

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

(21) International Application Number:

PCT/EP2011/060648

(22) International Filing Date:

24 June 2011 (24.06.2011)

(25) Filing Language:

English

(26) Publication Language:

English

(71) Applicant (for all designated States except US): **FUJITSU TECHNOLOGY SOLUTIONS INTELLECTUAL PROPERTY GMBH** [DE/DE]; Mies-van-der-Rohe-Straße 8, 80807 München (DE).

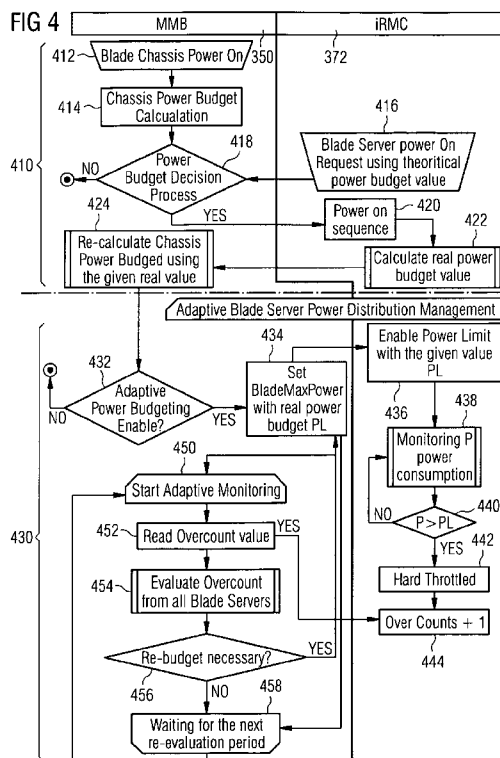
(72) Inventors; and

(75) Inventors/Applicants (for US only): **LUHUR, Julius** [ID/—]; 11F, 459 Zhongxiao East Road Section 5, Xinyi District, Taipei, 110 (TW). **SCHRÄDER, Bernhard** [DE/DE]; Untereichen 10, 33129 Delbrück (DE).(74) Agent: **EPPING HERMANN FISCHER PATENTANWALTSGESELLSCHAFT MBH**; Ridlerstraße 55, 80339 München (DE).

(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, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM,

[Continued on next page]

(54) Title: **BLADE SERVER SYSTEM AND METHOD FOR POWER MANAGEMENT**

(57) Abstract: The invention relates to a blade server system (300), comprising an enclosure (310), at least one power supply unit (332, 334, 336), at least one management blade (350) having at least one controller (352), and at least one blade server (342, 344, 346, 348) having at least one processor (362, 364, 366, 368) and a remote management controller (372, 374, 376, 378). The controller (352) of the management blade (350) is configured to query the at least one power supply unit (332, 334, 336) to determine a total available power for the blade server system (300), to determine a power limit for the at least one blade server (342, 344, 346, 348). The remote management controller (372, 374, 376, 378) of the at least one blade server (342, 344, 346, 348) is configured to increase a counter associated with the blade server if an associated power monitoring means indicates that a current power consumption of the at least one blade server exceeds the stored power consumption limit. The controller (352) of the management blade (350) is further configured to determine a new power consumption limit for the at least one blade server (342, 344, 346, 348) based on the total available power and the value of the counter associated with the at least one blade server (342, 344, 346, 348) after a predetermined monitoring interval. The invention further relates to a method for power management and a data processing system (100).

WO 2012/175144 A1



TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, **Published:**
ML, MR, NE, SN, TD, TG).

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

Description

Blade server system and method for power management

- 5 The present application relates to data processing systems in general. More specifically, the present invention relates to devices, systems and methods for power management within data processing systems such as blade server systems.
- 10 Data processing systems are used for many applications and in many context. Depending on their use, the required processing power of data processing systems vary greatly. Furthermore, over time, more processing power may be required to satisfy the need of a specific application. For example, if a busi-
- 15 ness grows, server computers used to store and process business critical data may need to be upgraded in order to process the increased amount of data. As a consequence, many data processing systems are designed in a modular fashion.
- 20 One example of such a modular design of a data processing systems are so-called blade server systems. A blade server system typically comprises a number of so-called blade servers, each of which comprises at least one processor for processing data and associated working memory. Individual blade
- 25 servers may be plugged into mounting positions typically arranged in a front portion of a so-called blade enclosure in order to increase the processing power of the overall blade server system.
- 30 Furthermore, a blade server system typically comprises a number of so-called infrastructure modules, such as power supply units, network switches, storage modules and the like. The infrastructure modules supply common infrastructure services

for all or at least some of the blade servers. Typically, at least one power supply unit and one network connection module are placed in mounting positions arranged in a back portion of a blade server enclosure.

5

Furthermore, known blade server systems typically also comprises a so-called blade management module for monitoring, configuring and administering the various plug-in modules of the blade server system. Finally, all of the modules of a blade server system are typically connected by a so-called midplane arranged in the blade enclosure.

In a conventional blade server system, the blade management module determines a current system configuration, typically during a start-up of the blade server system. For example, the blade management module may query the number and output power of all power supply units arranged in a blade server system. Furthermore, the blade management module may query the number and type of blade servers and other plug-in modules of the blade server system in order to determine their overall power consumption. The power consumption of individual modules may be provided by the modules themselves, may be determined by the blade management unit from a database storing module-specific power consumption values or may be specified manually by means of a user interface.

The blade management unit may then determine how many of the blade servers and other modules of the blade server system can be powered up using the available power supply units and activate them accordingly. If there is insufficient power provided by the power supply units in order to supply all blade servers, one or more server modules may not be activated.

The described approach to power management in modular data processing systems may result in an inefficient use of system resources. For example, one or more blade server modules may be deactivated because other blade server modules specify
5 power requirements that exceed their actual power consumption. Furthermore, modification of the allocation of a system power to individual components may result in the need for a manual configuration. A manual configuration of the power allocation within an electronic data processing system may be
10 difficult and error prone and therefore may lead to further inefficiencies.

Therefore, there exists a need for improved apparatuses, systems and methods for power management in data processing systems.
15

According to a first aspect of the invention, a blade server system is described. The blade server system comprises an enclosure having a plurality of first mounting positions to accommodate a plurality of power supply units and a plurality
20 of second mounting positions to accommodate a plurality of blade modules. The blade server system further comprises at least one power supply unit accommodated in one of the plurality of first mounting positions, at least one management blade and at least one blade server accommodated in one of
25 the plurality of second mounting positions. The management blade comprises at least one controller and the at least one blade server comprises at least one processor, a memory accessible to the at least one processor and a remote management controller. The controller of the management blade is
30 configured to query the at least one power supply unit to determine a total available power for the blade server system. The controller of the management blade is further configured

to determine a power limit for the at least one blade server and to provide the determined power limit to the remote management controller of the at least one blade server. The remote management controller of the at least one blade server is configured to store the power limit provided by the controller of the management blade and to increase a counter associated with the blade server if an associated power monitoring means indicate that a current power consumption of the blade server exceeds the stored power consumption limit. The controller of the management blade is further configured to determine a new power consumption limit for the at least one blade server based on the total available power and the value of the counter associated with the at least one blade server after a predetermined monitoring interval.

The described blade server system allows an adaptive and dynamic reallocation of total power of a blade enclosure to different blade servers. Therefore, the power allocation within the blade server system gracefully approaches a demand driven power requirements without manual interference. Furthermore, the described system results in a very low overhead with respect to the management blade.

According to an embodiment, the at least one blade server comprises a power controller, wherein the power monitoring means are comprised in the power controller. Furthermore, the power controller may be adapted to provide a throttle signal to the processor.

According to a second aspect of the present invention, a method for power management in a data processing system comprising a plurality of processing units and at least one power supply unit is provided. The method comprises the steps

of determining a total available power provided by the at least one power supply unit, determining and storing an initial power limit for each one of the plurality of processing units, repeatedly comparing the current power consumption of each one of the plurality of processing units with the respective stored power limit, increasing a counter associated with one of the plurality of processing units if the current power consumption of said one of the plurality of processing units exceeds the respective stored power limit, and determining and storing a new power limit for at least one of the plurality of processing units after a predetermined period of time based on the total available power and the associated counter value of said one of the plurality of processing units.

In one embodiment, the method further comprises the steps of throttling one of the plurality of processing units if a current power consumption of said one of the plurality of processing units exceeds the respective stored power limit.

In one embodiment of the described method, the step of determining and storing a new power limit comprises increasing the new power limit for said one of the plurality of processing units with respect to the previously stored power limit if the associated counter value exceeds a predetermined first threshold.

In another embodiment of the described method, in the step of determining and storing a new power limit, new power limits for at least a first and a second one of the plurality of processing units are determined and stored. The new power limit for the first one of the processing units is increased with respect to the associated previously stored power limit

and the new power limit for the second one of the processing units is decreased with respect to the associated previously stored power limit, if the counter value associated with the first one of the processing units exceeds a predetermined
5 first threshold and the counter value associated with the second one of the processing unit lies below a second threshold. In such an embodiment, power previously allocated to a second processing unit can be transferred to a first processing unit.

10

In another embodiment, the step of determining and storing a new power limit comprises a first and a second phase, wherein in the first phase, the amount of power available for distribution is determined based on the total available power and,
15 in a second phase, the determined power available is distributed between the plurality of processing units based on the counter values associated with the plurality of processing units.

20 In one embodiment, in the second phase, the following two steps of are performed repeatedly while the available power is greater than zero. In a first step, the power available is reduced by an additional power budget and in a second step, a new power limit for one processing unit having an associated
25 counter value exceeding a first threshold value is determined based on a sum of the current power limit of that processing unit and the additional power budget.

In a further embodiment, in the second phase, new power limits for the plurality of processing units are determined in a
30 sequence determined by a priority level associated with the respective processing unit. According to different embodiments, the priority level for each processing unit may be

provided using a user interface of a management unit or may be determined in order of the associated counter values of the plurality of processing units by a management unit.

5 In a further embodiment, in the first phase, the power available is determined based on the difference between the total available power and the sum of all power limits associated with the plurality of processing units. In a further embodiment, a preset power consumption of other components of the
10 data processing system is subtracted from the total available power.

