J Supercomput (2012) 62:1263–1283
DOI 10.1007/s11227-010-0504-1

GreenCloud: a packet-level simulator of energy-aware cloud computing data centers

Dzmitry Kliazovich · Pascal Bouvry · Samee Ullah Khan

Published online: 9 November 2010

Abstract Cloud computing data centers are becoming increasingly popular for the provisioning of computing resources. The cost and operating expenses of data centers have skyrocketed with the increase in computing capacity. Several governmental, industrial, and academic surveys indicate that the energy utilized by computing and communication units within a data center contributes to a considerable slice of the data center operational costs.

In this paper, we present a simulation environment for energy-aware cloud computing data centers. Along with the workload distribution, the simulator is designed to capture details of the energy consumed by data center components (servers, switches, and links) as well as packet-level communication patterns in realistic setups.

The simulation results obtained for two-tier, three-tier, and three-tier high-speed data center architectures demonstrate the effectiveness of the simulator in utilizing power management schema, such as voltage scaling, frequency scaling, and dynamic shutdown that are applied to the computing and networking components.

Keywords Energy efficiency · Next generation networks · Cloud computing simulations · Data centers

D. Kliazovich (✉) · P. Bouvry
University of Luxembourg, 6 rue Coudenhove Kalergi, Luxembourg, Luxembourg
e-mail: dzmitry.kliazovich@uni.lu

P. Bouvry
e-mail: pascal.bouvry@uni.lu

S.U. Khan
North Dakota State University, Fargo, ND 58108-6050, USA
e-mail: samee.khan@ndsu.edu

1 Introduction

Over the last few years, cloud computing services have become increasingly popular due to the evolving data centers and parallel computing paradigms. The notion of a cloud is typically defined as a pool of computer resources organized to provide a computing function as a utility. The major IT companies, such as Microsoft, Google, Amazon, and IBM, pioneered the field of cloud computing and keep increasing their offerings in data distribution and computational hosting [28].

The operation of large geographically distributed data centers requires considerable amount of energy that accounts for a large slice of the total operational costs for cloud data centers [6, 25]. Gartner group estimates energy consumptions to account for up to 10% of the current data center operational expenses (OPEX), and this estimate may rise to 50% in the next few years [10]. However, computing based energy consumption is not the only power-related portion of the OPEX bill. High power consumption generates heat and requires an accompanying cooling system that costs in a range of $2 to $5 million per year for classical data centers [23].

Failure to keep data center temperatures within operational ranges drastically decreases hardware reliability and may potentially violate the Service Level Agreement (SLA) with the customers. A major portion (over 70%) of the heat is generated by the data center infrastructure [26]. Therefore, optimized infrastructure installation may play a significant role in the OPEX reduction.

From the energy efficiency perspective, a cloud computing data center can be defined as *a pool of computing and communication resources organized in the way to transform the received power into computing or data transfer work to satisfy user demands*. The first power saving solutions focused on making the data center hardware components power efficient. Technologies, such as Dynamic Voltage and Frequency Scaling (DVFS), and Dynamic Power Management (DPM) [14] were extensively studied and widely deployed. Because the aforementioned techniques rely on power-down and power-off methodologies, the efficiency of these techniques is at best limited. In fact, an idle server may consume about $2/3$ of the peak load [3].

Because the workload of a data center fluctuates on the weekly (and in some cases on hourly basis), it is a common practice to overprovision computing and communicational resources to accommodate the peak (or expected maximum) load. In fact, the average load accounts only for 30% of data center resources [20]. This allows putting the rest of the 70% of the resources into a sleep mode for most of the time. However, achieving the above requires central coordination and energy-aware workload scheduling techniques. Typical energy-aware scheduling solutions attempt to: (a) concentrate the workload in a minimum set of the computing resources and (b) maximize the amount of resource that can be put into sleep mode [18].

Most of the current state-of-the-art research on energy efficiency has predominantly focused on the optimization of the processing elements. However, as recorded in earlier research, more than 30% of the total computing energy is consumed by the communication links, switching and aggregation elements. Similar to the case of processing components, energy consumption of the communication fabric can be reduced by scaling down the communication speeds and cutting operational frequency along with the input voltage for the transceivers and switching elements [29]. However, slowing the communicational fabric down should be performed carefully and

based on the demands of user applications. Otherwise, such a procedure may result in a bottleneck, thereby limiting the overall system performance.

