Cloud Services, Networking and Management, Edited by Nelson L. S. da Fonseca and Raouf Boutaba.
ISBN 0-471-XXXXX-X

Chapter 0

Energy Consumption Optimization in Cloud Data Centers

Dzmitry Kliazovich, Pascal Bouvry, Fabrizio Granelli, Nelson L. S. da Fonseca

Abstract. Cloud computing data centers are becoming increasingly popular for providing computing resources. However, the expenses of these data centers has skyrocketed with the increase in computing capacity with large percentage of the operational expenses due to energy consumption, especially in data centers that are used as backend computing infrastructure for cloud computing. This chapter emphasizes the role of the communication fabric in energy consumption and presents solutions for energy efficient network aware resource allocation in clouds.

Keywords: cloud computing, energy efficiency, resource allocation.

1. Introduction

Cloud computing has entered our lives and is dramatically changing the way people consume information. It provides platforms enabling the operation of a large variety of individually-owned terminal devices. There are about 1.5 billion computers [1] and 6

billion mobile phones [2] in the world today. Next generation user devices, such as Google glasses [3], offer not only constant readiness for operation, but also constant information consumption. In such an environment, computing, information storage and communication become a utility, and cloud computing is one effective way of offering easier manageability, improved security, and a significant reduction in operational costs [4].

Cloud computing relies on the data center industry, with over 500 thousand data centers deployed worldwide [5]. The operation of such widely distributed data centers, however, requires a considerable amount of energy, which accounts for a large slice of the total operational costs [6-7]. Interactive Data Corporation (IDC) [8] reported that, in 2000, on average the power required by a single rack was 1 kW, although in 2008, this had soared to 7.4 kW. The Gartner group has estimated that energy consumption accounts for up to 10% of the current data center operational expenses (OPEX), and with this estimate possibly rising to 50% in the next few years [9]. The cost of energy for running servers may already be greater than the cost of the hardware itself [10], [11]. In 2010, data centers consumed about 1.5% of the world's electricity [12], with this percentage rising to 2% for The United States of America. This consumption accounts for more than 50 million metric of tons of CO2 emissions annually.

Energy efficiency has never been a goal in the information technology (IT) industry. Since the 1980s, the only target has been to deliver more and faster; this has been traditionally achieved by packing more into a smaller space, and running processors at a higher frequency. This consumes more power, which generates more heat, and then requires an accompanying cooling system that costs in the range of $2 to $5 million per

year for corporate data centers [9]. These cooling systems may even require more power than that consumed by the IT equipment itself [13], [14].

Moreover, in order to ensure reliability, computing, storage, power distribution and cooling infrastructures tends to be overprovisioned. To measure this inefficiency, the Green Grid Consortium [15] has developed two metrics: the Power Usage Effectiveness (PUE) and Data Center Infrastructure Efficiency (DCIE) [16], which measures the proportion of power delivered to the IT equipment relative to the total power consumed by the data center facility. PUE is the ratio of total amount of energy used by a computer data center facility to the energy delivered to computing equipment while DCIE is the percentage value derived, by dividing information technology equipment power by total facility power. Currently, roughly 40% of the total energy consumed is related to that consumed by information technology (IT) equipment [17]. The consumption accounts approximately, while the power distribution system accounts the other 15%.

There are two main alternatives for reducing the energy consumption of data centers: (a) shutting down devices or (b) scaling down performance. The former alternative, commonly referred to as Dynamic Power Management (DPM) results in greatest savings, since the average workload often remains below 30% in cloud computing systems [18]. The latter corresponds to Dynamic Voltage and Frequency Scaling (DVFS) technology, which can adjust the performance of the hardware and consumption of power to match the corresponding characteristics of the workload.

In summary, energy efficiency is one of the most important parameters in modern cloud computing datacenters in determining operational costs and capital investment, along with the performance and carbon footprint of the industry. The rest of the chapter is

organized as follows: Section 2 discusses the role of communication systems in cloud computing. Section 3 presents energy efficient resource allocation and scheduling solutions. Finally, Section 4 concludes the paper.

2. Energy Consumption in Data Centers: Components and Models

This section introduces the energy consumption of computing and communication devices, emphasizing how efficient energy consumption can be achieved, especially in communication networks.

2.1 Energy Consumption of Computing Servers and Switches

Computing servers account for the major portion of energy consumption of data centers. The power consumption of a computing server is proportional to the utilization of the CPU utilization. Although an idle server still consumes around two-thirds of the peak-load consumption just to keep memory, disks, and I/O resources running [48], [49]. The remaining one-third increases almost linearly with an increase in the load of the CPU [6], [49]:

$$P_s(l) = P_{fixed} + \frac{\left(P_{peak} - P_{fixed}\right)}{2}(1 + l - e^{-\frac{l}{a}}), \quad (1)$$

where P_{fixed} is idle power consumption, P_{peak} is the power consumed at peak load, l is a server load, and a is the level of utilization at which the server attains power consumption which varies linearly with the offered load. For most CPUs, a ϵ [0.2, 05].

There are two main approaches for reducing energy consumption in computing servers: (a) DVFS [29] and (b) DPM [60]. The former scheme adjusts the CPU power (consequently the level of performance) according to the load offered. The power in a chip decreases proportionally to $V^2 f$, where V is a voltage, and f is the operating frequency. The scope of this DVFS optimization is limited to the CPUs, so that the computing server components, such as buses, memory, and disks continue functioning at the original operating frequency. On the other hand, the DPM scheme can power down computing servers but including all of their components, which makes it much more efficient, but if a power up (or down) is required, considerably more energy must be consumed in comparison to the DVFS scheme. Frequency downshifts can be expressed as follow (Eq. 1):

$$P_s(l) = P_{fixed} + \frac{\left(P_{peak} - P_{fixed}\right)}{2}(1 + l^3 - e^{-\frac{l^3}{a}}), \quad (2)$$

Figure 1 plots the power consumption of computing server.

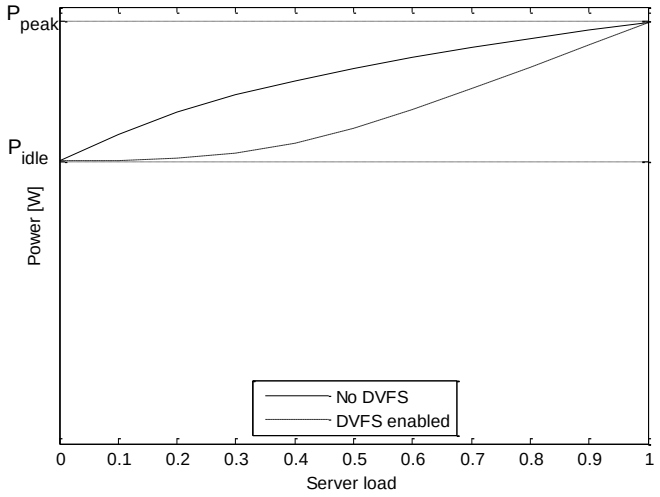

Fig. 1. Computing server power consumption.

Network switches form the basis of the interconnection fabric used to deliver job requests to the computing servers for execution. The energy consumption of a switch depends on various factor: (a) type of switch, (b) number of ports, (c) port transmission rates, and (d) employed cabling solutions; these can be expressed by the following [19]:

$$P_{switch} = P_{chassis} + n_c * P_{linecard} + \sum_{r=1}^{R} n_p^r * P_p^r * u_p^r, \quad (3)$$

where $P_{chassis}$ is the power related to the switch chassis, $P_{linecard}$ is the power consumed by a single line card, n_c is the number of line cards plugged into the switch, P_p^r is the power consumed by a port running at rate *r*, n_p^r is the number of ports operating at rate *r* and $u_p^r \in [0,1]$ is a port utilization, which can be defined as follows:

$$u_p = \frac{1}{T}\int_t^{t+T} \frac{B_p(t)}{C_p} dt = \frac{1}{T * C_p}\int_t^{t+T} B_p(t)dt, \quad (4)$$

where $B_p(t)$ is an instantaneous throughput at the port's link at the time t, C_p is the link capacity, and T is the time interval between measurements.

