Computing (2017) 99:597–615
DOI 10.1007/s00607-016-0501-1

Toward sustainable data centers: a comprehensive energy management strategy

Jordi Guitart[1]

Received: 6 August 2015 / Accepted: 21 June 2016 / Published online: 28 June 2016

Abstract Data centers are major contributors to the emission of carbon dioxide to the atmosphere, and this contribution is expected to increase in the following years. This has encouraged the development of techniques to reduce the energy consumption and the environmental footprint of data centers. Whereas some of these techniques have succeeded to reduce the energy consumption of the hardware equipment of data centers (including IT, cooling, and power supply systems), we claim that sustainable data centers will be only possible if the problem is faced by means of a holistic approach that includes not only the aforementioned techniques but also intelligent and unifying solutions that enable a synergistic and energy-aware management of data centers. In this paper, we propose a comprehensive strategy to reduce the carbon footprint of data centers that uses the energy as a driver of their management procedures. In addition, we present a holistic management architecture for sustainable data centers that implements the aforementioned strategy, and we propose design guidelines to accomplish each step of the proposed strategy, referring to related achievements and enumerating the main challenges that must be still solved.

Keywords Green computing · Energy efficiency · Energy management · Energy measurement · Sustainability · Resource management · Data centers

This work is supported by the Ministry of Science and Technology of Spain under contract TIN2015-65316-P, by the Generalitat de Catalunya under contract 2014-SGR-1051, and by the European Commission under FP7-SMARTCITIES-2013 contract 608679 (RenewIT) and FP7-ICT-2013-10 contracts 610874 (ASCETiC) and 610456 (EuroServer).

✉ Jordi Guitart
jordi.guitart@bsc.es

[1] Barcelona Supercomputing Center, Universitat Politecnica de Catalunya, 08034 Barcelona, Spain

Mathematics Subject Classification 68U01 · 68M01 · 68-02

1 Introduction

Energy used by data centers worldwide increased by about 56 % from 2005 to 2010, accounting for 1.5 % of total energy use in 2010 according to Koomey's study [1]. Greenpeace [2] estimates that data centers energy use can grow up to 1012 billion kWh by 2020, which is a 3× increment regarding their energy consumption in 2007. The cost of this enormous amount of energy has turned into the primary cost driver for data centers. In particular, Belady [3] estimates that the annual amortized energy costs in a data center for a single server exceeded the cost of the server itself in 2008, and that the combined cost of the cooling infrastructure and energy would be 75 % of the cost in 2014, while the IT equipment would be only 25 %. That is a significant shift from the 20–80 % ratio of the early 90s. In addition to the energy cost burden, the energy consumption of the data centers contributes also to the climate change by increasing the CO_2 emissions to the atmosphere. In particular, The Climate Group [4] states that worldwide data centers emitted 116 million metric tons of CO_2 ($MtCO_2$) in 2007, slightly more than the entire country of Nigeria, and claims that this figure could increase to 259 $MtCO_2$ by 2020, even considering the recent advances in virtualization, cooling, and power supply that are being introduced in data centers.

For these reasons, there has been recently a great interest in the development of techniques to reduce the energy consumption and the environmental footprint of data centers. They range from proposals to enhance specific aspects of data centers design and operation (for instance, usage of low-power components, energy-aware resource management, and free-cooling solutions) to overall strategies to achieve sustainable data centers, as proposed by Google [5] and HP [6]. This position paper follows the line of those latter approaches and proposes a comprehensive strategy to reduce the carbon footprint of data centers, by using the energy as a driver of the management procedures of the data center. In addition, we present a holistic management architecture for sustainable data centers that implements the aforementioned strategy.

Given the scenario described in the previous paragraphs, we claim that sustainable data centers will be only possible if the problem is faced by means of a holistic approach that includes not only solutions to reduce the energy consumption of the hardware equipment of data centers (including IT, cooling, and power supply systems) but also intelligent and unifying solutions that enable a synergistic and energy-aware management of the hardware equipment and the software infrastructure. In particular, we envision the following steps to come up with a strategy that can enable sustainable data centers:

1. Enable awareness of energy impact and carbon footprint.
2. Increase the energy efficiency of the IT equipment.
3. Increase the energy efficiency of the cooling and power supply subsystems.
4. Increase the use of renewable energy.
5. Exploit opportunities in the energy markets.
6. Customize this energy strategy to the data center reality.

In the following sections, we present the management architecture to implement such strategy and we describe in detail the several steps of our strategy and their interaction. For each step, we propose design guidelines to accomplish that step, we present related achievements and cite relevant works, and we enumerate the main challenges that must be still solved and the areas that need further research.

2 Architecture

Figure 1 presents an architecture for managing sustainable data centers that implements the proposed strategy. The management subsystem is led by the *energy-aware manager* and comprises several configurable controllers to manage the various subsystems in the data center. As described in Sect. 8, the controlling capabilities of the data center determine what management controllers are required. In addition, the *energy-aware manager* is in charge of customizing the management strategy to fulfill the objectives and constraints during operation of the whole data center and configuring and coordinating the management controllers accordingly.

The *monitoring and metrics* and *modeling and prediction* components are responsible for the energy and carbon awareness of the data center. As described in Sect. 3, the former monitors the operation of the data center and calculates relevant metrics that provide an indication of its sustainability. The latter uses this monitored information to build models that capture the system behavior and uses them to forecast that behavior in the future.

The *IT equipment controller* manages the IT equipment as described in Sect. 4, deciding about the energetic status and the operating frequency of the physical hosts in the data center, as well as the configuration parameters of the networking elements (i.e. routers, switches, …).

The *workload shaping and task scheduling* component is responsible for the execution of the IT workload in the data center. As described in Sects. 4 and 6, it decides about what tasks are accepted for execution, when such tasks will be executed, and their placement in the physical hosts of the data center.

The *remote controller* selects the most appropriate data center where to deploy a workload when remote execution in the ecosystem is expected to provide a better outcome regarding energy efficiency or carbon footprint, as described in Sects. 4.3 and 6.

The *cooling controller* configures the cooling subsystem over different operating points to maintain a target temperature while reducing the power consumption, as described in Sect. 5.

The *power supply controller* manages the power supply subsystem as described in Sect. 6, deciding for each time interval the amount of energy to be generated on-site, the amount of energy to be taken from the power grid, and the amount of energy to be stored for later use, and also how to distribute the available energy among the rest of subsystems of the data center.

