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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
29 April 2010 (29.04.2010)

PCT

(10) International Publication Number
WO 2010/048316 A1

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

(21) International Application Number:
PCT/US2009/061521

(22) International Filing Date:
21 October 2009 (21.10.2009)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
61/107,172 21 October 2008 (21.10.2008) US

(71) Applicant *(for all designated States except US)*: **RARITAN AMERICAS, INC.** [US/US]; 400 Cottontail Lane, Somerset, NJ 08873 (US).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: **MALIK, Naim** [US/US]; 404 Coventry Lane, Somerset, NJ 08873 (US). **PAETZ, Christian** [DE/DE]; Gert-Froebe-Str. 6a, 08064 Zwickau (DE). **WEINSTOCK, Neil** [US/US]; 11 Cedar Ridge Lane, Randolph, NJ 07869 (US). **YANG, Allen** [US/US]; 699 John Christian Drive, Bridgewater, NJ 08807 (US). **ONYSHKEVYH, Vsevolod** [US/US]; 69 Bayberry Road, Princeton, NJ 08540 (US). **SOMASUNDARAM, Siva** [US/US]; 356 Ridge Road, Apt. E-11, Dayton, NJ 08810 (US).

(74) Agents: **EPSTEIN, William** et al.; Gibbons P.C., One Gateway Center, Newark, NJ 07102-5310 (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, 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, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (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, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

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

(54) **Title**: METHODS OF ACHIEVING COGNIZANT POWER MANAGEMENT

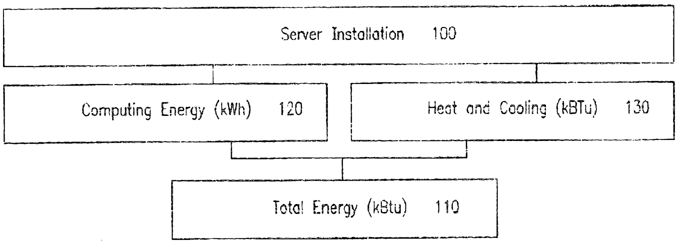

FIG. 1

(57) **Abstract**: A system and method of increasing the efficiency of overall power utilization in data centers by integrating a power management approach based on a comprehensive, dynamic model of the data center created with integrated environmental and computational power monitoring to correlate power usage with different configurations of business services utilization, with the techniques of CPU level power management.

WO 2010/048316 A1

# METHODS OF ACHIEVING COGNIZANT POWER MANAGEMENT

## FIELD OF THE INVENTION:

[0001] The invention relates to power management. In particular, a system and method is presented for instantiating optimized power management of computational and electrical appliances provisioned in data centers and server installations.

## BACKGROUND OF THE INVENTION:

[0002] Most business organizations today rely on computing power for their business services, including data analysis, supply chain management, inventory tracking, online transactions and customer support. This computing power comes in the form of Web services, Web portals and other open-source and proprietary applications hosted in either leased or owned data centers. These data centers have become a significant use of electrical power go through the data center copy paste appliances and indirectly through the humidity and thermal conditioners. Recent data shows that almost 50% of power delivered to a server farm is spent on cooling infrastructure, while less than 50% is actually utilizing server consumption. The amount select for power use during the competition a activity inside the server translates into a thermal load. The amount of electric power stands to maintain the operational temperature is also dependent on the server air flow characteristics and the relative location of the server hardware within the rack and many other parameters is described later in this disclosure. Even though there is a direct relationship between the computational power utilized by the data centers and supplied electric power, the factors affecting that relationship are many, and the instrumentation and analysis needed to quantify them to the required precision for effective control is challenging. Existing power control mechanisms do not attempt to correlate such utilization with given electrical supply units and hence fall short of global optimization of the power utilization and data centers and server installations. The above cross-referenced application describes a systematic procedure apparatus to achieve such monitoring and control using collaborative server copy digital power measurement and electrical power units consumed under different environmental operational conditions. The above cross-referenced application describes a method to provide the necessary adaptive learning required to address diverse data center server farms at infrastructural installations. The heuristics used in the previously described approach take into account the server hardware from what requirements and their locations inside the server rack remote locations within the data centers.

**[0003]** While a number of methods are described in the above cross-referenced application, a class of techniques used in low-level power management but which have not been applied in monitoring and controlling the direct relationship between the computational power utilized by the data centers and supplied electric power are described in the present application. Specifically, techniques for controlling the amount of energy consumption at the motherboard and CPU level may also be used in the monitoring and control of power utilized by servers in data centers

**[0004]** Heretofore, such capabilities have been applied in a piecemeal manner primarily to power management in laptop and other mobile systems. There remains a need in the art to apply this technique to systematic way such as described in the above cross-referenced application, to the balance of the computational power utilized by the data centers versus supplied electric power.

## BRIEF SUMMARY OF THE INVENTION:

**[0005]** A system and method of increasing the efficiency of overall power utilization in data centers by integrating a power management approach based on a comprehensive, dynamic model of the data center created with integrated environmental and computational power monitoring to correlate power usage with different configurations of business services utilization, with the techniques of CPU level power management

## BRIEF DESCRIPTION OF THE DRAWINGS:

For a better understanding of the present invention, reference is made to the following description and accompanying drawings, while the scope of the invention is set forth in the appended claims:

**[0006]** Fig. 1 illustrates major categories of energy consumption;

[0007] Fig. 2 is an exemplary architecture for power management in a server farm or data center environment.

[0008] Fig. 3 is an exemplary framework related to the architecture shown in Fig.

[0009] Fig. 4 is an exemplary architecture for power management in a virtual machine environment.

[0010] Fig. 5 is an exemplary top level flowchart of the method.

## DETAILED DESCRIPTION OF THE INVENTION

[0011] Figure 1 depicts the major sections of energy consumption in server installations 100. Total energy 110 is a summation of the power used for computational purposes 120 and the power used for heating and cooling functions 130. Environmentally based power management uses these major consumption areas to minimize the total energy usage in an adaptive manner. The iterative analysis begins by computing the initial measure of total electrical power consumed by the server and network hardware configuration from the asset information. The assets denote the electrical and computing units in the infrastructure that covers server, routers, switches, cooling units, racks, sensors and power distribution panels. This base computation is then supplemented by the electrical power delivered to the cooling and other infrastructure elements. This electrical side power consumption is tracked against the computational power consumed by business services and correlating it with real time electrical metering at the server level. The asset information and other configuration information can either be supplied directly to the power management system or imported from existing IT management systems, which is common in an enterprise level installation. The environmentally conscious power management system and method increases the efficiency of overall power utilization by using coordinated power monitoring at the electrical power distribution side as well as in the consumption side of server installations using simplified common management interfaces (e.g. simple network management protocol (SNMP) web services for environmental and server monitoring. The system and method improves the efficiency of power utilization per computational unit and provides policies for dynamic load allocation capabilities based on application loads and environmental contexts, e.g. rack configurations, cooling units and thermal characteristics of the data center room. The dynamic computational load balancing policies can be adopted by (i) positioning the servers at environmentally optimized locations, (ii) scheduling the computational tasks at physically diverse locations (ii) VMotion (utility for migration of virtual machines between physical servers offered by VMware) in the virtual environments leveraging virtualization technology. The proposed approach addresses modifications needed to measure the electrical power requirements in such virtualization environments.

