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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
29 December 2011 (29.12.2011)

PCT

(10) International Publication Number
WO 2011/163261 A1

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

(21) International Application Number:
PCT/US2011/041291

(22) International Filing Date:
21 June 2011 (21.06.2011)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
12/819,777 21 June 2010 (21.06.2010) US

(71) Applicant *(for all designated States except US)*: **ADVANCED MICRO DEVICES, INC.** [US/US]; One AMD Place, P.O. Box 3453, Sunnyvale, California 94088 (US).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **NAFFZIGER, Samuel, D.** [US/US]; 3749 Ashmount Drive, Fort Collins, Colorado 80522 (US). **PETRY, John, P.** [US/US]; 10334 Barrywood Way, San Diego, California 92131 (US). **HUGHES, William, A.** [GB/US]; 565 Brookes Avenue, San Jose, California 95125 (US).

(74) Agent: **RANKIN, Rory D.**; Meyertons, Hood, Kivlin, Kowert & Goetzel, P.C., P.O. Box 398, Austin, TX 78767-0398 (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, RO, RS, RU, 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, 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,

[Continued on next page]

(54) Title: MANAGING MULTIPLE OPERATING POINTS FOR STABLE VIRTUAL FREQUENCIES

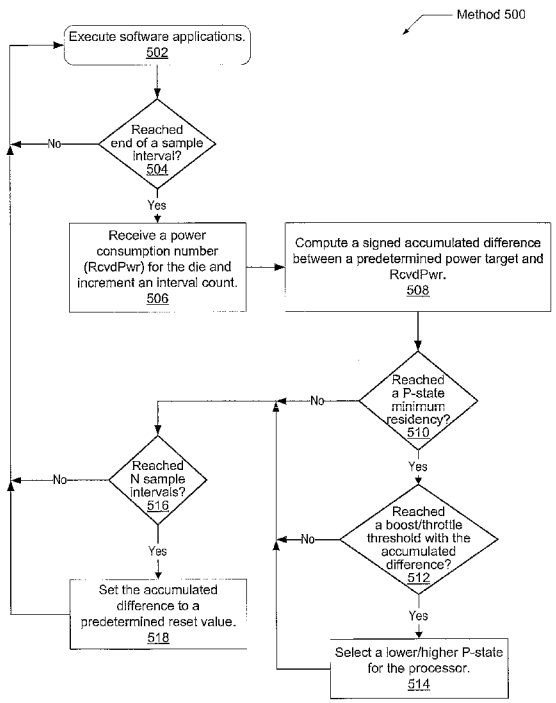

FIG. 5

(57) Abstract: A system and method for managing multiple discrete operating points to create a stable virtual operating point. One or more functional blocks within a processor produces data corresponding to an activity level associated with the respective functional block. A power manager determines a power consumption value based on the data once every given sample interval. In addition, the power manager determines a signed accumulated difference over time between a thermal design power (TDP) and the power consumption value. The power manager selects a next power-performance state (P-state) based on comparisons of the signed accumulated difference and given thresholds. Transitioning between P-states in this manner while the workload does not significantly change causes the processor to operate at a virtual operating point between supported discrete operating points.

WO 2011/163261 A1

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))*

TITLE: MANAGING MULTIPLE OPERATING POINTS FOR STABLE VIRTUAL FREQUENCIES

BACKGROUND OF THE INVENTION

Field of the Invention

[0001] This invention relates to computing systems, and more particularly, to efficient management of processor discrete operating points.

Description of the Relevant Art

[0002] The power consumption of modern integrated circuits (IC's) has become an increasing design issue with each generation of semiconductor chips. As power consumption increases, more costly cooling systems such as larger fans and heat sinks are utilized to remove excess heat and prevent IC failure. However, cooling systems increase the system cost. The IC power dissipation constraint is not only an issue for portable computers and mobile communication devices, but also for high-performance superscalar microprocessors, which may include multiple processor cores, or cores, and multiple pipelines within a core.

[0003] The power consumption of IC's, such as modern complementary metal oxide semiconductor (CMOS) chips, is proportional to at least the expression fV^2. The symbol f is the operational frequency of the chip. The symbol V is the operational voltage of the chip. In modern microprocessors, both parameters f and V may be varied during operation of the IC. For example, during operation, modern processors allow users to select one or more intermediate power-performance states between a maximum performance state and a minimum power state. The maximum performance state includes a maximum operating frequency and the minimum power state includes a minimum operating frequency. The intermediate discrete power-performance states (P-states) include given scaled values for a combination of the operating frequency and the operational voltage.

[0004] Software, such as an operating system or firmware, or hardware may select a particular P-state based on at least a projected time to change states, a selected power limit, workload characteristics, and inputs from on-chip power monitors corresponding to a current workload. However, a computed combination of operational frequency and operational voltage typically does not match a combination corresponding to a discrete given P-state. Therefore, a close-matching given P-state is chosen. Typically, this chosen P-state may correspond to a power

consumption lower than a computed power limit. Accordingly, the performance of the chosen P-state is lower than a computed power limit. If several more discrete P-states are added to a processor to provide finer grain combinations of operational frequency and voltage, then the design and test costs of the processor increase.

[0005] In view of the above, efficient methods and mechanisms for management of processor discrete operating points are desired.

[0006] Systems and methods for managing multiple discrete operating points to create a stable virtual operating point are contemplated.

[0007] In one embodiment, a processor comprises several functional blocks and a power manager. Each of the functional blocks produces data corresponding to an activity level associated with the respective functional block. The power manager determines a power consumption value based on the data once every given sample interval. In addition, the power manager determines a signed accumulated difference over time between a given power target and the power consumption value. In one embodiment, the power target may correspond to a thermal design power (TDP) for the processor. In various embodiments, hysteresis may be used to avoid unproductive P-state transitions. In such embodiments, the power manager selects a lower power-performance state (P-state) than a current P-state if the signed accumulated difference is less than a negative threshold by more than a given delta. The power manager selects a higher P-state than a current P-state if the signed accumulated difference is greater than a positive threshold by more than a given delta. Also contemplated are embodiments where a minimum residency in a P-state may be required before a transition to another P-state is permitted.

[0008] These and other embodiments will be further appreciated upon reference to the following description and drawings.

BRIEF DESCRIPTION OF THE DRAWINGS

[0009] FIG. 1 is a generalized block diagram of one embodiment of power-performance state transitions for a semiconductor chip.

[0010] FIG. 2 is a generalized block diagram of one embodiment of core power management.

[0011] FIG. 3 is a flow diagram of one embodiment of a method for managing multiple discrete operating points to create a stable virtual operating point.

[0012] FIG. 4 is a generalized block diagram of one embodiment of a core power management system.

[0013] FIG. 5 is a flow diagram of one embodiment of a method for managing multiple discrete operating points to create a stable virtual operating point.

[0014] While the invention is susceptible to various modifications and alternative forms, specific embodiments are shown by way of example in the drawings and are herein described in detail. It should be understood, however, that drawings and detailed description thereto are not intended to limit the invention to the particular form disclosed, but on the contrary, the invention is to cover all modifications, equivalents and alternatives falling within the spirit and scope of the present invention as defined by the appended claims.

DETAILED DESCRIPTION

[0015] In the following description, numerous specific details are set forth to provide a thorough understanding of the present invention. However, one having ordinary skill in the art should recognize that the invention might be practiced without these specific details. In some instances, well-known circuits, structures, and techniques have not been shown in detail to avoid obscuring the present invention.

[0016] Referring to FIG. 1, one embodiment of power-performance state transitions 100 for a semiconductor chip is shown. Two curves are shown in the diagram illustrating non-linear (e.g., cubic or quadratic) relationships between power versus voltage and frequency versus voltage. Five discrete power-performance states (P-states) are shown in the diagram denoted as P0 to P4. A small number of discrete P-states are shown to simplify the diagram. Although only five discrete P-states are shown, other numbers of discrete P-states may be supported.