2.2 Energy Efficiency

In an ideal data center, all the power would be delivered to the IT equipment executing user requests. This energy would then be divided between the communication and the computing hardware. Several studies have mistakenly considered the communication network as overhead, required only to deliver the tasks to the computing servers. However, as will be seen later in this section, communications is at the heart of task execution, and the characteristics of the communication network, such as bandwidth

capacity, transmission delay, delay jitter, buffering, loss ratio, and performance of communication protocols, all greatly influence the quality of task execution.

Mahadevan at el. [19] present power benchmarking of the most common networking switches. With current network switch technology, the difference in power consumption between peak consumption and idle state is less than 8%; turning off an unused port saves only 1-2 watts [20]. The power consumption of a switch is composed of three components: (a) power consumed by the switch base hardware (the chassis), (b) power consumed by active line cards, and (c) power consumed by active transceivers. Only the last component scales with the transmission rate, or the presence of the forwarded traffic, while the former two components remain constant, even when the switch is idle. This phenomenon is known as energy proportionality, and describes how energy consumption increases with an increase in workload [20].

Making network equipment energy proportional is one of the main challenges faced by the research community. Depending on the data center load level, the communication network can consume between 30 and 50% of the total power used by the IT equipment [21], [51] with 30% being typical for highly loaded data centers, whereas 50% is common for average load levels of 10-50% [22]. As with computing servers, most solutions for energy-efficient communication equipment depend on downgrading the operating frequency (or transmission rate) or powering down the entire device or its components in order to conserve energy. One solution, first studied by Shang at el. [21] and Benini at el. [23] in 2003, proposed a power-aware interconnection network utilized Dynamic Voltage Scaling (DVS) links [21], and this, DVS technology was later combined with Dynamic Network Shutdown (DNS) to further optimize energy consumption [25]. The following

papers review the challenges and some of the most important solutions for optimization of energy consumption and the use of resources [54], [55], [56], [57], [58], [59].

The design of these power-aware networks when on/off links are employed is challenging. There are issues with connectivity, adaptive routing, and potential network deadlocks [27]. Because a network always remains connected, such challenges are not faced when using DVS links. Some recent proposals combined traffic engineering with link shutdown functionality [28], but most of these approaches are reactive, and may perform poorly in the event of unfavorable traffic patterns. A proactive approach is necessary for on/off procedures. A number of studies have demonstrated that simple optimization of the data center architecture and energy-aware scheduling can lead to significant energy savings of up to 75% based on traffic management and workload consolidation techniques [29].

2.3 Communication Networks

Communication systems have rarely been extensively considered in cloud computing research. Most of the cloud computing techniques evolved from the fields of cluster and grid computing which are both designed to execute large computationally intensive jobs, commonly referred as High-Performance Computing (HPC) [30]. However, cloud computing is fundamentally different: Clouds satisfy the computing and storage of millions of users at the same time, yet each individual user request is relatively small. These users commonly need merely to read an email, retrieve an HTML page, or watch an online video. Such tasks require only limited computation to be performed yet their performance is determined by the successful completion of the communication requests but

communications involves more than just the data center network; the data path from the data center to the user also constitute an integral part for satisfying a communication request. Typical delays for processing users´ requests, such as search, social networks and video streaming, are less than a few milliseconds, and we sometimes even measured on the level of microsecond. Depending on the user location, these delays are as large as 100 milliseconds for intercontinental links and up to 200 milliseconds if satellite links are involved [31]. As a result, a failure to consider the communication characteristics on an end-to-end basis can mislead the design and operational optimization of modern cloud computing systems.

Optimization of cloud computing systems and cloud applications will not only significantly reduce energy consumption inside data centers, but also globally, in the wide-area network. The World hosts around 1.5 billion Internet users [1] and 6 billion mobile phone users [2], and all of them are potential customers for cloud computing applications. On an average, there are 14 hops between a cloud provider and end users on the Internet [24], [32]. This means that there are 13 routers involved in forwarding the user traffic, each consuming from tens of watts to kilowatts [19]. According to Nordman [33], Internet-connected equipment accounts for almost 10% of the total energy consumed in the United States. Obviously, optimization of the flow of communication between the data center providers and end users can make a significant difference. For example, a widespread adoption of the new Energy-Efficient Ethernet standard IEEE 802.3az [34] can result in savings of 1 billion Euro [35].

At the cloud user end, energy is becoming an even greater concern: More and more cloud users use mobile equipment (smart phones, laptops, tablet PCs) to access cloud

services. The only efficient way for these battery-powered devices to save power is to power off most of the main components, including the central processor, transceivers and memory, while also configuring sleeping cycles appropriately [36]. The aim is to decrease request processing time so that user terminals will consume less battery power. Smaller volumes of traffic arranged in bursts will permit longer sleeping times for the transceivers, and faster replies to the cloud service requests will reduce the drain on batteries.

3. Energy Efficient System-level Optimization of Data Centers

This section addresses issues related to scheduling, load balancing, data replication, virtual machine placement and networking that can be capitalized on to reduce the energy consumption in data centers.

3.1 Scheduling

Job scheduling is at the heart of the successful power management in data centers. Most of the existing approaches focus exclusively on the distribution between of jobs computing servers [37], the targeting of energy efficiency [38] or thermal awareness [39]. Only a few approaches consider the characteristics of the data center network [40-42], such as DPM-like power management [18].

Since energy savings result from such DPM-like power management procedures [18], job schedulers tend to adopt a policy of workload consolidation maximizing the load on the operational computing servers and increasing the number of idle servers that can be put into the “sleep” mode. Such a scheduling policy works well in systems that can be

treated as a homogenous pool of computing servers, but data center network topologies require special policies. For example, the most widely used data center architecture [43], fat tree architecture presented in Fig. 2, blindly concentrates scheduling and may end up grouping all of the highly loaded computing servers on a few racks, yet this creates a bottleneck for network traffic at a rack or aggregation switch.

Moreover, on a rack level, all servers are usually connected using Gigabit Ethernet (GE) interfaces. A typical rack hosts up to 48 servers, but has only two links of 10GE connecting them to the aggregation network. This corresponds to a mismatch of 48GE / 20GE =2.4 between the incoming and the outgoing bandwidth capacities. Implementation in a data center with cloud applications requiring communication means that the scheduler should tradeoff workload concentration with the load balancing of network traffic.

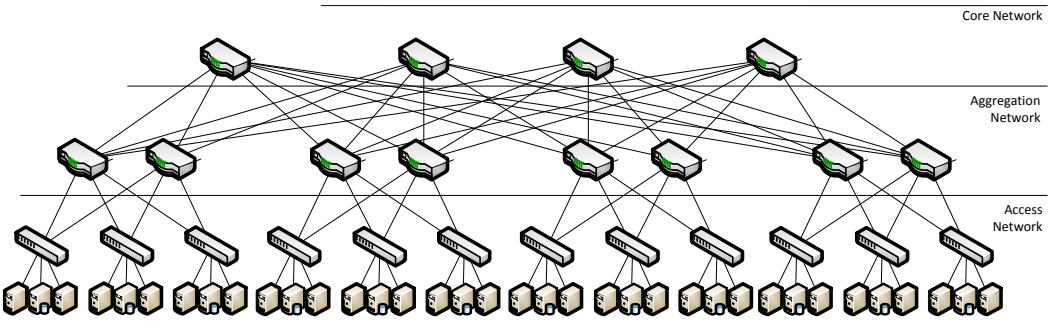

Fig. 2. Three-tier data center architecture.