[0012] A coordinated the framework is the most important ingredient to maximize the benefits of information gathered by the environmental and server measuring agents denoted as monitors in the figures illustrated here with. Practices address local electrical optimization which might not accurately determine the requirements. It environment where multiple business services or consolidated at the server level, they might help to reduce the power consumed of each individual server. However they do not account for the operating conditions and thus become ineffective in situations where those conditions are significant, such as when the servers are configured in a dense arrangement like server blades. In addition to consolidating multiple business services into a single or few servers for minimizing power consumption, it is also important to determine the optimum location of that server hardware based on environmental context. For example, an application running a server with rack mounted cooling unit would likely use less overall power than the same application running on an identical server in a less efficient cooling environment. Similarly, an application running on a server with a low-power hard drive would likely use less power than the same application running on a server with large power consumption or inefficient heat dissipation design. The selection of servers for load balancing is mostly ad hoc under current practices and does not involve a detailed analysis.

[0013] The most important challenge in power management is the lack of components to correlate the computational processing power related to services and electrical power consumption with the accuracy needed for globally optimized load-balancing and control. In addition to the challenges needed to monitor a single location environment, IT services, and today's enterprises are typically located geographically distant locations and better utilization of human, time and material resources. These geographic variables are not taken into account and present electrical power operational profiles due to the lack of a comprehensive solution to address such capabilities. As a business services are rendered for multiple geospatial locations, it is critical to include coordination among what power management and computational processing operations at the enterprise level within the global context of business services. The optimization of overall power utilization does requires a coordinated framework, systematic procedure and control elements to distribute the business computing load both at the physical location and logical locations (as applications and server clusters or virtual machines) by configurable and adaptive monitoring and continuous analysis of the environment for global power management decisions.

[0014] The server monitoring tools and appliances employ either agent or non---agent-based technologies to measure application load, starting with determining the number of instances of each application running on each server, using different mechanisms primarily depending on the operating system's hosted in server hardware. Typical monitoring interfaces available for monitoring and measurement include Windows management instrumentation (WMI) on Microsoft platforms, simple network management protocol (SNMP), and Web service management (WS-MAN).the granularity of the information necessary to associate the application loads of the electrical power users not directly computable from the standard interfaces. Thus it is useful to obtain the right metrics from these existing base metrics that are usable intellectual power computations. The environmentally-based power management framework addresses this constraint aggregate in the process level information from different mechanisms (including network computer monitoring devices such as Raritan's Command Center NOC, baseboard management controller such as Raritan's Kira, and power environmental monitoring devices such as Raritan Dominion PX) to improve the accuracy of the predicted electrical power unit consumption based on environmental factors, server hardware characteristics, operating system overheads and running application to provide business services. As shown below, a first level of power metrics is obtained from server hardware and power supply units installed within the servers and a second level of power metrics is obtained from operating system and applications executed within the servers. A third level of metrics is computed from the life monitoring of actual business utilization, configuration, topology, thermal activity and electrical power usage.

[0015] Another constraint in the power management operations that the server monitoring instrumentation and electrical environment monitoring instrumentation are installed separately and maintained independently, the snaking calibration and coordination of these measurements relatively difficult. The environmentally conscious and power management system and method integrates both server and environmental monitoring and provides Cooperative collection and processing of metrics, thus improving the scope of optimization at the global level. In order to improve the interoperability of measurement devices and distributed data collection, a common information model (CIM) is proposed, and management interfaces that support this CIM profiles are recommended. Intelligent platform management interface (IPMI) is one such standard and Raritan Bay sport management controller (KIRA) and dominion PX are a few examples of monitoring and management devices that support IPMI interface making it suitable for power optimization applications.

[0016] The environmentally cognizant power management framework also applies to both virtual and physical server environments. In particular, virtualization technology provides dynamic load-balancing operations that help simplify collaborative power management. As described later, the proposed and adopted a physical server installations as well as virtual server installations.

[0017] In addition to the instrumentation necessary to collect a server computational power metrics and electoral power measurements had decided faces during programmed intervals; the synchronization of this information needs to be clearly determined for accurate relation of and submission of both utilization supply. The management controller and power optimization (330) system supports the necessary components to order and group the data set collected over multiple interfaces to various time intervals. The information part compute the long-term trending is different from local and short bursts of measurements. In such cases, the management controller configures the necessary data collection devices to collect the measurements at different sampling intervals between maintenance cycles as necessary.

[0018] Referring now to Figure 2, there is shown a system 200 that uses coordinated power management server installations, data centers and other such constructs. System 200 includes servers 205 and 210 that are connected to a management and monitoring devices to 15. Servers 215 and 210 maybe any computing device in any hardware and software configuration that may include for example, server applications, database applications, web applications and the operating system that hosted business applications. Management and monitoring devices to 15 maybe Raritan's Command Center NOC or any other similar device. Servers 205 and 210 are also connected to power management devices 220 and 225, respectively. Power management devices to 220 to 225 may be Raritan's Dominion PX or any other intelligent power management device. Management and monitoring device 215 and power management devices 225 further coupled to the monitoring control and analysis device 230 that is accessible by the user by a client interface 235.

[0019] The environmentally conscious and power management system and method uses the above management components as building blocks for power management framework. In particular, the management and monitoring device 215 determines and real-time the operating system and applications running a server. In addition, the management and monitoring devices 215 determines and real-time the operating system and applications monitoring on each server. In addition, the management and monitoring devices 215 monitor server hardware parameters using baseboard management controller's hardware that may be available as embedded modules or parts, such as example, in a Raritan's Kira device. These parameters, including the power cycle events and CPU temperatures, provide an additional level of power metrics to correlate to the CPU utilization and the computational load created by the business applications. Power management devices to 225 gather information of electric power used by each connected device as was in the environmental data (primarily temperature, air flow and humidity) by wired and wireless sensor devices. This environmental monitoring integration grew as the energy requirements at the cooling unit to the computing load and ambient conditions. The confrontational mode affects electric power draw which creates thermal energy which in turn affects the lexical power consumed by the cooling supply. The ambient conditions within the data centers affect the efficiency of cooling units, thus necessitating the need for including the localization parameters into the dynamic model of the data center. The external ambient conditions also impact the energy needs to operate the clinic indenting units of the desired operating temperatures and should be incorporated of his model as well.

[0020] The interaction of these components and power optimization is detailed in figure 2. In this scenario, the business applications are distributed according to operating profiles computed by the centralized processing engine (main controller) running and control and analysis device 230. That is, business applications are scheduled to be executed at a particular server based on its efficiency and power and thermal profiles and its position within the racket, location within the server room or facility, as well as environmental conditions. The processing logic in device 230 gathers environmental conditions from the environmental sensors which are coupled to a common monitor (or integrated into power management devices 220 and 225) and the measurement electric power usage that are directly available from power management devices 220 and 225. Derive the optimize operating profile for the IT services heuristically by 230, the framework takes into account the following:

First, the current drawn efficiency of the power supply units which convert alternating current (AC) power into direct current (DC) power used inside the server hardware. This information is acquired from the configuration management database (CMDB.),known as asset data in IT terminology. The acid information includes the nameplate data that describes the maximum power drawn by the hardware and the inlet and outlet temperatures necessary for proper operating conditions based on its physical design and functional characteristics. In addition to the component level information, the physical location of the wrecks, the relative distance of servers from the cooling systems, and orientation of racks are also fed into the system either from the CMDB, if available, or from other data sources.