A number of studies demonstrate that often a simple optimization of the data center architecture and energy-aware scheduling of the workloads may lead to significant energy savings. The authors of [21] demonstrate energy savings of up to 75% that can be achieved by traffic management and workload consolidation techniques.

This article presents a simulation environment, termed GreenCloud, for advanced energy-aware studies of cloud computing data centers in realistic setups. GreenCloud is developed as an extension of a packet-level network simulator Ns2 [31]. Unlike few existing cloud computing simulators such as CloudSim [1] or MDCSim [19], GreenCloud extracts, aggregates, and makes information about the energy consumed by computing and communication elements of the data center available in an unprecedented fashion. In particular, a special focus is devoted to accurately capture communication patterns of currently deployed and future data center architectures. The GreenCloud simulator is currently available upon request sent to the authors.

The rest of the paper is organized as follows. Section 2 surveys the most demanded data center architectures outlining the reasons for their choice through the analysis of their physical components; Sect. 3 presents the main simulator components and related energy models; Sect. 4 focuses on the thorough evaluation of the developed simulation environment; Sect. 5 concludes the paper providing the guidelines for building energy-efficient data centers and outlining directions for future work on the topic.

2 Data center architectures

The pool of servers in today's data centers overcomes 100,000 hosts with around 70% of all communications performed internally [21]. This creates a challenge in the design of interconnected network architecture and the set of communication protocols.

Given the scale of a data center, the conventional hierarchical network infrastructure often becomes a bottleneck due to the physical and cost-driven limitations of the used networking equipment. Specifically, the availability of 10 Gigabit Ethernet (GE) components and their price defined the way the data center architectures evolved. The 10 GE transceivers are still too expensive and probably offer more capacity than needed for connecting individual servers. However, their penetration level keeps increasing in the backbone networks, metro area networks, and data centers.

Two-tier data center architectures follow the structure depicted in Fig. 1. In this example, computing Servers (S) physically arranged into racks form the tier-one network. At the tier-two network, Layer-3 (L3) switches provide full mesh connectivity using 10 GE links.

The Equal Cost Multi-Path (ECMP) routing [30] is used as a load balancing technology to optimize data flows across multiple paths. It applies load balancing on TCP and UDP packets on a per-flow basis using express hashing techniques requiring almost no processing from a switch's CPU. Other traffic, such as ICMP [24], is typically not processed by ECMP and forwarded on a single predefined path.

The two-tier architecture worked well for early data centers with a limited number of computing servers. Depending on the type of switches used in the access network,

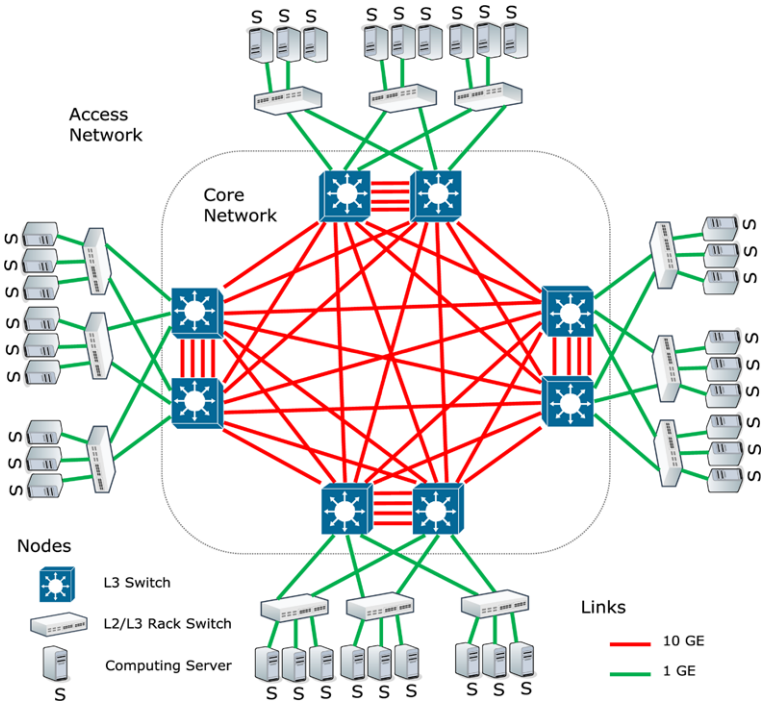

Fig. 1 Two-tier data center architecture

the two-tier data centers may support up to 5500 nodes [4]. The number of core switches and capacity of the core links defines the maximum network bandwidth allocated per computing server.