Any of the data center switches may become congested in either the uplink or downlink direction or both. In the downlink direction, congestion occurs when the capacity of individual ingress links surpasses that of egress links. In the uplink direction, the mismatch in bandwidth is primarily due to the bandwidth oversubscription ratio, which

occurs when the combined capacity of server ports overcomes a switch aggregate uplink capacity.

Congestion (or hotspots) may severely affect the ability of a data center network to transport data. The Data Center Bridging Task Group (IEEE 802.1) [44] specifies layer-2 solutions for congestion control in IEEE 802.1Qau standard. This standard introduces a feedback loop between data center switches to signal the presence of congestion. Such feedback allows overloaded switches to backpressure heavy senders by notifying them if the congestion. Such technique can avoid some of the congestion-related losses and keep the data center network utilization high. However, it does not address the problem adequately since as it is more efficient to assign data-intensive jobs to different computing servers so that those jobs can avoid sharing common communication paths. To benefit from such spatial separation in the three-tiered architecture (Fig. 2), these jobs must be distributed among the computing servers in proportion to job communication requirements. However, such approach contradict the objectives of energy-efficient scheduling, which tries to concentrate all of the active workloads on a minimum set of servers and involve a minimum number of communication resources.

Another energy efficient approach would be the DENS methodology, which takes the potential communication needs of the components of the data center into consideration along with the load level to minimize the total energy consumption when selecting the best-fit computing resource for job execution. Communicational potential is defined as the amount of end-to-end bandwidth provided to individual servers or group of servers by the data center architecture. Contrary to traditional scheduling solutions that model data centers as a homogeneous pool of computing servers [37], the DENS methodology

develops a hierarchical model consistent with the state of the art of topology of data centers. For a three-tier data center (see Fig. 2), DENS metric M is defined as a weighted combination of server-level (f_s), rack-level (f_r), and module-level (f_m) functions:

$$M = \alpha \cdot f_s + \beta \cdot f_r + \gamma \cdot f_m \quad (5)$$

where α, β, and γ are weighted coefficients that define the impact of the corresponding components (servers, racks, and/or modules) on the metric behavior. Higher α values favor the selection of highly loaded servers in lightly loaded racks. Higher β values will give priority to computationally loaded racks with low network traffic activity. Higher γ values favor the selection of loaded modules.

The selection of computing servers combines the server load $L_s(l)$ and the communication potential $Q_r(q)$ corresponding to the fair share of the uplink resources on the top of the rack ToR switch. This relationship is given as:

$$f_s(l, q) = L_s(l) \cdot \frac{Q_r(q)^{\varphi}}{\delta_r} \quad (6)$$

where $L_s(l)$ is a factor depending on the load of the individual servers l, $Q_r(q)$ defines the load at the rack uplink by analyzing the congestion level in the switch's outgoing queue q, δ_r is a bandwidth over provisioning factor at the rack switch, and φ is a coefficient defining the proportion between $L_s(l)$ and $Q_r(q)$ in the metric. Given that both $L_s(l)$ and $Q_r(q)$ must be within the range $[0, 1]$ higher φ values will decrease the importance of the traffic-related component $Q_r(q)$.

The fact that the energy consumption of an idle server consumes merely two-third of that at peak consumption [48], suggests that an energy-efficient scheduler must consolidate data center jobs on the minimum possible set of computing servers. On the

other hand, keeping servers constantly running at peak loads may decrease hardware reliability and consequently affect job execution deadlines [52]. These issues are addressed with DENS load factor, the sum of two sigmoid functions:

$$L_s(l) = \frac{1}{1 + e^{-10\left(l - \frac{1}{2}\right)}} - \frac{1}{1 + e^{-\frac{10}{\varepsilon}\left(l - \left(1 - \frac{\varepsilon}{2}\right)\right)}}. \quad (7)$$

The first component in Eq. (8) defines the shape of the main sigmoid, while the second serves to encourage convergence towards the maximum server load value (see Fig. 3). The parameter ε defines the size and the inclination of this falling slope and he server load l is within the range $[0, 1]$.

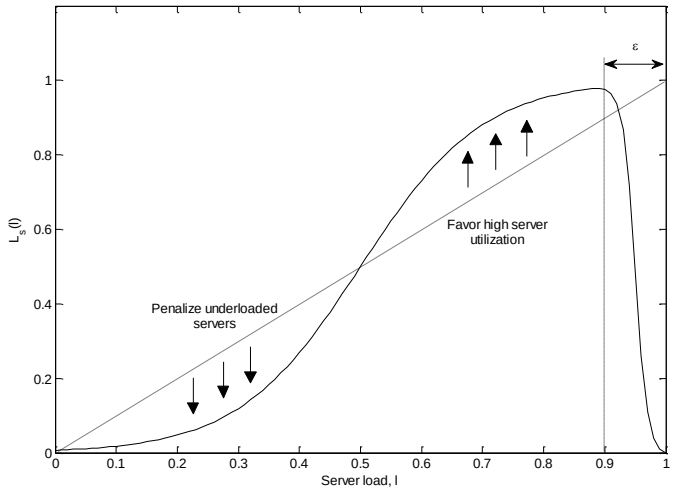

Fig. 3. DENS metric selection of computing server.

Fig. 4 presents the combined server load and queue-size related components. The bell-shaped function obtained favors the selection of servers with a load level above average located in racks with little or no congestion.

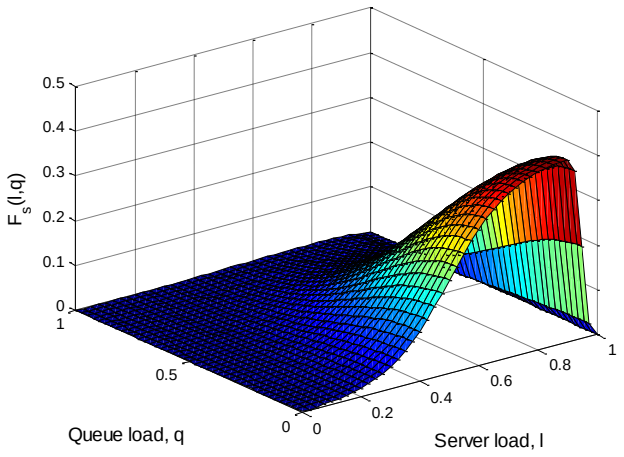

Fig. 4. Server selection according to load and communication potential.

3.2 Load Balancing

Enabling the sleep mode in idle computing servers and network hardware is the most efficient method of avoiding unnecessary power consumption. Consequently, load balancing becomes the key enabler for saving energy.

