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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
01 April 2021 (01.04.2021)

WIPO | PCT

(10) International Publication Number
WO 2021/056418 A1

(51) International Patent Classification:
G06F 1/3206 (2019.01)

(21) International Application Number:
PCT/CN2019/108559

(22) International Filing Date:
27 September 2019 (27.09.2019)

(25) Filing Language: English

(26) Publication Language: English

(71) Applicant: **ALIBABA GROUP HOLDING LIMITED** [—/CN]; Fourth Floor, One Capital Place, P.O. Box 847, George Town, Grand Cayman (KY).

(72) Inventors: **SONG, Jun**; 1752 Pointe Woodworth Dr NE, Tacoma, WA 98422 (US). **CHENG, Lin**; Aliyun Feitian Park, Intersection Of Dingshan Road And Shilong Road, Xihu District, Hangzhou, 310024 (CN). **LU, Yijun**; Building 1, No.969 West Wen Yi Road, Yu Hang District, Hangzhou 311121 (CN). **FENG, Youquan**; Wangjing Block A, Alibaba Center No. 9th Wangjing East Garden 4th Area, Chaoyang District, Beijing, 100102 (CN). **ZHU, Hao**; Building 1, No.969 West Wen Yi Road, Yu Hang District, Hangzhou, 311121 (CN). **ZHANG, Xuegang**; No.3331 Keyuan South Road, Nanshan District, Shenzhen, Guangdong 518054 (CN). **HE, Lingfang**; Building 1, No.969 West Wen Yi Road, Yu, Hang District, Hangzhou, 311121 (CN). **WANG, Guan**; Wangjing Block A, Alibaba Center No. 9th Wangjing East Garden 4th Area, Chaoyang District, Beijing, 100102 (CN).

(74) Agent: **BEIJING TSINGYUANHUI INTELLECTUAL PROPERTY LAW FIRM**; Room 101, 1st Floor, Building 1, No. C-18, Zhichun Road, Haidian District, Beijing, 100190 (CN).

(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, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME,

(54) Title: POWER MANAGEMENT METHOD AND APPARATUS

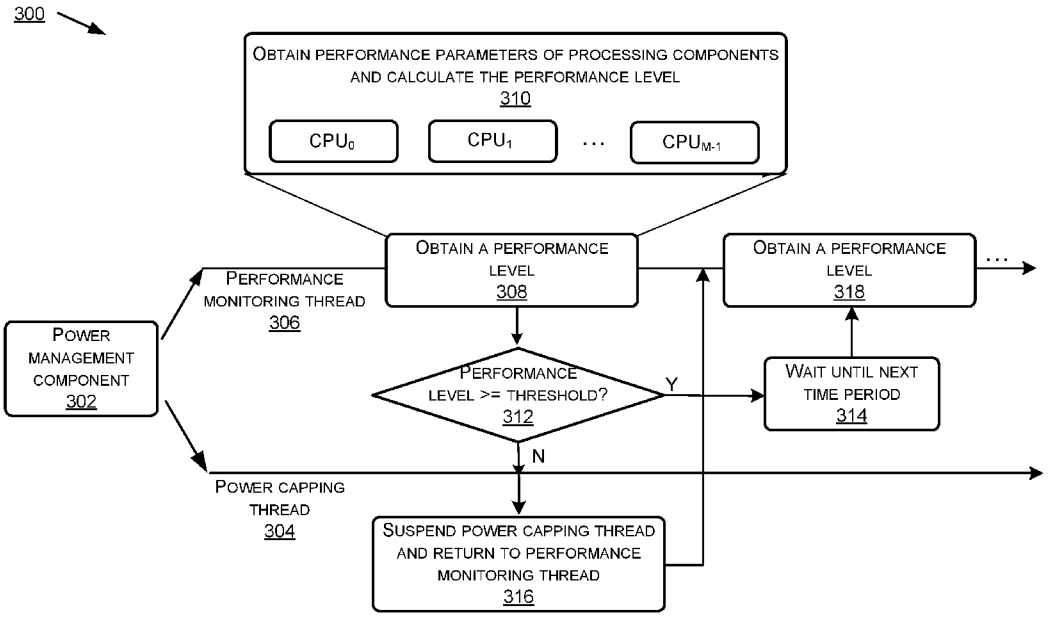

FIG. 3

(57) **Abstract:** Methods and apparatus are provided for improving power management. A power management component dynamically monitors a performance level of a computing system while performing a power capping process on the computing system. The power management component suspends the power capping process of the computing system when monitoring that the performance level of the computing system is lower than a threshold. An acceptable performance level of the computing system is ensured at which the performance downgrade of the computing system during the power capping process is under control or acceptable to the customers.

WO 2021/056418 A1

[Continued on next page]

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, 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))*

POWER MANAGEMENT METHOD AND APPARATUS

TECHNICAL FIELD

[0001] The present disclosure relates to the field of power management and, more particularly, relates to methods and apparatuses for power management of a computing system.

BACKGROUND

[0002] Power capping is a widely used technology in modern data centers (DCs) to increase on-rack compute density and avoid power outage. Most power capping technologies employ a proportional–integral–derivative (PID) control method to quickly find a set of parameters of hardware control knobs, e.g. operating frequencies of processing components, so that the system can reach a target power level. However, most PID controllers cause frequency overshooting and/or frequency undershooting during power capping, where the frequency undershooting could be very dangerous to cloud applications. If the frequencies of processing components are too low, a huge number of requests may not be processed in time. Moreover, an ill-designed back-pressure mechanism might spread the problem to upstream servers, lowering the overall performance of the system or even causing requests loss. Thus, performance downgrade occurs and may be disastrous to a cloud platform.

[0003] Hardware vendors such as Original Design Manufacturer (ODM) and Original Equipment Manufacturer (OEM) support power capping capability in the silicon or firmware. However, such capability does not come with minimal frequency protection. In other words, the performance level might be unacceptable during power capping procedure.

[0004] Software companies using power capping may suffer from performance problems. As investigated, some software companies try to avoid performance problems by conducting conservative power capping, leaving a big margin to a power-tripping level, and resulting in poor resource usage optimization.

[0005] In view of the above, improving the power management of a computing system is necessary.

SUMMARY

[0006] This summary is not intended to identify essential features of the claimed subject matter, nor is it intended for use in limiting the scope of the claimed subject matter.

[0007] The following describes example implementations of power management methods and apparatuses. In implementations, a power management component performs power capping on the computing system and dynamically monitors the performance level of the computing system. The power management component suspends the power capping process when monitoring that the performance level of the computing system is lower than a

threshold. Thus, an acceptable performance level of the computing system can be ensured where the performance downgrade of the computing system during power capping is under control or acceptable to the customer. The acceptable performance level of the computing system and the performance downgrade of the computing system acceptable to the customer can be determined based on a Service Level Agreement (SLA) between the service provider and the customer. Therefore, the power management of the computing system is improved.

BRIEF DESCRIPTION OF THE DRAWINGS

[0008] The detailed description is set forth with reference to the accompanying figures. In the figures, the left-most digit(s) of a reference number identifies the figure in which the reference number first appears. The use of the same reference numbers in different figures indicates similar or identical items or features.

[0009] FIGs. 1A and 1B illustrate example block diagrams of a scenario where performance downgrade occurs during power capping.

[0010] FIG. 2 illustrates an example flowchart of a process of the power management of a computing system.

[0011] FIG. 3 illustrates an example block diagram of a mechanism of the power management of a computing system.

[0012] FIG. 4 illustrates an example flowchart of a process for calculating the average value of frequencies of the processing components in the Performance Critical Zone in the computing system.

[0013] FIG. 5 illustrates an example block diagram of an apparatus for implementing the processes and methods described above.

DETAILED DESCRIPTION

[0014] Terminologies as used herein are denoted as follows. Power capping refers to the practice of limiting how much power a computing system can consume. Power cap refers to the power limit that the power consumption of the computing system cannot exceed. Nominal frequency refers to the guaranteed highest frequency within a Thermal Design Power (TDP) envelope. PID controller refers to a proportional–integral–derivative controller (or three-term controller), which is a control loop feedback mechanism widely used in industrial control systems and a variety of other applications requiring continuously modulated control.