In a further embodiment, in the first phase, for each processing unit having an associated counter value falling below
15 a second threshold, a released power budget is added to the power available, and, in the second phase, a new power limit for said processing units having an associated counter value falling below a second threshold value is based on the current power limit of the respective processing unit reduced by
20 the released power budget.

According to a third aspect of the present invention, an electronic data processing system is described. The electronic data processing system comprises a plurality of processing units, each processing unit comprising at least one
25 processor. The system further comprises at least one power supply unit for providing at least one supply voltage, and a management unit, the management unit comprising at least one controller operatively coupled with the plurality of processing units and the at least one power supply unit. In the
30 electronic data processing system, the management unit is adapted to determine a new power consumption limit for at least one of the plurality of processing units based on a to-

tal available power provided by the at least one power supply unit and a number of times a previously determined power consumption limit was exceeded by said one of the plurality of processing units in a predetermined monitoring interval.

5

The described data processing system allows an adaptive, dynamic and incremental reallocation of power. Therefore, the power allocation within the data processing system gracefully approaches a demand driven power requirements without manual interference. Furthermore, the described system results in a very low overhead with respect to the management unit.

In one embodiment, the data processing system further comprises a plurality of power monitoring means. Each one of the power monitoring means is configured to compare a current power consumption of an associated processing unit of the plurality of processing units with a current power consumption limit provided by the management unit for the associated processing unit.

20

According to further embodiments, the processing units are configured to throttle their power consumption if the associated power monitoring means indicate that the current power consumption of the processing unit exceeds the current power consumption limit.

According to different embodiments, the processing units may be configured to throttle their power consumption by asserting a throttling control signal for reducing the clock frequency of the at least one processor, by providing an ACPI control parameter and reducing the workload of a processor in response to the provided ACPI control parameter or by initiating a shutdown operation of a processing unit.

30

In a further embodiment, the management unit is adapted to allocate, in a first step, power to other infrastructure components of the electronic data processing system. The power management unit is further adapted to allocate, in a second step, the remaining available power to the plurality of processing units based on the number of times the previously determined power consumption limit was exceeded by a respective one of the plurality of processing units.

10

Further advantageous embodiments are described in the attached claims and the description of detailed embodiments below.

15 The invention will be further described using a detailed description of presently preferred embodiments with respect to the attached drawings.

Figure 1 shows an electronic data processing system in accordance with an embodiment of the invention.

20

Figure 2 shows a flow chart of a method for power management in accordance with an embodiment of the invention.

25 Figure 3 shows a schematic diagram of a blade server system in accordance with an embodiment of the invention.

Figure 4 shows a detailed flow chart of a method of power management for a blade server system.

30

Figure 5 shows the allocation of power within a blade server system before the maximum available power capacity of the blade server system is reached.

Figure 6 shows the allocation of power within a blade server system once the maximum available power capacity is reached.

5 Figure 7 shows a more detailed flow chart of a method for adaptive power budgeting.

Figure 8 shows flow chart of a method for power allocation in case the available power is unbounded.

10

Figure 9 shows a flow chart of a method for power allocation in case the available power is bounded.

Figure 1 shows an electronic data processing system 100 according to a first embodiment of the invention. The data
15 processing system 100 comprises two power supply units 112 and 114. Furthermore, the data processing system 100 comprises three processing units 122, 124 and 126. The power supply units 112 and 114 and the data processing units 122,
20 124 and 126 are connected by means of a power distribution system 130 comprising a power management unit 132. Power monitors 142, 144 and 146, arranged between the power distribution system 130 and the data processing units 122, 124 and 126, respectively, monitor the actual power consumption of
25 the respective data processing unit 122, 124 or 126.

The components shown in Figure 1 may be arranged on a common printed circuit board (PCB) or may be arranged on different printed circuit boards of individual modules of the data
30 processing system 100 such as a server computer. Some of the components shown in Figure 1 may be integrated with other components. For example, the power monitors 142, 144 and 146 may be build-in current monitors of a processor of the data

processing units 122, 124 and 126 or, alternatively, may be integrated into the power distribution system 130. The components of the data processing system 100 may be connected by one or more point-to-point connections, as indicated in Figure 1, or may be connected by means of appropriated power and communication busses. The power management unit 132 may retrieve component-specific power values stored in the power supply units 112 and 144, as well as the data processing units 122, 124 and 126. Alternatively, the power management unit 132 may be connected to a non-volatile memory system for storage of default power values for some or all components of the data processing system 100. The operation of the data processing system 100 in general and the power management unit 132 in particular will be described next with reference to the flow chart presented in Figure 2.

Figure 2 shows a schematic flow chart of a method of power management as performed in the data processing system 100. In a first step 210, initial power limits PL are determined. For example, the power management unit 132 may query the system configuration in order to determine the amount and type of power supply units and data processing units present in the data processing system 100. The individual components 112, 114, 122, 124 and 126 may either explicitly provide information about the power provided or consumed or may return a type identifier used by the power management unit 132 to identify corresponding power values in a table with stored system configuration information. The initial power limits PL determined in step 210 may be determined based on the minimum or maximum power consumption of the data processing units 122, 124 or 126. Alternatively, the total power provided by the power supply units 112 and 114 may be evenly distributed between the data processing units 122, 124 and 126. The de-

terminated initial power limits PL may be stored either by the power management unit 132 or may be provided to the power monitors 142, 144 and 146 for monitoring.

5 In a further step 220, the power monitors 142, 144 and 146 continuously monitor the power consumption P of the associated data processing units 122, 124 and 126, respectively. For example, the power monitors may compare an electric current provided to each one of the data processing units 122,
10 124 and 126 with a threshold value based on the initial power limit PL.

If a current power consumption P exceeds the respective power limit PL, in a step 230, a respective counter associated with
15 the respective data processing unit 122, 124 or 126 is increased, before the method continues with step 240. Otherwise, the method proceeds directly with method step 240.

In step 240, the power management unit 132 checks whether a
20 current evaluation period has expired. In particular, it may compare a time t elapsed since the last determination of power limits PL with an monitoring interval T between subsequent re-evaluations of power limits PL. If the monitoring interval has not yet elapsed, the method continues, after an
25 optional short delay of a few microsecond, with step 220. Otherwise, the method proceeds to step 250

In step 250 the total available power is determined. For example, the power management unit 142 may query the total
30 power available from the power supply units 112 and 114. In this way, the dynamic addition or removal, function or failure of individual power supply units may be detected even during operation of the data processing system 100. Alterna-

tively, in a simple version of the data processing system 100 without a so-called hot-plug capability, the total available power determined during boot-up of the data processing system may be used as total available power.

5

In a subsequent step 260, new power limits PL are determined based on the current power limits PL and the determined available power. For example, if the power monitor 142 detected that, in the system shown in Figure 1, the data processing unit 122 has exceeded its current power limit PL
10 twenty times within the monitoring interval while the power monitor 144 has only detected a single occurrence of the data processing unit 124 exceeding its previous power limit PL and the power monitor 146 has not detected any occurrence of ex-
15 ceeding a preset power limit PL, the power management unit 132 may re-allocate part of the power previously allocated to the data processing unit 126 to the data processing unit 122. In consequence, in the given example, the new power limit PL for the data processing unit 122 may be increased by a preset
20 amount, for example 20 watts, while the new power limit PL for the data processing unit 126 may be reduced by the equivalent amount of a power. In the described example, the power limit PL of the data processing unit 124 remains unchanged in the next monitoring interval. After the new power limits PL
25 have been determined in step 260, the method continues in a step 220.

In various embodiments of the invention, different means may be applied in order to enforce the set power limits PL. For
30 example, the individual data processing unit 122, 124 and 126 may react to the determined power limits PL by limiting their workload, reducing their operating frequency or the like. Alternatively, the power distribution system 130 may comprise

means for actively limiting the provided power to the preset power limit. For example, the power distribution system 130 may limit the amount of current available to the individual data processing units 122, 124 and 126 or may deactivate any-
5 one of the data processing units 122, 124 and 126 either temporarily or permanently in order to enforce the preset power limits. Alternatively, it may be acceptable for the data processing units 122, 124 and 126 to exceed their allocated power budget for a short time, e.g. until the end of the next
10 evaluation period T.

Figure 3 shows a blade server system 300 according to a second embodiment of the present invention. The blade server system 300 comprises a blade enclosure 310. A midplane 320 is
15 arranged in the enclosure 310. The midplane 310 comprises a system management bus 322 and a power distribution network 324. The enclosure 310 further comprises a number of first mounting positions for the accommodation of power supply units and other infrastructure components. In the system
20 shown in Figure 3, three power supply units 332, 334 and 336 are plugged into the enclosure 310. Furthermore, the enclosure 310 comprises a number of second mounting positions arranged on the opposite side of the midplane 320 for the connection of different blade units. In the system shown in Fig-
25 ure 3, four blade servers 342, 344, 346 and 348 are arranged in the blade server system 100. Furthermore, a management blade 350 is arranged in a further second mounting position and plugged into the midplane 320.

30 In the described embodiment, each blade server 342, 344, 346 and 348 comprises a processor 362, 364, 366 and 368, respectively. In addition, each blade server 342 to 348 comprises a remote management controller 372, 374, 376 and 378. Further-

more, each blade server 342, 344, 346 and 348 comprises a power controller 382, 384, 386 and 388, respectively. The blade server 342, 344, 346 and 348 may comprise further components, such as additional processors, memory modules, memory devices such as hard disks, and the like which are not shown in Figure 3 for reasons of representational simplicity.

The management blade 350 comprises a microcontroller 352 and a non-volatile memory 354. In the non-volatile memory 354, a system management table comprising configuration data of all components of the blade server system 300 is permanently stored for use by the microcontroller 352. The microcontroller 352 can communicate with the remote management controllers 372, 374, 376 and 378 of the blade server 342, 344, 346 and 348 and may also query the capabilities of the power supply units 332, 334 and 336.