Three-tier data center architectures are the most common nowadays. They include: (a) access, (b) aggregation, and (c) core layers as presented in Fig. 2. The availability of the aggregation layer facilitates the increase in the number of server nodes (to over 10,000 servers) while keeping inexpensive Layer-2 (L2) switches in the access network, which provides a loop-free topology.

Because the maximum number of ECMP paths allowed is eight, a typical three-tier architecture consists of eight core switches (only four are presented in Fig. 2). Such architecture implements an 8-way ECMP that includes 10 GE Line Aggregation Groups (LAGs) [15], which allow a network client to address several links and network ports with a single MAC address.

While the LAG technology is an excellent methodology to increase link capacities, its usage has several fundamental drawbacks that limit network flexibility and performance. LAGs make it difficult to plan the capacity for large flows and make it unpredictable in case of a link failure. In addition, several types of traffic patterns, such as ICMP and broadcast are usually routed through a single link only. Moreover, full mesh connectivity at the core of the network requires considerable amount of cablings.

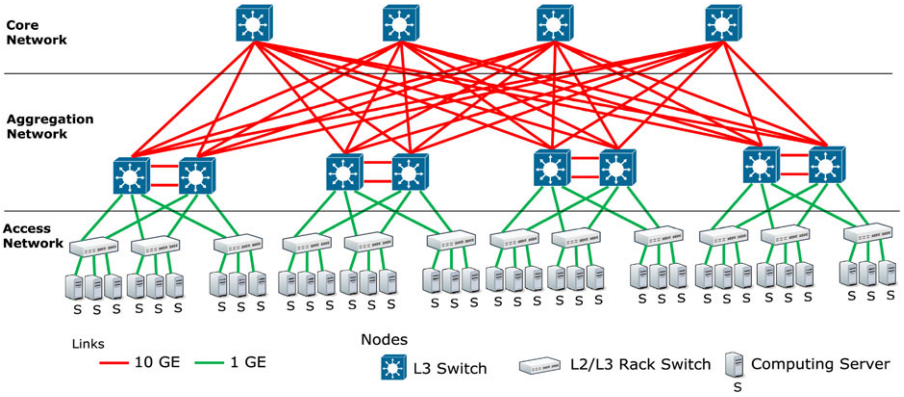

Fig. 2 Three-tier data center architecture

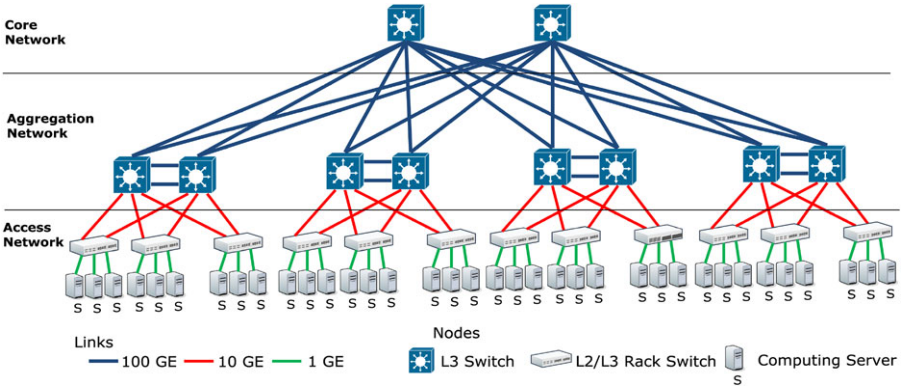

Fig. 3 Three-tier high-speed data center architecture

The aforementioned disadvantages have redirected the design choices for the next generation data centers to consider: (a) increasing the capacity of the core and (b) accessing parts of the network with beyond 10 GE links.

Three-tier high-speed data center architectures are designed to optimize the number of nodes, capacity of core, and aggregation networks that are currently a bottleneck, which limit the maximum number of nodes in a data center or a per-node bandwidth (see Fig. 3).

With the availability of 100 GE links (IEEE 802.3ba), standardized in June 2010 [16], between the core and aggregation switches, reduces the number of the core switches, avoids the shortcomings of LAG technology, reduces cablings, and considerably increases the maximum size of the data center due to physical limitations [9]. Fewer ECMP paths will lead to the flexibility and increased network performance.