[0017] In the diagram shown, the P-state P_4 may correspond to a discrete state with a lowest performance of all the supported discrete states and comprises the lowest operational frequency. In addition, the P-state P_4 may correspond to a discrete state with a lowest power consumption of all the supported discrete states and comprises the lowest operational voltage. On the other hand, the P-state P_0 may correspond to a discrete state with a highest performance of all the supported discrete states and comprises the highest operational frequency. In addition, the P-state P_0 may correspond to a discrete state with a highest power consumption of all the supported discrete states and comprises the highest operational voltage. Typically, the endpoint discrete states represented by P-states P_0 and P_4 define a region of predictable performance. Therefore, configuring a processor to support multiple P-states, or operating points, along the non-linear frequency versus voltage curve may provide stable, optimal utilization of power and delivery of performance for the semiconductor chip, such as a processor. The management of the P-states

may conform to an industry standard such as the Advanced Configuration and Power Interface (ACPI) standard, originally developed by Intel Corp., Microsoft Corp., and Toshiba Corp.

[0018] As shown in the diagram, a power $target_1$ (e.g., a desired power consumption level) may be chosen for the chip. In one embodiment, the selected power $target_1$ may correspond to a thermal design power (TDP) of the chip. The thermal design power (TDP), which may also be referred to as a thermal design point, represents a maximum amount of power a cooling system in a computer is able to dissipate. A cooling system for a laptop processor may be designed for a 20 watt TDP. Therefore, it has been determined that the cooling system is able to dissipate 20 watts without exceeding the maximum junction temperature for transistors within the processor. The TDP value may differ depending on the chip manufacturer producing the chip. For example, one manufacturer may define the TDP value as a power value measured at a default voltage level under given worst-case temperature conditions. Another manufacturer may define the TDP value as a maximum power value measured over a given interval as the chip executes typical applications versus high-power virus applications. Other measurement definitions are possible and contemplated.

[0019] In one embodiment, a power model executed on a pre-silicon model of the die 102 may perform a power measurement. Later in a design cycle, power measurements may be performed on actual fabricated silicon dies during a testing phase and debugging phase. In one embodiment, a peak power value for a chip may be defined by functional failure of the chip executing a high-power application on the core. The TDP value is typically less than the peak power value. The TDP value may be used to set the operational voltage and the operational frequency of a chip for binning purposes.

[0020] The value power $target_1$ in FIG. 1 may represent an assigned TDP value. As shown in FIG. 1, the power $target_1$ corresponds to a data point A on the power versus voltage non-linear curve. Data point A corresponds to an operating voltage V_2. Projecting data point A onto the non-linear frequency versus voltage curve with respect to the operating voltage V_2 provides data point A`. The data point A` corresponds to an operating frequency F_2. The operating point represented by the combination of the operating voltage V_2 and the operating frequency F_2 may provide an optimal utilization of power and delivery of performance for the chip.

[0021] As described above and shown in the diagram, an operating point for power $target_1$ is identified by data point A`. However, this operating point is not represented by a discrete P-state on the power versus frequency curve. The data point A` is located between the P-states P_1 and P_2. In order to reduce power consumption, the P-state P_2 may be chosen as an initial operating

point for the corresponding chip. A corresponding combination of the operating voltage V_1 and the operating frequency F_1 may be the resulting chosen operating point. This operating point corresponds to a lower power consumption value than the value power $target_1$. The value Power P_2 indicates the lower power consumption value of the operating point corresponding to the P-state P_2.

[0022] A chip, such as a processor, may continue processing workloads utilizing an initially assigned P-state until either (i) the workload significantly changes which causes a significant change in a reported activity level, or (ii) the initial TDP value changes, such as being adjusted by a power monitoring software or firmware, which changes the power target value shown in the diagram. For example, if a processor is executing instructions for a workload that is halved at a given point in time, the resulting total drawn current and thermal energy will be significantly reduced. In one embodiment, a power manager, which may be located within the processor, may detect this condition and accordingly choose a different P-state corresponding to a higher power-performance operating point. For example, the power manager may determine to increase, or boost, the current P-state P_2 to the higher performance P-state P_1. For purposes of discussion, higher performance P-states may have a lower number. For example, P_0 may represent a higher power-performance state that P_1. However, designations could be reversed such that P_0 is used to represent a lower power-performance state than P_1. The chosen approach for designations may simply be a matter of preference.

[0023] A "throttle" of a P-state includes reducing the currently selected P-state by one (or more) P-state(s) to a lower power consumption P-state. In contrast, a "boost" of a P-state includes increasing the currently selected P-state by one (or more) P-state(s) to a higher performance P-state. Throttling the P-state P_2 includes transitioning the currently selected P-state P_2 to the lower power-performance state P_3. A simple illustration of boosting and throttling a given P-state, such as the P-state P_2 as an example, is shown in FIG. 1. In some embodiments, each boost operation and each throttle operation may cause a currently selected P-state to transition by two or more P-states when the logic supports this added complexity. The P-state transitions may be controlled by logic within a processor, and thereby is a self-contained system. However, power management software running on the processor or a rack controller located external to the processor, may alter the TDP value for the processor, which changes the power target value in the diagram.

[0024] Although the operating point represented by the P-state P_2 in FIG. 1 consumes less power than the power $target_1$ value, the operating point represented by P_2 also has less

performance. Rather than maintain a lower performance P-state until a significant change in a measured activity level, a processor may toggle between two discrete P-states in order to achieve an average "virtual" operating point for the current workload. For example, a power manager may determine for a same workload an amount of time to utilize P-state P_2 before boosting to P-state P_1. Similarly, the power manager may determine for the same workload an amount of time to utilize P-state P_1 before throttling to P-state P_2. This process may occur multiple times until the workload significantly changes. An average of the times spent in a particular P-state coupled with the operating voltage and frequency may have an effect the processor was utilizing a "virtual" operating point represented by data point A' in FIG. 1. Such a method would enable flexible power management with arbitrary power limit settings to achieve stable virtual operating points (or virtual P-state) for any workload by building on already existing discrete P-states.

[0025] Turning now to FIG. 2, one embodiment of a core power management system 200 is shown. Changes to an operational voltage 210 and a measured activity level 220 over time are shown. In addition, a power target 235 and an average power versus a power target ratio 230 is shown over time. As shown in FIG. 2, there are P-state transitions both when an activity level experiences a significant change and when the activity level is constant. In the diagram, the P-state transitions correlate with changes in the operating voltage 210. Some P-state values are labeled in the diagram. For example, with a constant activity level 220, a P-state P_1 transitions to P-state P_2, which has a lower operating voltage as shown in the diagram. Toggling between P-state values when the measured activity level is constant may be referred to as P-state dithering. The P-state dithering may be used to maintain a ratio close to unity between an average power consumed on the chip and a power target, such as a TDP value. By maintaining such a ratio close to unity, the chip may seek to maximize performance while still consuming a desired amount of power (e.g., an amount of power dissipated by a corresponding cooling system).

[0026] Turning now to FIG. 3, one embodiment of a method 300 for managing multiple discrete operating points to create a relatively stable virtual operating point is shown. For purposes of discussion, the steps in this embodiment and subsequent embodiments of methods described later are shown in sequential order. However, some steps may occur in a different order than shown, some steps may be performed concurrently, some steps may be combined with other steps, and some steps may be absent in another embodiment.