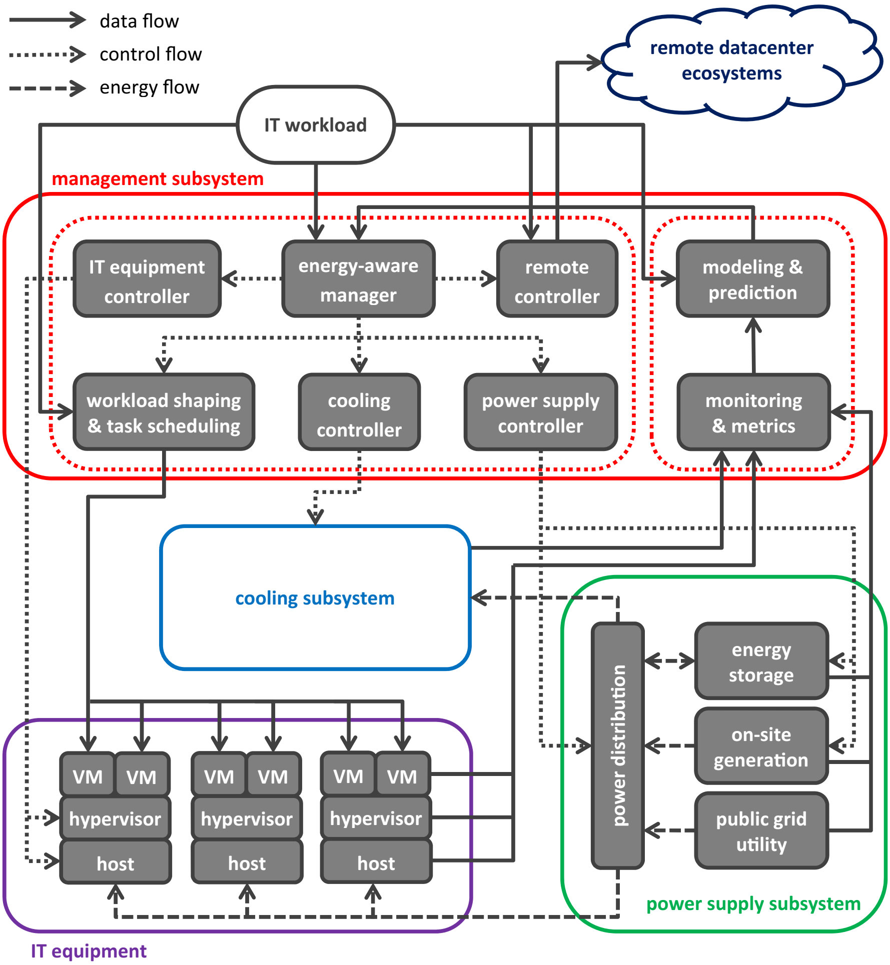

Fig. 1 Architecture for the management of sustainable data centers

3 Energy and carbon awareness

Energy and carbon awareness refers to the ability of data centers to measure their energy consumption and carbon footprint. It is a fundamental capability because "consumers and businesses can't manage what they can't measure", as stated by The Climate Group [4].

3.1 Metrics

First, we must define what to measure. Usually, data centers account for metrics that provide an indication of their sustainability degree, and to this end, they measure parameters such as greenhouse gas emissions, power consumption, temperatures, humidity, etc. [7]. For instance, a common metric is the Power Usage Effectiveness (PUE) [8], which is the ratio of the total amount of energy used by a data center (including the power supply and cooling subsystems in addition to the IT equipment) to the energy delivered to the IT equipment. A similar rationale is used to calculate the Carbon Usage Effectiveness (CUE) [9], but using the total CO_2 emissions caused by the total amount of energy used by a data center in the numerator.

PUE is useful as an indicator on how efficiently the data center uses the energy in the long-term (in fact, it is generally reported from annual information aggregates), but it does not capture well the dynamic nature of the energy usage in data centers. PUE Scalability [8] has been suggested to show how well a data center total energy consumption scales with dynamic changes in its IT equipment loads. Furthermore, PUE is not intended to assess how energy is being used to generate efficiently the useful work of the data center. Data Center energy Productivity (DCeP) [10] has been suggested to quantify the useful work output that a data center provides in relation to the amount of energy expended to produce it. This metric allows the continuous monitoring of the work product as a function of energy consumed during data center operation, but it can be sometimes difficult to implement, since the useful work in a data center can take many forms. This hardens also the comparison of the productivity of different data centers by using this metric. In addition, some authors claim that metrics such as DCeP can favor algorithms that achieve high sustained performance, which are also power hungry algorithms, and propose an alternative metric (so-called FTTSE), which prioritizes the total energy reduction and the minimization of time to solution [11].

There is an ongoing effort trying to harmonize the metrics to be used to measure data centers energy and ecological efficiency. Whereas some consensus has been achieved regarding the metrics [12], their adoption as drivers of the energy management strategies of data centers is still in the initial stages. An important challenge for the standardization of this kind of metrics comes up because of the different actors involved in the data center operation (i.e. owners, administrators, users), who have different requirements regarding the information they want to obtain and about the granularity of such information (per function, per service, per host, per rack, or the whole data center) [7].

3.2 Measurements versus models

Second, we must define how to measure. The total amount of energy used by a data center can be monitored with Metered Power Distributions Units (PDU), which provide the ability to measure power usage of a system at the socket. Modern IT equipment also allows to measure power usage at the server by means of the Intelligent Platform Management Interface (IPMI) [13]. Some servers can even report the power consumed by some individual components (CPU, RAM, etc.) if the server comes equipped with the appropriate sensors. Alternatively, Intel Sandy Bridge processors include an onboard energy and power metering capability for processor packages and DRAM so-called Running Average Power Limit (RAPL), which estimates energy usage by using hardware performance counters and a software power model [14].

The amount of carbon emissions must be determined for the actual mix of energy delivered to the data center, including on-site and grid-based energy sources, by using the carbon emission factor ($kgCO_2/kWh$) associated to each energy source [15]. Whereas it is easy to identify energy sources when energy is generated on-site, when the energy consumed comes from the grid the data center must rely on the information provided by the corresponding energy supplier. Generally, they provide aggregate values for a given region.

The information about energy consumption and carbon footprint should be integrated in the existing monitoring platforms of data centers (e.g., Ganglia [16]), allowing in that way a seamless access to all the information that is relevant to understand the data center behavior.

Power metering mechanisms described above provide real measures of the energy consumption of individual hosts. However, they cannot account for the energy consumption of individual virtual machines (VMs) or services running on each host, which is required for instance to implement effective power-aware scheduling policies. This can be accomplished by using power models, which correlate the resource usage and the power consumption of individual VMs based on monitored resource usage information [17,18]. Models also allow to estimate the current power consumption of hosts and VMs when direct measurement is not feasible (due to lack of scalability in large clusters, performance overhead of measurements, and cost of devices [19]) and to forecast the power consumption in the future (see Sect. 3.3).

Models are generated offline by collecting power measurements and resource usage indicators periodically while the system runs a special training workload that includes micro-benchmarks that selectively stress each resource at varying levels of utilization. Given the heterogeneity of applications running in current data centers [20], the training workload must be generic, since including only a given kind of applications [21] will result in inaccurate estimations for the rest, and able to capture the essentials of the power behavior of the modeled host. To this end, power models must consider the impact on energy consumption of several resources, including not only the processor (as models for traditional HPC applications did in the past), but also the memory, the disk, and the network. Those subsystems apart from the processor have been reported to make up 40–60 % of the total power consumption depending on the workload [22]. Similarly, power models must also support heterogeneous virtualization hypervisors, as they have been reported to impact differently the performance and energy consumption of the applications [23].