Next, the operation of the blade server system 300 is explained with respect to the flow chart of Figure 4. In Figure 4, the operations performed by the management blade 350 are shown on the left-hand side. Furthermore, the operations performed by an individual remote management controllers, for example the remote management controllers 372 of the blade servers 342, are shown on the right-hand side.

In an initial phase 410, the enclosure 310 of the blade server system 300 is powered on in a first step 412. In a subsequent step 414, an initial chassis power budget calculation will be performed for the enclosure 310. For example, in step 414, the number, type and/or output power of the available power supply units 332, 334 and 336 may be determined. In addition, a power redundancy mode configuration may be checked by the management blade 350. For example, the blade

server system 300 may be configured in a so-called 2+1-configuration, wherein two power supply units 332 and 334 are used to provide the power required by the blade server system 300 while the power supply units 336 is kept as hot replacement unit in case either one of the power supply units 332 or 334 fails during operation. In this case, assuming that each one of the power supply units 332, 334 and 336 has a nominal output power capacity of 500 W, a total available power of 1000 W is determined in step 414.

Optionally, in step 414 the power consumed by unmanaged infrastructure units of the blade server system 300 is subtracted from the total available power. For example, I/O components such as network switches or storage area network (SAN) adapters plugged into the midplane 320, which have a more or less constant power consumption, are identified. Based on the number and type of identified components, the management blade 350 may determine an amount of power consumed by the unmanaged components of the blade server 300.

After the infrastructure units of the blade server system 300 including the power supply units 332 to 336 have been powered up, the management blade 350 will query some or all blade server 342, 344, 346 and 348 arranged in the second mounting positions. For each identified blade server, the respective remote management controllers 372 to 378 will report its initial power consumption value based on a factory-set value. The values provided in step 416 may correspond to a minimum power requirement of the blade server 342 required to boot up. For example, the reported initial power consumption value may correspond to the operation of a single processor of a multi processor blade server required to start up the blade server 342.

In step 418, the management blade 350 compares the received initial power consumption values with the total available power budget. If the total available power budget is sufficient in order to activate the blade server 342, it will send a start command to the blade server 342. Otherwise, the management blade 350 does not issue a start command to the blade server 342. Furthermore, the management blade 350 may continue the process for the next blade server 344. Meanwhile, the blade server 342 may be held in a standby state.

In a step 420, after receiving a start command, the remote management controller of the blade server 342 initiates the power on sequence by activating a single processor of the blade server 342 in the given minimum configuration. Once the power on sequence is completed, in a step 422, the remote management controller 372 may calculate a real power budget value based on the actual configuration of the respective blade server 342. For example, the remote management controller 372 may detect that a second processor is present in the respective blade server 342 and update its real power budget value correspondingly. The calculated real power budget value is provided back to the management blade 350 which recalculates its current power budget using the given real power budget value in a step 424.

The steps 416 to 424 are repeated for each blade server detected in the blade server system 300. If the power budget does not suffice to power on all detected blade servers 342, 344, 346 and 348, some of the blade servers 342, 344, 346 and 348 may remain in a standby state and may only be activated once sufficient power becomes available for them.

In a subsequent operational phase 430, the power allocated to individual blade servers 342, 344, 346 and 348 may be adapted dynamically.

5 In a first step 432, the management blade 350 reads a configuration setting determining whether adaptive power budgeting is enabled or not. If no adaptive power budgeting is enabled, the initially calculated power budget remains fixed in the operating phase 330. Otherwise, in a step 434, a maximum
10 blade power limit PL for each blade server 342, 344, 346 and 348 is determined based on the initially computed power budget.

In a step 436, the value determined by the management blade
15 350 in step 434 for the blade server 342 is stored by the remote management controller 372 or power controller 382. From then on, in a step 438, the power controller 382 monitors the power consumption of the blade server 342 independently from the management blade 350 in a continuous loop.

20

In a step 440 the current power consumption P of the blade server 342 is compared with the provided power limit PL. If the current power P exceeds the stored power limit PL, in a step 442, a control signal for throttling the processor 362
25 or one or more further processors of the blade server 342 is provided. For example, the so-called PROCHOT# pin of an Intel processor can be asserted which will force the respective processor to reduce its current power consumption with almost immediate effect. Alternatively, other means of reducing the
30 power consumption of the blade server 342 can be employed. For example, the processor 362 may be switched into a higher P-State using the ACPI-Interface. Furthermore, in step 444 a counter representing the number of occasions on which throt-

thling had to be activated is increased. Otherwise, if the current power consumption P lies below the preset power limit PL , monitoring continues in step 438.

5 Using a hardware implemented power controller 382 to implement the power monitoring and throttling operations has the advantage of fast reaction times. In particular, the power consumption of the blade server 342 may be controlled within a few microseconds, e.g. 50 μ s. In this way, an overload
10 situation can be avoided for the blade server system 300.

Meanwhile, the microcontroller 452 of the management blade 350 runs an adaptive power monitoring loop for the entire blade server system 300.

15

In a first step 450, the adaptive monitoring starts by initializing the monitoring loop. For example, the length of a monitoring interval T , e.g. 20ms, may be queried from the non-volatile memory 354.

20

Then, in a next step 452, the counter values provided for each blade server 342, 344, 346 and 348 are read. The counter values may be stored in a volatile memory of the remote management controllers 372 to 378, within the power controller
25 382 to 388 or may be stored in a memory or register of the management blade 350. Accordingly, the values may either be read locally or retrieved over the system management bus 322.

In a subsequent step 454, the read counter values are evaluated. For example, the management blade 350 may determine
30 which one of the blade servers 342, 344, 346 and 348 exceeded their preset power limit PL most often.

Based on the evaluation, in a subsequent step 436, the management blade 350 decides whether re-budgeting of the available power is necessary. For example, if a number of blade servers 342, 344, 346 and 348 exceeded their respective power limit PL within the given monitoring interval T by a certain threshold value, new power limits PL for some or all of the blade servers 342, 344, 346 and 348 may be determined in a step 434 and provided to the blade servers 342, 344, 346 and 348. If no re-budgeting of the available power is required, for example because none of the blade servers 342, 344, 346 and 348 exceeded its preset power limit PL, in a step 458, the management blade 350 waits for the next re-evaluation period and then continues the method in step 450.

Figure 5 shows the development of the power distribution of a blade server system over time. In a first re-evaluation period 510, each one of eight blade servers BL#1 to BL#8 start with an initial, minimum power setting required to boot the respective blade server. Furthermore, a fixed amount of power is allocated to the remaining infrastructure units. In the example shown in Figure 5, there remains an additional, unused power headroom 520 between the maximum power available for the given enclosure and the combined infrastructure units' power consumption and the initial power consumption of the blade servers BL# 1 to BL#8.

During operation of the blade server system, the power consumption of the individual blade servers is re-evaluated at regular intervals. In one re-evaluation period, the blade servers BL#2, BL#6 and BL#8 have exceeded their preset power limit. In response, the management blade has increased their power limit for the re-evaluation period 530 by allocating them more power from the available power headroom 520. In a

further re-evaluation period 550, also the blade servers BL#1, BL#4 and BL#6 have their power allocation increased due to increased power requirements.

5 Figure 6 shows the power distribution of the same blade server system after the total power consumption of the blade servers BL#1 to BL#8 and the general infrastructure units has reached the maximum power available for the enclosure of the blade server system.

10

In an evaluation period 610 the blade servers BL#1, BL#3, BL#6 and BL#8 have a relatively large allocated power budget. In contrast, the blade servers BL#2 and BL#5 have a relatively small power budget. The remaining blade servers BL#4, 15 and BL#7 have an medium-sized power budget. The power budget 620, which exceeds the basic power consumption of the infrastructure units may be redistributed in each subsequent evaluation period 630 and 650. For example, between the re-evaluation period 610 and the subsequent re-evaluation period 20 630, the power budget of the blade server BL#1 was reduced while the power budget of the blade server BL#2 was increased. Similarly, between the re-evaluation periods 630 and 650, the power budget of the blade server BL#1 was further reduced, while the power budget of the blade server BL#3 was 25 increased.

The subsequent flow charts of Figures 7 to 9 show the allocation of power to the individual blade servers in more detail.

30

In a first step 710, a current power budget of a blade server system is initially set to the power consumption of unmanaged infrastructure components including storage blades, connec-

tion blades, fan modules and the like. Furthermore, some initial counter values used for the evaluation of the power budget are set to zero.

In a step 720 a loop starts by selecting a first detected
5 blade server from all detected blade server in order to determine the initial power budget for the first blade server.

In step 730, the current power budget is increased by the current power budget stored for the first blade server. For
10 example, the value used in step 730 may be provided by a remote management controller of the first blade server and may correspond to the initial power required for booting the first blade server or a previously set power limit PL.

15 In a step 740, the method checks whether adaptive power budgeting is available for the current blade server. This may be determined by the presence of an appropriate power controller, a configuration setting stored by a management blade, or a combination of both. If no adaptive power management is
20 available, the initially provided power budget value is used for that particular blade server and the loop continues with the next detected blade server in step 720.

Otherwise, the method continuous for the first blade server
25 in step 750. In step 750, an over count value indicating the number of times that the current power budget was exceeded by the first blade server in the last monitoring period is determined. Furthermore, a counter is increased, which indicates the number of blade servers included in the adaptive
30 power budgeting.

In a further step 760, the management blade checks whether a active priority table is set for the blade server system. A

present priority table may be used to manually give a higher priority to blade servers running particularly important applications. If an active table is set in a further step 770, the table entry containing the over count values retrieved in step 750 is updated in the active priority table. Alternatively, in a step 780, a corresponding table containing the number of over counts for the particular blade server as entry is created. This table is ordered by the number of over counts itself.

10

In a step 790, either the next available server blade is processed starting from step 720, or the loop ends if no further blade servers are to be considered.

15 Then, in a step 795, a decision is made whether the currently computed power budget exceeds the maximum power budget available for the enclosure of the blade server system or not. The maximum power budget available may either be the combined output of the active power supply units or correspond to a manually set power limit for the given blade enclosure. If the currently computed power budget does not exceed the total available maximum power, an unbounded reallocation of a power is performed as shown in Figure 5 and further detailed in the flow chart of Figure 8. Otherwise, if insufficient power is available to operate all blade servers according to their current needs, the bounded case as shown in Figure 6 and explained in more detail with respect to Figure 9 is performed.