[0027] In block 302, a power usage target for the die of a chip is initialized. Any of a variety of methods for selecting a power usage target, including those described earlier, may be used. In block 304, an initial discrete power-performance state (P-state) for the die is determined at a

given workload. Software, such as firmware, and/or hardware may determine the P-state. A process such as that depicted in FIG. may be used to determine a P-state. Average power consumption for the die is then measured in block 306. Further details of such a measurement are provided later. In block 308, the measured average power consumption is compared to a power target and a difference determined. In one embodiment, this difference may be accumulated with other determined differences. For example, an accumulated difference value may be maintained. While a transition to another P-state could be initiated in response to detecting a difference between the target and the measured power, such an approach may cause unproductive transitions between P-states. Therefore, in one embodiment, various techniques are utilized to prevent such unwanted transitions. In one embodiment, a given delta is used for purposes of determining when an accumulated difference is sufficient to cause a P-state transition. For example, if the measured average power exceeds the target power by more than the given delta, then a transition to a higher P-state may be initiated or otherwise permitted. Similarly, if the measured average power falls below the target power by at least a given delta, then a transition to a lower P-state may be initiated or otherwise permitted. It is noted that the above described deltas could be utilized as values which must be exceeded or simply met as desired. In some cases, a given delta for both a transition to a higher or lower state may be the same (in terms of absolute value). In other embodiments, a different delta value could be used for transitions to a higher state than transitions to a lower state. All such embodiments are contemplated. Additionally, the delta values may or may not be programmable in various embodiments. In addition to the above, further conditions may be utilized to determine whether a P-state transition may occur. For example, in some embodiments a minimum residency in a P-state may be required before a transition is permitted. Further discussion of such a minimum residency will be provided in the discussion of FIG. 5.

In block 310, when the accumulated difference reaches a given delta, a transition to another P-state may occur. For example, a power manager may select a lower power-performance state (P-state) than a current P-state if a signed accumulated difference corresponding to the above comparisons over time falls below a given threshold. The lower P-state may generally correspond to a power consumption value that is less than the power target. Similarly, the power manager may select a higher P-state than a current P-state if the signed accumulated difference corresponding to the above comparisons over time exceeds the given delta.

[0028] Referring again to FIG. 2, the activity level 220 may track the workload. As shown in the embodiment of FIG. 2, P-state transitions occur when the ratio of average power to power

target varies from unity. In various embodiments, a given threshold variance may be used to determine when a P-state transition occurs. Alternatively, an accumulated signed difference may be found between the measured average power and the power target, rather than a ratio. Before further details are provided, one embodiment of measuring an activity level to track the workload is described.

[0029] Referring to FIG. 4, one embodiment of a core power management 400 is shown. Here, core 102 may be any integrated circuit (IC). In one embodiment, core 102 may be a processor core. A processor core may have an on-die instruction and data cache. The processor core may be a superscalar processor with a single pipeline or multiple pipelines. In another embodiment, core 102 may be an application specific IC (ASIC). Any transistor family may be used to implement core 102. Examples include metal oxide semiconductor field effect transistors (MOSFETs) and bipolar junction transistors (BJTs).

[0030] A functional block 110 may include transistors configured to perform logic functions, data storage, or other. For power management purposes, functional block 110 may be divided into units 132a-132d. As used herein, elements referred to by a reference numeral followed by a letter may be collectively referred to by the numeral alone. For example, units 132a-132d may be collectively referred to as units 132. In one embodiment, units 132 may not correspond to functional components of a processor, such as a reorder buffer, a memory management unit, an execution unit, and so forth. Rather, units 132 may be selected based on the types of signals to be sampled for power management purposes. For example, in one embodiment, signals selected to be sampled include clock enable signals routed to local clock distribution blocks.

[0031] The selection of which signals to sample during a particular clock cycle may correspond to how well the selection correlates to the amount of switching node capacitance within units 132. The selected signals to be sampled, such as clock enable signals, may overlap functional blocks in the floorplan. Therefore, the division separating, for example, unit 132a and 132b may not correspond to a division in the floorplan. Units 132 are units that consume power and this power is to be measured in real-time. The activity level of the die associated with a current workload may correspond to values, or weights, associated with selected signals to be sampled.

[0032] In one embodiment, Power Monitor 130 may be used to collect data from units 132, such as the logic values of all the given sampled signals. In one embodiment, the values of the sampled signals may be scanned out in a serial manner. Therefore, the selected signals may be sampled in a single clock cycle from each of Units 112 and serially scanned out before the next

sample is performed. After collecting the data, Power Monitor 130 may calculate a power consumption estimate. One Monitor Control 132 may correspond to each Unit 132. In alternative embodiments, a Monitor Control 132 may collect data for two or more Units 112 and calculate total power consumption estimation for those Units 112. In yet another embodiment, one Monitor Control 132 (i.e. Control 132a) may have a signal interface with one or more other Monitor Controls 132 (i.e. Controls 132b-132d) in order to collect data from the one or more Monitor Controls 132 (i.e. Controls 132b-132d). Then a power consumption estimate for the one or more Monitor Controls 132 may be calculated.

[0033] The signals Sample 120 and Dataout 122 may be control and data signals used for power management purposes. The interface signals between Power Monitor 110 and Functional Block 130 may comprise any necessary number of signals and communication protocols. In one embodiment, the control signal Sample 120 may be asserted for a single clock cycle only during a chosen repeating interval, such as every 100 clock cycles. In one embodiment, at a given number of clock cycles after the control signal Sample 120 is asserted, the data signal Dataout 122 may begin providing a logic value for a different sampled signal each clock cycle. In other words, the data signal Dataout 122 may be used to scan out a chain of values comprising the logic values of the sampled signals at a particular cycle. Also, in other embodiments, there may not be a single pair of signals between each Monitor Control 132 and Unit 112 pair. In an alternative embodiment, additional signals may be included in order for a Monitor Control 132 to poll a Unit 112, for a Unit 112 to acknowledge to a Monitor Control 132 that it is ready to convey output data.

[0034] A multiple number of samples may be taken during a given time interval. The determination of the number of intermittent clock cycles to use before computing an activity level may depend on the desired accuracy and confidence of the sampled data. A spreadsheet, or a look-up table, may be generated using both statistical analysis and measurements of both the real power consumption of an application and estimated power consumption from a sampling. A confidence level and an error rate may be chosen to further develop the statistical analysis. An example of a real-time power estimation method includes Application Serial No. 12/101,598, filed April 11, 2008, entitled "Sampling Chip Activity for Real Time Power Estimation", which is incorporated herein by reference in its entirety.

[0035] When the Power Monitor 130 calculates a power consumption estimate from the data received from Functional Block 110 over repeated intervals, the Power Monitor 130 has determined a power profile of the currently running application(s). This determination is

conveyed to the Power Manager 140. The Power Manager 140 may alter an operating point of functional block 110 in order to decrease (or increase) power if the application is above (below) a threshold limit. For example, the Power Manager 140 may cause a boost or a throttle of a current P-state to transition to another given P-state.

[0036] In one embodiment, during the specified time period named above, the Power Manager 140 may compute a signed running accumulated difference between the power profile provided by the Power Monitor 130 and the power target. Again, the power target may be a thermal design point (TDP). The accumulated difference may be calculated at the end of each given time interval as AccTdpDelta = AccTdpDelta + (TDP – RcvdPwr). Here, the variable AccTdpDelta is the signed running accumulated difference. The variable TDP is the assigned thermal design power, or an equivalent number of thermal credits. The variable RcvdPwr is the power consumption estimation received from the Power Monitor 130. This value may track the activity level of the die by measuring the sampled signals in the functional blocks 110.

[0037] If the measured activity level represented by the variable RcvdPwr is higher than the TDP, then the accumulated value AccTdpDelta drifts toward a negative value. When the accumulated value reaches a negative given threshold, the power manager may determine to throttle the current P-state. Referring again to FIG. 1, an example of throttling a current P-state would be to transition from P-state P_1 to P-state P_2. Such a condition may occur when the activity level is high within the core. If the activity level remains at a high value, over time the power manager may continue to throttle the current P-state.