However, changes in the power mode introduce considerable delays. Moreover, the inability of instantaneous wake up of a sleeping server means that a pool of idle servers must be available to be able to accommodate incoming loads in the short term and prevent QoS degradation. It should be remembered that data centers are required to provide a specific level of quality of service, defined as Service Level Agreements (SLAs), even at peak loads. Therefore, they tend to overprovision computing and communication resources. In fact, on average, datacenters are functioning at only 30% of their capacity. The load in data centers is highly correlated with region and time of the day since more users are active during the daytime hours; the number of users during the day is almost double that at night. Moreover, user arrival rate is not constant, but can spike due to the

crowd effect. Most of the time almost 70% of data center servers, switches, and links remain idle, although during peak periods, this usage can reach 90%. However, idle servers still need to run OS software, maintain virtual machines, and power on both peripheral devices and memory. As a result, even when being idle, servers still consume around two thirds of the peak power consumption. In switches, this ratio is even higher with the energy consumed being shared by the switch chassis, the line cards, and the transceiver ports. Moreover, various Ethernet standards require the uninterrupted transmission of synchronization symbols in the physical layer to guarantee the synchronization required prevents the downscaling of the consumption of energy, even when no user traffic is transmitted.

An energy-efficient scheduler for cloud computing applications with traffic load balancing can be designed to optimize energy consumption of cloud computing data centers, like e-STAB proposed in [47]. One of these is the e-STAB scheduler, which gives equal treatment to communicational demands and computing requirements of jobs. Specifically, e-STAB aims at (a) balancing the communication flows produced by jobs and (b) consolidating jobs using a minimum of computing servers. Since network traffic can be highly dynamic and often difficult to predict [45], the e-STAB scheduler analyzes both load on the network links and occupancy of outgoing queues at the network switches. This queuing analysis helps prevent a buildup of network congestion. This scheduler is already involved in various transport-layer protocols [46] estimating buffer occupancy of the network switches and can react before congestion-related losses occur.

The e-STAB scheduling policy involves the execution of the following two steps for each incoming cloud computing data center job:

Step 1: Select a group of servers S connected to the data center network with the highest available bandwidth, if at least one of the servers in S can accommodate the computational demands of the scheduled job. The available bandwidth is defined as the unused capacity of the link or a set of links connecting the group of servers S to the rest of the data center network.

Step 2: Within the selected group of servers, S, select a computing server with the least available computing capacity, but sufficient to satisfy the computational demands of the scheduled task.

One of the main goals of the e-STAB scheduler is to achieve load balanced network traffic as well as to prevent network congestion. A helpful measure is the available bandwidth per computing node within the data center. However, such a measure does not capture the dynamics of the system, such as sudden increase in the transmission rate of cloud applications.

To provide a more precise measure of network congestion, e-STAB adjusts scales the available bandwidth to the component related to the size of the bottleneck queue (see Fig. 5). This favors empty queues or queues with minimum occupancy and penalizes highly loaded queues that are on the threshold of buffer overflow (or on the threshold of losing packets).

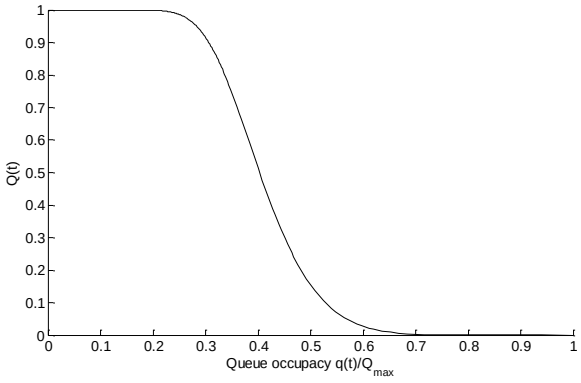

Fig. 5. Queue-size related component of the STAB scheduler.

By utilizing available bandwidth with the component $Q(t)$ metric, the available per-server bandwidth can be computed for modules and individual racks as

$$Fr_j(t) = \frac{1}{T}\int_t^{t+T} \left(\frac{(Cr_j - \lambda r_j(t)) \cdot e^{-(\frac{\rho \cdot qr_j(t)}{Q_{r_j \cdot max}})^{\varphi}}}{Sr_j} \right) dt \qquad (8)$$

where $Qr_j(t)$ is the weight associated with occupancy levels of the queues, $qr_j(t)$ is the size of the queue at time t, and $Qr_j.max$ is the maximum size of the queues allowed at the rack j.

Figure 6 presents the evolution of $Fr_i(t)$ with respect to different values of the network traffic and buffer occupancy. The function is insensitive to the level of utilization of the network links for highly loaded queues, while for lightly loaded queues, the links with the lighter load are preferred to the heavily utilized ones.

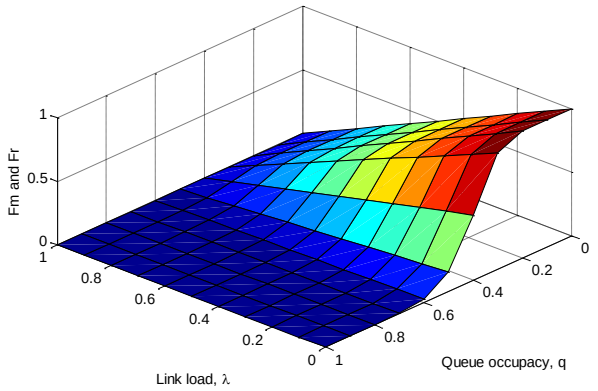

Fig. 6. Selection of racks and modules by the STAB scheduler.

Having selected a proper module and a rack based on their traffic load and congestion state indicated by the queue occupancy, we must select a computing server for the job execution. To do so, we must analyze energy consumption profile of the servers.

Once the energy consumption of a server is known, it is possible to derive a metric to be used by the e-STAB scheduler for server selection, as follows:

$$Fs_k(t) = \frac{1}{T}\int_t^{t+T} \left(\frac{1}{1+e^{-\frac{10}{\varepsilon}\left(l_k(t)-\frac{\varepsilon}{2}\right)}} - \frac{1}{2}\left(1-\frac{P_{idle}}{P_{peak}}\right)\right. \\ \left.\left(1+l_k(t)^3 - e^{-\left(\frac{l_k(t)}{\tau}\right)^3}\right)\right) dt, \tag{9}$$

where $l_k(t)$ is the instantaneous load of server k at time t and T is an averaging interval. While the second summand under the integral in Eq. (9) is a reverse normalized version of Eq. (2), the first summand is a sigmoid designed to penalize selection of idle servers for job execution. The parameter ε corresponds to the CPU load of an idle server required to keep the operating system and virtual machines running. Figure 7 presents a chart for $Fs_k(t)$.

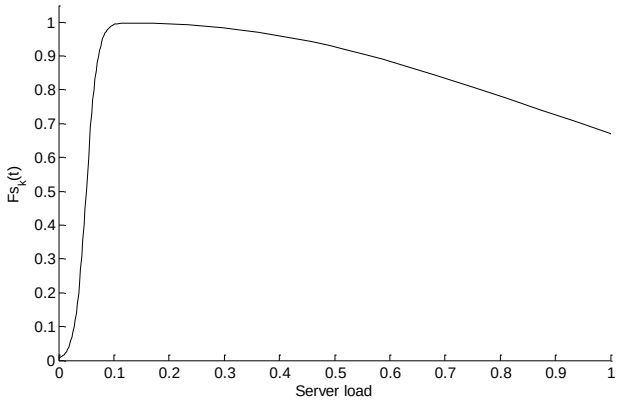

Fig. 7. Selection of computing servers by the STAB scheduler.

3.3 Data Replication

The performance of cloud computing applications, such as gaming, voice and video conferencing, online office, storage, backup, and social networking, depends largely on the availability and efficiency of high-performance communication resources. For better reliability and low latency service provisioning, data resources can be brought closer (replicated) to the physical infrastructure, where the cloud applications are running. A large number of replication strategies for data centers have been proposed in the literature [62]-[66]. These strategies optimize system bandwidth and data availability between geographically distributed data centers. However, none of them focuses on energy efficiency and replication techniques inside data centers.