30 In the unbounded case, in a first step 810, at first the available power is computed as the difference between the maximum available power of the enclosure and the current power budget. Then, in a loop, power is allocated to each blade server. For this purpose, a blade server having the

highest priority according to the tables prepared in either step 770 or step 780 is selected in step 820.

In a step 830, an additional power budget for the presently selected blade server is computed as ten percent of its cur-

5 rent power budget. In a step 840, the method checks whether the additional power budget lies below the available power computed in step 810. If this is the case, the power budget of the respective blade server is increased by the additional power budget. Furthermore, the current power budget of the
10 entire enclosure is also increased by the additional power budget. The new power limit is then communicated to the remote management controller of the respective blade server and the available power is reduced by the additional power budget. Otherwise, if the additional power budget exceeds the
15 available power, in a step 860, the remaining available power is granted to the respective blade server in a similar way and the remaining variables are updated accordingly.

In a step 870, the loop continues by selecting the blade
20 server having the next lower priority until all blade servers have been reallocated a new power budget or no further power is available for allocation. Then, the method ends the current re-evaluation period in step 880.

25 Figure 9 shows the corresponding flow chart for the case the current power budget has already reached the maximum power available for the enclosure of the blade server system.

In a step 910, the number of blade servers having an over
30 count value of zero is determined. Furthermore, the available power for redistribution is set to an initial value of zero.

In a further step 920, the management blade 350 checks if the sum of all blade servers having an over count of zero is larger than zero. If this is not the case, i.e. if all blade servers have at least once exceeded their preset power limit, there is no power to be redistributed and the method ends immediately in step 980.

However, if at least one blade server has not exceeded its previously set power limit, in a first phase 930, the amount of power available for redistribution is computed. In a step 932 a first blade server is selected in order to retrieve its associated over count value. In a step 934, it is checked whether the over count value of the respective blade server equals zero. If that is the case, in a step 936, ten percent of the previously allocated budget of that specific blade server is removed from the power budget of the selected server blade and is added to the budget of available power. If the counter of the associated blade server is larger than zero, the respective blade server is already using all of its associated power budget and the loop continues in step 938 with the selection of the next blade server.

In a second phase 940, the power budget available for reallocation is distributed among the managed blade servers. For this purpose, in a step 942, a second loop starts in the order of the priority of the respective blade servers stored in the previously determined priority tables.

In a step 944, the management blade 350 checks whether the current blade server has a over count value that exceeds zero. If not, the current blade server does not require any further power budget and the next blade server can be selected in step 942. However, if the counter value of the cur-

rent blade server exceeds zero, in a step 946, an additional power budget of ten percent of the current power budget of the respective blade server is computed.

In step 948, the management blade 350 checks whether the additional power budget exceeds the available power budget. If

the additional power budget does not exceed the available power budget, in a step 950, a new power limit for the respective blade server is computed based on the previously set power limit and the additional power budget. As described

with respect to step 850, the remaining variables relating to the available power and the current power limit of the respective blade server are updated. If the additional power budget exceeds the available power to be reallocated, the remaining power to be reallocated is allocated to the current

blade server in a step 960. The allocation in step 960 corresponds to the allocation described above with respect to step 860.

In a step 970, the reallocation of any remaining power is repeated for the blade server having the next lower priority until all available power is distributed or no further blade servers need to be considered. Then, the method ends in a step 980.

The method and systems described above with reference to Figures 1 to 9 describe various embodiments of particularly advantageous and currently preferred embodiments of the present invention. However, as it will be obvious to a person skilled in the art on data processing system, individual elements of different embodiments may be combined in order to carry out the invention.

Furthermore, particular details given above with respect to the algorithms for determining and redistributing the available power may be changed in accordance with a policy setting provided by the manufacturer or operator of the blade server system. For example, rather than a proportional increase and decrease of power budgets, a fixed amount of additional power may be added or removed from any blade server.

In addition, instead of the described hardware throttling using a reduction in a operating frequency of a processor, other means of reducing the current power consumption of a component of a blade server system and may be employed. For example, a so-called soft throttling may be employed by updating a P-state of a processor using the so-called advanced configuration and power interface (ACPI) of the blade servers in order to reduce their current power consumption. Furthermore, a fully software controlled throttling may be performed by means of an interface to an operating system running on an individual blade server.

Claims

1. A blade server system (300), comprising
- an enclosure (310) having a plurality of first mounting
5 positions to accommodate a plurality of power supply
units (332, 334, 336) and a plurality of second mounting
positions to accommodate a plurality of blade modules;
 - at least one power supply unit (332, 334, 336) accommo-
10 dated in one of the plurality of first mounting posi-
tions;
 - at least one management blade (350), the management
blade (350) comprising at least one controller (352);
 - at least one blade server (342, 344, 346, 348) accommo-
15 dated in one of the plurality of second mounting posi-
tions, the at least one blade server (342, 344, 346,
348) comprising at least one processor (362, 364, 366,
368), a memory accessible to the at least one processor
(362, 364, 366, 368) and a remote management controller
(372, 374, 376, 378);
- 20 wherein
- the controller (352) of the management blade (350) is
configured to query the at least one power supply unit
(332, 334, 336) to determine a total available power for
25 the blade server system (300), to determine a power
limit for the at least one blade server (342, 344, 346,
348), and to provide the determined power limit to the
remote management controller (372, 374, 376, 378) of the
at least one blade server (342, 344, 346, 348);
 - the remote management controller (372, 374, 376, 378) of
30 the at least one blade server (342, 344, 346, 348) is
configured to store the power limit provided by the con-
troller (352) of the management blade (350) and to in-
crease a counter associated with the blade server if an

associated power monitoring means indicates that a current power consumption of the at least one blade server exceeds the stored power consumption limit; and

- the controller (352) of the management blade (350) is further configured to determine a new power consumption limit for the at least one blade server (342, 344, 346, 348) based on the total available power and the value of the counter associated with the at least one blade server (342, 344, 346, 348) after a predetermined monitoring interval.

2. The blade server system (300) of claim 1, wherein

- the at least one blade server (342, 344, 346, 348) comprises a power controller (382, 384, 386, 388), the associated power monitoring means is comprised in the power controller (382, 384, 386, 388) and the power controller (382, 384, 386, 388) is adapted to provide a throttling signal to the processor (362, 364, 366, 368).

3. A method for power management in a data processing system (100) comprising a plurality of processing units (122, 124, 126) and at least one power supply unit (112, 114), the method comprising the steps:

- determining a total available power provided by the at least one power supply unit (112, 114);
- determining and storing an initial power limit for each one of the plurality of processing units (122, 124, 126);
- repeatedly comparing the current power consumption of each one of the plurality processing units with the respective stored power limits;
- increasing a counter associated with one of the plurality of processing units (122, 124, 126) if the current

power consumption of said one of the plurality of processing units (122, 124, 126) exceeds the respective stored power limit; and

- determining and storing a new power limit for at least one of the plurality of processing units (122, 124, 126) after a predetermined period of time based on the total available power and the associated counter value of said one of the plurality of processing units (122, 124, 126).

10

4. The method of claim 3, further comprising

- throttling one of the plurality of processing units (122, 124, 126) if a current power consumption of said one of the plurality of processing units (122, 124, 126) exceeds the respective stored power limit.

15

5. The method of claim 3 or 4, wherein, in the step of determining and storing a new power limit, the new power limit for said one of the plurality of processing units is increased with respect to the previously stored power limit if the associated counter value exceeds the predetermined first threshold.

20

6. The method of claim 3 or 4, wherein, in the step of determining and storing a new power limit, new power limits for at least a first and a second one of the plurality of processing units (122, 124, 126) are determined and stored; and the new power limit for the first one of the processing units (122) is increased with respect to the associated previously stored power limit and the new power limit for the second one of the processing units (126) is decreased with respect to the associated previously stored power limit, if the counter value associ-

25

30

ated with the first one of the processing units (122) exceeds the predetermined first threshold and the counter value associated with the second one of the processing units (126) lies below a second threshold.

5

7. The method of claim 3 or 4, wherein, in the step of determining and storing a new power limit comprises a first and a second phase, wherein,

10 - in the first phase (930), the amount of power available for distribution is determined based on the total available power; and,

15 - in a second phase (940), the determined power available is distributed between a plurality of processing units (122, 124, 126) based on the counter values associated with the plurality of processing units (122, 124, 126).

8. The method of claim 7, wherein, in the second phase (940), the following steps are performed repeatedly while the available power is greater than zero:

20 - reducing the power available by an additional power budget; and

25 - determine a new power limit for one processing unit (122, 124, 126) having an associated counter value exceeding a first threshold value based on the sum of the current power limit of that processing unit (122) and the additional power budget.

30 9. The method of claim 7 or 8, wherein, in the second phase (940), new power limits for the plurality of processing units (122, 124, 126) are determined in a sequence determined by a priority level associated with the respective processing unit (122, 124, 126).

10. The method of claim 9, wherein the priority level for each processing unit is stored by a management unit (132) and may be adjusted by using a user interface of the management unit (132).

5

11. The method of claim 9, wherein the priority level for each processing unit (122, 124, 126) is determined by a management unit (132) in order of the associated counter values of the plurality of processing units (122, 124, 126).

10

12. The method of any one of claims 7 to 11, wherein in the first phase (930), the power available is determined based on the difference between the total available power and the sum of all power limits associated with the plurality of processing units (122, 124, 126).

15

13. The method of any one of claims 7 to 12, wherein in the first phase (930), in the determination of the power available, a preset power consumption of other components of the data processing system (100) is subtracted from the total available power.

20

14. The method of any one of claims 7 to 13, wherein,
- in the first phase (930), in the determination of the power available, for each processing unit (122, 124, 126) having an associated counter value below a second threshold value, a released power budget is added to the power available; and
- a new power limit for said processing units (122, 124, 126) having an associated counter value below a second threshold value is based on the current power limit

25

30

of the respective processing unit reduced by the released power budget.