Second, the base level power garments with the operating systems and applications running on servers. This is dependent on the utilization of central processing unit or CPU, random access memory, hard disk, and other resources in addition to the operating system implementation. This information is typically imported from the IT infrastructure systems are automatically collected to the baseboard management controller (BMC). Third, the actual utilization of business computing services within the environmental and chronological context. This process level metrics are obtained by monitoring appliances to 15 through WMI and SNMP since interfaces at here at intervals as computed from the main controller.

[0021] In order to improve the accuracy of correlating electrical power requirements with the information from various appliances, the measurement data are synchronized to the clock server that runs inside device 230. The sequencing operation of the measurement data acquired from any device is validated based on the time of arrival of the relative event occurrences before storage for further processing. In addition to the synchronization capabilities, the Web services capability automates the discovery and communication between the information provider services and the processing engine which runs in device 230. Before the start of monitoring and data collection, all measurement devices must be discovered improperly registered in order to successfully interpret the events with its context. The context of the operating conditions is important to the measurement. For example, the electrical power usage of Sea RAC computer room air conditioner units in a lightly loaded environment is to be treated differently from the measurement of highly loaded operating conditions. The data received throughout the entire system is thus interpreted within the operational context of the global level. The monitor data are now accessible to the receivers, which, in the usual scenario, is the main controller in the system. If there is a failure to receive data from one device multiple devices, the management controller adapts the next level metrics or derived metrics and previous context or earlier data acquisition to the system is back to normal operation. The standard interface of using Web services fit well into scalable reconfigurable enterprise platform and the proposed approach can be easily adopted into its existing infrastructure. The prior art power management for information technology services does not provide the capabilities of coordinated operations over the network and thus cannot benefit from load-balancing across infrastructure and computational equipment in power utilization. The inventive system and method address this constraint by provisioning network aware services in each appliance.

This facilitates real-time synchronization between different units in the migration processing across multiple units. A network enabled Web services technology supports both reactive and proactive power monitoring, which is valuable to complex environments where the behavior of applications and short term failures or irreparable crashes. Web services can provide configuration of dates in enable reconfiguration by subscription methods between peer-to-peer systems, which is in this case are the data measurement devices, data receivers for processing the data.

[0022] Where the power (and hence cooling) is of critical importance, server level power monitoring is highly beneficial for information technology management, specifically for auditing, capacity planning, thermal assessment, and performance improvement. These scenarios essentially require power usage and active of monitoring on servers to determine the need for rescheduling the load based on predefined and/or dynamic power policies. The dynamic power management policies are either stored in the system or referred from an external network management system. Basic level of power profiles are supported at the operating system level and advanced configuration and power interface (ACPI) provided a way to control this aspect. The proposed approach applies to any power profiles including ACPI and it uses appropriate metrics for different kinds of profiles. Current power policies that utilized ACPI focus on reliable power feed to the computing units based on the availability of electoral power from alternate sources like uninterruptible power supply generators.

[0023] Referring now to figure 3, there is shown a framework of an intelligent power monitoring system 300 for server farms or data centers in accordance with the invention. In a first phase 301, system 300 as a power management appliance 305 that acts as a power distribution unit (PDU) to target servers 310 enables users to measure power at the outlet level providing an accurate view of power consumption and the target server 310 or a PDU level (total power distributed). Most of the existing power distribution units measure a rack level or branch circuit level power measurements, which might not be sufficient for power monitoring at the server level. In a second phase of 302, management and monitoring device 315 collects IT service utilization information from target servers the intent using WMI, SNMP or WS-MAN client interfaces and feeds it to a data application layer 325 of a monitor

331. In addition, an IPMI client interface 320 gathers information from power management appliances 305 feet of the data acquisition layer 325. This is the orthogonal metric and provides the assessment of maximum, minimum and average electrical consumption. Monitor 331 is also coupled to a configuration and scripting interface 333 that inputs threshold and alerts to the monitor turned 331. In this description, the monitor refers to the software component that is responsible for monitoring the measurements from the devices obtaining server-side metrics (e.g. operating systems, CPU temperature, application instances) as well as those obtaining the electoral site metrics (e.g. power usage, temperature, air flow and humidity). This is the data interface layer of the system and is configurable from the main controller 330. Main controller 330 can define the polling interval, event receiving mechanisms and device topology and interface for communication.

[0024] Data acquisition layer 325 in turn feeds it to a database 340. Database 340 is coupled in a feedback arranged made arrangement with an analysis and 45, which applies processed information for user interface (visualization and interaction) layer 335 that is coupled to an application server or Web server 350. In a third phase 303, the application server or Web server 350 exchanges the processed information, which includes trend analysis, capacity analysis and efficiency metrics, with a Java client or Web Bowser 355 at the client side.

[0025] The reporting and publishing logic of this framework includes 2 kinds of information dissemination to either the form of alerts or reports for energy audits. As depicted in figure 3, the derived metrics from the framework ranges from capacity data collection 336 to estimates based on rigorous optimization 345. The reporting of the output of the analysis engine can provide data center efficiency metrics including individual server efficiency and relation to the applications that are running in the data center and the environmental context. The output also provides a service that the base model of data center power utilization for tracking the trend and adaptive learning. These outputs cane be used to optimize power consumption taking into account the desired business services, server hardware and software terrorist characteristics and the data center environment parameters, including cooling and ventilating facilities. The strength of this framework is a configurable power monitoring logic and retrieval of IT business service dependencies based on heuristic analysis.

[0026] Referring now to figure 4, there is shown a system 400 uses coordinated power management in a virtual machine implementation. System 400 includes servers 405 and 410 that are coupled to a control and analysis device 430 that is accessible by a user via a client interface 135. Servers 400 510 may be computing devices that implement a virtual machine environment such as VMWare or Xen. In an exemplary embodiment, a provider 407 (412) manages a plurality of virtual machines 409 (414). Servers 405 and 410 rolls connected with power management devices 424 and 25, respectfully. Each server 405 and 410 has a power management endpoint appliance 415 (417) that monitor all the running processes (computer power) and their resource utilization at the application level in terms of memory, disk and network usage. This is achieved by monitoring the hypervisor APIs supplied by the virtualization server for hardware resources. Device 430 collects this information and from devices 405 and 410 and additional information from devices 420 and 425.

[0027] In both scenarios illustrated in figures 2 and 4, energy consumption metric is normalized to the power required to operate the core services like operating systems, network, memory, storage and infrastructure. In an exemplary embodiment, application load may be distributed intelligently among the different physical and logical bins to minimize the political power utilization in terms of computing and cooling.

[0028] The data driven framework facilitates monitoring and provides interface for dynamic control of electrical and computing power management. This method reduces the cost of equivalent computing power needed by providing the basis for an optimal distribution of total computing load needed by the business at any given time in general, figure 5 shows the stages involved in the power optimization method, which analyzes the present operating conditions in the services context to compute the operational computing electric power profiles for effective electric power utilization for any given computing load. Decision rules may be established according to the criticality of business services, which is configurable by the IT administrators and infrastructure providers. For example, the administrator could decide to place an absolute limit on the aggregate rate of power consumption in the data center and identify certain applications whose execution is to be deferred if the power consumption limit has been met. The output of the stage will be delivered to the control logic states allocate the application of it among the service in order to minimize total power utilization.