[0015] FIGs. 1A and 1B illustrate example block diagrams of scenario 100 where performance downgrade occurs during power capping.

[0016] Referring to FIG. 1A, at block 102, there are 16 instances A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, and P, where the data generation and consumption are tightly connected among the instances. The arrows between the instances represent the dependencies between the instances. For example, after instance A is performed, instance B is performed, and so on. The output of instance A is provided to instance F, and the output of instance B is provided to instance E, and so on.

[0017] Referring to block 104, the 16 instances are mapped to 12 nodes on 4 racks. For example, instance J runs on rack_3 node_5 at 2.5 GHz+. Instance K runs on rack_4, node_7 at 2.5 GHz+. Instances A through I run on rack_11, node_6 through node_10 at 2.5 GHz. Instances L through P run on rack_15, node_11 through node_15 at 2.5 GHz.

[0018] Referring to block 106, when power capping is on, the total power drawn by the 12 nodes at the 4 racks is capped to not exceed a power limit/cap. The power limit/cap may be set and/or adjusted dynamically based on actual needs. The frequencies of the nodes are dynamically adjusted to satisfy the power limit/cap. For example, rack 3 runs in a failsafe mode (S5) due to the power capping. The failsafe mode is a design feature or practice that in the event of failure, the rack/node runs in a minimum responsive way (almost suspended). Rack 4 runs at 1.2 GHz. Rack 11 runs at 2.0 GHz. Instances L through P at rack 15, node 11 through 15, run at 2.0 GHz. Thus, the frequencies of the 12 nodes at the 4 racks are lowered due to the power capping.

[0019] Referring to FIG. 1B, block 108 shows the consequences at a time T. For example, a buffer running on node_5 at rack_3 is blown (stops responding) because of the low frequency due to the power capping. As a result, instance J running on node_5 at rack_3 is crashed (stops responding).

[0020] Block 110 shows the consequences at a time T+1. Because the buffer blowing problem is contagious, after one node fails to process requests in time, another node will do so. For example, another instance K also stops responding.

[0021] Block 112 shows the consequences at a time T+2, where four other nodes are severely impacted. Instances A, C, F, and G respond slowly or even stop responding. As a result, the performance downgrade can be observed. In that case, an avalanche occurs. The service stops, and the customers are surprised.

[0022] FIG. 2 illustrates an example flowchart of a process 200 of the power management of a computing system.

[0023] At block 202, a power management component dynamically monitors a performance level of a computing system while a power capping process is performed on the computing system. The computing system includes a plurality of processing components. In implementations, the computing system includes one or more nodes in a distributed system. The distributed system may be a large-scale distributed computing system.

[0024] In implementations, the power management component is a component that monitors and manages the power consumption of the computing system. The power management component may be implemented with software, hardware, firmware, or any combination thereof. The power management component performs the power capping process on the computing system such that the power consumption of the computing system does not exceed a power limit/cap. The power limit/cap may be set and/or adjusted dynamically based on actual needs.

[0025] In implementations, the power management component obtains a plurality of performance parameters of a group of processing components among the plurality of processing components in the computing system. A respective

performance parameter in the plurality of performance parameters is associated with a respective processing component in the group of processing components. The power management component calculates the performance level based on the plurality of performance parameters. The performance level comprises an average value of the plurality of performance parameters. In some embodiments, the average value is an instantaneous value. The plurality of performance parameters of the group of processing components includes a plurality of frequencies of the group of processing components.

[0026] At block 204, the power management component suspends/stops the power capping process of the computing system when monitoring that the performance level of the computing system is lower than a threshold.

[0027] In implementations, the power management component obtains a plurality of frequencies of a group of processing components in the computing system. The power management component calculates the performance level based on the plurality of frequencies of the group of processing components in the computing system.

[0028] In implementations, the performance of the group of processing components has a significant impact on the overall performance of the computing system. In some embodiments, the group of processing components runs one or more instances that require a latency under a first threshold, for example, 1μs, 5μs, and so on. In some embodiments, the group of processing components runs one or more instances that require an instruction execution rate above a second threshold, for example, 2 billion instructions per second, 5 billion

instructions per second, and so on. The first and second thresholds may be set and/or adjusted dynamically based on actual needs. In some embodiments, the group of processing components is not fixed, and the processing components in the group can be changed dynamically.

[0029] In implementations, the threshold represents the minimal/lowest performance level of the computing system at which the downgrade of the performance level of the computing system during the power capping process is under control or acceptable to the customers. The minimal/lowest performance level of the computing system and the performance downgrade acceptable to the customer can be determined based on the SLA between the service provider and the customer. The threshold may be set and/or adjusted dynamically based on actual needs.

[0030] In implementations, the power management component automatically resumes the power capping process after suspending the power capping process for a period of time, for example, 50ms, 2s, 1min, and so on. In some embodiments, the power management component resumes the power capping process when conditions are met, for example, when the power consumption of the computing system is over an upper limit. The upper limit may be set and/or adjusted dynamically based on actual needs.

[0031] With the above example process 200, the performance level of the computing system is dynamically monitored during power capping. When the performance level is going below the threshold, the power capping is suspended/stopped such that further performance downgrade of the computing

system can be prevented. Thus, an acceptable performance level of the computing system is ensured at which the performance downgrade of the computing system during power capping is under control or acceptable to the customer. Therefore, the power management of the computing system is improved.

[0032] FIG. 3 illustrates an example block diagram of a mechanism 300 of the power management of a computing system.

[0033] Referring to FIG. 3, a power management component 302 creates a power capping thread 304 and a performance monitoring thread 306. In the power capping thread 304, the power management component 302 performs power capping on the computing system (not shown) to ensure the power consumption of the computing system to not exceed the power limit/cap. The power limit/cap may be set and/or adjusted dynamically based on actual needs.

[0034] In the performance monitoring thread 306, the power management component 302 monitors the performance of the computing system.

[0035] In implementations, the performance of one or more processing components may be critical to or have a significant impact on the overall performance of the computing system. Thus, Performance Critical Zone (PCZ) is defined herein as a group of processing components whose performance is critical to or has a significant impact on the overall performance of the computing system. In some embodiments, the Performance Critical Zone is not fixed, and the processing components in the Performance Critical Zone can be changed dynamically.

[0036] In implementations, the computing system includes N processing components, for example, N CPUs, where N is a positive integer. Among the N CPUs, there are M CPUs in the Performance Critical Zone, where M is a positive integer. In other words, the performance of M CPUs has a significant impact on the overall performance of the computing system. In some embodiments, M CPUs in the Performance Critical Zone run latency sensitive workloads/instances. The latency sensitive workloads/instances require a latency under a first threshold, for example, 10μs, 5μs, and so on. In some embodiments, M CPUs in the Performance Critical Zone run throughput sensitive workloads/instances. The throughput sensitive workloads/instances require an instruction execution rate above a second threshold, for example, 2 billion instructions per second, 5 billion instructions per second, and so on. The first and second thresholds may be set and/or adjusted dynamically.

[0037] At block 308, the power management component 302 obtains the performance level of the computing system dynamically. In implementations, the power management component 302 obtains the performance level of the computing system periodically, for example, every X milliseconds, where X may be set and/or adjusted based on actual needs. In some embodiments, X may be dozens to hundreds of milliseconds.

[0038] At block 310 the power management component 302 obtains the performance parameters of the processing components in the Performance Critical Zone and calculates the performance level.

[0039] In implementations, the power management component 302 may collect telemetry data using performance monitoring units (PMUs) based mechanism to assist performance downgradation sensing. Telemetry is the automatic recording and transmission of data from remote or inaccessible sources to an IT system in a different location for monitoring and analysis. For example, the performance parameters are frequencies of the processing components. The power management component 302 reads Mode Specific Registers (MSRs) including APERF and MPERF through dev/CPUx/MSR interface in each of the M CUPs in the Performance Critical Zone. In more details, the power management component 302 reads $APERF_0$ and $MPERF_0$ in CPU_0, $APERF_1$ and $MPERF_1$ in CPU_1, …, and $APERF_{M-1}$ and $MPERF_{M-1}$ in CPU_{M-1}. The power management component 302 calculates an average value F_{avg} of frequencies of M CPUs in the Performance Critical Zone based on the reading results. In some embodiments, the average value F_{avg} is an instantaneous value. Details regarding the algorithm of calculating the average value F_{avg} are described hereinafter with reference to FIG. 4.