While the fat-tree topology is the most widely used in modern data centers other more advanced architectures have been proposed. For example, architectures such as DCell [12] or BCube [13] implement server centric approach relying on mini-switches for interconnection. Both architectures do not rely on the core or aggregation

layers and offer scalability to millions of servers. The routing is performed by the servers themselves, requiring a specific routing protocol to ensure fault tolerance. However, due to the fact that both architectures are only recent research proposals which have not been tested in real data centers and unveil their advantages in very large data centers, we leave their performance evaluation out of the scope of this paper, focusing on more widely used architectures.

3 Simulation of energy-efficient data center

3.1 Energy efficiency

Only a part of the energy consumed by the data center gets delivered to the computing servers directly. A major portion of the energy is utilized to maintain interconnection links and network equipment operations. The rest of the electricity is wasted in the power distribution system, dissipates as heat energy, and used up by air-conditioning systems. In light of the above discussion, in GreenCloud, we distinguish three energy consumption components: (a) computing energy, (b) communicational energy, and (c) the energy component related to the physical infrastructure of a data center.

The efficiency of a data center can be defined in terms of the performance delivered per watt, which may be quantified by the following two metrics: (a) Power Usage Effectiveness (PUE) and (b) Data Center Infrastructure Efficiency (DCiE) [27]. Both PUE and DCiE describe which portion of the totally consumed energy gets delivered to the computing servers.

3.2 Structure of the simulator

GreenCloud is an extension to the network simulator Ns2 [31] which we developed for the study of cloud computing environments. The GreenCloud offers users a detailed fine-grained modeling of the energy consumed by the elements of the data center, such as servers, switches, and links. Moreover, GreenCloud offers a thorough investigation of workload distributions. Furthermore, a specific focus is devoted on the packet-level simulations of communications in the data center infrastructure, which provide the finest-grain control and is not present in any cloud computing simulation environment.

Figure 4 presents the structure of the GreenCloud extension mapped onto the three-tier data center architecture.

Servers (S) are the staple of a data center that are responsible for task execution. In GreenCloud, the server components implement single core nodes that have a preset on a processing power limit in MIPS (million instructions per second) or FLOPS (floating point operations per second), associated size of the memory/storage resources, and contain different task scheduling mechanisms ranging from the simple round-robin to the sophisticated DVFS- and DNS-enabled.

The servers are arranged into racks with a Top-of-Rack (ToR) switch connecting it to the access part of the network. The power model followed by server components is dependent on the server state and its CPU utilization. As reported in [3, 8],

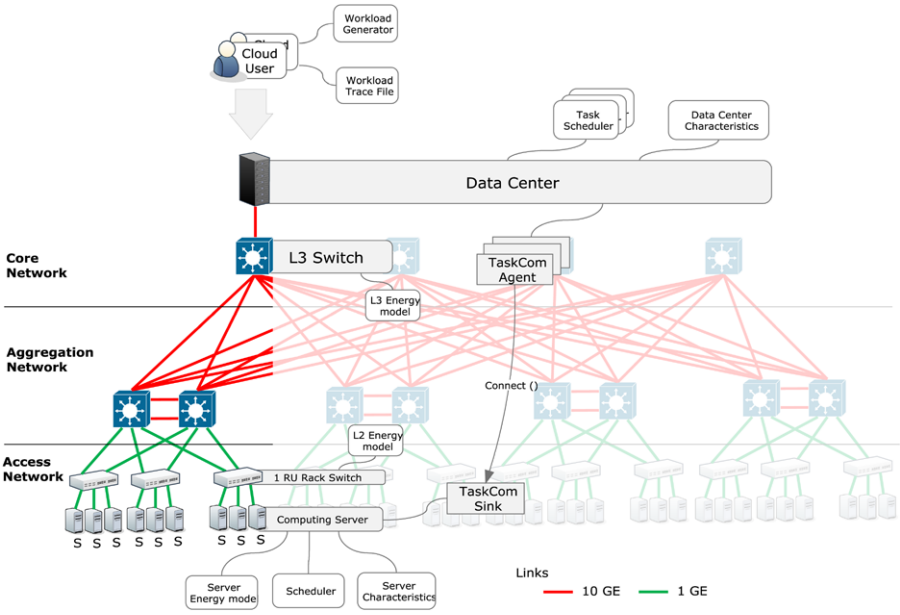

Fig. 4 Architecture of the GreenCloud simulation environment

an idle server consumes about 66% of energy compared to its fully loaded configuration. This is due to the fact that servers must manage memory modules, disks, I/O resources, and other peripherals in an acceptable state. Then, the power consumption linearly increases with the level of CPU load. As a result, the aforementioned model allows implementation of power saving in a centralized scheduler that can provision the consolidation of workloads in a minimum possible amount of the computing servers.