15. An electronic data processing system (100), comprising:
- 5 - a plurality of processing units (122, 124, 126), each processing unit comprising at least one processor;
- at least one power supply unit (112, 114) for providing at least one supply voltage; and
- a management unit (132) comprising at least one controller operatively coupled with the plurality of processing
- 10 units (122, 124, 126) and the at least one power supply unit (112, 114);
- wherein the management unit (132) is adapted to determine a new power consumption limit for at least one of
- 15 the plurality of processing units (122, 124, 126) based on a total available power provided by the at least one power supply unit (112, 114) and the number of times a previously determined power consumption limit was exceeded by said one of the plurality of processing units
- 20 (122, 124, 126) in a predetermined monitoring interval.
16. The electronic data processing system (100) of claim 15, further comprising:
- a plurality of power monitoring means (142, 144, 146),
- 25 each one of the power monitoring means configured to compare a current power consumption of an associated processing unit (122, 124, 126) of the plurality of processing units (122, 124, 126) with a current power consumption limit provided by the management unit (132)
- 30 for the associated processing unit (122, 124, 126).
17. The electronic data processing system (100) of claim 16, wherein the processing units (122, 124, 126) are config-

ured to throttle their power consumption if the associated power monitoring means (142, 144, 146) indicate that the current power consumption of the processing unit (122, 124, 126) exceeds the current power consumption limit.

18. The electronic data processing system (100) of claim 17, wherein the processing units (122, 124, 126) are configured to throttle their power consumption by asserting a throttling control signal, and wherein the at least one processor of the respective processing units (122, 124, 126) is configured to reduce a clock frequency in response to the assertion of the throttling control signal.

19. The electronic data processing system (100) of claim 17, wherein the processing units (122, 124, 126) are configured to throttle their power consumption by provision of an ACPI control parameter, and wherein the at least one processor of the respective processing units (122, 124, 126) is configured to reduce its workload in response to the provided ACPI controlled parameter.

20. The electronic data processing system (100) of claim 17, wherein the processing units (122, 124, 126) are configured to throttle their power consumption by initiation of a shutdown operation of the processing unit (122, 124, 126).

21. The electronic data processing system (100) of any one of claims 15 to 20, wherein the management unit (132) is adapted to allocate, in a first step, power to other infrastructure components of the electronic data process-

ing system and to allocate, in a second step, a remain-
ing available power to the plurality of processing units
(122, 124, 126) based on the number of times the previ-
ously determined power consumption limit was exceeded by
5 a respective one of the plurality of processing units.

FIG 1

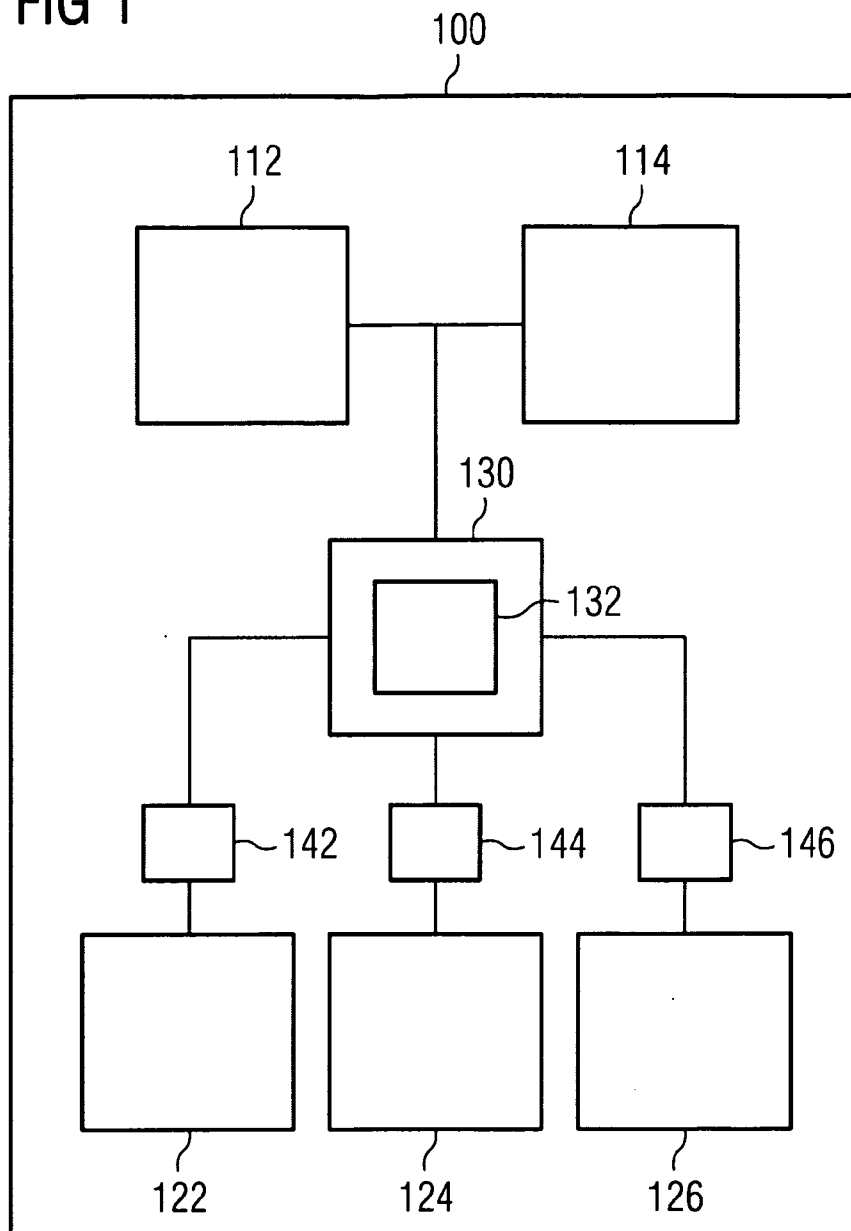


FIG 2

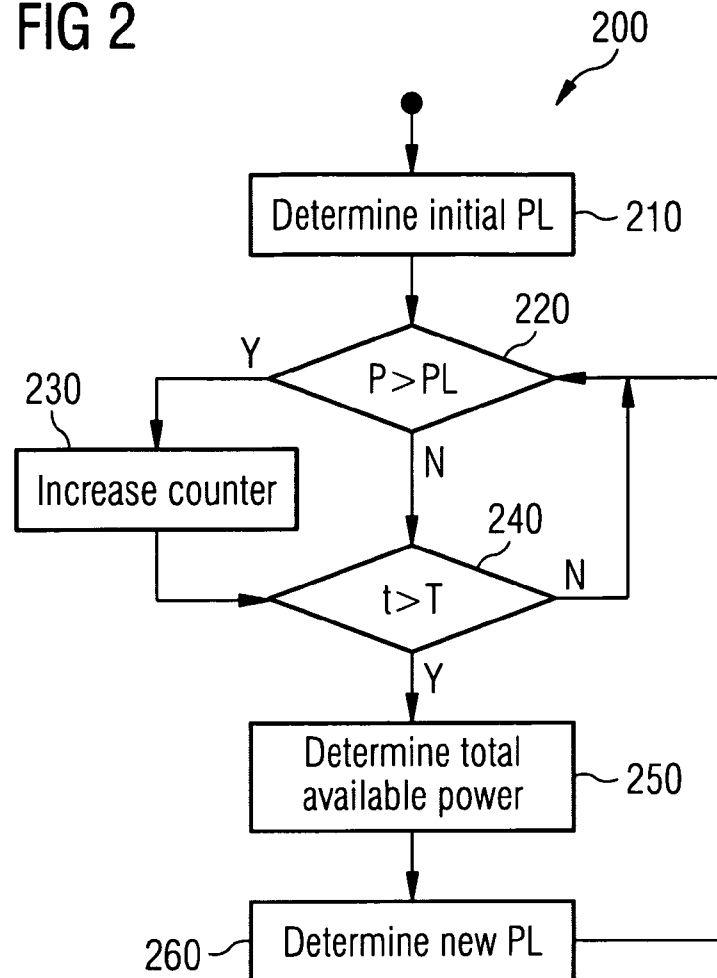
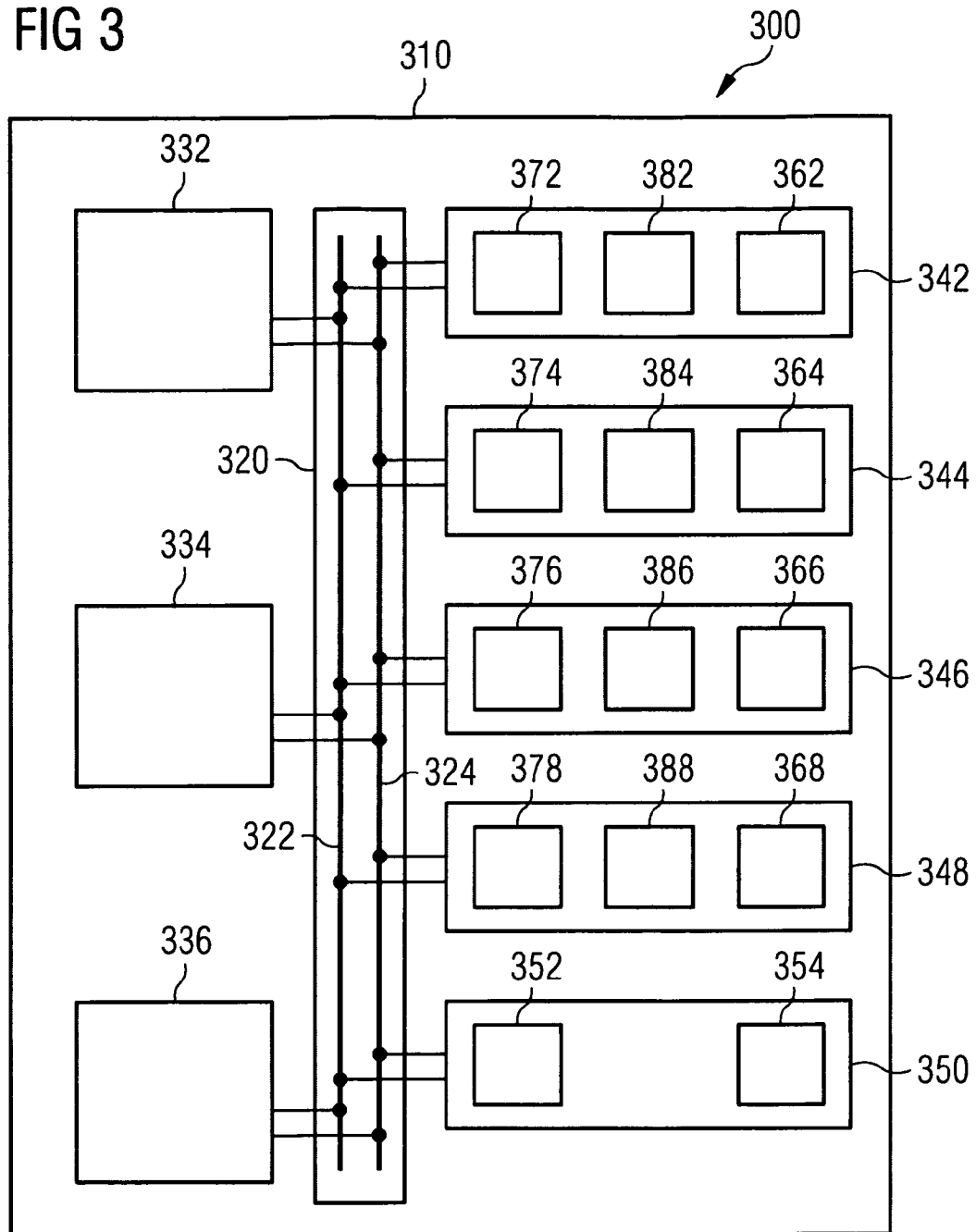