[0038] If the measured activity level is lower than the TDP, then the accumulated value AccTdpDelta drifts toward a positive value. When the accumulated value reaches a positive given threshold, the Power Manager 140 may determine to boost the current P-state. Referring again to FIG. 1, an example of boosting a current P-state would be to transition from P-state P_2 to P-state P_1. Such a condition may occur when the activity level is low within the core. If the activity level remains at a low value, over time the power manager may continue to boost the current P-state.

[0039] The Power Manager 140 may be able to provide quicker responses to potential thermal problems in core 102 when the information sent from the Power Monitor 130 corresponds to actual activity levels and power consumption within core 102 and not temperature information. Analog or digital thermal sensors placed throughout the die of a semiconductor chip die may determine a temperature waveform over time. The thermal sensors provide information as to when the die heats up in a particular area due to increased compute activity. However, these

sensors respond to each change in thermals, whether it's driven by a compute-related increase in power consumption in the core 102 or by an external environmental factor, such as a rise in ambient temperature. For example, surrounding servers in a rack system in a data center may cause a rise in ambient temperature. The amount of switching capacitance within a particular core may not change over a time interval, but the sensors may report higher thermal energy consumption due to the rise in ambient temperature. In addition, there is a time delay between a compute-related increase in power consumption and a temperature increase. Therefore, while attempting to maintain a ratio of unity between average power consumption and a power target, measurements associated with an activity level and switching capacitance within a core versus measurements of temperature may provide better results.

[0040] Turning now to FIG. 5, another embodiment of a method 500 for managing multiple discrete operating points to create a stable virtual operating point is shown. For purposes of discussion, the steps in this embodiment and subsequent embodiments of methods described later are shown in sequential order. However, some steps may occur in a different order than shown, some steps may be performed concurrently, some steps may be combined with other steps, and some steps may be absent in another embodiment.

[0041] In block 502, a semiconductor chip executes instructions of one or more software applications. If the end of a given sampling interval is reached (conditional block 504), then in block 506, a power consumption estimate is determined and conveyed to a power manager. A power consumption estimate may be found by sampling selected signals in functional blocks and associating corresponding weights to the sampled signals as described earlier. A counter corresponding to a count of determined power estimation values may be incremented. In block 508, the power manager may compute a signed accumulated difference between a given power target, such as a TDP, and the received power consumption value. The computation as described earlier may be AccTdpDelta = AccTdpDelta + (TDP – RcvdPwr), wherein the variable RcvdPwr represents the received power consumption value.

[0042] It is noted when a P-state transition occurs, the signed accumulated difference, AccTdpDelta, may still exceed a given threshold by the time another sample interval occurs. For example, a boost from a P-state P_2 to a P-state P_1 may occur due to the value AccTdpDelta is greater than a positive boost threshold. After the P-state transition, at the next time interval, the signed accumulated difference, AccTdpDelta, may still be greater than the positive boost threshold. There may not have been sufficient time for the measured power consumption value to exceed the TDP value. Therefore, it's possible to rapidly continue boosting the current P-state

before it is determined the current P-state provides the best power-performance operating point at a given time.

[0043] In order to avoid rapid P-state transitions as described above, a following P-state transition may not be allowed to occur for a given time after a current P-state transition. The given time may be referred to as a minimum residency. In one embodiment, a counter may be used to determine whether a permissible amount of time has elapsed following a current P-state transition. The counter value may be compared to a given threshold. Alternatively, the counter may be loaded with the given threshold and decremented to a value of zero. When this permissible amount of time has elapsed, another P-state transition may occur.

[0044] In addition, value aging for the accumulated difference AccTdpDelta may be used. Value aging may aid in preventing overheating of the chip die. An aged accumulated difference may remain transitioning, or dithering, between high power P-states for a long time. The accumulated difference, AccTdpDelta, may continue accumulating when a P-state is throttled. After a period of time, the accumulated difference may no longer represent an actual thermal energy headroom available for boost. Therefore, from time to time, the accumulated difference AccTdpDelta may be set to a reset value at the end of a given time period. The reset value may vary from a fraction of the current value of the accumulated difference AccTdpDelta to zero. A value stored in a configuration register may be used to determine the reset value. For example, a first stored value may correspond to a reset value equal to the accumulated difference AccTdpDelta. A second stored value may correspond to a reset value equal to the accumulated difference AccTdpDelta shifted (e.g., divide by 2, divide by 4, and so forth). A third stored value may correspond to a reset value equal to zero.

[0045] A counter, which may be decrementing in one example, may set the given time period referred to above. The counter may load a value N, which is stored in a configuration register. After N samples occur, wherein the accumulated difference AccTdpDelta is updated at the end of each sample, or time interval, accumulated difference AccTdpDelta may be reset. The counter loaded with the value N may be reset each time the accumulated difference AccTdpDelta changes sign.

[0046] Referring again to method 500 in FIG. 5, if a P-state minimum residency time period has not been reached (conditional block 510), then a check is performed regarding whether N samples have occurred. At the end of each sample interval, a new power consumption value for the die is determined and a new value is computed for the accumulated difference AccTdpDelta.

If a count of N samples has occurred (conditional block 516), then in block 518, accumulated difference AccTdpDelta is set to a given reset value as described above.

[0047] If a P-state minimum residency time period has been reached (conditional block 510), then a check is performed regarding the boost and throttling thresholds. A comparison may be performed to determine whether the signed accumulated difference AccTdpDelta exceeds a threshold. For example, the accumulated difference AccTdpDelta may be greater than a positive boost threshold. Alternatively, the accumulated difference AccTdpDelta may be less than a negative throttle threshold. If the accumulated difference AccTdpDelta exceeds a threshold (conditional block 512), then in block 514, a corresponding next P-state is selected for the chip die. For example, if the signed accumulated difference AccTdpDelta is greater than a positive (boost) threshold, then a transition to a higher power-performance P-state than the current P-state may occur. Alternatively, if the signed accumulated difference AccTdpDelta is less than a negative throttling threshold, then a transition to a lower power-performance P-state than the current P-state may occur.

[0048] In one embodiment, if a processor comprises multiple cores, a power consumption estimate may be computed within each core. In addition, each core may determine a signed accumulated difference, AccTdpDelta. When any of the multiple cores exceeds a boost or throttling threshold, the P-state for the entire processor may transition accordingly. Then control flow of method 500 moves to conditional block 516.

[0049] It is noted that the above-described embodiments may comprise software. In such an embodiment, program instructions and/or a database (both of which may be referred to as "instructions") that represents the described systems and/or methods may be conveyed or stored on a computer readable medium. Generally speaking, a computer accessible storage medium may include any storage media accessible by a computer during use to provide instructions and/or data to the computer. For example, a computer accessible storage medium may include storage media such as magnetic or optical media, e.g., disk (fixed or removable), tape, CD-ROM, or DVD-ROM, CD-R, CD-RW, DVD-R, DVD-RW, or Blu-Ray. Storage media may further include volatile or non-volatile memory media such as RAM (e.g. synchronous dynamic RAM (SDRAM), double data rate (DDR, DDR2, DDR3, etc.) SDRAM, low-power DDR (LPDDR2, etc.) SDRAM, Rambus DRAM (RDRAM), static RAM (SRAM), etc.), ROM, Flash memory, non-volatile memory (e.g. Flash memory) accessible via a peripheral interface such as the Universal Serial Bus (USB) interface, etc. Storage media may include microelectromechanical

systems (MEMS), as well as storage media accessible via a communication medium such as a network and/or a wireless link.