Another option for power management is Dynamic Voltage/Frequency Scaling (DVFS) [29] which introduces a tradeoff between computing performance and the energy consumed by the server. The DVFS is based on the fact that switching power in a chip decreases proportionally to $V^2 \cdot f$, where V is voltage, and f is the switching frequency. Moreover, voltage reduction requires frequency downshift. This implies a cubic relationship from f in the CPU power consumption. Note that server components, such as bus, memory, and disks, do not depend on the CPU frequency. Therefore, the power consumption of an average server can be expressed as follows [2]:

$$P = P_{\text{fixed}} + P_f \cdot f^3, \tag{1}$$

where P_{fixed} accounts for the portion of the consumed power which does not scale with the operating frequency f, while P_f is a frequency-dependent CPU power consumption.

Figure 5 presents the server power consumption model implemented in GreenCloud. The curve is built for a typical server running an Intel Xeon processor [17]. It consumes 301 W of energy with around 130 W allocated for peak CPU power consumption and around 171 W allocated for other peripheral devices.

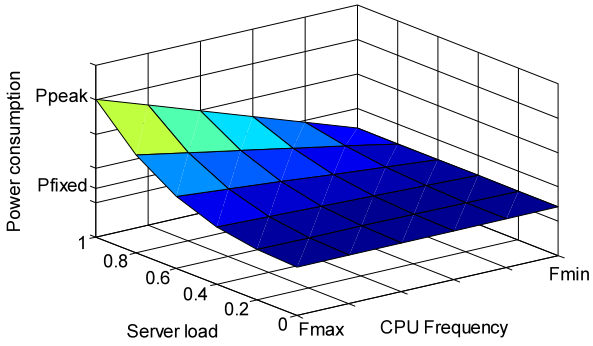

Fig. 5 Computing server power consumption

The scheduling depends on the server load level and operating frequency, and aims at capturing the effects of both of the DVFS and DPM techniques.

Switches and Links form the interconnection fabric that delivers workload to any of the computing servers for execution in a timely manner. The interconnection of switches and servers requires different cabling solutions depending on the supported bandwidth, physical and quality characteristics of the link. The quality of signal transmission in a given cable determines a tradeoff between the transmission rate and the link distance, which are the factors defining the cost and energy consumption of the transceivers.

The twisted pair is the most commonly used medium for Ethernet networks that allows organizing Gigabit Ethernet (GE) transmissions for up to 100 meters with the consumed transceiver power of around 0.4 W or 10 GE links for up to 30 meters with the transceiver power of 6 W. The twisted pair cabling is a low cost solution. However, for the organization of 10 GE links it is common to use optical multimode fibers. The multimode fibers allow transmissions for up to 300 meters with the transceiver power of 1 W [9]. On the other hand, the fact that multimode fibers cost almost 50 times of the twisted pair cost motivates the trend to limit the usage of 10 GE links to the core and aggregation networks as spending for the networking infrastructure may top 10–20% of the overall data center budget [11].

The number of switches installed depends on the implemented data center architecture as previously discussed in Sect. 2. However, as the computing servers are usually arranged into racks, the most common switch in a data center is the Top-of-Rack (ToR) switch. The ToR switch is typically placed at the top unit of the rack unit (1RU) to reduce the amount of cables and the heat produced. The ToR switches can support either gigabit (GE) or 10 gigabit (10 GE) speeds. However, taking into account that 10 GE switches are more expensive and that current capacity of aggregation and core networks is limited, gigabit rates are more common for racks.

Similar to the computing servers early power optimization proposals for interconnection network were based on DVS links [29]. The DVS introduced a control element at each port of the switch that depending on the traffic pattern and current levels of link utilization could downgrade the transmission rate. Due to the comparability requirements, only few standard link transmission rates are allowed, such as for GE links 10 Mb/s, 100 Mb/s, and 1 Gb/s are the only options.

