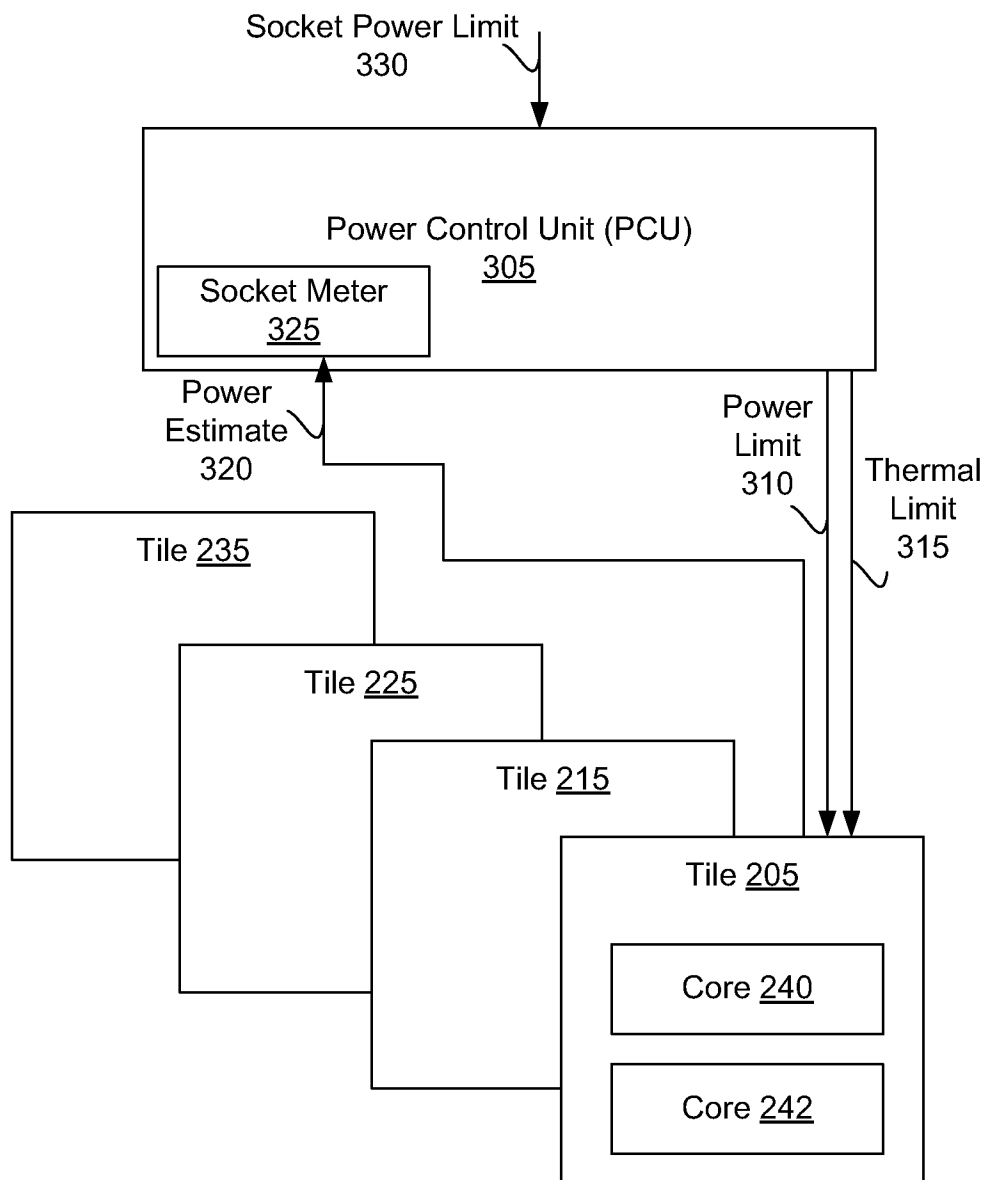
26. The system of claim 25, wherein the frequency of the first core is to be modulated based on the estimated power requirement of the first core.

27. The system of claim 25, wherein the frequency of the first core is to be modulated based on a comparison between the power limit assigned to the first core and the estimated power requirement of the first core.

- 5 28. The system of any of the claims 24 to 27, wherein the frequency of the first core is to be modulated independently of a frequency of the second core.