In [62], an energy efficient data replication scheme have been proposed for datacenter storage. Underutilized storage servers can be turned off to minimize energy consumption, although one of the replica servers must be kept for each data object to guarantee availability. In [63], dynamic data replication in a cluster of data grids is proposed. This approach creates a policy maker, which is responsible for the replica

management. It periodically collects information from the cluster heads, with significance determined by a set of weights selected according to the age of the reading. The policy maker further determines the popularity of a file based on the access frequency. To achieve load balancing, the number of replicas for a file is computed in relation to the access frequency of all other files in the system. This solution follows a centralized design approach, however, leaving it vulnerable to a single point of failure.

Other proposals have concentrated on replication strategies between multiple data centers. In [64], power consumption in the backbone network is minimized by linear programming to determine the optimal points of replication on the basis of data center traffic demands and the popularity of data objects. This linear relation of the traffic load to power consumption at aggregation ports is linear and, consequently, optimization approaches that consider the traffic demand can bring significant power savings.

Another proposal for replication is designed to conserve energy by replicating data closer to consumers to minimize delays. The optimal location for replicas of each data object is determined by periodically processing a log of recent data accesses. The replica site is then determined by employing a weighted *k*-means clustering of user locations and deploying the replica closer to the centroid of each cluster. Migration will take place from one site to another if the gain in quality of service from migration is higher than a predefined threshold.

Another approach is cost-based data replication [66]. This approach analyzes failures in data storage and the probability of data loss probability, which are directly

related to each other, and builds a reliability model. Time points for replica creation are then determined from the data storage reliability function.

The approach presented in [67] is different from all the others replication approaches discussed above due to (a) the scope of the data replication, which is implemented both within a single data center and between geographically distributed data centers, and (b) the optimization target, which takes into account system energy consumption, network bandwidth and communication delay to define the replication strategy to be employed.

Large-scale cloud computing systems are composed of data centers geographically distributed around the globe data centers (see Fig. 8). The central database (Central DB) is located in the wide-area network and hosts all the data required by the cloud applications. To speed up database access and reduce access latency, each data center hosts a local database, called a data center database (Datacenter DB), which is used to replicate the most frequently used data items from the central database. Moreover, each rack hosts at least one server capable of running a local rack-level database (Rack DB), which is used for subsequent replication from the datacenter database.

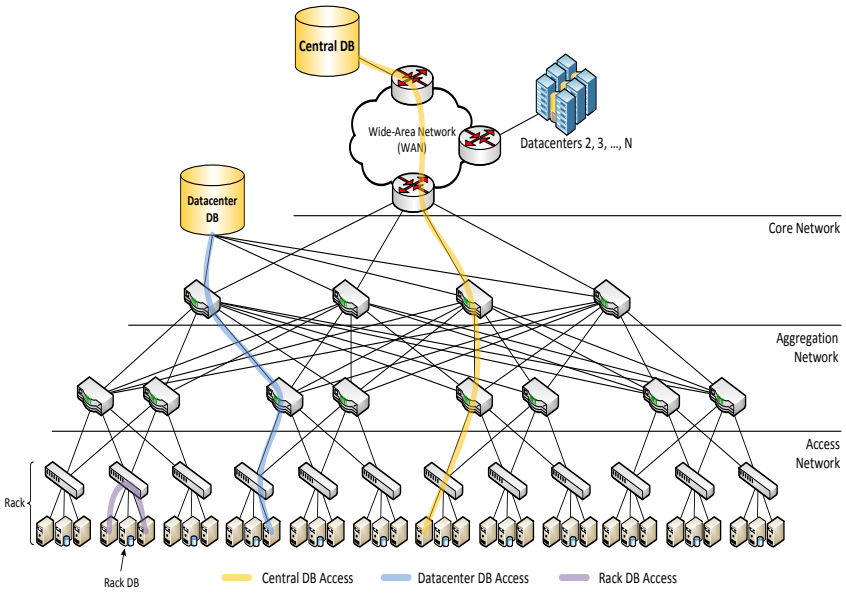

Figure 8. Replication in cloud computing data centers. All database requests produced by the cloud applications running on computing servers are first directed to the rack-level database server. Rack DB either replies with the requested data or forwards the request to the Datacenter DB. In a similar fashion, the Datacenter DB either satisfies the request or forwards it up to the Central DB.

When data is requested, the information about requesting server, rack, and datacenter is stored. Moreover, the statistics showing the number of accesses and updates are maintained for each data item. The access rate (or popularity) is measured as the number of access events per period of time. While accessing data items, cloud applications can also modify them. Such modifications must be sent back to the database so that all replica sites will be updated.

A module located at the central database, the replica manager, periodically analyzes data access statistics to identify what items are the most suitable for replication and at

which replication sites. The availability of these access and update statistics makes it possible to project data center bandwidth usage and energy consumption.

Figure 9 presents the requirements of downlink bandwidth. Since it is proportional to both the size of a data item and the rate of update, the bandwidth consumption grows rapidly and easily overtakes the corresponding capacities of the core, aggregation and access segments of the datacenter network requiring replication.

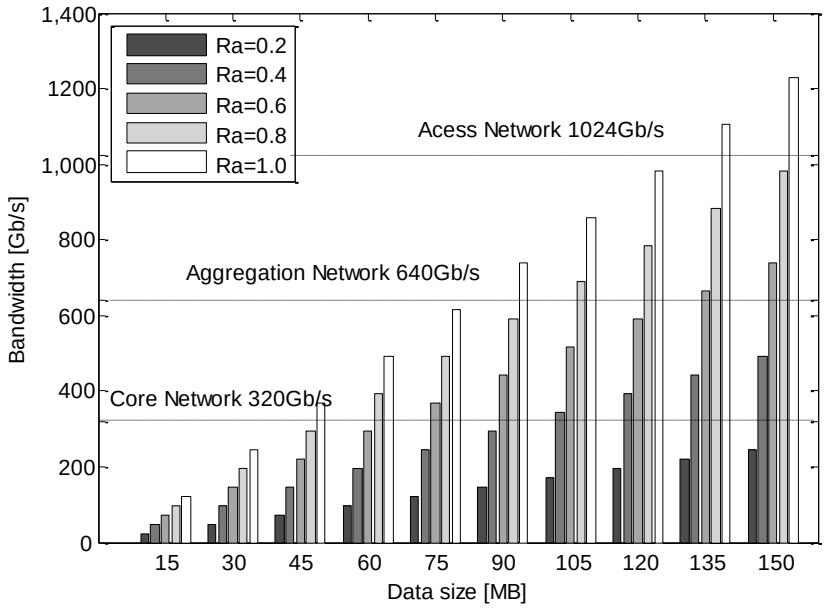

Figure 9. Downlink bandwidth requirements.

Figure 10 reports the tradeoff between datacenter energy consumption, including the consumption of both the servers and network switches, and the downlink residual bandwidth. For all replication scenarios, the core layer reaches saturation firs since it is the smallest of the datacenter network segments and has capacity of only 320 GB/s. The residual bandwidth for all network segments generally decreases with increase in load, except for the gateway link, for which the available bandwidth remains constant for both

Datacenter DB and Rack DB replication scenarios, since data queries are processed at the replica databases and only data updates are routed from the Central DB to the Datacenter DB. The benefit of Rack DB replication is two-fold: on one hand network, traffic can be restricted to the access network, which has lower nominal power consumption and higher network capacity, while on the other, data access becomes localized, thus improving performance of cloud applications.