Only appropriate indicators to account for the usage of each resource must be selected. The indicator must clearly represent the usage intensity of the resource (i.e. if the usage increases, the value of the indicator should increase correspondingly) and it must be unequivocally correlated with the energy consumption incurred by that resource. For instance, the utilization is not the best indicator of the processor usage regarding its correlation with energy consumption, because applications with the same utilization can have different processor energy consumption depending on what instructions they are executing [17]. According to this, the utilization cannot be the unique determinant of the processor power consumption, though it could be used to further refine a model that considers, for instance, the number of operations executed.

The collected resource usage indicators are fitted to the power measurements by using machine learning techniques [24], which include mainly traditional auto-regressive methods (e.g. linear regression) [19] but also more novel methods such as artificial neural networks [25]. Those are very powerful techniques to capture data correlations, but they might require driving the mining process to achieve better accuracy and reduce the computational complexity. According to this, the modeling methodology should systematically filter the relevant indicators and include derivatives that capture better the existing relations. Models have assumed often a linear relationship between the power consumption and the resource usage. Linear models provided a reasonable accuracy with low computational complexity in traditional platforms, but they are weak in modern multicore platforms [26], where non-linear models provide better accuracy, although they incur also higher computational complexity.

Power models must support heterogeneous platforms (e.g. low-power such as Intel Atom vs. high-performance such as Intel Xeon), which are the norm in current data centers [20], by incorporating a specific model for each type of platform, as they can present very different energy consumption patterns [27]. One could think that this would lead to different models for each application-platform pair, which would entail a noticeable effort to generate those models. However, we can reduce the complexity incurred due to platform heterogeneity by deriving approximate models of new platforms using the model of existing platforms with the same architecture, thus avoiding the benchmarking of the new platforms. Similarly, we can reduce the complexity incurred due to application heterogeneity by using the platform model generated with the training workload and, if needed, refining it online for each application (or group of applications sharing common resource usage patterns) using captured data.

Finally, power models must consider the impact of co-locating several VMs in the same resources. Sharing resources generates an interference between the VMs, which induces some overhead [28]. This overhead has an obvious impact on the performance of the VMs, but also on the energy consumption.

3.3 Forecasts

As power models correlate the resource usage and the power consumption, if one can predict the resource usage of a host or a VM in the future, those models could be then used to forecast the energy consumption of that host or VM [29]. Resource usage predictions can be obtained using time series analysis techniques [30]. Note that this approach supports predictions of the energy consumption at different periods ahead, from short-term to long-term, provided that the model is supplied with the corresponding resource usage predictions. Contrariwise, many other model-based predictors only allow for very short-term forecasts, as they feed the model with current measurements of resource usage instead of predictions.

It should be also helpful if power models can be used to forecast the energy consumption for potential actions. For instance, predicting the energy consumption of the data center upon a VM deployment, VM migration, or VM cancellation before the actual action takes place [31]. The system scheduler could use these forecasts to guide its resource management decisions. Predicting the future resource usage is harder with this kind of forecasts because the effect of the potential action must be incorporated to the predicted resource usage in the data center. The complexity increases considerably when forecasting the energy impact of a combination of potential actions.

The amount of carbon emissions can be predicted from the expected energy consumption and the carbon emission factor associated to each energy source, using the same rationale described in Sect. 3.2. This requires predicting the expected workload, but also the expected energy mix to be delivered to the data center, as the energy mix will vary over time when using intermittent energy sources such as solar and wind energy.

4 Energy efficiency of the IT equipment

Energy efficiency describes the ratio between the produced outcomes and the energy used. Something is more energy efficient if it delivers more services using the same amount of energy, or the same services using less energy. IT equipment can be more energy efficient by means of energy-driven management strategies that reduce its energy consumption and carbon footprint.

4.1 Energy-related actuators

Energy-driven management strategies require actuators that can modify the behavior of the IT equipment regarding their energy footprint. For instance, data centers can take profit of virtualization technology to consolidate workloads from different customers into a smaller number of physical hosts. This allows turning off (or suspending) unused hosts (see ACPI C-states [32]) and therefore reducing energy consumption. Data centers can apply also elasticity of VMs regarding the use of resources to control the number of VMs in each physical host. Moreover, they can use also Dynamic Voltage and Frequency Scaling (DVFS) to adjust the voltage and/or frequency of the processors on the fly (see ACPI P-states [32]) aiming to reduce power consumption (dynamic power consumption changes linearly with frequency and in a quadratic manner with voltage) or to reduce the amount of heat dissipated. Note that running an application at lower frequency increases its execution time, which could lead to higher energy consumption depending on the relationship between the reduction of power consumption and the increase in the execution time for a given frequency reduction [33]. Smart management solutions that trade off these variables to reduce the overall energy consumption are needed here.

4.2 Energy-driven management

Energy-awareness and energy-related actuators allow putting energy-driven management strategies for the IT equipment into operation, as they can be the basis for management algorithms that optimize their operation according to energy-based objectives (e.g. maximize the energy productivity of the data center). On one side, management algorithms can decide the placement of VMs in the physical hosts of the data center, as well as the energetic status of those hosts (on(P-state)/suspended(C-state)/off) [34]. Each VM must be allocated with enough resources of each type to fulfill its demand. This complicates the allocation problem as several dimensions have to be considered. The impact of co-locating several VMs in the same physical host must be also considered as this can degrade their performance due to interferences accessing the resources [35]. As the energy consumption of the data center is correlated with the number of active physical hosts, a common strategy to improve energy efficiency is to consolidate all the VMs running in the data center in the minimum number of hosts (the necessary to fulfill the desired performance) while suspending the rest of hosts [36]. These idle hosts could be turned on again if they were needed when a peak load occurs. However, management algorithms can achieve better results if, instead of the number of active hosts, they consider the dynamic power consumption of the physical hosts (which depends on their load) that would result from each configuration [37]. For that, they will use the power models described in the previous section to forecast the energy consumption of VMs and physical hosts according to their predicted resource usage and the status of the infrastructure. In addition, DVFS and elasticity of VMs can be used to adapt the computing capacity of the hosts and the VMs to the intensity of the workloads they have to execute [38].

On the other side, management algorithms can decide the configuration parameters of the networking elements (i.e. routers, switches, …), as well as their energetic status, to optimize the energy needed to route messages within the data center. This can have a noticeable impact, especially in big data centers, which organize in complex hierarchical switching topologies. According to this, common strategies to improve energy efficiency consist of adapting dynamically the operating data transmission rate of communication links [39], finding routes that minimize the total number of switches and turning off unused ones [40], and exploiting low energy modes in physical hosts and networking devices together in a single management algorithm [41].

The outcome of those optimization algorithms can be significant when applied in heterogeneous environments [27]. On one side, data centers can have heterogeneous platforms (e.g. low-power vs. high-performance), which have very different energy consumption profiles. Low-power platforms offer reduced power consumption in the idle state and mild increments with resource usage, but they provide also reduced performance. For this reason, heterogeneous platforms present interesting trade-offs for the management algorithms, i.e. run briefly in a high-performance high-power platform and suspend this platform for longer vs. run for a longer time in a low-performance low-power platform and only suspend the platform briefly.