[0050] Additionally, the instructions may comprise behavioral-level descriptions or register-transfer level (RTL) descriptions of the hardware functionality in a programming language such as C, or a design language (e.g., HDL) such as Verilog, VHDL, or a database format such as GDS II stream format (GDSII). These instructions may then be read and used to fabricate hardware comprising the system (or portions of the system). In some cases the description may be read by a synthesis tool (e.g., program code running on a computing device) to form an implementation of the design. For example, such a tool may be used to synthesize the description to produce a netlist comprising a list of gates from a synthesis library. The netlist may generally comprise a set of gates which also represent the functionality of the hardware comprising the system. The netlist may then be placed and routed to produce a data set describing geometric shapes to be applied to masks. The masks may then be used in various semiconductor fabrication steps to produce a semiconductor circuit or circuits corresponding to the system. Alternatively, the instructions on the computer accessible storage medium may be the netlist (with or without the synthesis library) or the data set, as desired. Additionally, the instructions may be utilized for purposes of emulation by a hardware based type emulator, such as those from vendors Cadence®, EVE®, and Mentor Graphics®. For example, in such an embodiment the instructions may be utilized to configure FPGA based hardware to perform according to the design. Numerous such embodiments are possible and are contemplated.

[0051] Although the embodiments above have been described in considerable detail, numerous variations and modifications will become apparent to those skilled in the art once the above disclosure is fully appreciated. It is intended that the following claims be interpreted to embrace all such variations and modifications.

WHAT IS CLAIMED IS

1. A computing system comprising:

one or more functional blocks, each configured to produce data corresponding to an activity level of a respective block; and

a power manager, wherein the power manager is configured to:

determine an average power consumption during a given time interval for a functional block of the one or more functional blocks based at least in part on said data;

select a lower power-performance state (P-state) than a current P-state in response to determining the average power consumption exceeds a desired power consumption;

select a higher P-state than a current P-state in response to determining the average power consumption value is less than the desired power consumption.

2. The computing system as recited in claim 1, wherein the desired power consumption corresponds to a virtual P-state, said virtual P-state being lower than the higher P-state and higher than the lower P-state, and wherein the power manager is configured to alternately select the higher P-state and/or lower P-state in order to produce an average power consumption over time that more closely corresponds to the desired power consumption than either the higher P-state or the lower P-state.

3. The computing system as recited in claim 1, wherein the power manager is configured to:

select the lower P-state in further response to determining the average power consumption exceeds the desired power consumption by at least a first delta amount, the first delta amount having an absolute value greater than zero; and

select the higher P-state in further response to determining the average power consumption is less than the desired power consumption by at least a second delta amount, the second delta amount having an absolute value greater than zero.

4. The computing system as recited in claim 3, wherein the desired power consumption corresponds to a thermal design power value for the one or more functional blocks.

5. The computing system as recited in claim 2, wherein the power manager is further configured to select a P-state two or more states away from the current P-state based on a rate of reaching a corresponding threshold.

6. The computing system as recited in claim 1, wherein prior to selecting a new P-state, the power manager is further configured to determine the one or more functional blocks have been operating in a current P-state for a given amount of time.

7. The computing system as recited in claim 1, wherein the power manager is further configured to:

determine a signed accumulated difference over time between the desired power consumption and a plurality of functional block power consumption values;

select the lower P-state in further response to determining the signed accumulated difference is greater than a first delta value, the first delta value having an absolute value greater than zero; and

select the higher P-state in further response to determining the signed accumulated difference is less than a second delta, an absolute value of the second delta value being greater than zero.

8. The computing system as recited in claim 7, wherein the time interval comprises a maximum count of N functional block sample intervals, and wherein the power manager is further configured to reset the count responsive to detecting the signed accumulated difference has changed sign or the count has reached N, whichever occurs first.

9. A method for managing multiple discrete operating points to create a stable virtual operating point, the method comprising:

producing data corresponding to activity levels of one or more functional blocks;

determining an average power consumption during a given time interval for a functional block of the one or more functional blocks based at least in part on said data;

selecting a lower power-performance state (P-state) than a current P-state in response to determining the average power consumption exceeds a desired power consumption; and

selecting a higher P-state than a current P-state in response to determining the average power consumption value is less than the desired power consumption value by at least a second delta value.

10. The method as recited in claim 9, wherein the desired power consumption value corresponds to a virtual P-state, said virtual P-state being lower than the higher P-state and higher than the lower P-state, and wherein the method further comprises alternately selecting the higher P-state and/or lower P-state in order to produce an average power consumption level over time that more closely corresponds to the desired power consumption level than either the higher P-state or the lower P-state.

11. The method as recited in claim 9, further comprising:

selecting the lower P-state in further response to determining the average power consumption exceeds the desired power consumption by at least a first delta amount, the first delta amount having an absolute value greater than zero; and

selecting the higher P-state in further response to determining the average power consumption is less than the desired power consumption by at least a second delta amount, the second delta amount having an absolute value greater than zero.

12. The method as recited in claim 11, wherein the desired power consumption value corresponds to a thermal design power value for the one or more functional blocks.

13. The method as recited in claim 10, further comprising selecting a P-state two or more states away from the current P-state based on a rate of reaching a corresponding threshold.

14. The method as recited in claim 9, wherein prior to selecting a new P-state, the method further comprises determining the one or more functional blocks have been operating in the current P-state for a given amount of time.

15. The method as recited in claim 9, further comprising:

determining a signed accumulated difference over time between the desired power consumption value and a plurality of functional block power consumption values;

selecting the lower P-state in further response to determining the signed accumulated difference is greater than a first delta value, the first delta value having an absolute value greater than zero; and

selecting the higher P-state in further response to determining the signed accumulated difference is less than a second delta, an absolute value of the second delta value being greater than zero.

16. The method as recited in claim 15, wherein the time interval comprises a maximum count of N functional block sample intervals, and wherein the method further comprises resetting the count responsive to detecting the signed accumulated difference has changed sign or the count has reached N, whichever occurs first.

17. A computer readable storage medium storing program instructions operable to manage multiple discrete operating points to create a stable virtual operating point, wherein the program instructions are executable to:

produce data corresponding to activity levels of one or more functional blocks;

determine an average power consumption during a given time interval for a functional block of the one or more functional blocks based at least in part on said data;

select a lower power-performance state (P-state) than a current P-state in response to determining the average power consumption exceeds a desired power consumption; and

select a higher P-state than a current P-state in response to determining the average power consumption value is less than the desired power consumption value by at least a second delta value.

18. The storage medium as recited in claim 17, wherein the desired power consumption value corresponds to a virtual P-state, said virtual P-state being lower than the higher P-state and higher than the lower P-state, and wherein the method further comprises alternately selecting the higher P-state and/or lower P-state in order to produce an average power consumption level over time that more closely corresponds to the desired power consumption level than either the higher P-state or the lower P-state.

19. The storage medium as recited in claim 17, wherein the program instructions are further executable to:

select the lower P-state in further response to determining the average power consumption exceeds the desired power consumption by at least a first delta amount, the first delta amount having an absolute value greater than zero; and

select the higher P-state in further response to determining the average power consumption is less than the desired power consumption by at least a second delta amount, the second delta amount having an absolute value greater than zero.

20. The storage medium as recited in claim 19, wherein prior to selecting a new P-state the program instructions are further executable to determine the one or more functional blocks have been operating in the current P-state for a given amount of time.