On the other hand, the power efficiency of DVS links is limited as only a portion (3–15%) of the consumed power scales linearly with the link rate. As demonstrated

by the experiments in [22], the energy consumed by a switch and all its transceivers can be defined as:

$$P_{\text{switch}} = P_{\text{chassis}} + n_{\text{linecards}} + P_{\text{linecard}} + \sum_{i=0}^{R} n_{\text{ports},r} + P_r \tag{2}$$

where P_{chassis} is related to the power consumed by the switch hardware, P_{linecard} is the power consumed by any active network line card, P_r corresponds to the power consumed by a port (transceiver) running at the rate r. In (2), only the last component appears to be dependent on the link rate while other components, such as P_{chassis} and P_{linecard} remain fixed for all the duration of switch operation. Therefore, P_{chassis} and P_{linecard} can be avoided by turning the switch hardware off or putting it into sleep mode.

The proposed GreenCloud simulator implements energy model of switches and links according to (2) with the values of power consumption for different elements taken in accordance as suggested in [21]. The implemented powers saving schemes are: (a) DVS only, (b) DNS only, and (c) DVS with DNS.

Workloads are the objects designed for universal modeling of various cloud user services, such as social networking, instant messaging, and content delivery. In grid computing, the workloads are typically modeled as a sequence of jobs that can be divided into a set of tasks. The tasks can be dependent, requiring an output from other tasks to start execution, or independent. Moreover, due to the nature of grid computing applications (biological, financial modeling, or climate modeling), the number of jobs available prevail the number of computing resources available. While the main goal is the minimization of the time required for the computing of all jobs which may take weeks or months, the individual jobs do not have a strict completion deadline.

In cloud computing, incoming requests are typically generated for such applications like web browsing, instant messaging, or various content delivery applications. The jobs tend to be more independent, less computationally intensive, but have a strict completion deadline specified in SLA. To cover the vast majority of cloud computing applications, we define three types of jobs:

- *Computationally Intensive Workloads (CIWs)* model High-Performance Computing (HPC) applications aiming at solving advanced computational problems. CIWs load computing servers considerably, but require almost no data transfers in the interconnection network of the data center. The process of CIW energy-efficient scheduling should focus on the server power consumption footprint trying to group the workloads at the minimum set of servers as well as to route the traffic produced using a minimum set of routes. There is no danger of network congestion due to the low data transfer requirements, and putting the most of the switches into the sleep mode will ensure the lowest power of the data center network.
- *Data-Intensive Workloads (DIWs)* produce almost no load at the computing servers, but require heavy data transfers. DIWs aim to model such applications like video file sharing where each simple user request turns into a video streaming process. As a result, the interconnection network and not the computing capacity becomes a bottleneck of the data center for DIWs. Ideally, there should be a continuous feedback implemented between the network elements (switches) and the

central workload scheduler. Based on such feedback, the scheduler will distribute the workloads taking current congestion levels of the communication links. It will avoid sending workloads over congested links even if certain server's computing capacity will allow accommodating the workload. Such scheduling policy will balance the traffic in the data center network and reduce average time required for a task delivery from the core switches to the computing servers.

– *Balanced Workloads (BWs)* aim to model the applications having both computing and data transfer requirements. BWs load the computing servers and communication links proportionally. With this type of workloads the average load on the servers equals to the average load of the data center network. BWs can model such applications as geographic information systems which require both large graphical data transfers and heavy processing. Scheduling of BWs should account for both servers' load and the load of the interconnection network.

The execution of each workload object in GreenCloud requires a successful completion of its two main components: (a) computing and (b) communicational. The computing component defines the amount of computing that has to be executed before a given deadline on a time scale. The deadline aims at introducing Quality of Service (QoS) constraints specified in SLA. The communicational component of the workload defines the amount and the size of data transfers that must be performed prior, during, and after the workload execution. It is composed of three parts: (a) the size of the workload, (b) the size of internal, and (c) the size of external to the data center communications. The size of the workload defines the number of bytes that after being divided into IP packets are required be transmitted from the core switches to the computing servers before a workload execution can be initiated. The size of external communications defines the amount of data required to be transmitted outside the data center network at the moment of task completion and corresponds to the task execution result. The size of internal to the data center communications defines the amount of data to be exchanged with another workload that can be executed at the same or a different server. This way the workload interdependencies are modeled. In fact, internal communication in the data center can account for as much as 70% of total data transmitted [21]. In current version of the GreenCloud simulator, internal communication is performed with a randomly chosen workload. However, in the next version inter-workload communication patterns will be defined at the moment of the workload arrival and communication-aware scheduling will be studied.