FIG 3



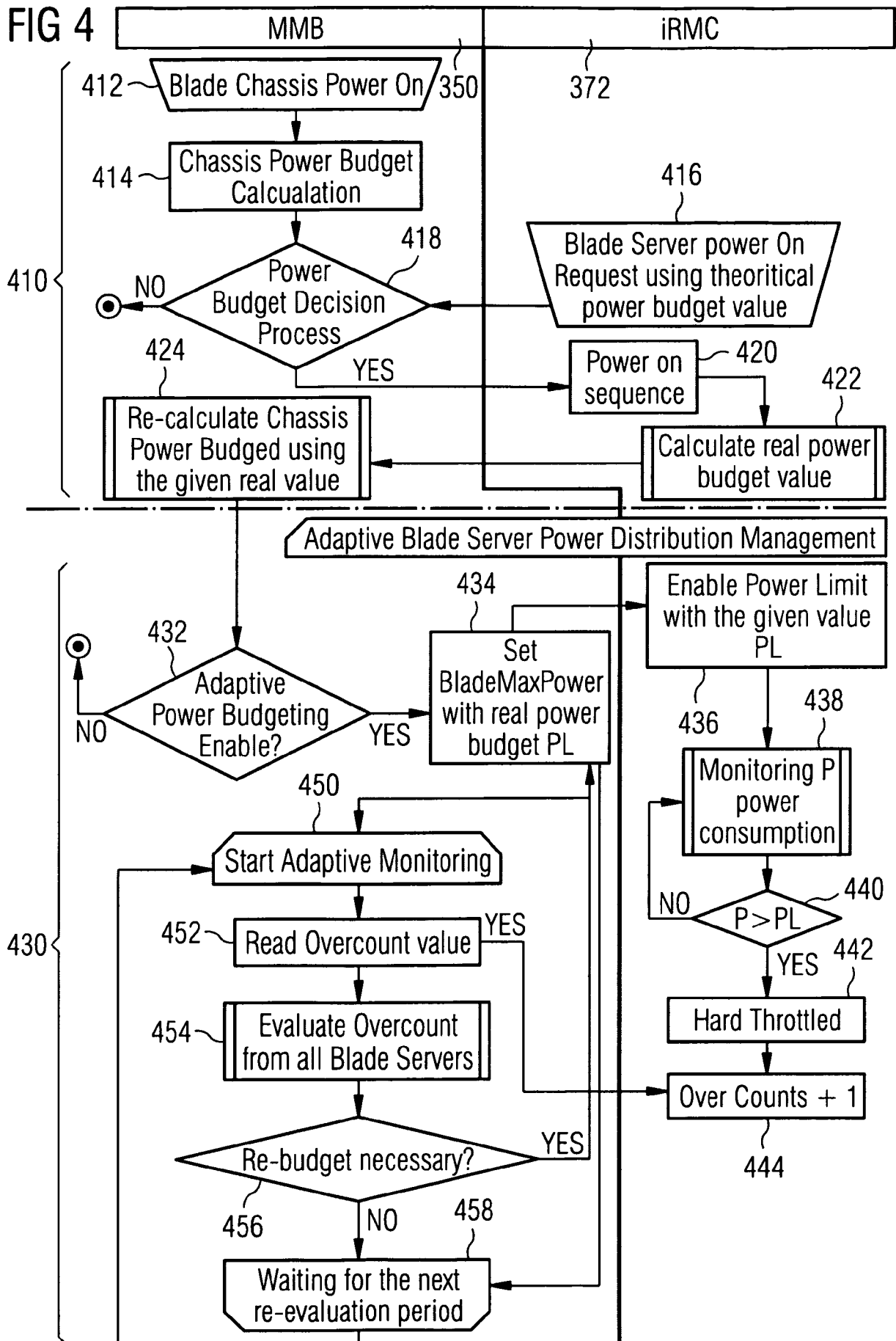


FIG 5

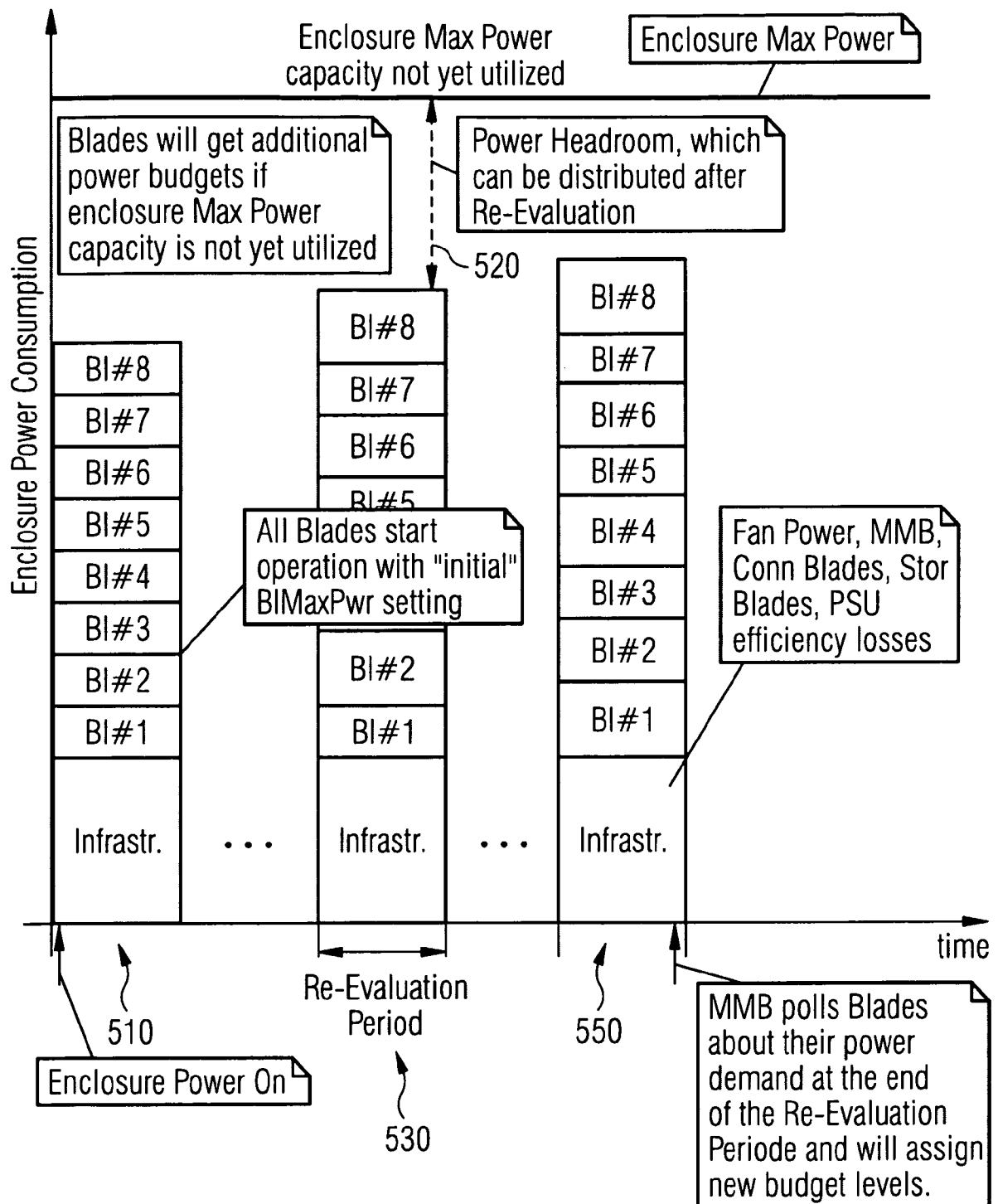
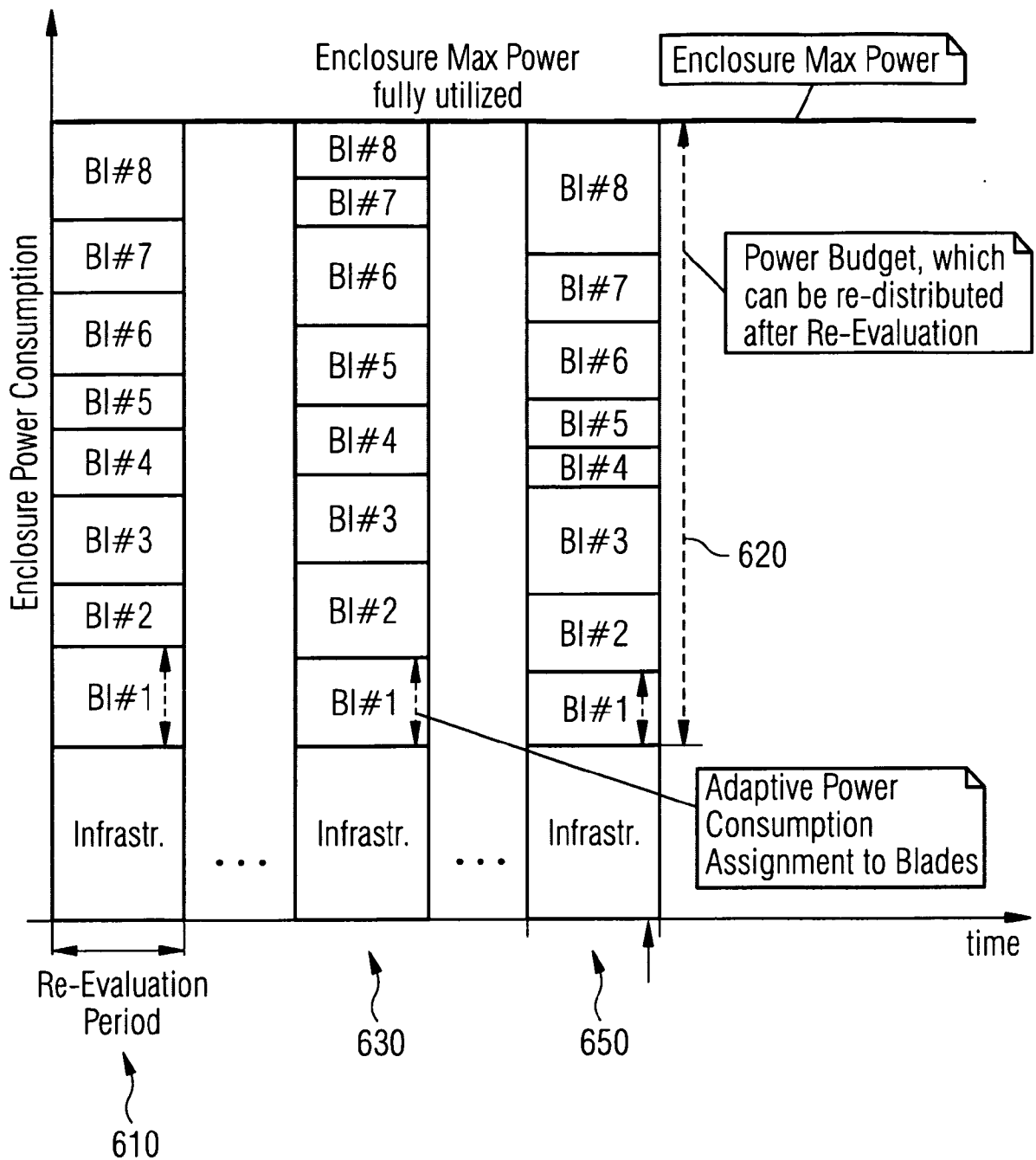