[0029] Referring specifically to figure 5, a top-level flowchart of the inventive method is shown. Environmental, application server power data is input to a filter 505. Since the volume of data to be collected through device 230 in figure 2 and device 430 in figure 4 is quite large, filter 505 is admitted at the front end of device 230 at 430. Filter 505 provides the capability to prioritize information related to political power consumption computer CPU power metrics that are adequate to correlate with the power data received from, for example, devices 220 and 225 in figure 2. The filter supplies the short-term details as well as the long-term trends from the data collected across many and nodes and aggregation devices. The CPU usage, main memory, disk I/O, and the CPU time slice shared among many virtual machines are examples of metrics that have major impact on electoral power consumption on that server hardware. A function of the filter 5 5 is to cumulate such resource utilization into a representation that facilitates mapping between the company showed an electoral power. In a smaller datacenter with especially low level of complexity, the filter may not be required, and the resource utilization information be used as is.

[0030] Filtered data is then input into behavior modeling module 510. Workload or business service character they characterization is performed in advance for common business services and common server platforms to create basic model for data center power usage. This is stored in knowledge base 515, providing a second input module 510. The base modeling of combined characteristics of server and application combinations also helps to reduce the data needed to process elect will power you are the utilization at any given time server operation. The workload characteristics were normalized to the hardware parameters acquired from the asset database. A third input to module 510 is database hundred and 20 which provides the current dynamic model information in parameters.

[0031] Module 510 sends a database 520 the information about the current datacenter configuration and computational loads which is used updated databases if necessary. The current state information is also pastor power metric computation module 525if module 525 which uses it to complete the power usage and environmental metrics (e.g. ambient temperature) predicted by the dynamic model. The power consumption measure module 530 gets information about the existing distribution of application loads among the various servers from module 510 and also gets predicted power environment metrics from module 525. Module 530 acquires information on the actual article power and thermal and other environmental conditions to the monitoring inputs from, for example, devices 220 and 225 in a time justice and correlated manner. If the predicted data is within defined tolerance limits of the actual data, the dynamic model is validated if not, the information is output to a trend analysis module 535 which provides feedback for updating and defining dynamic modeling algorithm parameters in database 520.

[0032] As a final step the dynamic model is applied to the current application to determine if a relaxation of that load among services required for power usage optimization. Given the models predict power usage based on the application of common environmental factors and server hardware and software characteristics, there are numerous methods for making that determination known in the art. For example, a simple approach would be to rank order the servers based on the predicted incremental power draw for an increment of computational load under existing environmental conditions and rank order the applications based on the anticipated schedule of computational load. Starting with the application with a large load this process would allocate instances of the applications first to the most efficient server and, as a server's maximum competition was reached, proceed to the next most efficient server until done. This approach will yield reasonable results and has the advantage of being simple and fast, it will often result in a sub optimal allocation. A better approach of the preferred embodiment develops fewest rules using a process similar to the back propagation learning algorithms used in home networks. No networks are similar to artificial intelligence methods, trying to model the human learning and intelligence using computer algorithms with a set of neurons and interconnections between themselves based on the learning. In the event that the analysis concludes that they realize allocation of application loads or adjustment to the environmental condition infrastructure is required, appropriate instructions are sent to 1) heating and cooling control 540 and 2) server and application load balance control 545 in order to implement the desired changes.

**[0033]** In general the major components in the framework for mining the environment for energy conservation include the main controller, monitor, database and analysis engine. The metadata collected to monitor engine would become unmanaged the large and may not be fully utilized for energy data collection. The framework provides an intelligent configuration module that adapts the customer environment after initial run-in facilitates capture and analysis of only selected but relevant information pertaining to the electrical computing energy calculations. The framework is adapted to collect the measurements at the necessary intervals with the precision and sampling rate. This module is designed in such a way that the business will services and its impact recently supplemented add to the acquisition and analysis logic. The overall framework either includes logic together as information of the environment or could be integrated to an external repository. This information is used in the heuristic information for adaptively tracking the association between the electrical and computing power entities. The framework gathers the services and the infrastructure either through the dynamic learning or through a static import. This information is then used as the base for analysis of the services in addition to the typical in our mental behavior. This increases efficiency in terms of deployment effort and automated learning thereafter. The synchronization of all individual components in the framework is configurable to an internal time server on external network time service.

**[0034]** The approach described creates global profiles for allegedly computing utilization 545 by systematic processing of power matters collected using multiple mechanisms for configuration and characterization of various power utilization heuristics. The operational profile specified amount of getting loaded each server, the location of server to host that computing load and the electrical units delivered to the looming cooling unit on required schedules. The method models the energy consumption at the process level to maximize the accuracy of power optimization across the entire environment. The Web services-based approach works well for distributed enterprises were the communication framework is extended hierarchically through multiple levels of aggregation control to gather the dynamic load conditions and load-balancing operations at locations remote from the central power management system.

[0035] Is understood that figure 5 is illustrative only and that other program entry and exit points, timeout functions, error checking routines and the like (not shown) would normally be implemented in a typical system software. It is also understood that the systems offer can be implemented to run continuously in an embedded system. Accordingly start blocks and blocks are intended to indicate logical beginning and ending points of a portion of code that can be integrated into the main program and called as needed to support continuous system operation. Implementation of these aspects of the invention is readily apparent and well within the grasp of those skilled in the art based on the disclosure herein.

[0036] Although an exemplary network environment as described above, in the network of interconnected computers, servers, applications and other devices are applicable and can be used with respect to the method described above. Computers commonly operate in a networked environment using logical connections to one or more peers, the computers used in conjunction with the method may be a personal computer 8, a server, a router, a network PC, a peer device or other common network node, and typically includes many or all the elements described above. The connections include but are not limited to local area network (LAN), wide area network (WAN) and other such networking environments are commonplace in offices, enterprise wide computer networks, intranets, and the Internet. It will be appreciated that the network connections shown are exemplary and other means of establishing communications links between the computers may be used. The purpose of illustration, programs and other executable program components such as the operating system are illustrated herein as discrete blocks, although it is recognized that such programs and components reside at various times in different storage components of the computer, and are executed by the data processors of the computer. Different combinations of hardware and software can be used to carry out the teachings of the present invention. The computer or computing device typically includes a processor. The processor typically includes a CPU such as a microprocessor. A CPU general includes an arithmetic logical unit ALU which performs arithmetic and logical operations, and a control unit, which extracts instructions (e.g. code) from memory and decodes and executes them, calling the ALU as necessary. Memory as used herein refers to one or more memory (RAM), read-only memory (ROM), programmable read-only memory (PROM), erasable programmable read-only memory (EPROM), or electrically erasable programmable read-only memory (EEPROM) chips, by way of further non-limiting example only. Memory may be internal or external to an integrated unit including a processor. Memory preferably stores a computer program, *e.g.,* code or sequence of instructions being operable by a processor.