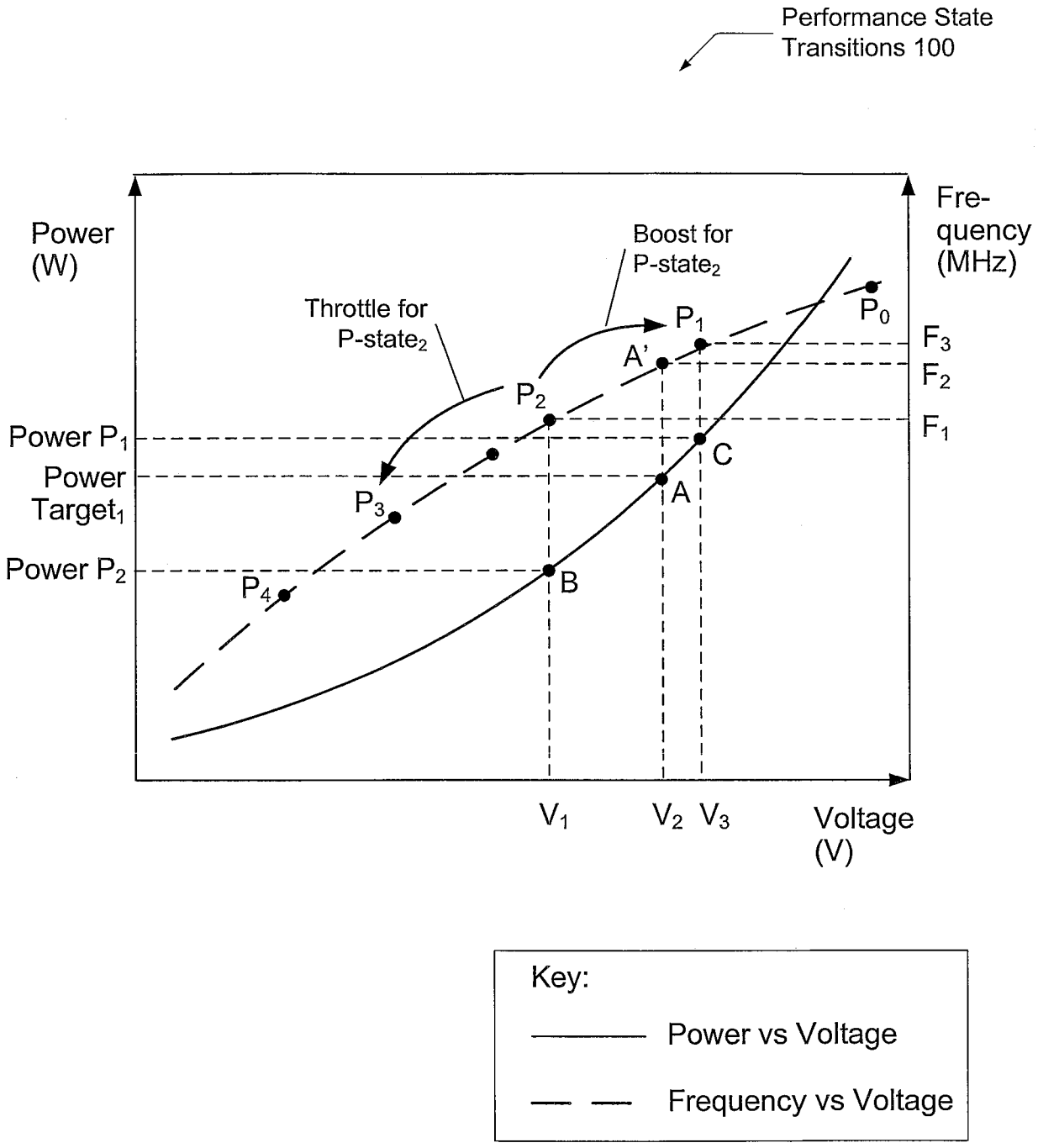
Performance State Transitions 100
Power (W)
Fre-quency (MHz)
Boost for P-state2
Throttle for P-state2
P0
P1
P2
P3
P4
A'
A
B
C
F3
F2
F1
Power P1
Power Target1
Power P2
V1
V2
V3
Voltage (V)
Key:
Power vs Voltage
Frequency vs Voltage

FIG. 1

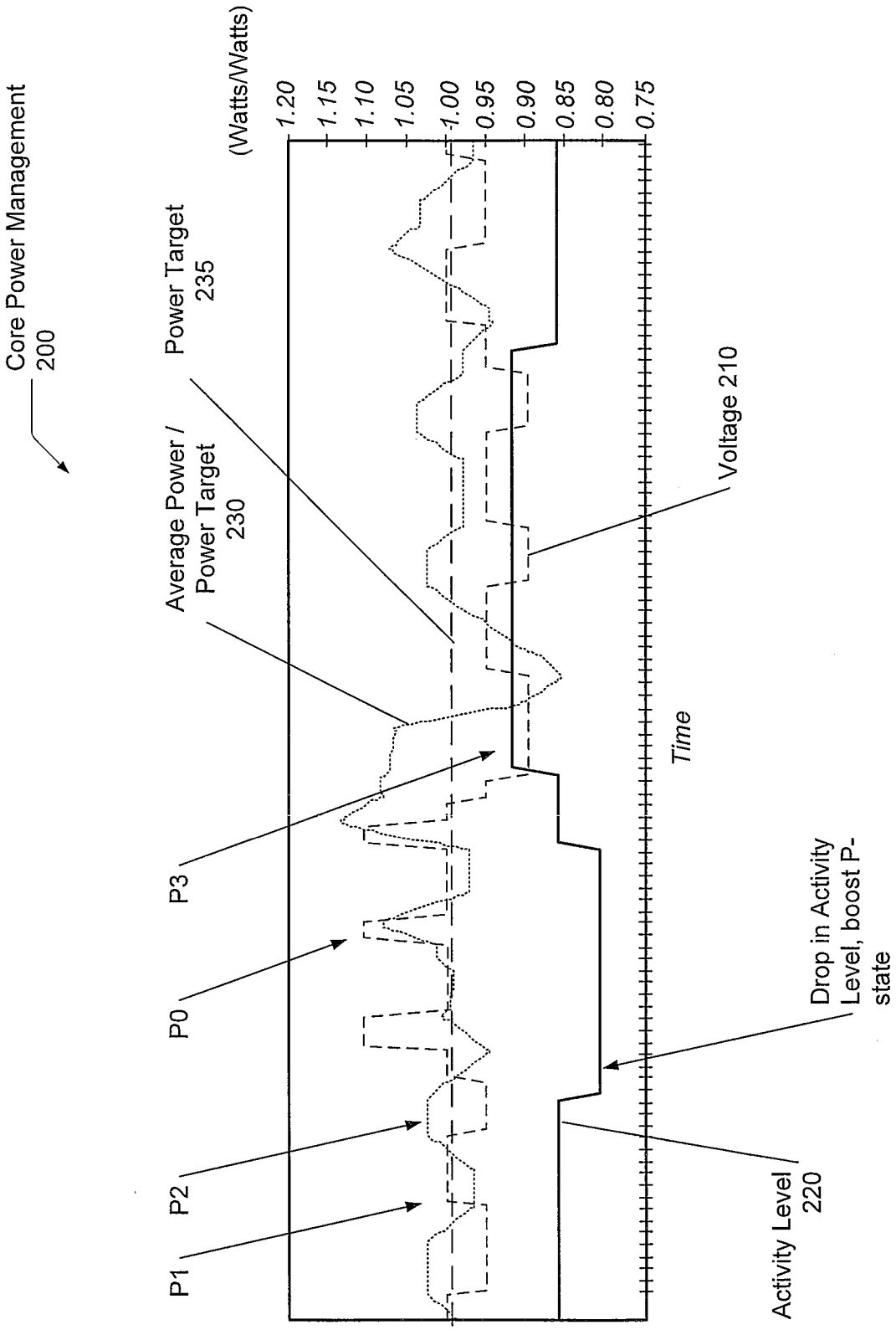
Core Power Management
200
(Watts/Watts)
1.20
1.15
1.10
1.05
1.00
0.95
0.90
0.85
0.80
0.75
Power Target
235
Average Power /
Power Target
230
Voltage 210
P3
P0
P2
P1
Time
Drop in Activity
Level, boost P-
state
Activity Level
220