**FIG. 1**

**FIG. 2**

**FIG. 3**

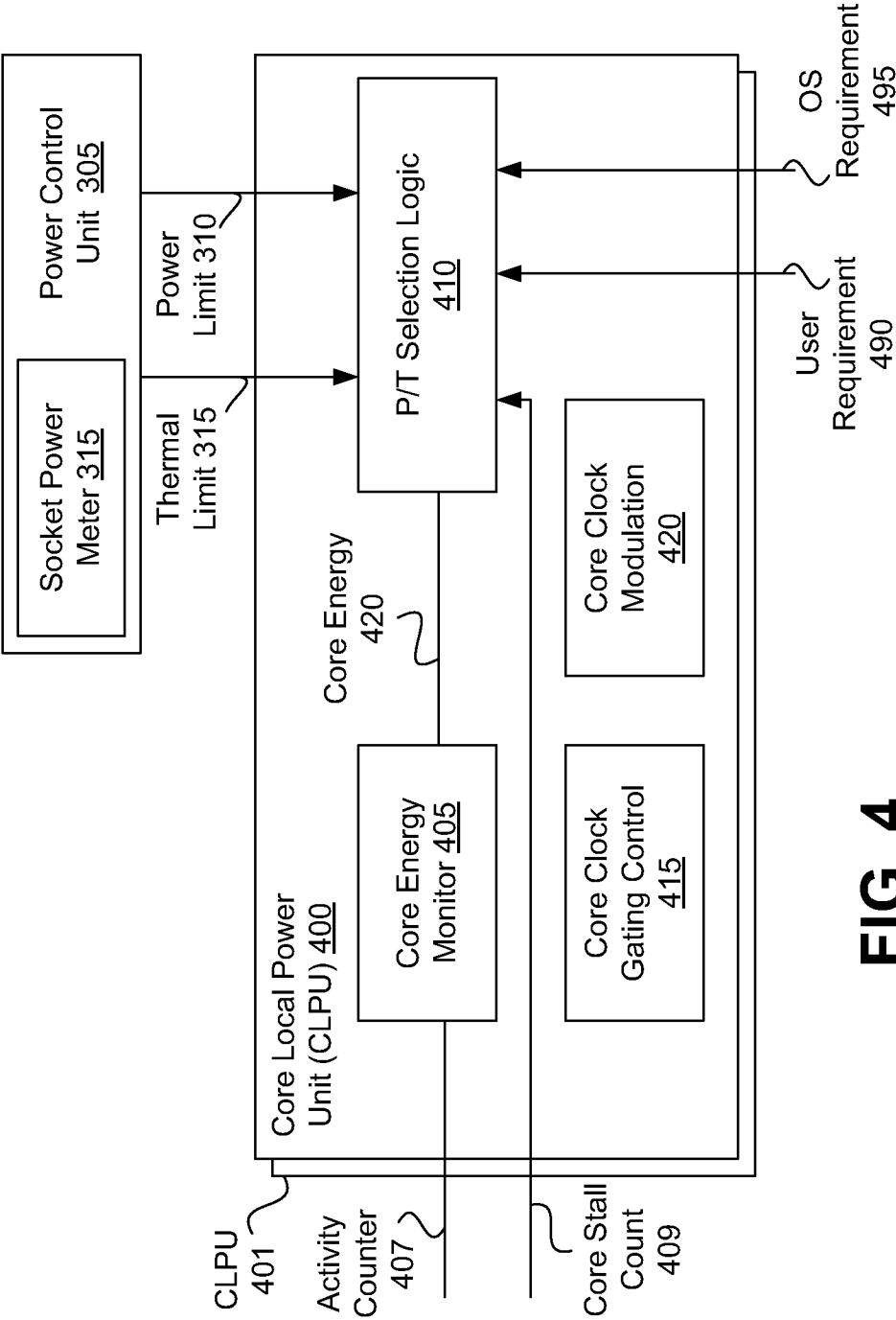
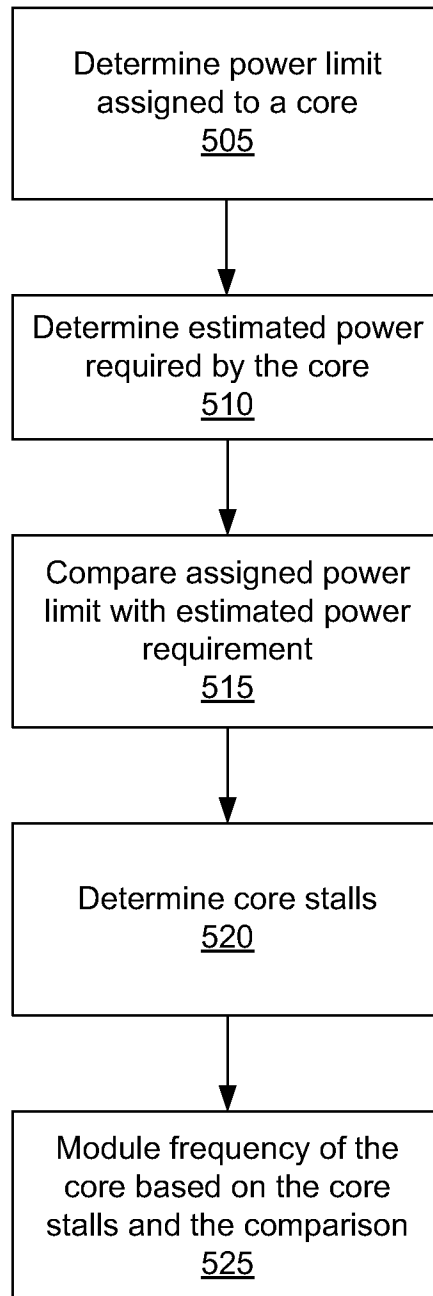