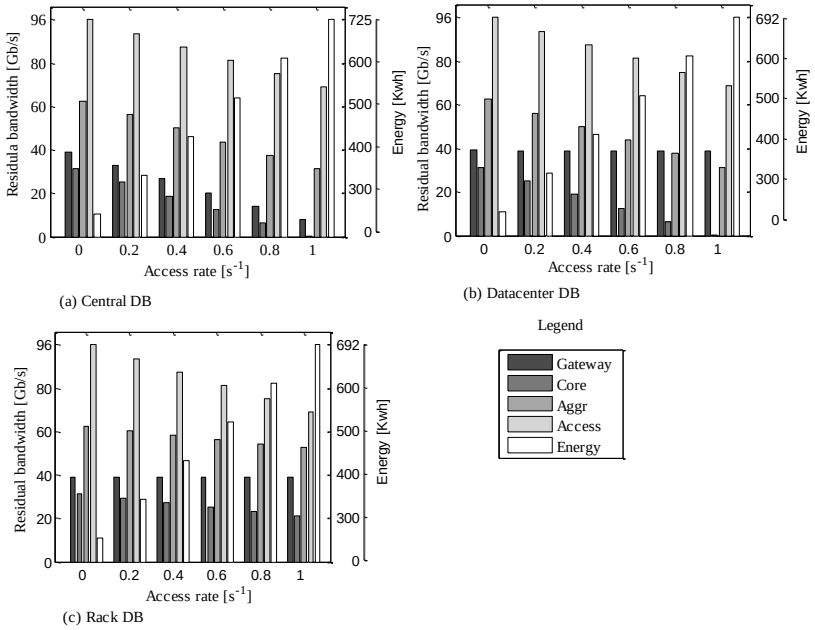

Figure 10. Energy and residual bandwidth for (a) Central DB, (b) Datacenter DB, and (c) Rack DB replication scenarios.

3.4 Placement of Virtual Machines

Virtualization represents a key technology for efficient operation of cloud data centers. Energy consumption n virtualized data centers can be reduced by appropriate decision on which physical server a virtual machines should be placed. Virtual machine

consolidation strategies try to use the lowest possible number of physical machines to host a certain number of virtual machines. Some proposed strategies are described next.

In [50], the authors developed a strategy for traditional three-tier data center architectures which takes into consideration the energy consumption of both servers and network switches. The proposed strategy analyzes the load of each network switch to avoid overloading them. It tries to compromise load balancing of data center network traffic and consolidation of virtual machines. Such compromise is important to the operation of data centers running jobs that impose low computational load but produce heavy traffic streams.

The problem of virtual machine placement has been addressed by different formulations of the bin-packing problem. The proposal in [38] employs a variation of the best fit decreasing algorithm. Although, in this case, only the energy consumption of servers is considered, results showed potential energy savings without a significant number of violation of service level agreements. In [70], a heuristic is proposed to achieve server utilization close to an optimal level determined by the computation of the Euclidean distance of the allocation state. A first fit decreasing strategy was employed in [71] for data centers processing web search and MapReduce applications. The consolidation approach is based on the analysis of CPU usage, and favors the placement of correlated virtual machines in distinct physical servers, to avoid overloading the servers.

The formulation of virtual machine problem presented in [69] includes active cooling control besides the traditional approaches such as DPM and DVFS. This work also does not take into account the contribution of network switches to the energy

consumption of a data center and it shown that active cooling control result in small, but relevant, gains.

The work in [72] promotes energy reduction by consolidating network flows instead of virtual machines; only the consumption of network switches are considered. Correlated flows are analyzed and assigned to network paths in a greedy way. This approach employs link rate adaptation and shutting down of switches with low utilization. Results derived using simulations based on real traces of Wikipedia traffic demonstrated that this approach can in fact reduce energy consumption.

3.5 Communications Infrastructure

The energy efficiency of a data center also depends on the underlying communication infrastructure. Indeed, at the average load level of a data center, the communication network consumes between 30% and 50% of the total power used by the IT equipment; this in turn represents roughly 40% of the total energy budget.

Moreover, an analysis of the distribution of data traffic in clouds suggests that the majority of the traffic is transferred within the data center itself (around 75%), with rest being split between communication with users (18%) and data center to data center exchanges (7%) [68].

Based on these facts, it is clear the need to develop energy efficient solutions for communication technologies and architectures to interconnect the servers in data centers. Since high-speed and high capacity are required, the most suitable communication technology for cloud data centers is optical. In the remainder of this section, some possible

architectures addressing energy efficient solutions for internal communications in data centers are presented.

Optical interconnection networks are a novel alternative technology to provide high bandwidth, low latency and reduced power consumption. Up until recently, such optical technology has been used only for point-to-point links to connect the electrical switches (fiber optics) thus reducing noise and leaving smaller footprints. However, since the switches operate in the electrical domain, power hungry electrical-to-optical (E/O) and optical-to-electrical (O/E) transceivers are required.

New modules connecting the silicon chip directly with optical fibers have been developed, thus enabling switching to be performed in the optical domain.

Optical interconnections can be based on circuit switching or packet switching, each generating different trade-off in terms of energy vs performance. Solely in terms of energy efficiency, optical circuit switching represents the most efficient solution, but it leads to high reconfiguration times due to the nature of circuit switching. On the other side, packet switching, although less energy efficient, potentially the source of greater latency, achieves better performance, since its reconfiguration time is lower and its scalability higher.

One recent alternative is the usage of optical OFDM. Optical OFDM distributes the data on a large number of low data rate subcarriers and can thus provide fine-granularity capacity to connections by the elastic allocation of subcarriers according to connection demands.

The use of optical OFDM as a bandwidth-variable and highly spectrum-efficient modulation format can provide scalable and flexible sub- and super-wavelength granularity, compared to the conventional, fixed-bandwidth fixed-grid WDM network. However, this new concept poses new challenges for the routing and wavelength assignment algorithms. Indeed, traditional algorithms for routing and wavelength assignment will no longer be directly applicable for such new kinds of communication infrastructure.

4. Conclusions and Open Challenges

Costs and operating expenses have become a growing concern in the cloud computing industry, with energy consumption accounting for a large percentage of the operational expenses in the data centers used as backend computing infrastructure. This chapter emphasizes the role of communications and network awareness of this consumption and presents suggested solutions for energy efficient resource allocation in clouds.

The challenge of energy efficiency will largely determine the future of cloud computing systems, at present experiencing unprecedented growth. Most of the existing energy-efficient and performance optimization solutions in the IT domain focus on computing, with communications-related processes relegated to a secondary role or unaccounted for. In reality, however, communications are at the heart of cloud systems, and network characteristics, such as bandwidth capacity, transmission delay, delay jitter, buffering, loss rate and performance of communication protocols, often determine the quality of task execution. However, most current research is restricted to processes inside

data centers, yet the models must also account for communication dynamics in the wide-area network, and at the user end.

Open research challenges are essentially related to improving the energy scalability of cloud computing. The previous sections have underlined the need for the joint optimization of computing and communication while maintaining an appropriate balance between performance and energy consumption for the overall architecture.

The following specific research challenges have been identified:

- Integration of novel and more efficient energy consumption models for the different components of the cloud computing architecture. As the concept of energy-proportional computing is emerging in the design of computing hardware and software infrastructures, it is also becoming relevant in the design of communication equipment. These emerging models will drive the need for improved and innovative approaches for the joint optimization and balancing of performance and energy consumption in cloud computing.
- The concept of Mobile Cloud, deriving from the clear trend towards user mobility (and the "always on" paradigm) and the availability of ever more powerful devices in the hands of the cloud services' users is shaping the possibility of even more pervasive usage of the cloud computing infrastructure. Users' request for 24/7 availability of cloud services even in sparsely "covered" areas, will lead to a redefinition or least an evolution, of the cloud architecture, which will involve the need for efficient dissemination of both information and services across the Internet, whether

in data centers, on users´ devices, or somewhere in between. This is sure to have an impact on the way data is replicated and services are provided.