[0040] At block 312, the power management component 302 determines whether the performance level is more than or equal to a threshold. In implementations, the power management component 302 determines whether the average value F_{avg} is more than or equal to a threshold frequency F_{min}. The threshold frequency F_{min} represents a minimal/lowest frequency at which the performance level of the computing system is acceptable and the performance downgrade of the computing system during power capping is under control or

acceptable to the customer. The threshold frequency F_{min} can be determined based on the customer's requirement, machine learning, empirical value, experimental data, etc. The threshold frequency F_{min} may be set and/or adjusted dynamically. The performance lever and performance downgrade of the computing system that are acceptable to the customer can be determined based on the SLA between the service provider and the customer.

[0041] If the power management component 302 determines that the performance level is greater than or equal to the threshold at block 312, the power management component 302 waits until the next period at block 314. In implementations, if the power management component 302 determines that the average value F_{avg} is greater than or equal to the threshold frequency F_{min}, the power management component 302 waits until the next period.

[0042] If the power management component 302 determines that the performance level is not greater than or equal to the threshold at block 312, the power management component 302 suspends/stops the power capping thread 304 at block 316. In implementations, if the power management component 302 determines that the average value F_{avg} is lower than the threshold frequency F_{min}, the power management component 302 suspends/stops the power capping thread 304, and returns to the performance monitoring thread 306 immediately at block 316.

[0043] Additionally or alternatively, if the power management component 302 determines that the performance level is lower than the threshold, the power management component indicates a scheduler to migrate instances running on

the computing system to other computing systems (not shown) as soon as possible. However, such migration may impose pressure on the scheduler and some instances may not be migrated.

[0044] At block 318, in a next period, the power management component 302 obtains the performance level of the computing system again as described above with reference to blocks 308 to 310, and details are not repeated here.

[0045] Additionally or alternatively, the power management component 302 automatically resumes the power capping process after suspending the power capping process for a period of time, for example, 50ms, 2s, 1min, and so on. In some embodiments, the power management component 302 resumes the power capping process when a condition is met, for example, when the power consumption of the computing system is over an upper limit. The upper limit may be set and/or adjusted dynamically based on actual needs.

[0046] With the above example mechanism 300, the performance level of the computing system is dynamically monitored during power capping. When the performance level is going below the threshold, the power capping is suspended/stopped such that further performance downgrade of the computing system can be prevented. Thus, the acceptable performance level of the computing system is guaranteed at which the performance downgrade of the computing system during power capping is under control or acceptable to the customer. Therefore, the power management of the computing system is improved.

[0047] FIG. 4 illustrates an example flowchart of a process 400 for calculating the average value of frequencies of the processing components in the Performance Critical Zone in the computing system. In implementations, the computing system includes N CPUs, where N is a positive integer.

[0048] At block 402, the power management component reads a nominal frequency F* of CPU_i, where i is a positive integer. The nominal frequency F* can be determined based on a specification, empirical value, experimental data, etc.

[0049] At block 404, the power management component determines whether CPU_i is in the Performance Critical Zone (PCZ).

[0050] If the power management component determines that CPU_i is in the PCZ at block 404, the power management component reads Mode Specific Registers, including APERF and MPERF, of CPU_i, where the reading results are referred as $APERF_i$ and $MPERF_i$ at block 406. If the power management component determines that CPU_i is not in the PCZ at block 404, the process 400 proceeds to block 412.

[0051] At block 408, the power management component calculates a first change value $delta\_APERF_i$, which is the change value between a current $APERF_i$ and a previous $APERF_{prv\_i}$ according to the following formula (1), and a second change value $delta\_MPERF_i$ which is the change value between a current $MPERF_i$ and a previous $MPERF_{prv\_i}$ according to the following formula (2):

[0052] $$delta\_APERF_i = APERF_i - APERF_{prv\_i} \quad (1)$$

[0053] $delta\_MPERF_i = MPERF_i - MPERF_{prv\_i}$ (2)

[0054] At block 410, the power management component calculates an average frequency F_{avg_i} of CPU_i according to the following formula (3). In some embodiments, the average frequency F_{avg_i} is an instantaneous value.

[0055] $F_{avg\_i} = F*(delta\_APERF/delta\_MPERF)$ (3)

[0056] At block 412, the power management component increases i by 1.

[0057] At block 414, the power management component determines whether i is greater than the total number of CPUs in the computing system, that is, whether i is greater than N.

[0058] If the power management component determines that i is greater than N at block 414, the power management component calculates the average value F_{avg} of frequencies of all CPUs in the PCZ at block 416. If i is determined not to be greater than N at block 414, the process 400 goes back to block 404.

[0059] FIG. 5 illustrates an example block diagram of an apparatus 500 for implementing the processes and methods described above.

[0060] The apparatus 500 includes one or more processors 502 and memory 504 communicatively coupled to the processor(s) 502. The processor(s) 502 executes one or more modules and/or processes to cause the processor(s) 502 to perform a variety of functions. In implementations, the processor(s) 502 may include a central processing unit (CPU), a graphics processing unit (GPU), both CPU and GPU, or other processing units or components known in the art. Additionally, each of the processor(s) 502 may possess its own local memory, which also may store program modules, program data, and/or one or more

operating systems. In implementations, the memory 504 may be volatile, such as RAM, non-volatile, such as ROM, flash memory, miniature hard drive, memory card, and the like, or some combination thereof.

[0061] The apparatus 500 may additionally include an input/output (I/O) interface 506 for receiving and outputting data. The apparatus 500 may also include a communication module 508 allowing the apparatus 500 to communicate with other devices (not shown) over a network (not shown). The network may include the Internet, wired media such as a wired network or direct-wired connections, and wireless media such as acoustic, radio frequency (RF), infrared, and other wireless media.

[0062] The memory 504 may include one or more computer-executable modules (modules) that are executable by the processor(s) 502. In implementations, the memory 504 may include, but not limited to, a monitoring module 510 and a suspending module 512.

[0063] The monitoring module 510 is configured to dynamically monitor a performance level of a computing system while a power capping process is performed on the computing system. The computing system including a plurality of processing components. In implementations, the computing system comprises one or more nodes in a distributed system. The distributed system may be a large-scale distributed computing system.

[0064] The monitoring module 510 is further configured to obtain a plurality of performance parameters of a group of processing components among the plurality of processing components in the computing system, and calculate the

performance level based on the plurality of performance parameters. A respective performance parameter in the plurality of performance parameters is associated with a respective processing component in the group of processing components. The performance level comprises an average value of the plurality of performance parameters. In some embodiments, the average value is an instantaneous value. The plurality of performance parameters of the group of processing components includes a plurality of frequencies of the group of processing components.

[0065] The suspending module 512 is configured to suspend/stop the power capping process of the computing system when monitoring that the performance level of the computing system is lower than a threshold.

[0066] The suspending module 512 is further configured to determine whether the average value is lower than a threshold frequency, and suspend/stop the power capping process of the computing system in response to determining that the average value is lower than the threshold frequency.

[0067] In implementations, the performance of the group of processing components or has a significant impact on the performance level of the computing system. In some embodiments, the group of processing components runs one or more instances that require a latency under a first threshold, for example, 10μs, 5μs, and so on. In some embodiments, the group of processing components runs one or more instances that require an instruction execution rate above a second threshold, for example, 2 billion instructions per second, 5 billion instructions per second, and so on. The first and second thresholds may be set

and/or adjusted dynamically based on actual needs. In some embodiments, the group of processing components is not fixed, and the processing components in the group can be changed dynamically.

[0068] In implementations, the threshold represents the minimal/lowest performance level of the computing system at which the performance downgrade of the computing system during the power capping process is under control or acceptable to the customers. The minimal/lowest performance level of the computing system and the performance downgrade acceptable to the customer can be determined based on the SLA between the service provider and the customer. The threshold may be set and/or adjusted dynamically based on actual needs.