**[0037]** As previously noted In the event that analysis concludes that a reallocation of application loads or an adjustment to the environmental conditioning infrastructure is required, appropriate instructions are sent to 1) heating and cooling control 540 and 2) server and application load balancing control 545 in order to implement the desired changes. APCI enabled Hardware and OS allow another method of reducing energy costs via efficient energy consumption. . Specifically, techniques for controlling the amount of energy consumption at the device, motherboard, chip set and CPU level may also be used in addition to steps 540 and 545 referred to above. Many technology companies are implementing compliance with ACPI As noted above, ACPI is the current open standard for implementing power management at the processor level. ACPI is, at its most basic, a method for describing hardware interfaces in terms abstract enough to allow flexible hardware implementations while allowing shrink-wrap OS code to use such hardware interfaces. Operating System-directed Power Management (OSPM) is a model of power (and system) management in which the OS plays a central role and uses global information to optimize system behavior for the task at hand. ACPI Hardware is computer hardware with the features necessary to support OSPM and with the interfaces to those features described using the ACPI Description Tables.

**[0038]** The operating system interacts with ACPI hardware by selecting various ACPI defined states – a mode of operation of either a system, processor or device which defines certain behaviors of the system, process or device. . A non-complete list of states is found below. These various states are for described in, for example, Advanced Configuration and Power Interface Specification (Revision 3.0b October 10, 2006). .

- Global System States which are entire system state which is visible to use, and divided into 4 states G0, G1, G2, G3;
- Sleeping States residing within global system state G1 (except for S5) and divided into S1, S2, S3, S4, S5;
- Device Power states usually invisible to the user divided into 4 states – D0, D1, D2, and D3 and which vary with each device;
- CPU Power states also known as CPU sleep states which reside within the global system state G) and is divided into C0, C1, C2, C3, and C4;
- CPU / Device Performance states which controls CPU / Device power management when it is still active, and including usually clock speed and voltage variance depending on workload, and which will have a CPI dependent number of states
- CPU Thermal Monitor which throttles the CPU to low performance state when temperature exceeds a threshold, wherein TM1 throttling is done by changing the duty cycle and in TM2, throttling is done by changing the clock speed or core voltage (P state)

**[0039]** For purposes of power management and determining the amount of electrical power used by a CPU or device, obviously establishing a G1 state or sleeping state is useful, with S1-S4 determining wake up latency of 2 seconds to 30 seconds. Further, the various CPU power states (C-states) also allow useful methods to have an operating system regulate power at the CPU levels. C states are processor power consumption and thermal management states. The C0 state is baseline - the processor in this state executes instructions. The C1 state has the lowest latency. The hardware latency in this state must be low enough that the operating software does not consider the latency aspect of the state when deciding whether to use it. Aside from putting the processor in a non-executing power state, this state has no other software-visible effects. This is done in actual execution by using the assembly instruction "halt" with a wake-up time having an order of magnitude of 10 ns. The C2 state offers improved power savings over the C1 state.

The worst-case hardware latency for this state is provided via the ACPI system firmware and the operating software can use this information to determine when the C1 state should be used instead of the C2 state. The transition time from C-2 to C-0 is on the order of magnitude of 100 nanoseconds, and the processes achieved on a processor level by gating the processor core clock and platform I/O buffers.. . Aside from putting the processor in a non-executing power state, this state has no other software-visible effects. The C3 state offers improved power savings over the C1 and C2 states. In the C-3 state, the bus clock and PLLs are dated. Wakeup time is now on the order of 50 μs. The worst-case hardware latency for this state is provided via the ACPI system firmware and the operating software can use this information to determine when the C2 state should be used instead of the C3 state. While in the C3 state, the processor's caches maintain state but ignore any snoops. The operating software is responsible for ensuring that the caches maintain coherency.

[0040] Even further, Device and Processor Performance states (Px states) are power consumption and capability states within the active/executing state C0 and D0. The Px states are briefly defined below. P0 -While a device or processor is in this state, it uses its maximum performance capability and may consume maximum power. P1 - In this performance power state, the performance capability of a device or processor is limited below its maximum and consumes less than maximum power. Pn - In this performance state, the performance capability of a device or processor is at its minimum level and consumes minimal power while remaining in an active state. State n is a maximum number and is processor or device dependent.

[0041] The P-states can be achieved by dynamic voltage scaling and dynamic frequency scaling. Dynamic voltage scaling is a power management technique in computer architecture, where the voltage used in a component is increased or decreased, depending upon circumstances. Reducing the voltage means increasing program runtime while reducing power consumption. Dynamic frequency scaling (also known as CPU throttling) is a technique in computer architecture where a processor is run at a less-than-maximum frequency in order to conserve power.

[0042] In this technique of dynamic voltage and frequency scaling (DVFS), the CPU core voltage, clock rate, or both, can be altered to decrease power consumption at the price of potential performance. This can be used to optimize the power-performance trade-off. DVFS technology includes dynamic voltage scaling and dynamic frequency scaling. Dynamic voltage scaling is a power management technique in computer architecture, where the voltage used in a component is increased or decreased depending on circumstances. Dynamic voltage scaling to decrease voltage is known as undervoltage. MOSFET-based digital circuits operate using voltages at circuit nodes to represent logical states. The voltage at these nodes switches between a high-voltage the low voltage during normal operation-when the inputs to a logic gate transition, the transistors making up that gave me toggle the gates output. At each node in a circuit is a certain amount of capacitance. This capacitance arises from various sources, mainly transistors through a capacitance of diffusion capacitance and wires coupling capacitance. Toggling a voltage at a circuit node requires charging or discharging the capacitance in that no, since cursor related to voltage, the time it takes depends on the voltage applied. By applying a higher voltage to device a circuit, the capacitance is her charge and discharge more quickly, resulting in a faster operation of the circuit and outing for a higher frequency operation. A switching power dissipated by a chip using static CMOS gates is determined by the equation $C^*V^2$ *f, where C is the capacitance being switched per clock cycle, V is the voltage and f is the switching frequency. While this formula is not exact, as modern chips are not implemented using only CMOS technology, it does show that overall power consumption decreases with voltage. Accordingly, dynamic voltage scaling is widely used as part of strategies to manage switching power consumption in battery-powered devices such as cell phones and laptop computers. Low voltage modes are used in conjunction with lower clock frequencies to minimize power consumption associated with such as CPUs and DSPs. Only when significant computational power is needed both voltage and frequency be raised. Note that many peripherals also support low-voltage operational modes. Please note that reducing the voltage means that circuit switch is slower, producing the maximum frequency at which that's circuit can run. This in turn, reduces the rate at which program instructions can be issued, which may increase runtime for program segments which are succeeded significantly CPU-bound. Thus there is a complex trade-off between power management and computational load. Also note that the efficiency of some logical components, such as the regulators, decreases with increasing temperature, so the power use may increase with temperature. Since increasing power use may increase the temperature, increasing the voltage or frequency may also increase system power demands even faster than the above CMOS formula indicates.

**[0043]** Dynamic frequency scaling is a technique into your architecture where processors run at a less than maximum frequency in order to conserve power. Again, the efficiency of some electrical components, such as the voltage regulators, decreases with increasing temperature, so the power use may increase with temperature. Since increasing power use may increase to the temperature, increases in voltage or frequency may increase a system power demand even further than the seamless formula indicates and vice versa.