On the other side, this platform heterogeneity fits very well with the heterogeneity of the applications running in today's data centers, which include interactive services, high performance computing jobs and big data applications. As these applications present very different resource usage patterns (and consequently, different energy consumption profiles), they allow for several placement options in the data center according to the required performance by each application and the energy consumption incurred to provide that performance [42]. Again, interesting trade-offs appear here, i.e. place all the applications of a given type in the platform where their energy productivity is better but foster their interference because they are using the same kind of resources vs. co-locate applications of different types in a given platform to minimize interference even if their energy productivity is not optimal.

4.3 Data center ecosystems

The aforementioned management algorithms can contribute to the energy efficiency of data centers when considered in isolation. However, current data centers normally work within ecosystems where they interact with other data centers. For example, Cloud data centers can be organized in federations where a given data center can lease capacity from other data centers to meet peaks in demand or to outsource less critical workloads. Service owners can also use multi-cloud scenarios to deploy their service across multiple data centers [43].

In such scenarios, energy-driven management strategies can also have a noticeable impact. In this case, management algorithms will aim to optimize the operation of data centers according to energy-based objectives by selecting the most appropriate data center where to deploy every workload [a.k.a. Geographical Load Balancing (GLB)]. For instance, GLB allows exploiting data centers located in different time zones and energy price variability [44] by enabling a follow-the-moon strategy [45], which deploys workloads in data centers by following the night-time, when data centers may be in an off-peak demand interval, the energy is cheaper, and lower outside temperature opens the possibility to use free cooling. GLB also allows exploiting data centers powered by renewable sources to reduce carbon emissions, as we will discuss in Sect. 6. In addition, GLB can be integrated with energy buffering management in order to shave peak power draw from the grid [46].

Whereas data center ecosystems offer new opportunities for energy-driven management, they also encompass new challenges that must be considered [34,46,47], such as the distant geographic locations of data centers, which have an impact in the migration of VMs (increasing the cost significantly (and the consumed energy) and frequently causing service level degradations for the affected customers) and in the data exchange among VMs located in different data centers (increasing the communication latency among them); the independent administrative domains involved in the ecosystem, which have frequently conflicting goals and do not generally disclose information about their energy consumption and energy mix, thus increasing the need for third parties to independently assess energy data of data centers and share this information within the ecosystem; and the importance of the prediction accuracy of the input data (e.g. workload, energy price, renewable energy), which depends on the predictability of each data source and the prediction window length, and can be a downgrading factor on the efficiency of the management algorithms.

4.4 Multi-objective management

Energy-related objectives are important for data centers administrators, but they are also interested in other aspects, such as availability, reliability, profit, and performance. For this reason, the management algorithms should optimize the operation of data centers according to multiple objectives. This complicates the optimization problem, especially when some objectives are mutually exclusive and both cannot be maximized at the same time. For instance, consolidating VMs onto a small number of physical hosts and turning off idle hosts is an effective way to reduce energy consumption, but this can cause heat imbalances and create hot spots, which may impact cooling costs and degrade host life and performance [48].

Multi-objective optimization problems can be solved via scalarization, that is, converting the original problem with multiple objectives into a single-objective optimization problem, for instance, by considering their impact on the data center profit [49]. The problem can be also solved as a single-objective problem when all but one of the objectives have a target value, as a constraint can be placed on those objectives. When a target value can be identified for all the objectives, the problem can be solved by means of goal-oriented adaptation, which is able to learn from monitored data the impact of repair actions on the value of the goals and apply the most convenient action when any of them deviates from its target value [50]. Another possibility for solving true multi-objective problems consists of computing all or a representative set of Pareto optimal solutions (those in which it is impossible to make any objective better off without making at least another one worse off), which are usually derived by means of evolutionary algorithms [48,51].

A similar issue arises when we consider not only the objectives of the data center, but also the objectives of its clients. Hopefully, a data center should fulfill the goals of its clients in addition to its own goals, but there could be incompatibilities between the goals of the data center and its clients or among the goals of the clients themselves. For instance, this occurs when data centers offer distinct quality of service categories (gold, silver, etc.) to their customers [52]. Again, trade-off solutions are required.

4.5 Optimization problem solving

All the described optimization algorithms, whether they are single- or multi-objective, must solve a packing problem [34]. A way to obtain optimal solutions consists of formulating the problem in terms of some mathematical program, for instance by using Integer Linear Programming, and using an existing solver to calculate the solution [53]. However, packing problems are known to be NP-hard. This means that brute-force methods to find the optimal solution become infeasible when the scenario starts to grow. According to this, construction heuristics and local search methods are required to allow good solutions (although maybe not optimal) to be decided in real time.

Given the similarity with the well-known bin-packing problem, simple bin-packing heuristics (such as First Fit, Best Fit, and First Fit Decreasing) have been adapted to the VM allocation problem [54]. Nevertheless, the VM allocation problem is more complicated than bin packing [55]. This obliges to analyze to what extent the bin-packing heuristics can be applied and what adaptations are required. The main differences of the VM allocation problem refer to its multi-dimensionality regarding the type of resources that determine the host capacities and VM sizes, the costs incurred by VM migrations, the performance degradation when a host is overloaded (or close to), the heterogeneity of the hosts, the impact of the load of the host in its energy consumption, the finite number of hosts in the data center which constraints the number of VMs that can be allocated, and the volatility of the system regarding the number of VMs to allocate and the number of available hosts [55].

5 Energy efficiency of the cooling and power supply subsystems

The Climate Group states that in 2002 only about half of the energy used by data centers powered the IT equipment, while the rest was needed to run backup, uninterruptible power supplies and cooling systems [4]. Despite the recent advances in cooling and power supply systems, The Climate Group estimates that in 2020 the energy footprint of these systems will still account for a 39 % of the total footprint of data centers (about 101 $MtCO_2$).

According to this, sustainable data centers must also aim to increase the energy efficiency of the power and cooling systems, or in other words, aim to reduce the PUE. This goal requires the usage of novel techniques to reduce the energy consumption of these subsystems (such as free cooling, which consists of using ambient air or water to cool data center space and equipment), but also including these systems in a comprehensive energy-aware management strategy that optimizes the IT equipment, as described in Sect. 4, while reducing the energy consumption of the cooling and energy supply subsystems.

On one side, this can be accomplished by means of smarter algorithms to manage the cooling subsystem, for instance, by adapting proactively the amount of cooling to the expected workload [56]. If we envision a low utilization in the data center in a near future, we can already throttle down the cooling subsystem, because less hosts will required and we will be able to suspend some of them (or use DVFS), thus reducing the amount of dissipated heat and the cooling needs. This differs from traditional reactive management approaches, which are triggered when the temperature surpasses predefined thresholds, and can result in delayed response to temperature changes. In addition, the optimal threshold range to avoid component damage while at the same time avoiding energy waste in unnecessary cooling is very hard to determine [56].