[0069] With the above example apparatus 500, the performance level of the computing system is dynamically monitored during power capping. When the performance level is going below the threshold, the power capping is suspended/stopped such that further performance downgrade of the computing system can be prevented. Thus, the acceptable performance level of the computing system is guaranteed at which the performance downgrade of the computing system during power capping is under control or acceptable to the customer. Therefore, the power management of the computing system is improved.

[0070] Processes and systems discussed herein may be implemented in, but not limited to, distributed computing environment, parallel computing

environment, cluster computing environment, grid computing environment, cloud computing environment, electrical vehicles, power facilities, etc.

[0071] Some or all operations of the methods described above can be performed by execution of computer-readable instructions stored on a computer-readable storage medium, as defined below. The term "computer-readable instructions" as used in the description and claims, include routines, applications, application modules, program modules, programs, components, data structures, algorithms, and the like. Computer-readable instructions can be implemented on various system configurations, including single-processor or multiprocessor systems, minicomputers, mainframe computers, personal computers, hand-held computing devices, microprocessor-based, programmable consumer electronics, combinations thereof, and the like.

[0072] The computer-readable storage media may include volatile memory (such as random access memory (RAM)) and/or non-volatile memory (such as read-only memory (ROM), flash memory, etc.). The computer-readable storage media may also include additional removable storage and/or non-removable storage including, but not limited to, flash memory, magnetic storage, optical storage, and/or tape storage that may provide non-volatile storage of computer-readable instructions, data structures, program modules, and the like.

[0073] A non-transient computer-readable storage medium is an example of computer-readable media. Computer-readable media includes at least two types of computer-readable media, namely computer-readable storage media and communications media. Computer-readable storage media includes volatile and

non-volatile, removable and non-removable media implemented in any process or technology for storage of information such as computer-readable instructions, data structures, program modules, or other data. Computer-readable storage media includes, but is not limited to, phase change memory (PRAM), static random-access memory (SRAM), dynamic random-access memory (DRAM), other types of random-access memory (RAM), read-only memory (ROM), electrically erasable programmable read-only memory (EEPROM), flash memory or other memory technology, compact disk read-only memory (CD-ROM), digital versatile disks (DVD) or other optical storage, magnetic cassettes, magnetic tape, magnetic disk storage or other magnetic storage devices, or any other non-transmission medium that can be used to store information for access by a computing device. In contrast, communication media may embody computer-readable instructions, data structures, program modules, or other data in a modulated data signal, such as a carrier wave, or other transmission mechanism. As defined herein, computer-readable storage media do not include communication media.

[0074] The computer-readable instructions stored on one or more non-transitory computer-readable storage media that, when executed by one or more processors, may perform operations described above with reference to FIGs. 1-5. Generally, computer-readable instructions include routines, programs, objects, components, data structures, and the like that perform particular functions or implement particular abstract data types. The order in which the operations are described is not intended to be construed as a limitation, and any number of the

described operations can be combined in any order and/or in parallel to implement the processes.

EXAMPLE CLAUSES

[0075] Clause 1. A method comprising: dynamically monitoring a performance level of a computing system while a power capping process is performed on the computing system, the computing system including a plurality of processing components; and suspending the power capping process of the computing system when monitoring that the performance level of the computing system is lower than a threshold.

[0076] Clause 2. The method of clause 1, wherein dynamically monitoring the performance level of the computing system comprises obtaining the performance level of the computing system periodically.

[0077] Clause 3. The method of clause 1, wherein dynamically monitoring the performance level of the computing system comprises: obtaining a plurality of performance parameters of a group of processing components among the plurality of processing components in the computing system, a respective performance parameter in the plurality of performance parameters being associated with a respective processing component in the group of processing components; and calculating the performance level based on the plurality of performance parameters.

[0078] Clause 4. The method of clause 3, wherein the performance level comprises an average value of the plurality of performance parameters.

[0079] Clause 5. The method of clause 4, wherein suspending the power capping process of the computing system when monitoring that the performance level of the computing system is lower than the threshold comprises: determining whether the average value is lower than a threshold frequency; and suspending the power capping process of the computing system in response to determining that the average value is lower than the threshold frequency.

[0080] Clause 6. The method of clause 3, wherein a performance of the group of processing components has a significant impact on the performance level of the computing system.

[0081] Clause 7. The method of clause 6, wherein the group of processing components runs one or more instances that require a latency under a first threshold.

[0082] Clause 8. The method of clause 6, wherein the group of processing components runs one or more instances that require an instruction execution rate above a second threshold.

[0083] Clause 9. The method of clause 3, wherein obtaining a plurality of performance parameters of the group of processing components comprises: reading a plurality of registers of the group of processing component to obtain the plurality of performance parameters.

[0084] Clause 10. The method of clause 3, wherein the plurality of performance parameters of the group of processing components comprises a plurality of frequencies of the group of processing components.

[0085] Clause 11. The method of clause 1, wherein the threshold represents a lowest performance level of the computing system at which a performance downgrade of the computing system during the power capping process is acceptable.

[0086] Clause 12. The method of clause 1, wherein the computing system comprises one or more nodes in a distributed system.

[0087] Clause 13. A computer-readable storage medium storing computer-readable instructions executable by one or more processors, that when executed by the one or more processors, cause the one or more processors to perform operations comprising: dynamically monitoring a performance level of a computing system while a power capping process is performed on the computing system, the computing system including a plurality of processing components; and suspending the power capping process of the computing system when monitoring that the performance level of the computing system is lower than a threshold.

[0088] Clause 14. The computer-readable storage medium of clause 13, wherein dynamically monitoring the performance level of the computing system comprises obtaining the performance level of the computing system periodically.

[0089] Clause 15. The computer-readable storage medium of clause 13, wherein dynamically monitoring the performance level of the computing system comprises: obtaining a plurality of performance parameters of a group of processing components among the plurality of processing components in the computing system, a respective performance parameter in the plurality of

performance parameters being associated with a respective processing component in the group of processing components; and calculating the performance level based on the plurality of performance parameters.

[0090] Clause 16. The computer-readable storage medium of clause 15, wherein the performance level comprises an average value of the plurality of performance parameters.

[0091] Clause 17. The computer-readable storage medium of clause 18, wherein suspending the power capping process of the computing system when monitoring that the performance level of the computing system is lower than the threshold comprises: determining whether the average value is lower than a threshold frequency; and suspending the power capping process of the computing system in response to determining that the average value is lower than the threshold frequency.

[0092] Clause 18. The computer-readable storage medium of clause 15, wherein a performance of the group of processing components has a significant impact on the performance level of the computing system.

[0093] Clause 19. The computer-readable storage medium of clause 18, wherein the group of processing components runs one or more instances that require a latency under a first threshold.

[0094] Clause 20. The computer-readable storage medium of clause 18, wherein the group of processing components runs one or more instances that require an instruction execution rate above a second threshold.

[0095] Clause 21. The computer-readable storage medium of clause 15, wherein obtaining a plurality of performance parameters of the group of processing components comprises: reading a plurality of registers of the group of processing component to obtain the plurality of performance parameters.

[0096] Clause 22. The computer-readable storage medium of clause 15, wherein the plurality of performance parameters of the group of processing components comprises a plurality of frequencies of the group of processing components.

[0097] Clause 23. The computer-readable storage medium of clause 13, wherein the threshold represents a lowest performance level of the computing system at which a performance downgrade of the computing system during the power capping process is acceptable.

[0098] Clause 24. The computer-readable storage medium of clause 13, wherein the computing system comprises one or more nodes in a distributed system.

[0099] Clause 25. An apparatus comprising: one or more processors; and memory communicatively coupled to the one or more processors, the memory storing computer-executable modules executable by the one or more processors, the computer-executable modules including: a monitoring module, configured to dynamically monitor a performance level of a computing system while a power capping process is performed on the computing system, the computing system including a plurality of processing components; and a suspending module, configured to suspend the power capping process of the

computing system when monitoring that the performance level of the computing system is lower than a threshold.