**[0044]** Note that each of the above techniques, there is a reinforcing effect, where the production of either voltage or frequency applied to the CPU as whole or functional units within the CPU also reduces the amount of heat generated by the CPU or its components. As such, less cooling is the necessary 4 the motherboard, further reducing overall system energy consumption.

**[0045]** The above described techniques are well known in the art of power management for laptops and other such devices. These techniques have been caught codified in a number of ways, again most prominently the Advanced Configuration and Power Interface (ACPI). ACPI is an open industry specification within an established industry. Implementation of these aspects of the invention is readily apparent and well within the grasp of those skilled in the art based on the disclosure herein.

**[0046]** In one embodiment of the present invention, servers 205 and 210 a figure 2 include ACPI hardware and firmware to allow compliance with ACPI standards, and operating systems that are OSPM capable. Figure 5, a top-level flowchart of one method of the present invention as previously discussed, also includes dynamic frequency and voltage scaling control 548. The particulars of any set of available states among a heterogeneous collection of CPUs within a server and/or among a set of servers such as servers 210 and 215 are included in the knowledgebase 515, which shall update module 510 to allow algorithms of power metric computation module 525 and power consumption management module 530 to incorporate this information for enhanced power optimizations. Note that these added dimensions of CPU state will increase the computational complexity of these algorithms. In operation, an application such as the analysis engine shown in figure 5 will issue an APCI call to an operating system to change the CPU state. This operating system may then either call on a server such as server 210 or 215 or a hypervisor such as hypervisor 407 or 412. The hypervisor then interfaces with the hardware to cause the CPU state change.

**[0047]** In another embodiment of the present invention, the analysis engine shown in figure 5 may issue requests for changes to APCI state in any of a plurality of devices, central processing units, motherboards, or any other APCI compliant hardware.

**[0048]** While the foregoing description and drawings represent the preferred embodiments of the present invention, it will be understood that various changes and modifications may be made without departing from the spirit and scope of the present invention.

WHAT IS CLAIMED IS

1. A method for by readily cognizant power management in a distributed computing system, the distributing computing system having a number of processing units , comprising the steps of:
   a) gathering process level information from different mechanisms to improve accuracy of power metric obligation;
   b) gathering environmental metrics;
   c) generating a behavior model based on the process level information and the environmental metrics; and
   d) setting a state in at least one of the number of processing units based on applying the behavioral model to application, utilization and environmental contexts.

2. The method of claim1, wherein the a number of processing units includes at least one of a plurality of central processing units, a plurality of functional units within central processing units, and a plurality of hypervisors managing virtual machines.

3. The method of claim 2, wherein the state is one of a plurality of performance states.

4. The method of claim 3, wherein setting one of a plurality of performance states includes changing the frequency of a processing unit.

5. The method of claim 3, wherein setting one of a plurality of performance states includes changing the voltage of a processing unit.

6. The method of claim 3, wherein setting one of a plurality of performance states includes changing the frequency and the voltage of a processing unit.

7. The method of claim 2, wherein one of a plurality of performance states includes at least a first performance state and a second performance state, wherein the first performance state can bear a greater computational load than the second state.

8. The method of claim 1, wherein the state is one of a plurality of sleep states.

9. The method of claim 8, wherein one of a plurality of sleep states includes at least a first sleep state and second sleep state, wherein the first sleep state has a lower wake up time than the second sleep state.

10. A method for by readily cognizant power management in a distributed computing system, the distributing computing system having a number of processing units , comprising the steps of:

    a) gathering process level information from different mechanisms to improve accuracy of power metric obligation;
    b) gathering environmental metrics;
    c) generating a behavior model based on the process level information and the environmental metrics;
    d) selecting a state for at least one of the number of processing units based on applying the behavioral model to application, utilization and environmental contexts.
    e) issuing a software call to an operating system running on the at least one of the number of processing units to set a state in at least one of the number of processing units based on applying the behavioral model to application, utilization and environmental contexts.

11. The method of claim 10, wherein the a number of processing units includes at least one of a plurality of central processing units, a plurality of functional units within central processing units, and a plurality of hypervisors managing virtual machines.

12. The method of claim 11, wherein the state is one of a plurality of performance states.

13. The method of claim 12, wherein setting one of a plurality of performance states includes changing the frequency of a processing unit.

14. The method of claim 12, wherein setting one of a plurality of performance states includes changing the voltage of a processing unit.

15. The method of claim 12, wherein setting one of a plurality of performance states includes changing the frequency and the voltage of a processing unit.

16. The method of claim 12, wherein one of a plurality of performance states includes at least a first performance state and a second performance state, wherein the first performance state can bear a greater computational load than the second state.

17. The method of claim 11, wherein the state is one of a plurality of sleep states.

18. The method of claim 17, wherein one of a plurality of sleep states includes at least a first sleep state and second sleep state, wherein the first sleep state has a lower wake up time than the second sleep state.

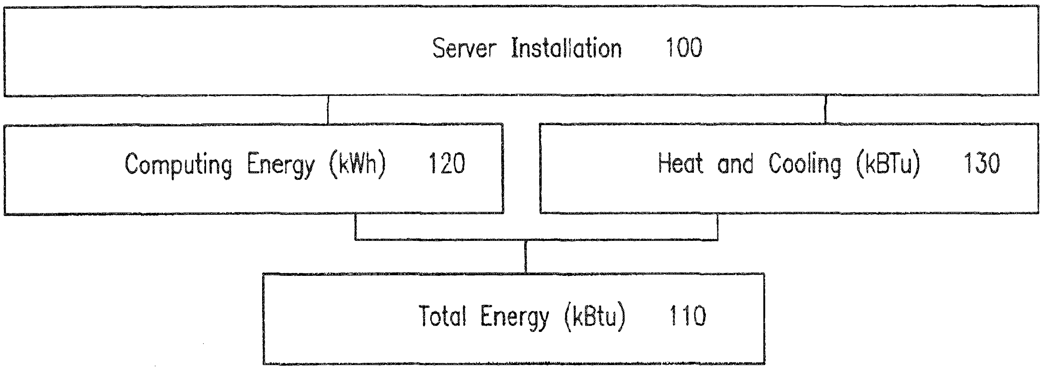
Server Installation 100
Computing Energy (kWh) 120
Heat and Cooling (kBTu) 130
Total Energy (kBtu) 110