An efficient and effective methodology to optimize energy consumption of interdependent workloads is to analyze the workload communication requirements at the moment of scheduling and perform a coupled placement of these interdependent workloads—a co-scheduling approach. The co-scheduling approach will reduce the number of links/switches involved into communication patterns.

The workload arrival rate/pattern to the data center can be configured to follow a predefined (within the simulator) distribution, such as Exponential or Pareto, or can be re-generated from traces log files. Moreover, different random distributions can be configured to trigger the time of a workload arrival as well as specify the size of the workload. The above flexibility provides ample provisions for users to thoroughly investigate network utilization, traffic load, and impact on various switching

components. Furthermore, the trace-driven workload generation is designed to simulate more realistic workload arrival process capturing also intraday fluctuations [20], which may influence simulated results greatly.

3.3 Comparison of cloud computing simulators

The number of simulation environments for cloud computing data centers available for public use is limited. The CloudSim simulator [1] is probably the most sophisticated among the simulators overviewed. It is evolved as a built up on top of the grid network simulator GridSim developed at the University of Melbourne, Australia in 2002. The MDCSim simulator [19] is, on the contrary, a relatively fresh data center simulator developed at the Pennsylvania State University in 2009. It is supplied with specific hardware characteristics of data server components such as servers, communication links and switches from different vendors and allows estimation of power consumption. Table 1 compares cloud computing simulators via comparison of their characteristics.

Platform (Language/Script): The proposed GreenCloud simulator is developed as an extension of the Ns2 network simulator [31] which is coded in C++ with a layer of OTcl libraries implemented on top of it. It is a packet level simulator, meaning that whenever a data message has to be transmitted between simulator entities a packet structure with its protocol headers is allocated in the memory and all the associated protocol processing is performed. On the contrary, CloudSim and MDCSim are event-based simulators. They avoid building and processing small simulation objects (like packets) individually. Instead, the effect of object interaction is captured. Such a method reduces simulation time considerably, improves scalability, but lacks in the simulation accuracy.

Table 1 Comparison of cloud computing simulators

Parameter	GreenCloud	CloudSim	MDCSim
Platform	Ns2	SimJava	CSIM
Language/Script	C++/OTcl	Java	C++/Java
Availability	Open source	Open source	Commercial
Simulation time	Tens of minutes	Seconds	Seconds
Graphical support	Limited (Network animator)	Limited (CloudAnalyst)	None
Application models	Computation, Data transfer, and Exec. deadline	Computation, Data transfer	Computation
Communication models	Full	Limited	Limited
Support of TCP/IP	Full	None	None
Physical models	Available using plug in	None	None
Energy models	Precise (servers + network)	None	Rough (servers only)
Power saving modes	DVFS, DNS, and both	None	None

Availability: Both GreenCloud and CloudSim simulators are released under open source GPL license. The MDCSim simulator is currently not available for public download, as is its platform CSIM [5] which is a commercial product.

Simulation time: The time required for the simulation depends on many factors such as the simulated scenario or the hardware used for running the simulator software. In general, CloudSim and MDCSim, being event-based simulators, are faster and scale to a larger number of data center nodes. Nevertheless, the GreenCloud simulator still achieves reasonable simulation times. They are in the order of tens of minutes for an hour of simulation time while simulating a typical data center with a few thousand of nodes. Apart of the number of nodes, the simulation duration is greatly influenced by the number of communication packets produced as well as the number of times they are processed at network routers during forwarding. As a result, a typical data center simulated in GreenCloud can be composed of thousands of nodes while the Java-based CloudSim and MDCSim can simulate millions of computers.

Graphical support: Basically, there is no simulator among the overviewed ones implementing advanced GUI. The GreenCloud may be enabled to produce a trace files recognized by the network animation tool Nam [31] which visualizes a simulated topology and a packet flow after the simulation is completed. However, no GUI tool is available to configure a simulation setup or display simulation graphs in a friendly way. In the case of the CloudSim simulator, an external tool CloudAnalyst [32] is developed. It visualizes only high-level simulation setup parameters targeting cloud applications at the globe scale. The MDCSim simulator does not supply any GUI.

Application models: All three simulators implement user application models as simple objects describing computational requirements for the application. In addition, GreenCloud and CloudSim specify communicational requirements of the applications in terms of the amount of data to be transferred before and after a task completion. The application model implemented in the CloudSim fits well with High-Performance Computing (HPC). HPC tasks, being computationally intensive and having no specific completion deadline, are the typical application for grid networks. In cloud computing, QoS requirements for the execution of user requests are defined in SLA. Therefore, GreenCloud extends the model of user application by adding a predefined execution deadline.