[00100] Clause 26. The apparatus of clause 25, wherein the monitoring module is further configured to: obtain a plurality of performance parameters of a group of processing components among the plurality of processing components in the computing system, a respective performance parameter in the plurality of performance parameters being associated with a respective processing component in the group of processing components; and calculate the performance level based on the plurality of performance parameters.

[00101] Clause 27. The apparatus of clause 26, wherein the performance level comprises an average value of the plurality of performance parameters.

[00102] Clause 28. The apparatus of clause 27, wherein the suspending module is further configured to: determine whether the average value is lower than a threshold frequency; and suspend the power capping process of the computing system in response to determining that the average value is lower than the threshold frequency.

[00103] Clause 29. The apparatus of clause 26, wherein a performance of the group of processing components has a significant impact on the performance level of the computing system.

[00104] Clause 30. The apparatus of clause 29, wherein the group of processing components runs one or more instances that require a latency under a first threshold.

[00105] Clause 31. The apparatus of clause 29, wherein the group of processing components runs one or more instances that require an instruction execution rate above a second threshold.

[00106] Clause 32. The apparatus of clause 26, wherein the plurality of performance parameters of the group of processing components comprises a plurality of frequencies of the group of processing components.

[00107] Clause 33. The apparatus of clause 25, wherein the threshold represents a lowest performance level of the computing system at which a performance downgrade of the computing system during the power capping process is acceptable.

[00108] Clause 34. The apparatus of clause 25, wherein the computing system comprises one or more nodes in a distributed system.

CONCLUSION

[00109] Although the subject matter has been described in language specific to structural features and/or methodological acts, it is to be understood that the subject matter defined in the appended claims is not necessarily limited to the specific features or acts described. Rather, the specific features and acts are disclosed as exemplary forms of implementing the claims.

CLAIMS

WHAT IS CLAIMED IS:

1. A method comprising:

dynamically monitoring a performance level of a computing system while a power capping process is performed on the computing system, the computing system including a plurality of processing components; and

suspending the power capping process of the computing system when monitoring that the performance level of the computing system is lower than a threshold.

2. The method of claim 1, wherein dynamically monitoring the performance level of the computing system comprises obtaining the performance level of the computing system periodically.

3. The method of claim 1, wherein dynamically monitoring the performance level of the computing system comprises:

obtaining a plurality of performance parameters of a group of processing components among the plurality of processing components in the computing system, a respective performance parameter in the plurality of performance parameters being associated with a respective processing component in the group of processing components; and

calculating the performance level based on the plurality of performance parameters.

4. The method of claim 3, wherein the performance level comprises an average value of the plurality of performance parameters.

5. The method of claim 4, wherein suspending the power capping process of the computing system when monitoring that the performance level of the computing system is lower than the threshold comprises:

determining whether the average value is lower than a threshold frequency; and

suspending the power capping process of the computing system in response to determining that the average value is lower than the threshold frequency.

6. The method of claim 3, wherein a performance of the group of processing components has a significant impact on the performance level of the computing system.

7. The method of claim 6, wherein the group of processing components runs one or more instances that require a latency under a first threshold.

8. The method of claim 6, wherein the group of processing components runs one or more instances that require an instruction execution rate above a second threshold.

9. The method of claim 3, wherein obtaining a plurality of performance parameters of the group of processing components comprises:

reading a plurality of registers of the group of processing component to obtain the plurality of performance parameters.

10. The method of claim 3, wherein the plurality of performance parameters of the group of processing components comprises a plurality of frequencies of the group of processing components.

11. The method of claim 1, wherein the threshold represents a lowest performance level of the computing system at which a performance downgrade of the computing system during the power capping process is acceptable.

12. The method of claim 1, wherein the computing system comprises one or more nodes in a distributed system.

13. A computer-readable storage medium storing computer-readable instructions executable by one or more processors, that when executed by the

one or more processors, cause the one or more processors to perform operations comprising:

dynamically monitoring a performance level of a computing system while a power capping process is performed on the computing system, the computing system including a plurality of processing components; and

suspending the power capping process of the computing system when monitoring that the performance level of the computing system is lower than a threshold.

14. The computer-readable storage medium of claim 13, wherein dynamically monitoring the performance level of the computing system comprises obtaining the performance level of the computing system periodically.

15. The computer-readable storage medium of claim 13, wherein dynamically monitoring the performance level of the computing system comprises:

obtaining a plurality of performance parameters of a group of processing components among the plurality of processing components in the computing system, a respective performance parameter in the plurality of performance parameters being associated with a respective processing component in the group of processing components; and

calculating the performance level based on the plurality of performance parameters.

16. The computer-readable storage medium of claim 15, wherein the performance level comprises an average value of the plurality of performance parameters.

17. The computer-readable storage medium of claim 18, wherein suspending the power capping process of the computing system when monitoring that the performance level of the computing system is lower than the threshold comprises:

determining whether the average value is lower than a threshold frequency; and

suspending the power capping process of the computing system in response to determining that the average value is lower than the threshold frequency.

18. The computer-readable storage medium of claim 15, wherein a performance of the group of processing components has a significant impact on the performance level of the computing system.

19. The computer-readable storage medium of claim 18, wherein the group of processing components runs one or more instances that require a latency under a first threshold.

20. The computer-readable storage medium of claim 18, wherein the group of processing components runs one or more instances that require an instruction execution rate above a second threshold.

21. The computer-readable storage medium of claim 15, wherein obtaining a plurality of performance parameters of the group of processing components comprises:

reading a plurality of registers of the group of processing component to obtain the plurality of performance parameters.

22. The computer-readable storage medium of claim 15, wherein the plurality of performance parameters of the group of processing components comprises a plurality of frequencies of the group of processing components.

23. The computer-readable storage medium of claim 13, wherein the threshold represents a lowest performance level of the computing system at which a performance downgrade of the computing system during the power capping process is acceptable.

24. The computer-readable storage medium of claim 13, wherein the computing system comprises one or more nodes in a distributed system.

25. An apparatus comprising:

one or more processors; and

memory communicatively coupled to the one or more processors, the memory storing computer-executable modules executable by the one or more processors, the computer-executable modules including:

a monitoring module, configured to dynamically monitor a performance level of a computing system while a power capping process is performed on the computing system, the computing system including a plurality of processing components; and

a suspending module, configured to suspend the power capping process of the computing system when monitoring that the performance level of the computing system is lower than a threshold.

26. The apparatus of claim 25, wherein the monitoring module is further configured to:

obtain a plurality of performance parameters of a group of processing components among the plurality of processing components in the computing system, a respective performance parameter in the plurality of performance parameters being associated with a respective processing component in the group of processing components; and

calculate the performance level based on the plurality of performance parameters.

27. The apparatus of claim 26, wherein the performance level comprises an average value of the plurality of performance parameters.

28. The apparatus of claim 27, wherein the suspending module is further configured to:

determine whether the average value is lower than a threshold frequency; and

suspend the power capping process of the computing system in response to determining that the average value is lower than the threshold frequency.

29. The apparatus of claim 26, wherein a performance of the group of processing components has a significant impact on the performance level of the computing system.

30. The apparatus of claim 29, wherein the group of processing components runs one or more instances that require a latency under a first threshold.

31. The apparatus of claim 29, wherein the group of processing components runs one or more instances that require an instruction execution rate above a second threshold.

32. The apparatus of claim 26, wherein the plurality of performance parameters of the group of processing components comprises a plurality of frequencies of the group of processing components.

33. The apparatus of claim 25, wherein the threshold represents a lowest performance level of the computing system at which a performance downgrade of the computing system during the power capping process is acceptable.