FIG. 1

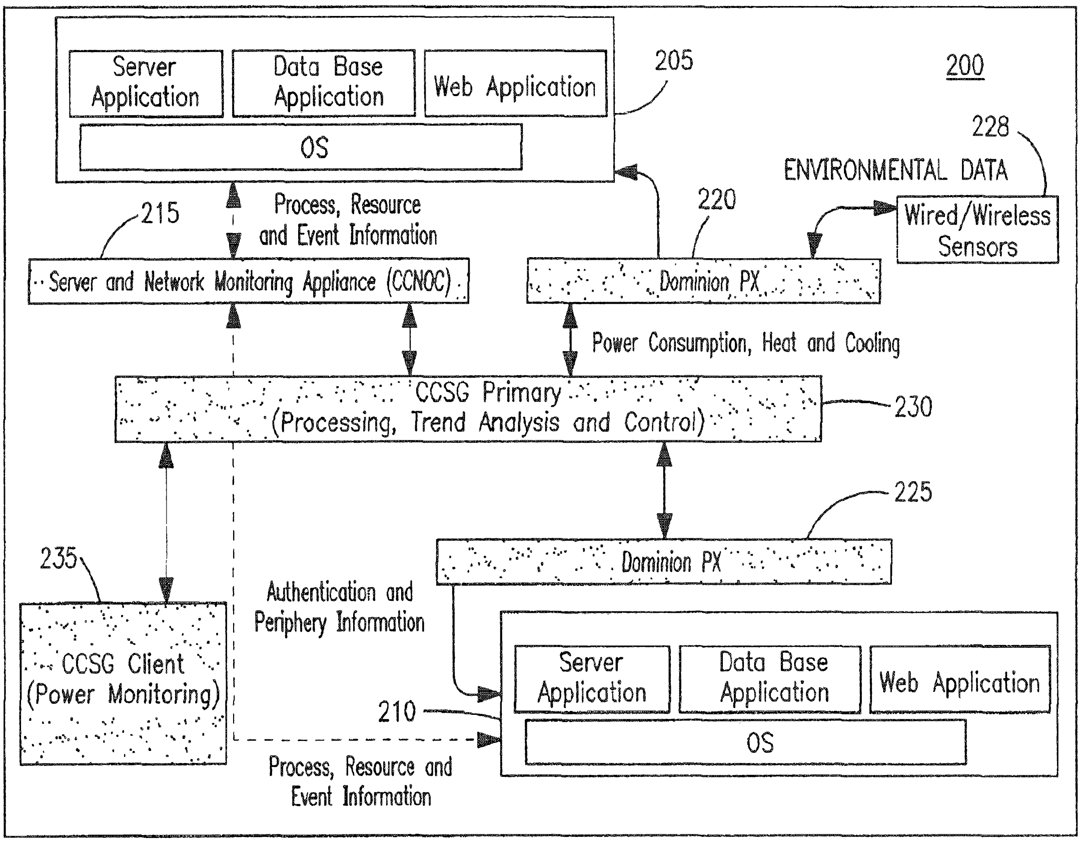
Server Application
Data Base Application
Web Application
OS
205
200
228
ENVIRONMENTAL DATA
Wired/Wireless Sensors
220
215
Process, Resource and Event Information
Server and Network Monitoring Appliance (CCNOC)
Dominion PX
Power Consumption, Heat and Cooling
CCSG Primary (Processing, Trend Analysis and Control)
230
225
Dominion PX
235
Authentication and Periphery Information
CCSG Client (Power Monitoring)
Server Application
Data Base Application
Web Application
210
OS
Process, Resource and Event Information

FIG. 2

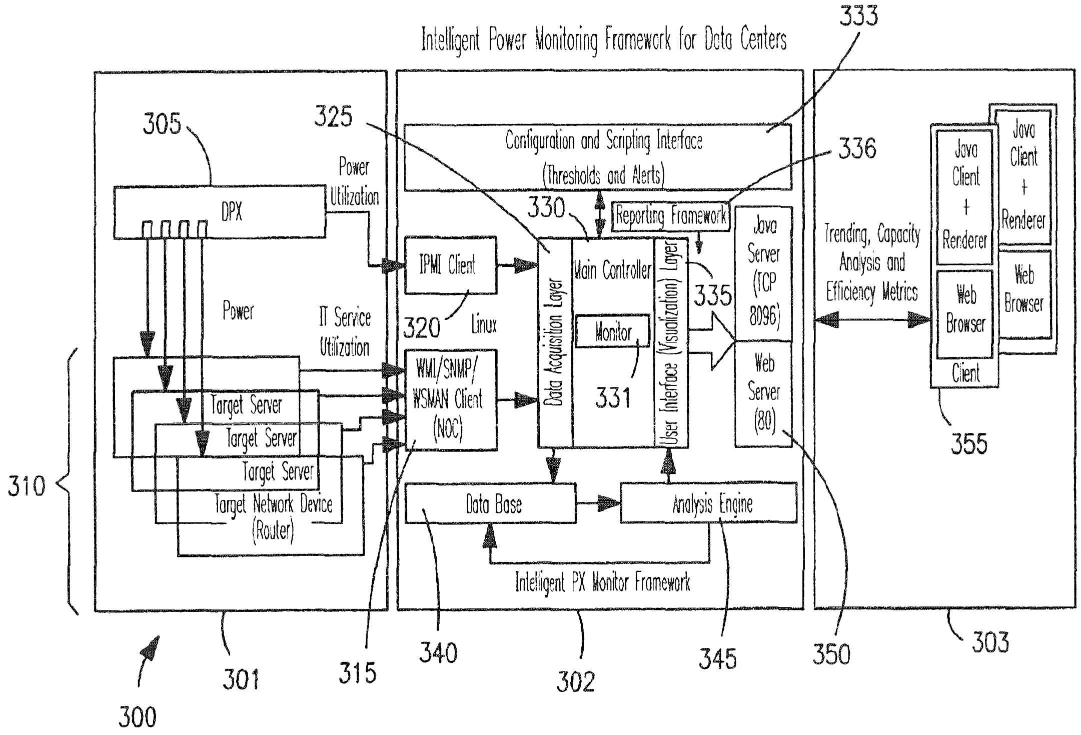
Intelligent Power Monitoring Framework for Data Centers
333
305
325
Configuration and Scripting Interface
(Thresholds and Alerts)
336
Power Utilization
DPX
330
Reporting Framework
IPMI Client
Main Controller
Java Server (TCP 8096)
Power
IT Service Utilization
320
Linux
335
Trending, Capacity Analysis and Efficiency Metrics
Java Client + Renderer
Java Client + Renderer
Web Browser
Web Browser
Client
Monitor
WMI/SNMP/ WSMAN Client (NOC)
Data Acquisition Layer
User Interface (Visualization) Layer
331
Web Server (80)
Target Server
Target Server
Target Server
Target Network Device (Router)
310
Data Base
Analysis Engine
Intelligent PX Monitor Framework
355
300
301
315
340
302
345
350
303