References

[1] Internet World Statistics, available at http://www.internetworldstats.com.
[2] "Forecast: Mobile Data Traffic and Revenue, Worldwide, 2010-2015," Market report, Gartner Inc., 2011.
[3] Google Glass project, available at https://plus.google.com/ 111626127367496192147.
[4] A. Weiss, "Computing in the clouds," netWorker, vol. 11, no. 4, pp. 16–25, 2007.
[5] "State of the Data Center 2011," Emerson Network Power, 2011.
[6] X. Fan, W.-D. Weber, and L. A. Barroso, "Power Provisioning for a Warehouse-sized Computer," In Proceedings of the ACM International Symposium on Computer Architecture, San Diego, CA, June 2007.
[7] R. Raghavendra, P. Ranganathan, V. Talwar, Z. Wang, and X. Zhu, "No "Power" Struggles: Coordinated Multi-level Power Management for the Data Center," APLOS 2008.
[8] Interactive Data Corporation, available at: http://www.interactivedata.com/.
[9] Gartner Group, available at: http://www.gartner.com/
[10] A. Vasan and A. Sivasubramaniam "Worth their watts?-an empirical study of datacenter servers," IEEE 16th International Symposium on High Performance Computer Architecture (HPCA), 2010.
[11] IDC, "Worldwide Server Power and Cooling Expense 2006-2010," Market Analysis, 2006.
[12] J. G. Koomey, "Growth in Data center electricity use 2005 to 2010," Analytics Press, 2011.
[13] "Reducing Data Center Cost with an Air Economizer," Intel IT@Intel Brief, August 2008.
[14] N. Rasmussen, "Calculating Total Cooling Requirements for Data Centers," White Paper #25, American Power Conversion, 2007.
[15] The Green Grid Consortium, available at http://www.thegreengrid.org/.
[16] "Green Grid Data Center Power Efficiency Metrics: PUE and DCIE," White paper #6, The Grid Grid, 2008.
[17] R. Brown et al. "Report to congress on server and data center energy efficiency: public law 109-431," Lawrence Berkeley National Laboratory, Berkeley, 2008.
[18] J. Liu, F. Zhao, X. Liu, and W. He, "Challenges Towards Elastic Power Management in Internet Data Centers", in Proceedings of the 2nd International Workshop on Cyber-Physical Systems (WCPS 2009), in conjunction with ICDCS 2009., Montreal, Quebec, Canada, June 2009.
[19] P. Mahadevan, P. Sharma, S. Banerjee, and P. Ranganathan, "A Power Benchmarking Framework for Network Devices," IFIP Networking, May 2009.
[20] B. Heller, S. Seetharaman, P. Mahadevan, Y. Yiakoumis, P. Sharma, S. Banerjee and N. McKeown, "ElasticTree: saving energy in data center networks," 7th USENIX conference on Networked systems design and implementation (NSDI). USENIX Association, Berkeley, CA, USA, 17-17, 2010.
[21] L. Shang, L.-S. Peh and K. N. Jha, "Dynamic voltage scaling with links for power optimisation of interconnection networks," International Symposium on High Performance Computer Architecture, 2003.
[22] Dennis Abts, Mike Marty, Philip Wells, Peter Klausler, Hong Liu, "Energy Proportional Datacenter Networks," Proceedings of the International Symposium on Computer Architecture, pp. 338 – 347, 2010.
[23] L. Benini and G. D. Micheli, "Powering networks on chips: Energy-efficient and reliable interconnect design for SoCs," International Symposium on Systems Synthesis, Oct. 2001, pp. 33–38.
[24] M. E. Crovella and R. L. Carter, "Dynamic Server Selection In The Internet," Third IEEE Workshop on the Architecture and Implementation of High Performance Communication Subsystems (HPCS), pp.158-162, 1995.
[25] J. S. Kim, M. B. Taylor, J. Miller, and D. Wentzlaff, "Energy characterization of a tiled architecture processor with on-chip networks," International Symposium on Low Power Electronics and Design, Aug. 2003, pp. 424–427.
[26] V. Soteriou and L.-S. Peh, "Design-Space exploration of power-aware on/off interconnection networks," IEEE International Conference on Computer Design, 2004, pp. 510-517.
[27] J. Duato, "A theory of fault-tolerant routing in wormhole networks," IEEE Transactions on Parallel and Distributed Systems, vol. 8, no. 8, pp. 790–802, Aug. 1997.

[28] G. Wei, J. Kim, D. Liu, S. Sidiropoulos, and M. Horowitz, "A variable frequency parallel I/O interface with adaptive power-supply regulation," Journal of Solid-State Circuits, vol. 35, no. 11, pp. 1600–1610, Nov. 2000.

[29] J. Pouwelse, K. Langendoen, and H. Sips, "Energy priority scheduling for variable voltage processors," International Symposium on Low Power and Design, 2001, pp. 28-33.

[30] S. K. Garg, Chee Shin Yeo, A. Anandasivam and R. Buyya, "Energy-Efficient Scheduling of HPC Applications in Cloud Computing Environments," CoRR, abs/0909.1146, 2009.

[31] B. Huffaker, D. Plummer and D. Moore and K. Claffy, "Topology discovery by active probing," Symposium on Applications and the Internet (SAINT), pp. 90-96, 2002.

[32] Xuan Chen, Ling Xing and Qiang Ma, "A distributed measurement method and analysis on Internet hop counts," 2011 International Conference on Computer Science and Network Technology (ICCSNT), pp.1732 - 1735, Dec. 2011.

[33] B. Nordman, "What the real world tells us about saving energy in electronics," Proceedings of 1st Berkeley symposium on energy efficient electronic systems (E3S), May 2009.

[34] IEEE Std 802.3az-2010, "Media Access Control Parameters, Physical Layers, and Management Parameters for Energy-Efficient Ethernet," pp.1-302, Oct. 27 2010.

[35] K. Christensen, P. Reviriego, B. Nordman, M. Bennett, M. Mostowfi and J. A. Maestro, "IEEE 802.3az: the road to energy efficient ethernet," IEEE Communications Magazine, vol.48, no.11, pp.50-56, November 2010.

[36] G. Y. Li, Zhikun Xu, Cong Xiong, Chenyang Yang, Shunqing Zhang, Yan Chen and Shugong Xu, "Energy-efficient wireless communications: tutorial, survey, and open issues," IEEE Wireless Communications, vol.18, no.6, pp.28-35, December 2011.

[37] Ying Song, Hui Wang, Yaqiong Li, Binquan Feng, and Yuzhong Sun, "Multi-Tiered On-Demand Resource Scheduling for VM-Based Data Center," IEEE/ACM International Symposium on Cluster Computing and the Grid (CCGRID), pp. 148 – 155, May 2009.

[38] A. Beloglazov, and R. Buyya, "Energy Efficient Resource Management in Virtualized Cloud Data Centers," IEEE/ACM International Conference on Cluster, Cloud and Grid Computing (CCGrid), pp. 826 – 831, May 2010.

[39] Qin Tang, S. K. S. Gupta, and G. Varsamopoulos, "Energy-Efficient Thermal-Aware Task Scheduling for Homogeneous High-Performance Computing Data Centers: A Cyber-Physical Approach," IEEE Transactions on Parallel and Distributed Systems, vol.19, no. 11, pp. 1458 – 1472, November 2008.