On the other side, the algorithms managing the IT workload can profile the tasks according to their impact on the temperature and consider the resulting temperature of the hosts as a criterion to decide the allocations. Again, a lower resulting temperature allows the cooling subsystem to be throttled down, further decreasing power consumption. For instance, the algorithms can apply this idea to decide when tasks must be executed, and run *cold* tasks (i.e. those with low impact on the processor temperature) after *hot* tasks to lower the resulting temperature (the order of execution of the tasks has an impact on the temperature) [57]. Similarly, they can also decide where tasks must be executed, and place *hot* tasks in the hosts of the data center that are easier to cool [58]. Naturally, both decisions on when and in which host to execute a task can be taken together to minimize the resulting temperature [59].

The described approaches to optimize the cooling subsystem must be tightly integrated with the optimization of the IT equipment. Isolated strategies to reduce the energy consumption of the IT equipment tend to consolidate the workload on fewer hosts and turn off the rest, which can cause heat imbalances and create hot spots, thus increasing the cooling needs. Isolated strategies to reduce the energy consumption of the cooling subsystem tend to spread the workload over all the hosts, thus increasing the idle power consumption of the IT equipment, as more hosts are turned on. A joint optimization that balances the number of active hosts to trade off the power needed by the cooling subsystem and the idle power needed by the IT equipment can reduce the total power consumption of the data center [60]. As the cooling subsystem could require some time to adapt changing conditions, joint approaches to optimize the cooling subsystem and the IT equipment can configure the cooling subsystem over different operating points to deal with long-term fluctuations and use thermal-aware workload placement to deal with short-term ones [61].

There is also room for the optimization of the power supply subsystem, especially when the data center can generate energy on-site (see Sect. 6).

6 Ecological efficiency of the IT equipment

Enhancements in energy efficiency, as described in previous sections, will contribute to the reduction of the energy consumed by a data center to accomplish a given amount of work. Although this is very important to reduce the carbon footprint of data centers, it is not enough. As the workload demand over data centers is expected to increase, their energy consumption will increase also, even if energy efficiency measures are applied. In fact, some argue that energy efficiency could encourage further increments in the workload demand due to lower energy costs (which is known as Jevons' Paradox [62]). According to this, sustainable data centers require not only to reduce their energy consumption but also to use as much as possible energy sources that do not contribute to carbon emissions, that is, renewable energy sources.

The use of renewable energy sources introduces new challenges into the management of data centers. One of them appears when using intermittent energy sources, that is, sources that are not always available, such as solar and wind energy. In these scenarios, the management algorithms must adapt the IT workload to the energy availability (i.e. workload shaping). On one side, workload shaping can be accomplished by using techniques such as DVFS or server power state tuning to adjust the power demand to the time-varying renewable power budget [63]. On the other side, workload shaping can be also based on task scheduling, for instance by postponing tasks when not enough renewable energy is available to execute them, or by bringing forward tasks if there is more renewable energy than needed [64].

Data center ecosystems offer more opportunities to increase the use of renewable energy (and reduce the carbon footprint), as they comprise data centers located in different geographical locations, and even in different time zones, which results in different local weather conditions at some particular time. This allows enabling a follow-the-renewables strategy [65], where workloads are deployed in data centers according to the availability of renewable energy in their geographical locations [66], or to the expected carbon emissions, which are calculated from the data center energy mix [67].

Co-location (data centers draw renewable energy directly from an existing nearby plant) and self-generation (data centers generate their own renewable energy) have been reported as the preferred methods (versus grid-centric approaches) for data centers to exploit renewable energy [68], as they allow decreasing the transmission and conversion losses (because the power is generated close to where it is consumed and DC power can be directly supplied to the IT equipment instead of AC power) and avoid the grid-transmission charges imposed by power utilities. Self-generation allows also the utilization of the power plant waste heat to generate cooling for the data center [69]. These features make self-generation an appealing option to obtain cheaper power than the grid once the important initial capital cost of installing the needed infrastructure has been amortized.

Self-generation offers also additional management opportunities because the management algorithms can control the power supply system. In particular, they can use a load following approach that adapts the amount of energy generated to the expected workload [70]. As energy generation systems could require some time to adapt the amount of energy they generate, load following is not adequate to fit the energy demands of the IT workload when this fluctuates in short time intervals. In these situations, load following can be combined with workload shaping. The former defines the amount of energy to be generated during the next (coarse) time period according to the estimated workload demand, which is calculated from past information, and the latter works within that time period to adapt the IT workload to fit its deviations to the available energy [71].

7 Exploit opportunities in the energy markets

Even after applying the previous steps, data centers can still have some carbon footprint. This can be further reduced by exploiting the opportunities that are available in the current energy markets, provided that the required expenses to reduce its carbon footprint by using these opportunities fit in the data center business strategy. For instance, Renewable Energy Certificates (RECs) (a.k.a. Renewable Energy Credits) [72] represent proof that one MWh of electricity was generated from a renewable energy source. RECs are tradable commodities in energy markets and they allow their owner to claim that the corresponding portion of its overall energy consumption comes from renewable sources.

A data center can reduce its carbon footprint by purchasing RECs [73,74]. Similarly, a data center that generates renewable energy on-site can offer its spare green energy in the market by issuing RECs, allowing other data centers to offset their carbon impact [74]. Note that, in addition to the benefit of a global reduction of the carbon impact, these strategies can offer also business opportunities to the data centers to increase their profit.

8 Customization of the energy strategy

Energy management strategies should be flexible in the sense that they must be easily customizable to each data center reality. On one side, they should adapt to the energy management mechanisms that the data center has. For instance, whether it can control the energy supply (because some energy is generated on-site), the operation point of the cooling subsystem, or the energetic status of the IT equipment. This requires a management architecture with independent controllers for each subsystem that can be plugged in/out as necessary and can work together to manage the data center.

On the other side, they should allow the definition of the objectives and constraints during operation of the whole data center (or any of the applications running there) and they should be able to fulfill them. For instance, ensuring a minimum performance (e.g. a minimum number of requests per second for a Web server), a maximum amount of energy that can be consumed (e.g. the amount of available green energy is limited), a maximum amount of carbon that can be emitted (e.g. renewable energy credits are limited), a maximum amount of power that can be provided (e.g. there is a power budget), or a maximum value for the operational costs (e.g. there is an economic budget).

The data center administrators define the high-level objectives and constraints according to the business interests of the provider. Those high-level objectives are commonly referred as Business Level Objectives (BLO). The Green Grid Data Center Maturity Model [75], which suggests best practices for energy efficient data centers, can be also used as reference to define constraints [50]. The energy-aware manager configures the management controllers in order to fulfill the BLOs. Some controllers support their configuration directly by means of BLOs and can adapt their behavior to fulfill them under changing conditions [76]. Other controllers are configured by means of low-level objectives related to its own domain or adjustable low-level policies [77]. In this case, the energy-aware manager must be able to translate the BLOs to low-level terms that the management controllers can understand [78].