FIG. 2

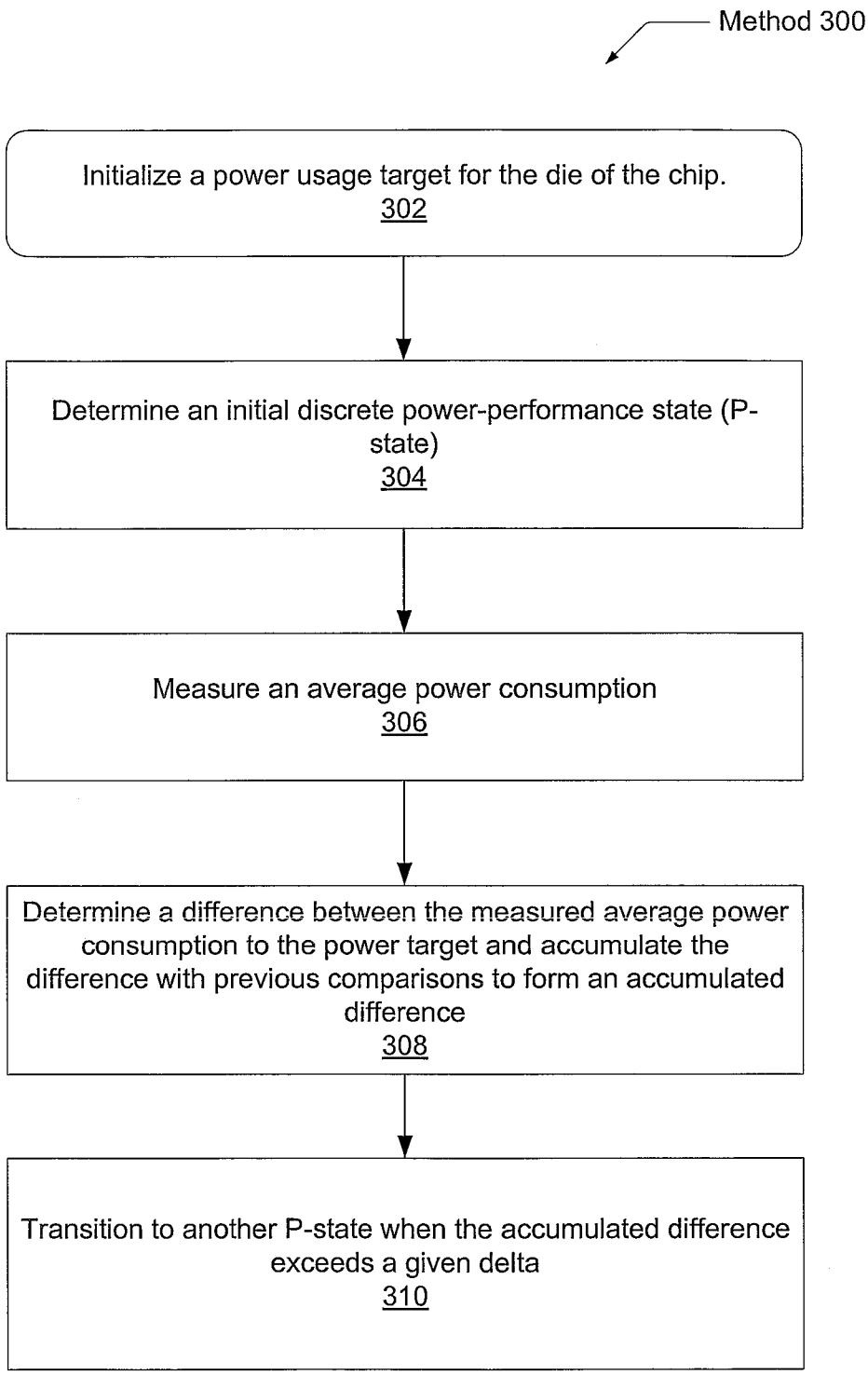
Method 300
Initialize a power usage target for the die of the chip.
302
Determine an initial discrete power-performance state (P-state)
304
Measure an average power consumption
306
Determine a difference between the measured average power consumption to the power target and accumulate the difference with previous comparisons to form an accumulated difference
308
Transition to another P-state when the accumulated difference exceeds a given delta
310

FIG. 3

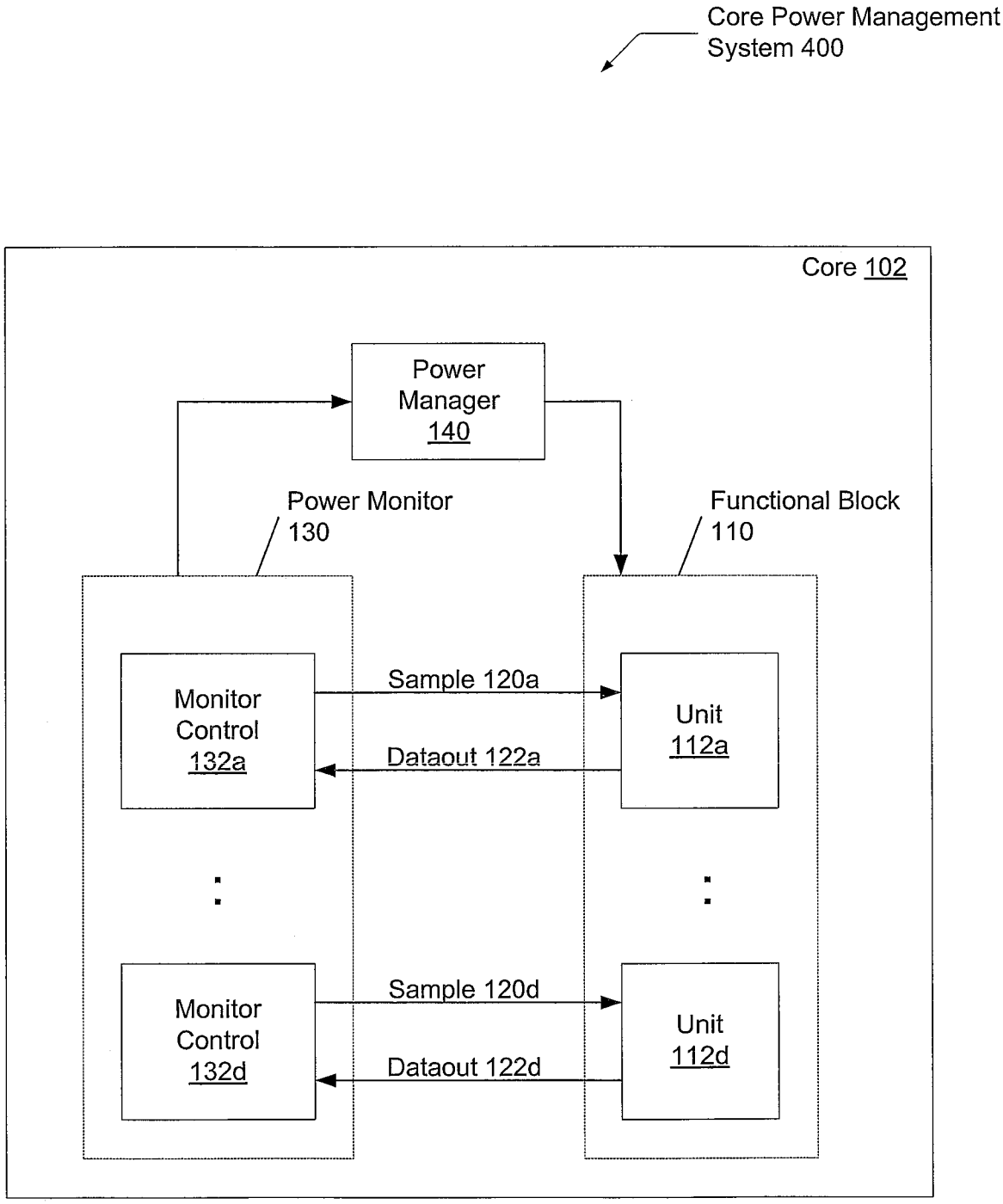
Core Power Management System 400
Core 102
Power Manager 140
Power Monitor 130
Functional Block 110
Monitor Control 132a
Sample 120a
Dataout 122a
Unit 112a
Monitor Control 132d
Sample 120d
Dataout 122d
Unit 112d