34. The apparatus of claim 25, wherein the computing system comprises one or more nodes in a distributed system.

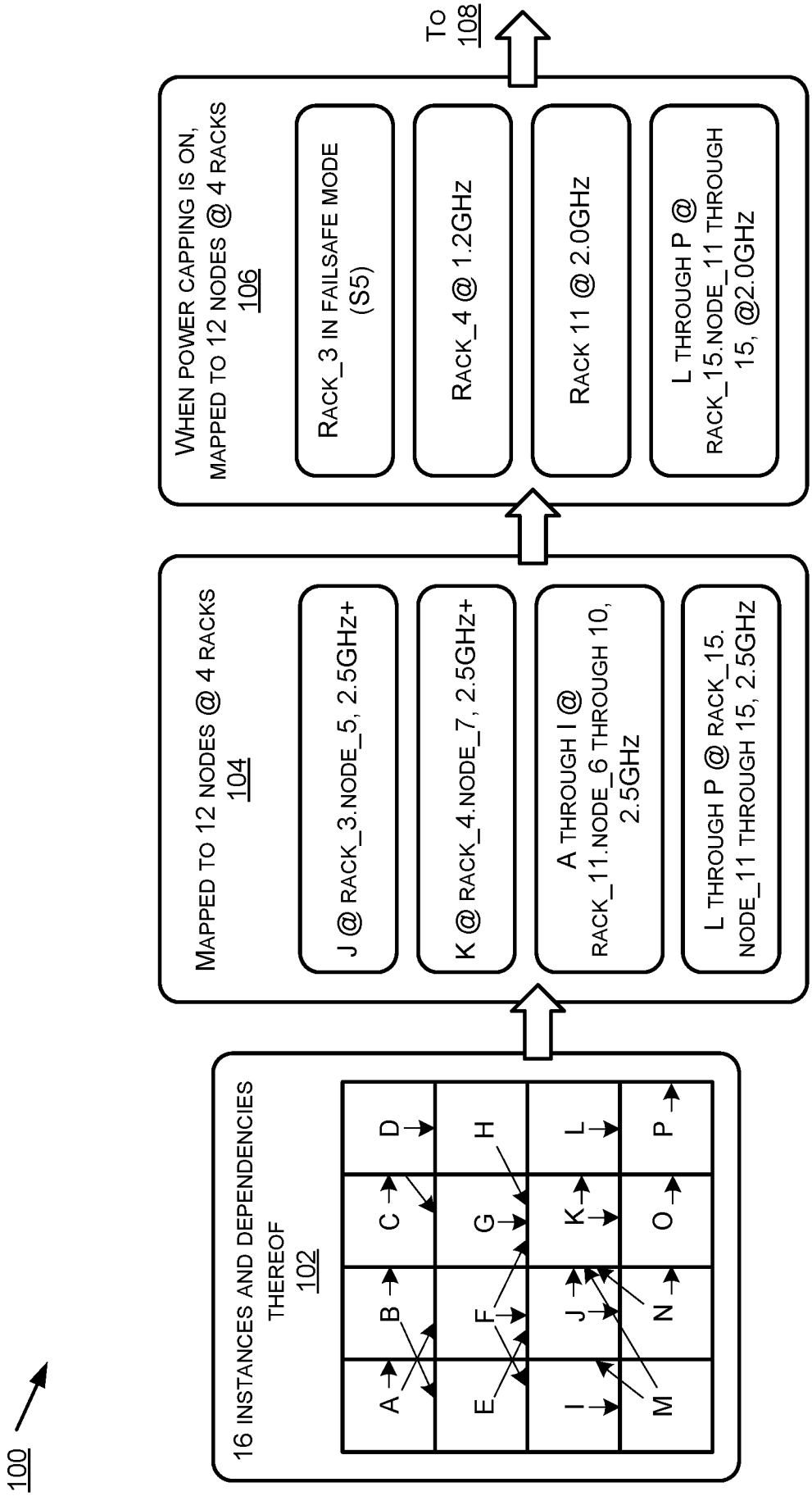
100
16 INSTANCES AND DEPENDENCIES THEREOF
102
A
B
C
D
E
F
G
H
I
J
K
L
M
N
O
P
MAPPED TO 12 NODES @ 4 RACKS
104
J @ RACK_3.NODE_5, 2.5GHz+
K @ RACK_4.NODE_7, 2.5GHz+
A THROUGH I @ RACK_11.NODE_6 THROUGH 10, 2.5GHz
L THROUGH P @ RACK_15. NODE_11 THROUGH 15, 2.5GHz
WHEN POWER CAPPING IS ON, MAPPED TO 12 NODES @ 4 RACKS
106
RACK_3 IN FAILSAFE MODE (S5)
RACK_4 @ 1.2GHz
RACK 11 @ 2.0GHz
L THROUGH P @ RACK_15.NODE_11 THROUGH 15, @2.0GHz
TO 108

FIG. 1A

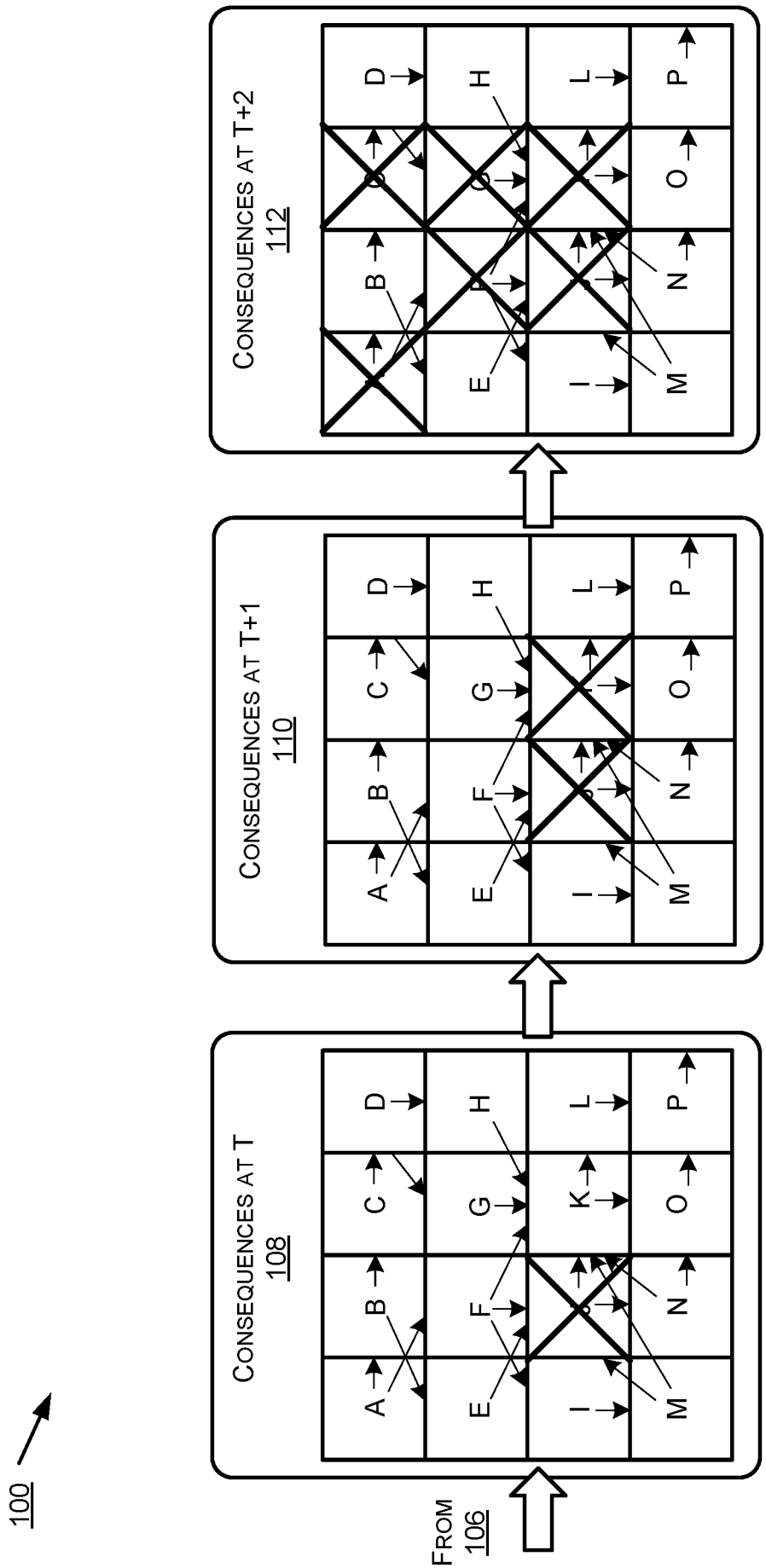
100
FROM 106
CONSEQUENCES AT T 108
CONSEQUENCES AT T+1 110
CONSEQUENCES AT T+2 112
A
B
C
D
E
F
G
H
I
K
L
M
N
O
P