[40] M. Al-Fares, S. Radhakrishnan, B. Raghavan, N. Huang, and A. Vahdat, "Hedera: Dynamic Flow Scheduling for Data Center Networks," in Proceedings of the 7th USENIX Symposium on Networked Systems Design and Implementation (NSDI '10), San Jose, CA, April 2010.

[41] A. Stage and T. Setzer, "Network-aware migration control and scheduling of differentiated virtual machine workloads," in Proceedings of the 2009 ICSE Workshop on Software Engineering Challenges of Cloud Computing, International Conference on Software Engineering. IEEE Computer Society, Washington, DC, May 2009.

[42] Xiaoqiao Meng, V. Pappas, and Li Zhang, "Improving the Scalability of Data Center Networks with Traffic-aware Virtual Machine Placement," IEEE INFOCOM, San Diego, California, March 2010.

[43] "Cisco Data Center Infrastructure 2.5 Design Guide," Cisco press, March 2010.

[44] IEEE 802.1 Data Center Bridging Task Group, available at: http://www.ieee802.org/1/pages/dcbridges.html.

[45] Aimin Sang and San-qi Li, "A predictability analysis of network traffic," Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM), vol. 1, pp.342-351, 2000.

[46] C. Barakat, E. Altman and W. Dabbous, "On TCP performance in a heterogeneous network: a survey," IEEE Communications Magazine, vol. 38, no. 1, pp. 40-46, Jan 2000.

[47] D. Kliazovich, S. T. Arzo, F. Granelli, P. Bouvry, and S. U. Khan, "e-STAB: Energy-Efficient Scheduling for Cloud Computing Applications with Traffic Load Balancing," IEEE International Conference on Green Computing and Communications (GreenCom), Beijing, China, pp. 7-13, 2013.

[48] G. Chen, W. He, J. Liu, S. Nath, L. Rigas, L. Xiao, and F. Zhao, "Energy-aware server provisioning and load dispatching for connection-intensive internet services," the 5th USENIX Symposium on Networked Systems Design and Implementation, Berkeley, CA, USA, 2008.

[49] Server Power and Performance characteristics, available on: http://www.spec.org/power_ssj2008/

[50] D. Kliazovich, P. Bouvry, and Samee U. Khan, "DENS: Data Center Energy-Efficient Network-Aware Scheduling," Cluster Computing, vol. 16, no. 1, pp. 65-75, 2013.

[51] D. Kliazovich, S. T. Arzo, F. Granelli, P. Bouvry, and S. U. Khan, "Accounting for Load Variation in Energy-Efficient Data Centers," IEEE International Conference on Communications (ICC), Budapest, Hungary, 2013.

[52] C. Kopparapu. "Load Balancing Servers, Firewalls, and Caches," John Wisely & Sons Inc., 2002.

[53] C. Fiandrino, D. Kliazovich, P. Bouvry, A. Y. Zomaya, "Performance and Energy Efficiency Metrics for Communication Systems of Cloud Computing Data Centers," submitted for publication to the IEEE Transactions on Cloud Computing, 2014.

[54] M. A. Sharkh, M. Jammal, A. Shami, A. Ouda, "Resource allocation in a network-based cloud computing environment: design challenges," IEEE Communications Magazine, vol. 51, no. 11, pp. 46-52, November 2013.

[55] X. Leon, L. Navarro, "Limits of energy saving for the allocation of data center resources to networked applications," IEEE INFOCOM, pp. 216-220, 10-15 April 2011.

[56] B. Guenter, N. Jain, C. Williams, "Managing cost, performance, and reliability tradeoffs for energy-aware server provisioning," IEEE INFOCOM, pp. 1332-1340, 10-15 April 2011.

[57] J. Doyle, R. Shorten, O'Mahony, D., "Stratus: Load Balancing the Cloud for Carbon Emissions Control," Cloud Computing, IEEE Transactions on , vol.1, no.1, pp.1,1, Jan.-June 2013.

[58] J. Doyle, R. Shorten, D. O'Mahony, "Stratus: Load Balancing the Cloud for Carbon Emissions Control," IEEE Transactions on Cloud Computing, vol. 1, no. 1, pp. 1-1, June 2013.

[59] Li Hongyou, Wang Jiangyong, Peng Jian, Wang Junfeng, Liu Tang, "Energy-aware scheduling scheme using workload-aware consolidation technique in cloud data centres," China Communications, vol. 10, no. 12, pp. 114-124, Dec. 2013.

[60] L. Benini, A. Bogliolo, and G. De Micheli, "A survey of design techniques for system-level dynamic power management," IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 8, no. 3, pp. 299 – 316, June 2000.

[61] Y. Audzevich, P. Watts, S. Kilmurray, and A. W. Moore, "Efficient photonic coding: a considered revision," SIGCOMM workshop on Green networking (GreenNets '11), ACM, New York, NY, USA, 2011.

[62] Bin Lin, Shanshan Li, Xiangke Liao, Qingbo Wu, and Shazhou Yang, "eStor: Energy efficient and resilient data center storage," 2011 International Conference on Cloud and Service Computing (CSC), pp. 366-371, December 2011.

[63] Ruay-Shiung Chang, Hui-Ping Chang, and Yun-Ting Wang, "A dynamic weighted data replication strategy in data grids," IEEE/ACS International Conference on Computer Systems and Applications, pp. 414-421, April 2008.

[64] Xiaowen Dong, T. El-Gorashi, and J. M. H. Elmirghani, "Green IP Over WDM Networks With Data Centers," Journal of Lightwave Technology, vol. 29, no. 12, pp. 1861-1880, June 2011.

[65] Fan Ping, Xiaohu Li, C. McConnell, R. Vabbalareddy, Jeong-Hyon Hwang, "Towards Optimal Data Replication Across Data Centers," International Conference on Distributed Computing Systems Workshops (ICDCSW), pp. 66-71, June 2011.

[66] Wenhao Li, Yun Yang, and Dong Yuan, "A Novel Cost-Effective Dynamic Data Replication Strategy for Reliability in Cloud Data Centres," International Conference on Dependable, Autonomic and Secure Computing (DASC), pp. 496-502, December 2011.

[67] D. Boru, D. Kliazovich, F. Granelli, P. Bouvry, A.Y. Zomaya, "Energy-Efficient Data Replication in Cloud Computing Datacenters," Springer Cluster Computing, 2014.

[68] Cisco, 2011, *Global cloud index: Forecast and methodology, 2011–2016* (Whitepaper). Cisco.

[69] D. G. d. Lago, E. R. M. Madeira, and L. F. Bittencourt, "Power-aware virtual machine scheduling on clouds using active cooling control and dvfs," in Proceedings of the 9th International Workshop on Middleware for Grids, Clouds and e-Science, ser. MGC'11. New York, NY, USA: ACM, 2011, pp. 2:1–2:6.

[70] S. Srikantaiah, A. Kansal, and F. Zhao, "Energy aware consolidation for cloud computing," in Proceedings of the 2008 conference on Power aware computing and systems, ser. HotPower'08. Berkeley, CA, USA: USENIX Association, 2008, pp. 10–10.

[71] J. Kim, M. Ruggiero, D. Atienza, and M. Lederberger, "Correlation-aware virtual machine allocation for energy-efficient datacenters," in Design, Automation Test in Europe Conference Exhibition (DATE), 2013, 2013, pp. 1345–1350.

[72] X. Wang, Y. Yao, X. Wang, K. Lu, and Q. Cao, "Carpo: Correlation-aware poweroptimization in data center networks," in INFOCOM, 2012 Proceedings IEEE, 2012, pp. 1125–1133.