FIG. 4

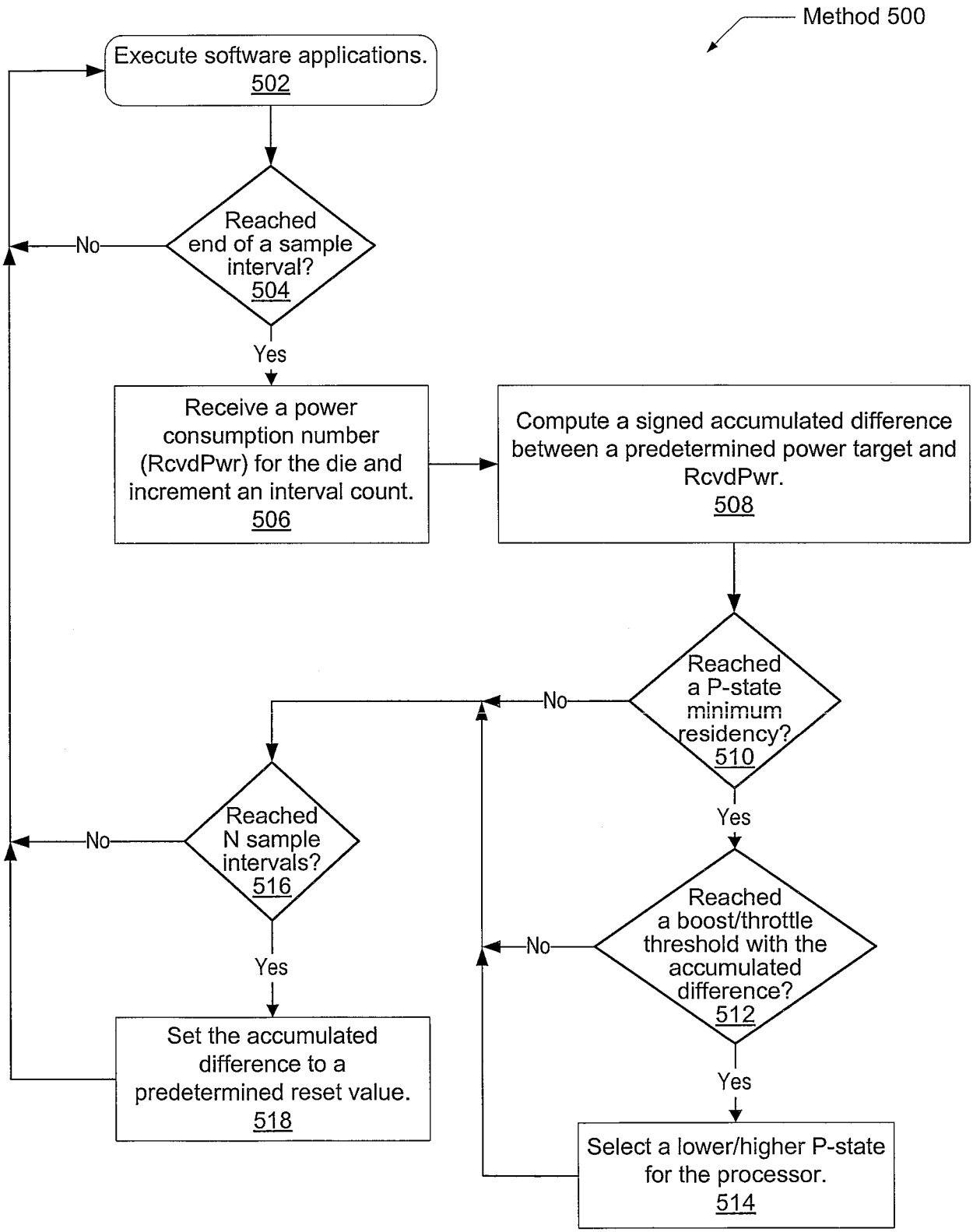
Method 500
Execute software applications.
502
Reached end of a sample interval?
504
No
Yes
Receive a power consumption number (RcvdPwr) for the die and increment an interval count.
506
Compute a signed accumulated difference between a predetermined power target and RcvdPwr.
508
Reached a P-state minimum residency?
510
No
Yes
Reached a boost/throttle threshold with the accumulated difference?
512
No
Yes
Select a lower/higher P-state for the processor.
514
Reached N sample intervals?
516
No
Yes
Set the accumulated difference to a predetermined reset value.
518

FIG. 5

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2011/041291

A. CLASSIFICATION OF SUBJECT MATTER
INV. G06F1/32
ADD.

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

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
G06F

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

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)
EPO-Internal, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	WO 2007/024396 A1 (APPLE COMPUTER [US]; CONROY DAVID G [US]; COX KEITH ALAN [US]; CULBERT) 1 March 2007 (2007-03-01) abstract; figures 1,2,3,8,9,10,12-14 paragraph [0045] - paragraph [0050] paragraph [0059] - paragraph [0078] paragraph [0116] - paragraph [0135] ----- -/--	1-20

[X] Further documents are listed in the continuation of Box C.

[X] See patent family annex.

* Special categories of cited documents :

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

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

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

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

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

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

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

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

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
21 September 2011	28/09/2011

Name and mailing address of the ISA/	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016	Rousset, Antoine

1

Form PCT/ISA/210 (second sheet) (April 2005)

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2011/041291

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CHARLES LEFURGY ET AL: "Server-Level Power Control", AUTONOMIC COMPUTING, 2007. ICAC '07. FOURTH INTERNATIONAL CONFERENCE O N, IEEE, PI, 1 June 2007 (2007-06-01), pages 1-10, XP031116521, ISBN: 978-0-7695-2779-6 abstract; figures 1,3,7 page 4, column 1, line 4 - page 7, column 1, line 10	1,2,9, 10,18
A	US 2003/125900 A1 (ORENSTIEN DORON [IL] ET AL) 3 July 2003 (2003-07-03) abstract; figure 1 paragraph [0016] - paragraph [0021]	1-20
A	US 2004/268166 A1 (FARKAS KEITH ISTVAN [US] ET AL) 30 December 2004 (2004-12-30) abstract; figures 2,3,9 paragraph [0038] - paragraph [0061] paragraph [0116] - paragraph [0118]	1-20

1

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/US2011/041291

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 2007024396 A1	01-03-2007	US 2007050646 A1	01-03-2007
		US 2007050650 A1	01-03-2007
		US 2007049133 A1	01-03-2007
		US 2007049134 A1	01-03-2007
		US 2009276651 A1	05-11-2009
		US 2011060932 A1	10-03-2011
		US 2011001358 A1	06-01-2011
US 2003125900 A1	03-07-2003	AU 2002360784 A1	30-07-2003
		CN 1526089 A	01-09-2004
		GB 2393008 A	17-03-2004
		TW I281608 B	21-05-2007
		WO 03060678 A2	24-07-2003
US 2004268166 A1	30-12-2004	NONE	