Management controllers operate independently and adhere to their specified local objectives by means of a self-adaptation loop. However, the impact of their decisions goes beyond their own domain, and can affect the rest of controllers. In fact, these decisions could be conflicting and their combined effect could fail to fulfill the data center BLOs. According to this, the energy-aware manager must implement a self-adaptive loop that monitors the system status, evaluates the BLOs fulfillment under those conditions, and aligns the configuration of the management controllers to ensure the fulfillment of the BLOs [77]. The effect of different configurations over the BLOs can be learned from available data using machine learning techniques [50].

9 Conclusions

In this paper, we have presented a comprehensive energy management strategy for data centers that aim for sustainability and a holistic architecture that implements such strategy. We claim that those data centers must first be aware of their energy impact and carbon footprint by measuring appropriate metrics and being able to forecast their value by using models. They must also increase the energy efficiency of their IT equipment by optimizing the placement of tasks in the data center physical hosts, as well as the energetic status of those hosts, while exploiting the opportunities given by heterogeneous platforms and applications and by data center ecosystems. They must increase the energy efficiency of the cooling and power supply subsystems by integrating them in the energy strategy applied with the IT equipment. As carbon footprint is directly impacted by the used energy sources, data centers must increase also the use of renewable energy by generating it on-site and optimizing its management depending on the available renewable energy, or by purchasing carbon credits to avoid accounting their emissions as real carbon impact. Finally, all the previous steps must fit in the frame of each data center and have to be customized accordingly.

References

1. Koomey J (2011) Growth in data center electricity use 2005 to 2010. Report, Analytics Press, Oakland
2. Greenpeace International (2010) Make IT green: cloud computing and its contribution to climate change. Report. http://www.greenpeace.org/international/en/publications/reports/make-it-green-cloudcomputing/
3. Belady CL (2007) In the data center, power and cooling costs more than the IT equipment it supports. Electronics cooling. http://www.electronics-cooling.com/2007/02/in-the-data-center-power-and-cooling-costs-more-than-the-it-equipment-it-supports
4. The Climate Group (2008) SMART 2020: enabling the low carbon economy in the information age. Report. http://www.smart2020.org/_assets/files/02_Smart2020Report.pdf
5. Kava J (2011) Google's three steps for a zero carbon green data center. Data center Dynamics London. http://www.greenm3.com/gdcblog/2011/12/6/googles-three-steps-for-a-zero-carbon-green-data-center.html
6. Arlitt M, Bash C, Blagodurov S, Chen Y, Christian T, Gmach D, Hyser C, Kumari N, Liu Z, Marwah M, Mcreynolds A, Patel R, Shah A, Wang Z, Zhou R (2012) Towards the design and operation of net-zero energy data centers. In: Proceedings of 13th IEEE intersociety conference on thermal and thermomechanical phenomena in electronic systems (ITherm'12), San Diego, USA, pp 552–561. doi:10.1109/ITHERM.2012.6231479
7. Wang L, Khan SU (2013) Review of performance metrics for green data centers: a taxonomy study. J Supercomput 63(3):639–656. doi:10.1007/s11227-011-0704-3
8. The Green Grid (2012) PUE: a comprehensive examination of the metric. Tech. Rep. White Paper n.49. https://www.thegreengrid.org/en/Global/Content/white-papers/WP49-PUEAComprehensiveExaminationoftheMetric
9. The Green Grid (2010) Carbon usage effectiveness (CUE): a green grid data center sustainability metric. Tech. Rep. White Paper n.32. http://www.thegreengrid.org/Global/Content/white-papers/Carbon_Usage_Effectiveness_White_Paper
10. The Green Grid (2008) A framework for data center energy productivity. Tech. Rep. White Paper n.13. http://www.thegreengrid.org/Global/Content/white-papers/Framework-for-Data-Center-Energy-Productivity
11. Bekas C, Curioni A (2010) A new energy aware performance metric. Comput Sci Res Dev 25(3–4):187–195. doi:10.1007/s00450-010-0119-z
12. The Green Grid (2014) Harmonizing global metrics for data center energy efficiency. Statement. http://www.thegreengrid.org/~/media/Regulatory/HarmonizingGlobalMetricsforDataCenterEnergyEfficiency.pdf
13. Intel Corporation: Intelligent Platform Management Interface (IPMI). http://www.intel.com/content/www/us/en/servers/ipmi/ipmi-home.html
14. Intel Corporation (2015) Intel 64 and IA-32 architectures software developer's manual. Volume 3B: system programming guide, Part 2. http://www.intel.es/content/dam/www/public/us/en/documents/manuals/64-ia-32-architectures-software-developer-vol-3b-part-2-manual.pdf
15. The Carbon Trust (2013) Energy and carbon conversion factors. Tech. Rep. CTL 153. http://www.carbontrust.com/media/18223/ctl153_conversion_factors.pdf
16. Ganglia Monitoring System. http://ganglia.sourceforge.net/
17. Kansal A, Zhao F, Liu J, Kothari N, Bhattacharya, AA (2010) Virtual machine power metering and provisioning. In: Proceedings of 1st ACM symposium on cloud computing (SoCC'10), Indianapolis, USA, pp 39–50. doi:10.1145/1807128.1807136
18. Yang H, Zhao Q, Luan Z, Qian D (2014) iMeter: an integrated VM power model based on performance profiling. Fut Gen Comput Syst 36:267–286. doi:10.1016/j.future.2013.07.008
19. Mobius C, Dargie W, Schill A (2014) Power consumption estimation models for processors, virtual machines, and servers. IEEE Trans Parall Distrib Syst 25(6):1600–1614. doi:10.1109/TPDS.2013.183