FIG. 3

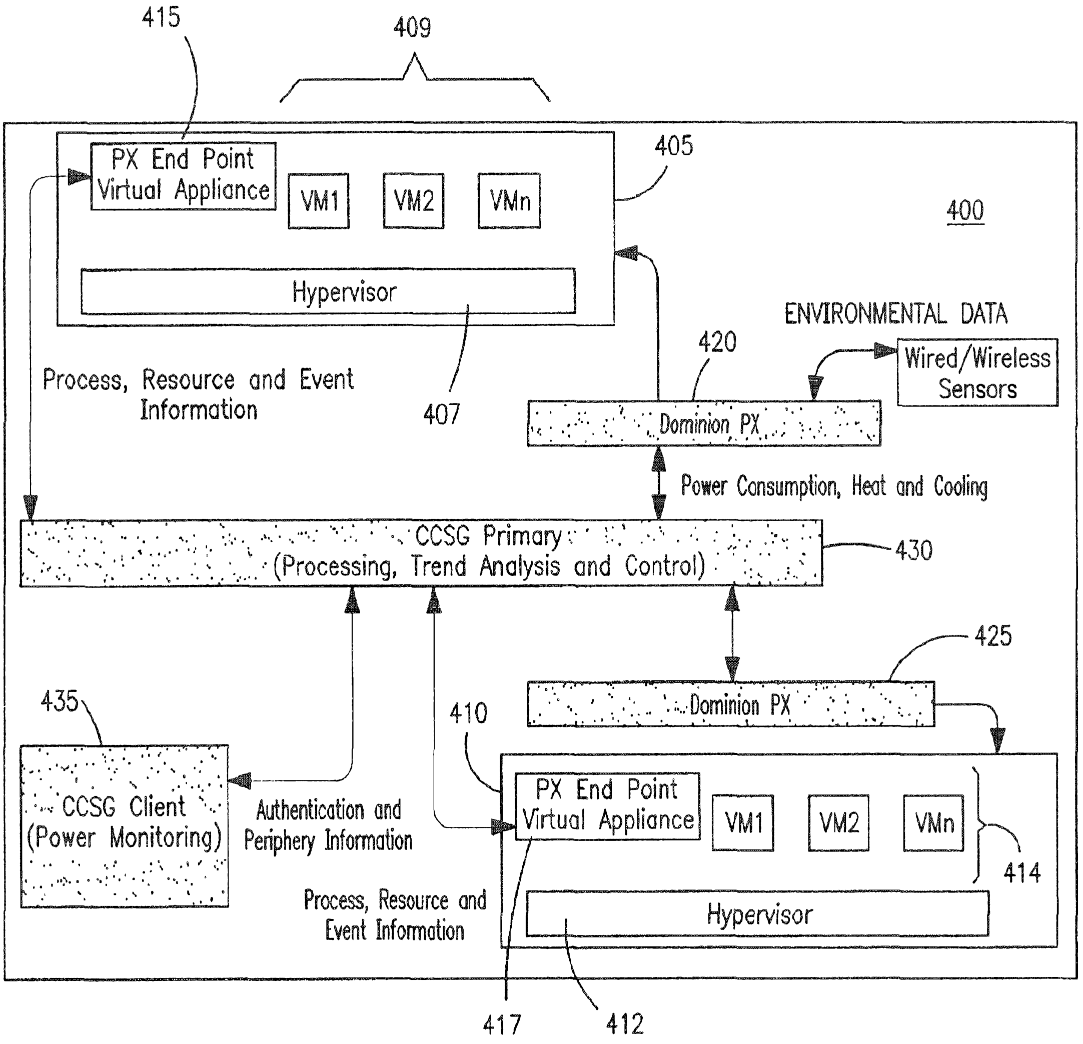
415
409
PX End Point
Virtual Appliance
VM1
VM2
VMn
405
400
Hypervisor
ENVIRONMENTAL DATA
420
Wired/Wireless Sensors
Process, Resource and Event Information
407
Dominion PX
Power Consumption, Heat and Cooling
CCSG Primary
(Processing, Trend Analysis and Control)
430
425
435
Dominion PX
410
CCSG Client
(Power Monitoring)
Authentication and Periphery Information
PX End Point
Virtual Appliance
VM1
VM2
VMn
414
Process, Resource and Event Information
Hypervisor
417
412

FIG. 4

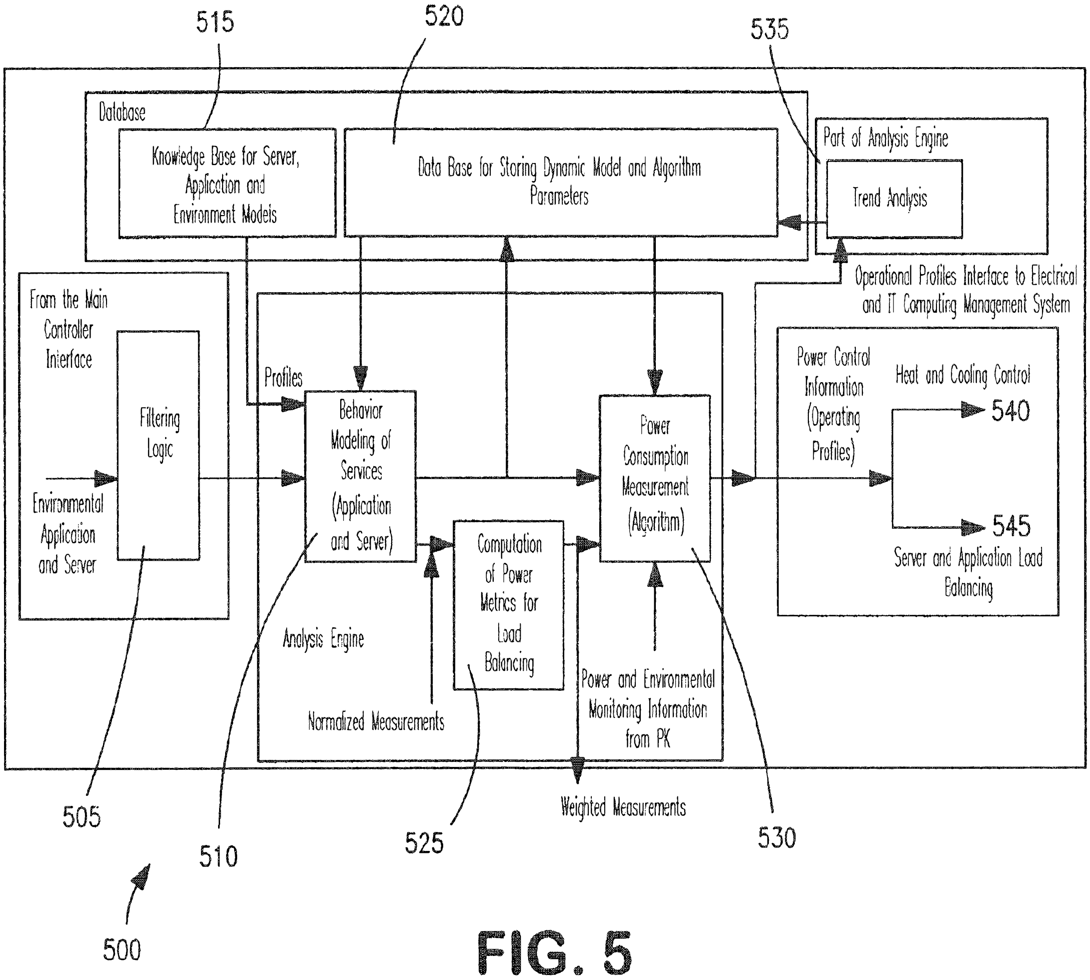
515
520
535
Database
Knowledge Base for Server, Application and Environment Models
Data Base for Storing Dynamic Model and Algorithm Parameters
Part of Analysis Engine
Trend Analysis
Operational Profiles Interface to Electrical and IT Computing Management System
From the Main Controller Interface
Filtering Logic
Environmental Application and Server
Profiles
Behavior Modeling of Services (Application and Server)
Power Consumption Measurement (Algorithm)
Power Control Information (Operating Profiles)
Heat and Cooling Control
540
545
Server and Application Load Balancing
Computation of Power Metrics for Load Balancing
Analysis Engine
Normalized Measurements
Power and Environmental Monitoring Information from PK
Weighted Measurements
505
510
525
530
500

FIG. 5