FIG 6



7/10

FIG 7

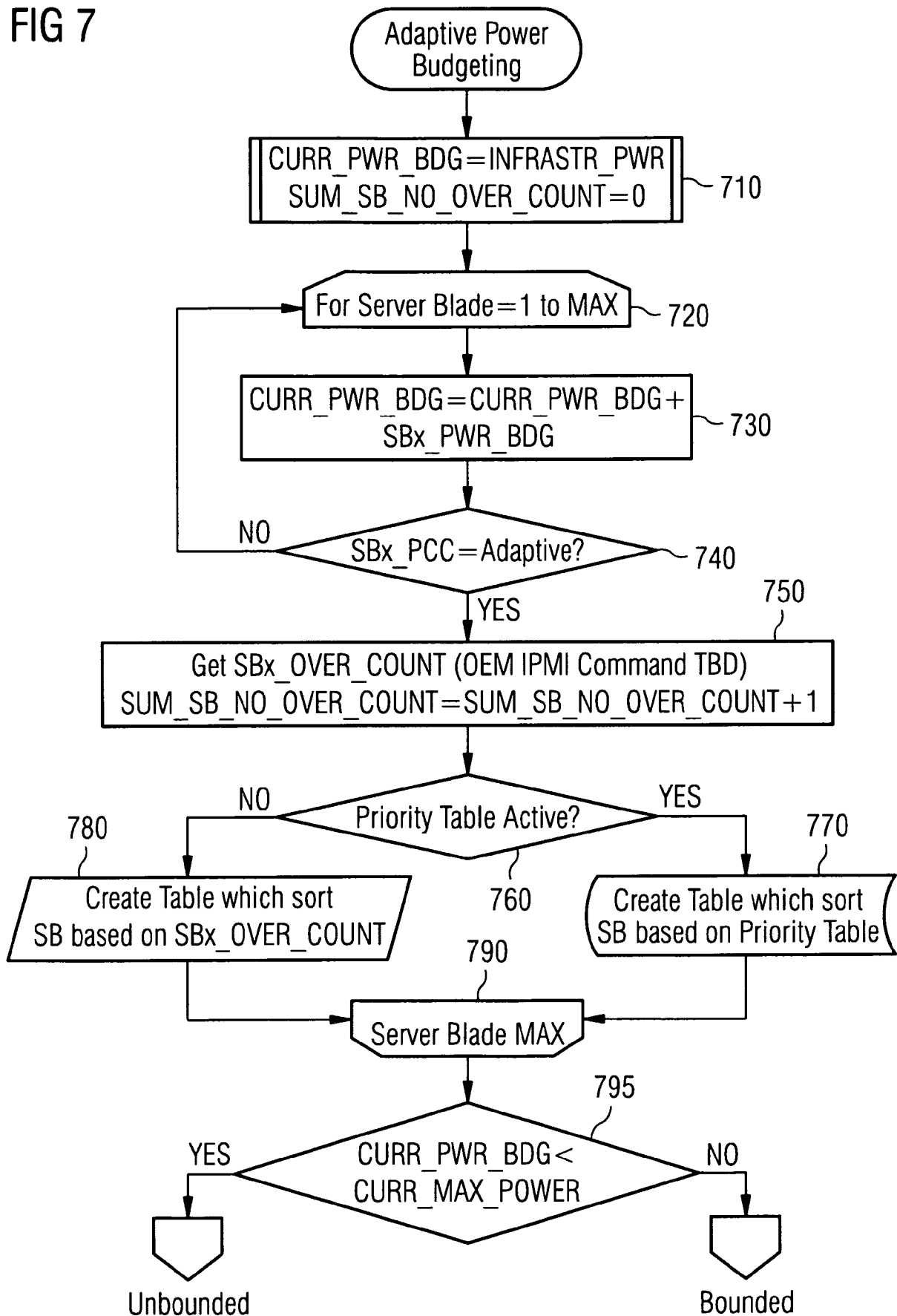


FIG 8

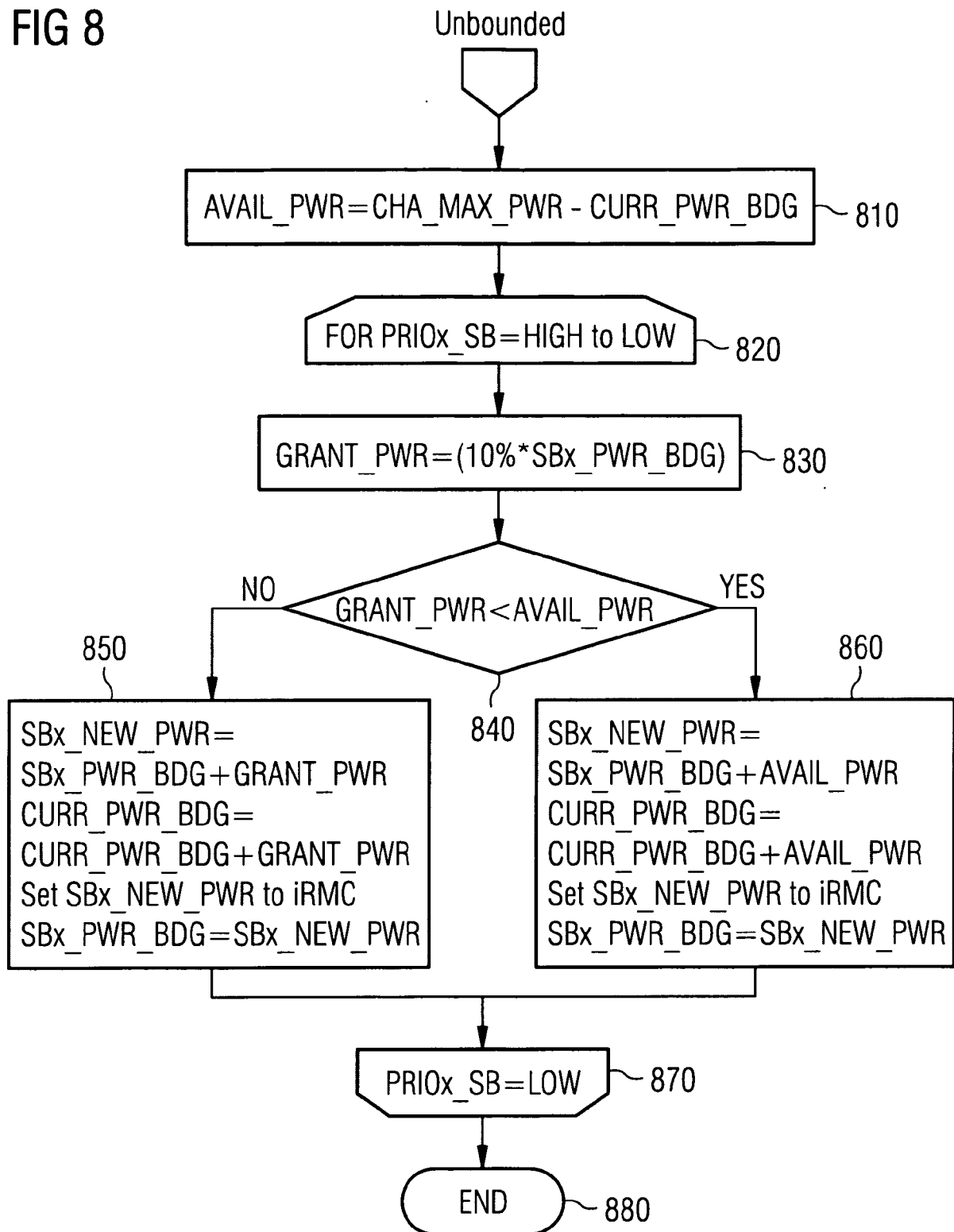
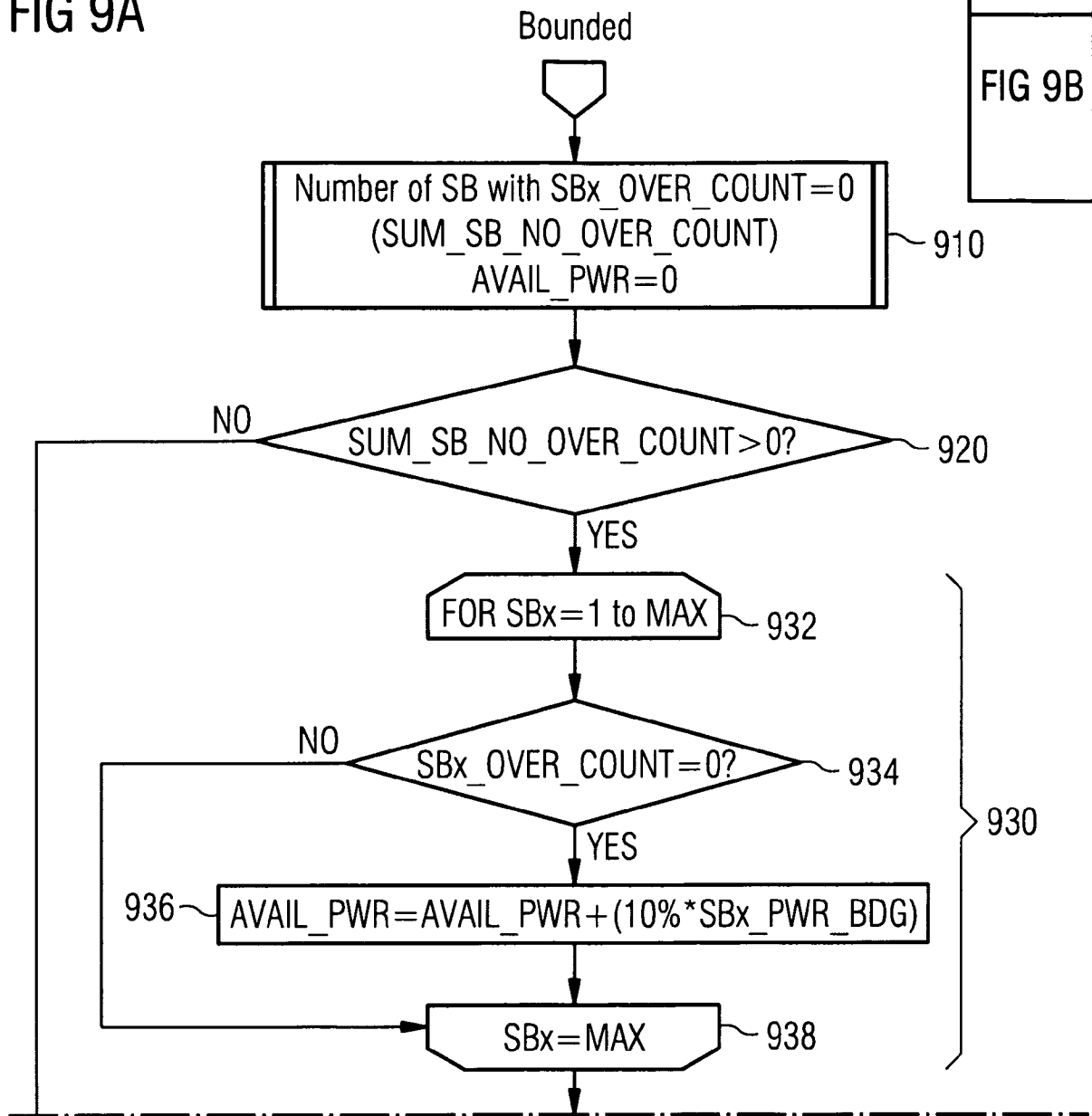


FIG 9

FIG 9A

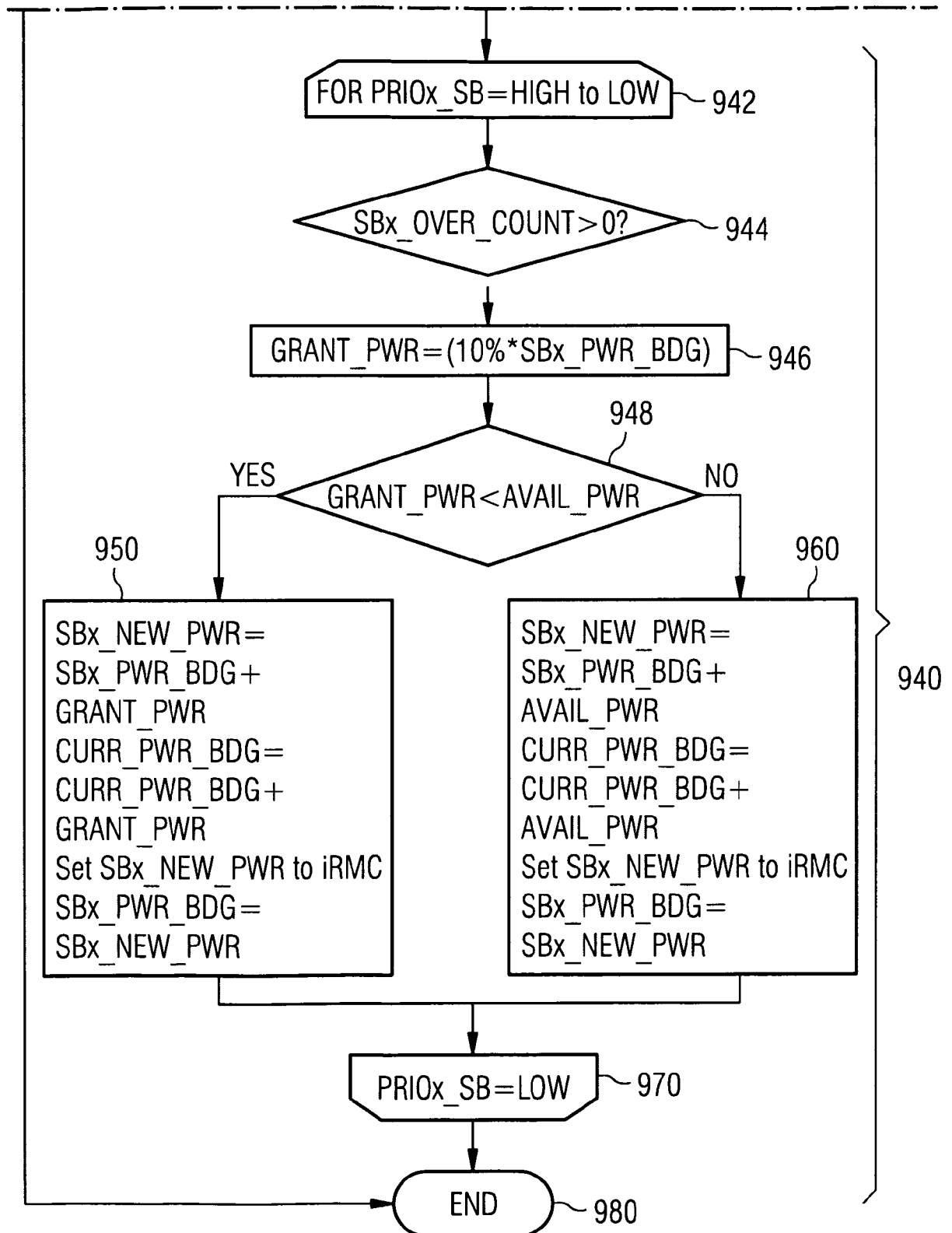
FIG 9B

FIG 9A



10/10

FIG 9B



INTERNATIONAL SEARCH REPORT

International application No
PCT/EP2011/060648

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

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2007/245161 A1 (SHAW MARK E [US] ET AL) 18 October 2007 (2007-10-18) the whole document	1-21
X	US 2009/125737 A1 (BREY THOMAS M [US] ET AL) 14 May 2009 (2009-05-14) paragraph [0018] - paragraph [0041]; figures 1-4	1-21
X	US 2009/077407 A1 (AKIMOTO SHUNSUKE [JP]) 19 March 2009 (2009-03-19) the whole document	1-21



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 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

29 March 2012

Date of mailing of the international search report

04/04/2012

Name and mailing address of the ISA/

European Patent Office, P.B. 5818 Patentlaan 2
NL - 2280 HV Rijswijk
Tel. (+31-70) 340-2040,
Fax: (+31-70) 340-3016

Authorized officer

Alonso Nogueiro, M

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No

PCT/EP2011/060648

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2007245161 A1	18-10-2007	JP 4489789 B2	23-06-2010
		JP 2007287150 A	01-11-2007
		US 2007245161 A1	18-10-2007

US 2009125737 A1	14-05-2009	CN 101430595 A	13-05-2009
		JP 2009116862 A	28-05-2009
		US 2009125737 A1	14-05-2009

US 2009077407 A1	19-03-2009	JP 4395800 B2	13-01-2010
		JP 2009070328 A	02-04-2009
		US 2009077407 A1	19-03-2009