20. Reiss C, Tumanov A, Ganger GR, Katz RH, Kozuch MA (2012) Heterogeneity and dynamicity of clouds at scale: google trace analysis. In: Proceedings 3rd ACM symposium on cloud computing (SoCC'12), San Jose, USA, pp 7:1–7:13. doi:10.1145/2391229.2391236
21. Jarus M, Oleksiak A, Piontek T, Weglarz J (2014) Runtime power usage estimation of HPC servers for various classes of real-life applications. Fut Gen Comput Syst 36:299–310. doi:10.1016/j.future.2013.07.012
22. Bircher WL, John LK (2012) Complete system power estimation using processor performance events. IEEE Trans Comput 61(4):563–577. doi:10.1109/TC.2011.47
23. Varrette S, Guzek M, Plugaru V, Besseron X, Bouvry P (2013) HPC performance and energy-efficiency of Xen, KVM and VMware Hypervisors. In: Proceedings of 25th international symposium on computer architecture and high performance computing (SBAC-PAD'13), Porto de Galinhas, Brazil, pp 89–96, IEEE (2013). doi:10.1109/SBAC-PAD.2013.18
24. Alpaydin E (2014) Introduction to machine learning, 3rd edn. The MIT Press, Cambridge
25. Cupertino L, Da Costa G, Pierson JM (2015) Towards a generic power estimator. Comput Sci Res Dev 30(2):145–153. doi:10.1007/s00450-014-0264-x
26. McCullough JC, Agarwal Y, Chandrashekar J, Kuppuswamy S, Snoeren AC, Gupta RK (2011) Evaluating the effectiveness of model-based power characterization. In: Proceedings 2011 USENIX annual technical conference (ATC'11), Portland, USA, pp 159–172
27. Zhang Q, Zhani MF, Boutaba R, Hellerstein JL (2014) Dynamic heterogeneity-aware resource provisioning in the cloud. IEEE Trans Cloud Comput 2(1):14–28. doi:10.1109/TCC.2014.2306427
28. Kousiouris G, Cucinotta T, Varvarigou T (2011) The effects of scheduling, workload type and consolidation scenarios on virtual machine performance and their prediction through optimized artificial neural networks. J Syst Softw 84(8):1270–1291. doi:10.1016/j.jss.2011.04.013
29. Lewis AW, Tzeng NF, Ghosh S (2012) Runtime energy consumption estimation for server workloads based on chaotic time-series approximation. ACM Trans Arch Code Optim 9(3):15:1–15:26. doi:10.1145/2355585.2355588
30. Brockwell P, Davis R (2002) Introduction to time series and forecasting, 2nd edition. Springer, New York. doi:10.1007/b97391
31. Subirats J, Guitart J (2015) Assessing and forecasting energy efficiency on cloud computing platforms. Fut Gen Comput Syst 45:70–94. doi:10.1016/j.future.2014.11.008
32. UEFI Forum (2013) Advanced configuration and power interface specification revision 5.0a. http://www.acpi.info/spec50a.htm
33. Etinski M, Corbalan J, Labarta J, Valero M (2012) Understanding the future of energy-performance trade-off via DVFS in HPC environments. J Parall Distrib Comput 72(4):579–590. doi:10.1016/j.jpdc.2012.01.006
34. Mann ZA (2015) Allocation of virtual machines in cloud data centers: a survey of problem models and optimization algorithms. ACM Comput Surv 48(1):11:1–11:34. doi:10.1145/2797211
35. Kim SG, Eom H, Yeom HY (2013) Virtual machine consolidation based on interference modeling. J Supercomput 66(3):1489–1506. doi:10.1007/s11227-013-0939-2
36. Xiao Z, Song W, Chen Q (2013) Dynamic resource allocation using virtual machines for cloud computing environment. IEEE Trans Parall Distrib Syst 24(6):1107–1117. doi:10.1109/TPDS.2012.283
37. Beloglazov A, Abawajy J, Buyya R (2012) Energy-aware resource allocation heuristics for efficient management of data centers for cloud computing. Fut Gen Comput Syst 28(5):755–768. doi:10.1016/j.future.2011.04.017
38. Tesfatsion S, Wadbro E, Tordsson J (2014) A combined frequency scaling and application elasticity approach for energy-efficient cloud computing. Sustain Comput Inf Syst 4(4):205–214. doi:10.1016/j.suscom.2014.08.007
39. Bilal K, Khan S, Madani S, Hayat K, Khan M, Min-Allah N, Kolodziej J, Wang L, Zeadally S, Chen D (2013) A survey on green communications using adaptive link rate. Cluster Comput 16(3):575–589. doi:10.1007/s10586-012-0225-8
40. Zhang Y, Ansari N (2015) HERO: hierarchical energy optimization for data center networks. IEEE Syst J 9(2):406–415. doi:10.1109/JSYST.2013.2285606
41. Addis B, Ardagna D, Capone A, Carello G (2014) Energy-aware joint management of networks and cloud infrastructures. Comput Netw 70:75–95. doi:10.1016/j.comnet.2014.04.011
42. Chun BG, Iannaccone G, Iannaccone G, Katz R, Lee G, Niccolini L (2010) An energy case for hybrid datacenters. SIGOPS Oper Syst Rev 44(1):76–80. doi:10.1145/1740390.1740408

43. Ferrer AJ, Hernandez F, Tordsson J, Elmroth E, Ali-Eldin A, Zsigri C, Sirvent R, Guitart J, Badia RM, Djemame K, Ziegler W, Dimitrakos T, Nair SK, Kousiouris G, Konstanteli K, Varvarigou T, Hudzia B, Kipp A, Wesner S, Corrales M, Forg N, Sharif T, Sheridan C (2012) OPTIMIS: a holistic approach to cloud service provisioning. Fut Gen Comput Syst 28(1):66–77. doi:10.1016/j.future.2011.05.022
44. Le K, Bianchini R, Martonosi M, Nguyen TD (2009) Cost- and energy-aware load distribution across data centers. In: Proceedings of 2009 SOSP workshop on power aware computing and systems (HotPower'09), Big Sky, USA, USENIX Association
45. Hatcher J (2013) Follow the moon. Data centre management magazine. http://datacentremanagement.com/news/view/follow-the-moon
46. Abbasi Z, Pore M, Gupta S (2013) Impact of workload and renewable prediction on the value of geographical workload management. In: Energy-efficient data centers: 2nd internationl on workshop, E2DC 2013. Revised selected papers, Lecture notes in computer science, Springer, vol 8343, pp 1–15. doi:10.1007/978-3-642-55149-9_1
47. Kecskemeti G, Kertesz A, Cs Marosi A, Nemeth Z (2014) Strategies for increased energy awareness in cloud federations. In: High-performance computing on complex environments, chap. 19, pp 365–382. Wiley, New York. doi:10.1002/9781118711897.ch19
48. Xu J, Fortes JAB (2010) Multi-objective virtual machine placement in virtualized data center environments. In: Proceedings of 2010 international conference on green computing and communications and international conference on cyber, physical and social computing, Hangzhou, China, IEEE 2010, pp 179–188. doi:10.1109/GreenCom-CPSCom.2010.137
49. Goiri I, Berral JL, Fitó JO, Julià F, Nou R, Guitart J, Gavaldà R, Torres J (2012) Energy-efficient and multifaceted resource management for profit-driven virtualized data centers. Fut Gen Comput Syst 28(5):718–731. doi:10.1016/j.future.2011.12.002
50. Vitali M, Pernici B, OReilly UM (2015) Learning a goal-oriented model for energy efficient adaptive applications in data centers. Inf Sci 319:152–170. doi:10.1016/j.ins.2015.01.023
51. Gao Y, Guan H, Qi Z, Hou Y, Liu L (2013) A multi-objective ant colony system algorithm for virtual machine placement in cloud computing. J Comput Syst Sci 79(8):1230–1242. doi:10.1016/j.jcss.2013.02.004
52. Wada H, Suzuki J, Yamano Y, Oba K (2012) E^3: a multiobjective optimization framework for SLA-aware service composition. IEEE Trans Serv Comput 5(3):358–372. doi:10.1109/TSC.2011.6
53. Guenter B, Jain N, Williams C (2011) Managing cost, performance, and reliability tradeoffs for energy-aware server provisioning. In: Proceedings on IEEE INFOCOM 2011, Shanghai, China, IEEE 2011, pp 1332–1340. doi:10.1109/INFCOM.2011.5934917
54. Shi L, Furlong J, Wang R (2013) Empirical evaluation of vector bin packing algorithms for energy efficient data centers. In: Proceedings 18th IEEE symposium on computers and communications (ISCC'13), Split, Croatia, IEEE 2013, pp 9–15. doi:10.1109/ISCC.2013.6754915
55. Mann ZA (2015) Approximability of virtual machine allocation: much harder than bin packing. In: Proceedings of 9th Hungarian-Japanese symposium on discrete mathematics and its applications, Fukuoka, Japan, pp 21–30
56. Lee EK, Kulkarni I, Pompili D, Parashar M (2012) Proactive thermal management in green datacenters. J Supercomput 60(2):165–195. doi:10.1007/s11227-010-0453-8
57. Chrobak M, Durr C, Hurand M, Robert J (2011) Algorithms for temperature-aware task scheduling in microprocessor systems. Sustain Comput Inf Syst 1(3):241–247. doi:10.1016/j.suscom.2011.05.009
58. Chen Y, Gmach D, Hyser C, Wang Z, Bash C, Hoover C, Singhal S (2010) Integrated management of application performance, power and cooling in data centers. In: Proceedings of 2010 IEEE network operations and management symposium (NOMS'10), Osaka, Japan, IEEE 2010,pp 615–622. doi:10.1109/NOMS.2010.5488433
59. Banerjee A, Mukherjee T, Varsamopoulos G, Gupta SK (2011) Integrating cooling awareness with thermal aware workload placement for HPC data centers. Sustain Comput Inf Syst 1(2):134–150. doi:10.1016/j.suscom.2011.02.003
60. Ahmad F, Vijaykumar TN (2010) Joint optimization of idle and cooling power in data centers while maintaining response time. In: Proceedings of 15th international conference on architectural support for programming languages and operating systems (ASPLOS'10), Pittsburgh, USA, ACM 2010, pp 243–256. doi:10.1145/1736020.1736048
61. Vasic N, Scherer T, Schott W (2010) Thermal-aware workload scheduling for energy efficient data centers. In: Proceedings of 7th internationl conference on autonomic computing (ICAC'10), Washington, USA, ACM 2010, pp 169–174. doi:10.1145/1809049.1809076