FIG. 1B

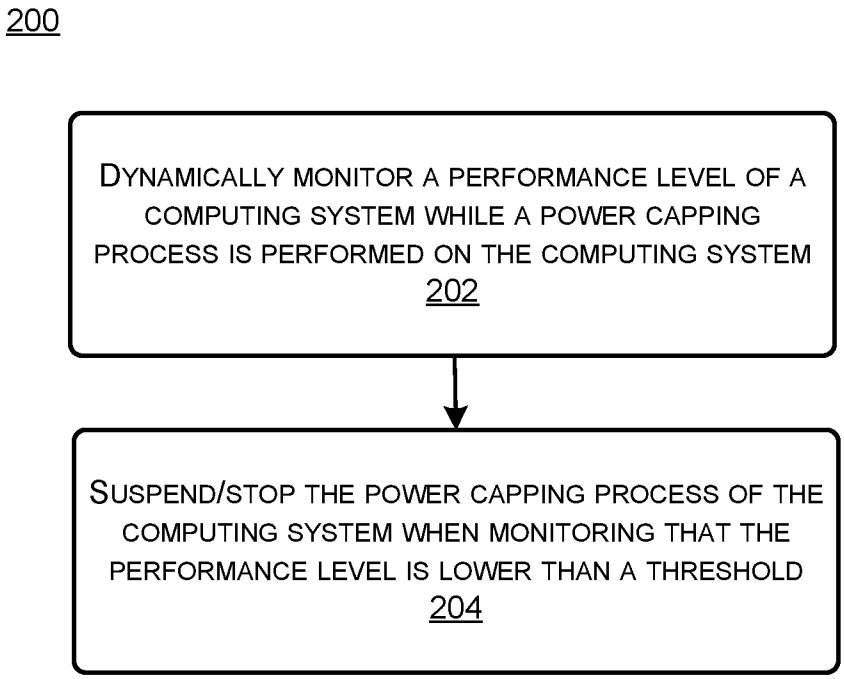
200
DYNAMICALLY MONITOR A PERFORMANCE LEVEL OF A COMPUTING SYSTEM WHILE A POWER CAPPING PROCESS IS PERFORMED ON THE COMPUTING SYSTEM
202
SUSPEND/STOP THE POWER CAPPING PROCESS OF THE COMPUTING SYSTEM WHEN MONITORING THAT THE PERFORMANCE LEVEL IS LOWER THAN A THRESHOLD
204

FIG. 2

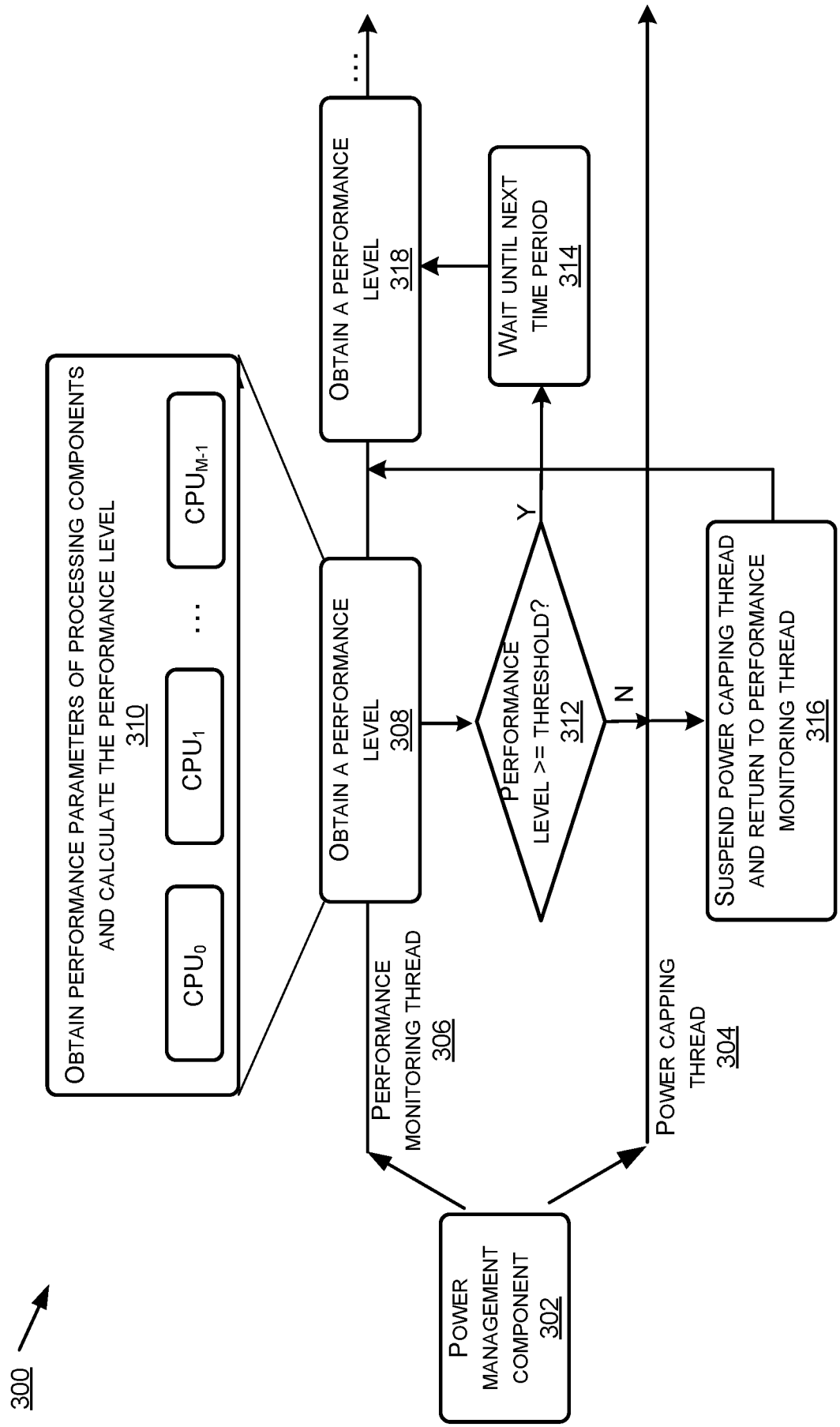
300
POWER MANAGEMENT COMPONENT 302
PERFORMANCE MONITORING THREAD 306
POWER CAPPING THREAD 304
OBTAIN PERFORMANCE PARAMETERS OF PROCESSING COMPONENTS AND CALCULATE THE PERFORMANCE LEVEL 310
CPU0
CPU1
...
CPUM-1
OBTAIN A PERFORMANCE LEVEL 308
PERFORMANCE LEVEL >= THRESHOLD? 312
Y
N
WAIT UNTIL NEXT TIME PERIOD 314
OBTAIN A PERFORMANCE LEVEL 318
...
SUSPEND POWER CAPPING THREAD AND RETURN TO PERFORMANCE MONITORING THREAD 316