FIG. 4

**FIG. 5**

A. CLASSIFICATION OF SUBJECT MATTER**G06F 1/32(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 1/00

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: multicore processor, power limit, stall count, modulate frequency, independently, phase locked loop(PLL)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 2011-0022865 A1 (GUNTHER STEPHEN H. et al.) 27 January 2011 See paragraphs [0020]-[0028], claims 1-3, and figures 1-3.	1-28
A	US 2006-0282692 A1 (JANG GEUN OH) 14 December 2006 See paragraphs [0030]-[0046], and figure 2.	1-28
A	US 2012-0054521 A1 (NAFFZIGER SAMUEL) 01 March 2012 See paragraphs [0050]-[0057], and figure 4.	1-28
A	US 2008-0201591 A1 (HU CHUNLING et al.) 21 August 2008 See abstract, and claims 1-5.	1-28
A	US 7849332 B1 (ALBEN JONAH M. et al.) 07 December 2010 See column 15, lines 43-58, and figure 2.	1-28



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 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

07 NOVEMBER 2012 (07.11.2012)

Date of mailing of the international search report

09 NOVEMBER 2012 (09.11.2012)

Name and mailing address of the ISA/KR

Korean Intellectual Property Office
189 Cheongsa-ro, Seo-gu, Daejeon Metropolitan
City, 302-701, Republic of Korea

Facsimile No. 82-42-472-7140

Authorized officer

Park Ji Eun

Telephone No. 82-42-481-5696



INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No.

PCT/US2012/031751

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2011-0022865 A1	27.01.2011	CN 101403944 A	08.04.2009
		DE 102007051841 A1	21.05.2008
		GB 0721314 D0	12.12.2007
		GB 2444597 A	11.06.2008
		GB 2444597 B	22.12.2010
		JP 2008-117397 A	22.05.2008
		KR 10-0988396 B1	18.10.2010
		KR 10-2008-0039824 A	07.05.2008
		KR 10-2010-0075806 A	05.07.2010
		US 2008-0104425 A1	01.05.2008
		US 2009-0313489 A1	17.12.2009
		US 2011-191607 A1	04.08.2011
		US 7949887 B2	24.05.2011
		US 8069358 B2	29.11.2011
US 2006-0282692 A1	14.12.2006	CN 100461072 C0	11.02.2009
		CN 1877492 A	13.12.2006
		EP 1736851 A2	27.12.2006
		EP 1736851 A3	07.10.2009
		KR 10-1108397 B1	30.01.2012
		KR 10-2006-0128469 A	14.12.2006
		US 7664971 B2	16.02.2010
US 2012-0054521 A1	01.03.2012	None	
US 2008-0201591 A1	21.08.2008	US 7730340 B2	01.06.2010
US 7849332 B1	07.12.2010	US 2011-0126056 A1	26.05.2011
		US 7882369 B1	01.02.2011
		US 7886164 B1	08.02.2011