62. Belady CL (208) Does efficiency in the data center give us what we need? Mission critical magazine. http://www.missioncriticalmagazine.com/articles/does-efficiency-in-the-data-center-give-us-what-we-need
63. Gmach D, Rolia J, Bash C, Chen Y, Christian T, Shah A, Sharma R, Wang Z (2010) Capacity planning and power management to exploit sustainable energy. In: Proceedings of 2010 international conference on network and service management (CNSM'10), Niagara Falls, Canada, IEEE 2010, pp 96–103. doi:10.1109/CNSM.2010.5691329
64. Goiri I, Haque ME, Le K, Beauchea R, Nguyen TD, Guitart J, Torres J, Bianchini R (2015) Matching renewable energy supply and demand in green datacenters. Ad Hoc Netw 25(Part B):520–534. doi:10.1016/j.adhoc.2014.11.012
65. Liu Z, Lin M, Wierman A, Low SH, Andrew LL (2011) Geographical load balancing with renewables. SIGMETRICS Perf Eval Rev 39(3):62–66. doi:10.1145/2160803.2160862
66. Zhang Y, Wang Y, Wang X (2011) GreenWare: greening cloud-scale data centers to maximize the use of renewable energy. In: Proceedings of 12th ACM/IFIP/USENIX international middleware conference, Lisbon, Portugal, Lecture notes in computer science, Springer, vol 7049, pp 143–164. doi:10.1007/978-3-642-25821-3_8
67. Cappiello C, Melia P, Pernici B, Plebani P, Vitali M (2014) Sustainable choices for cloud applications: a focus on CO2 emissions. In: Proceedings of 2nd international conference on ICT for sustainability (ICT4S'14), Stockholm, Sweden, Atlantis Press, pp 352–358. doi:10.2991/ict4s-14.2014.43
68. Goiri I, Katsak W, Le K, Nguyen TD, Bianchini R (2013) Parasol and greenswitch: managing datacenters powered by renewable energy. In: Proceedings of 18th international conference on architectural support for programming languages and operating systems (ASPLOS'13), Houston, USA, ACM 2013, pp 51–64. doi:10.1145/2451116.2451123
69. Erden HS, Khalifa HE (2012) Energy and environmental assessment of on-site power and cooling for data centers. HVAC & R Res 18(1–2):51–66. doi:10.1080/10789669.2011.585422
70. Kirby B, Hirst E (2000) Customer-specific metrics for the regulation and load-following ancillary services. Tech. Rep. CON-474, Oak Ridge National Laboratory (ORNL). http://web.ornl.gov/~webworks/cpr/rpt/105927.pdf
71. Li C, Zhou R, Li T (2013) Enabling distributed generation powered sustainable high-performance data center. In: Proceedings of 19th international symposium on high performance computer architecture (HPCA'13), Shenzhen, China, IEEE 2013, pp 35–46. doi:10.1109/HPCA.2013.6522305
72. United States Environmental Protection Agency (EPA) Renewable Energy Certificates (RECs). http://www.epa.gov/greenpower/gpmarket/rec.htm
73. Deng N, Stewart C, Gmach D, Arlitt M, Kelley J (2012) Adaptive green hosting. In: Proceedings of 9th international conference on autonomic computing (ICAC'12), San Jose, USA, ACM 2012, pp 135–144. doi:10.1145/2371536.2371561
74. Ren C, Wang D, Urgaonkar B, Sivasubramaniam A (2012) Carbon-aware energy capacity planning for datacenters. In: Proceedings of 20th international symposium on modeling, analysis and simulation of computer and telecommunication systems (MASCOTS'12), Arlington, USA, IEEE 2012, pp 391–400. doi:10.1109/MASCOTS.2012.51
75. The Green Grid (2015) Data center maturity model. White Paper n.56. http://www.thegreengrid.org/en/Global/Content/Tools/DataCenterMaturityModel
76. Macias M, Guitart J (2014) SLA negotiation and enforcement policies for revenue maximization and client classification in cloud providers. Fut Gen Comput Syst 41:19–31. doi:10.1016/j.future.2014.03.004
77. Sedaghat M, Hernandez F, Elmroth E (2011) Unifying cloud management: towards overall governance of business level objectives. In: Proceedings of 11th IEEE/ACM international symposium on cluster, cloud and grid computing (CCGrid'11), Newport Beach, USA, IEEE 2011, pp 591–597. doi:10.1109/CCGrid.2011.65
78. Kousiouris G, Menychtas A, Kyriazis D, Gogouvitis S, Varvarigou T (2014) Dynamic, behavioral-based estimation of resource provisioning based on high-level application terms in cloud platforms. Fut Gen Comput Syst 32:27–40. doi:10.1016/j.future.2012.05.009