FIG. 3

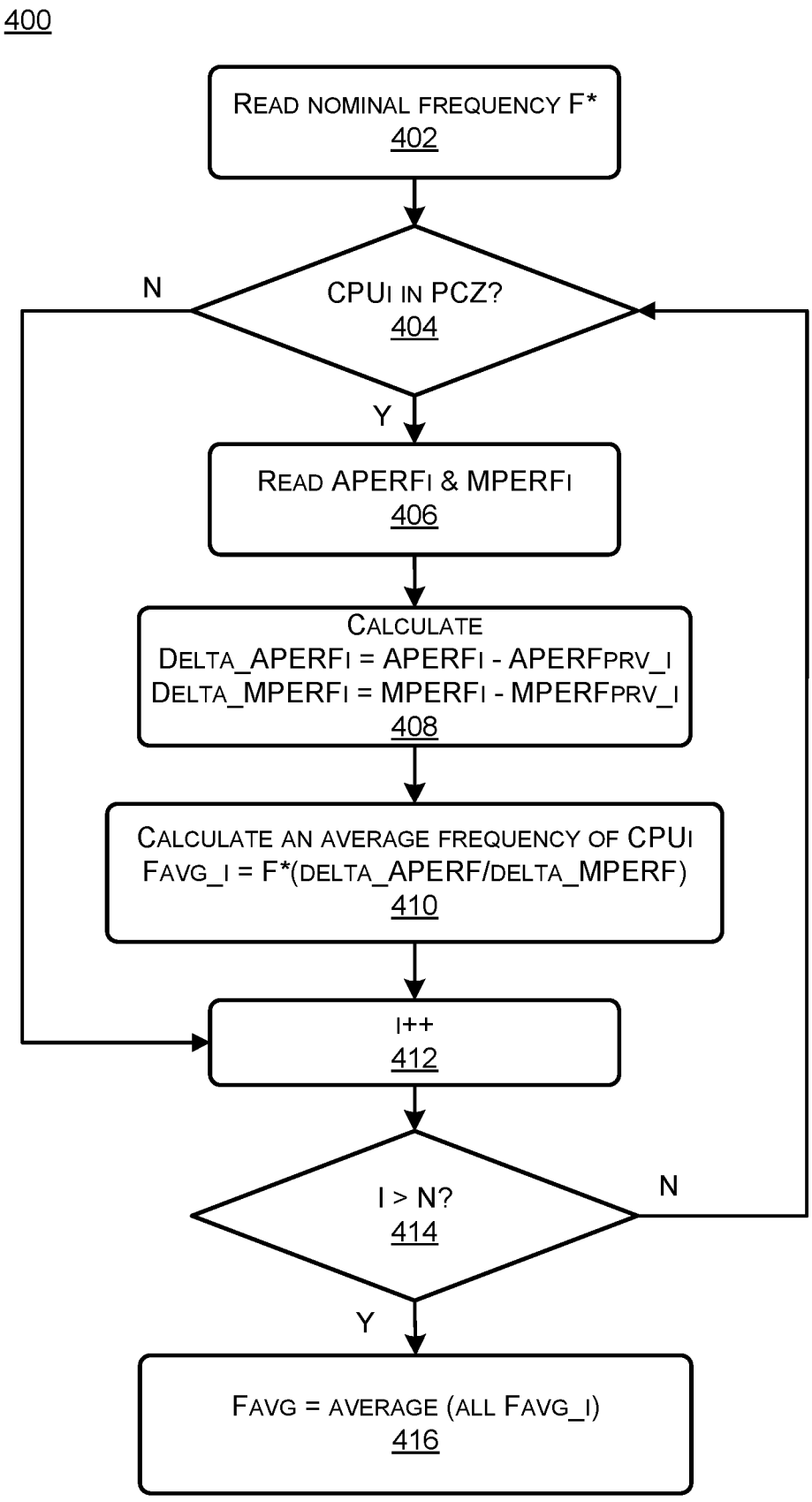
400
READ NOMINAL FREQUENCY F*
402
CPUI IN PCZ?
404
N
Y
READ APERFI & MPERFI
406
CALCULATE
DELTA_APERFI = APERFI - APERFPRV_I
DELTA_MPERFI = MPERFI - MPERFPRV_I
408
CALCULATE AN AVERAGE FREQUENCY OF CPUI
FAVG_I = F*(DELTA_APERF/DELTA_MPERF)
410
I++
412
I > N?
414
N
Y
FAVG = AVERAGE (ALL FAVG_I)
416

FIG. 4

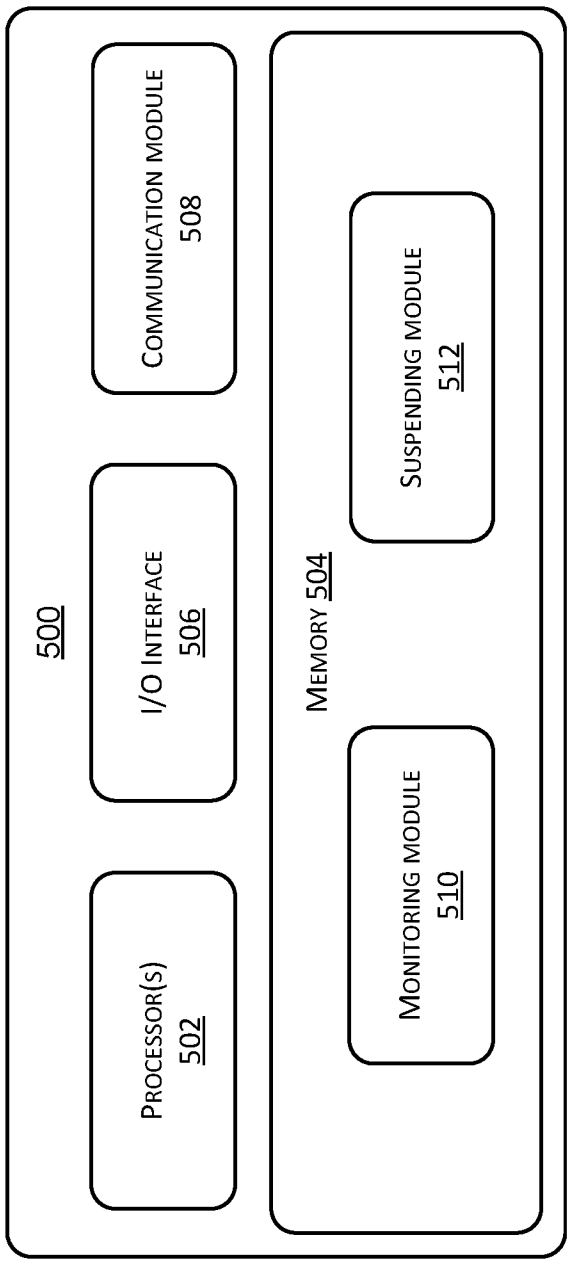
500
PROCESSOR(S)
502
I/O INTERFACE
506
COMMUNICATION MODULE
508
MEMORY 504
MONITORING MODULE
510
SUSPENDING MODULE
512

FIG. 5

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2019/108559

A. CLASSIFICATION OF SUBJECT MATTER

G06F 1/3206(2019.01)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

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

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

CNPAT, CNKI, WPI, EPODOC, IEEE: monitor, performance, power w capping, downgrade, frequency, suspend, threshold, parameter, level, lower

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2017017288 A1 (CISCO TECHNOLOGY, INC.) 19 January 2017 (2017-01-19) description, paragraphs [0002], [0014]-[0047], figures 1-6	1-34
A	US 2011144818 A1 (LI, Cong et al.) 16 June 2011 (2011-06-16) the whole document	1-34
A	US 2011178652 A1 (INTERNATIONAL BUSINESS MACHINES CORPORATION) 21 July 2011 (2011-07-21) the whole document	1-34
A	US 2012036385 A1 (FAASSE, Scott P. et al.) 09 February 2012 (2012-02-09) the whole document	1-34
A	US 2013111478 A1 (INTERNATIONAL BUSINESS MACHINES CORPORATION) 02 May 2013 (2013-05-02) the whole document	1-34

☐ 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	Date of mailing of the international search report
09 June 2020	**24 June 2020**
Name and mailing address of the ISA/CN **National Intellectual Property Administration, PRC** **6, Xitucheng Rd., Jimen Bridge, Haidian District, Beijing** **100088** **China** Facsimile No. **(86-10)62019451**	Authorized officer **LIU,Jian** Telephone No. **86-(10)-53961304**

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2019/108559

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
US 2017017288 A1	19 January 2017	None	
US 2011144818 A1	16 June 2011	EP 2360552 A2	24 August 2011
		CN 102096460 A	15 June 2011
		JP 2011123873 A	23 June 2011
		US 2013283068 A1	24 October 2013
US 2011178652 A1	21 July 2011	None	
US 2012036385 A1	09 February 2012	None	
US 2013111478 A1	02 May 2013	TW 201232416 A	01 August 2012
		WO 2012059294 A1	10 May 2012
		US 2012110588 A1	03 May 2012