Communication model/Support of TCP/IP: One of the main strengths of the GreenCloud simulator is in the details it offers while modeling communication aspects of the data center network. Being based on the platform implementing TCP/IP protocol reference mode in full, it allows capturing the dynamics of widely used communication protocols such as IP, TCP, UDP, etc. Whenever a message needs to be transmitted between two simulated elements, it is fragmented into a number of packets bounded in size by network MTU. Then, while routed in the data center network, these packets become a subject to link errors or congestion-related losses in network switches. Both CloudSim and MDCSim implement limited communication model mainly just accounting for the transmission delay and bandwidth. The CloudSim with its *network* package maintains a data center topology in the form of a directed graph. Each edge is assigned with the bandwidth and delay parameters. Whenever an edge is involved into transmission, its bandwidth component is reduced for transmission delay duration. However, no protocol dynamics are captured for the

congestion control, error recovery, or routing specifics. Similar simplified communication model is implemented in MDCSim.

Physical models: A degree of details the models of the simulated components capture the behavior of their physical analogs defines the precision of the simulator and the validity of the results. While there is no direct support for simulating physical processes implemented in the GreenCloud, it allows a plug-in insertion of the model derived from physical layer simulators. For example, a packet loss probability in the optical fiber depending on the transmission range can be obtained via simulation of signal propagation dynamics and inserted into GreenCloud.

Energy models: In GreenCloud, the energy models are implemented for every data center element (computing servers, core and rack switches). Moreover, due to the advantage in the simulation resolution, energy models can operate at the packet level as well. This allows updating the levels of energy consumption whenever a new packet leaves or arrives from the link, or whenever a new task execution is started or completed at the server. While the CloudSim does not account of the energy spent, the MDCSim performs only rough estimation. It uses special heuristics averaging on the number of the received requests in a given monitoring period. Such energy estimation is performed for computing servers only, and no communication-related consumptions are monitored.

Power saving modes: The GreedCloud is the only simulator with the support of different power saving modes. The following three algorithms are implemented: DVFS, DNS, and DVFS + DNS.

Summarizing, short simulation times are provided by CloudSim and MDCSim even for very large data centers due to their event-based nature, while GreenCloud offers an improvement in the simulation precision keeping the simulation time at the reasonable level. None of the tools offer user-friendly graphical interface. The CloudSim allows implementation of the complex scheduling and task execution schemes involving resource virtualization techniques. However, its workloads are more relevant in grid networks. The GreenCloud supports cloud computing workloads with deadlines, but only simple scheduling policies for single core servers are implemented. The MDCSim workloads are described with the computational requirements only and require no data to be transferred.

Communication details and the level of energy models support are the key strengths of the GreenCloud which are provided via full support TCP/IP protocol reference model and packet level energy models implemented for all data center components: servers, switches, and links.

4 Performance evaluation

In this section, we present case study simulations of an energy-aware data center for two-tier (2T), three-tier (3T), and three-tier high-speed (3Ths) architectures.

For comparison reasons, we fixed the number of computing nodes to 1536 for all three topologies, while the number and interconnection of network switches varied. Table 2 summarizes the main simulation setup parameters.

In contrast with other architectures, a 2T data center does not include aggregation switches. The core switches are connected to the access network directly using

Table 2 Simulation setup parameters

	Parameter	Data center architectures		
		Two-tier	Three-Tier	Three-tier high-speed
Topologies	Core nodes (C_1)	16	8	2
	Aggregation nodes (C_2)	–	16	4
	Access switches (C_3)	512	512	512
	Servers (S)	1536	1536	1536
	Link (C_1–C_2)	10 GE	10 GE	100 GE
	Link (C_2–C_3)	1 GE	1 GE	10 GE
	Link (C_3–S)	1 GE	1 GE	1 GE
	Link propagation delay	10 ns		
Data Center	Data center average load	30%		
	Task generation time	Exponentially distributed		
	Task size	Exponentially distributed		
	Simulation time	60 minutes		

1 GE links (referred as C_2–C_3) and interconnected between them using 10 GE links (referred as C_1–C_2).

The 3Ths architecture mainly improves the 3T architecture with providing more bandwidth in the core and aggregation parts of the network. The bandwidth of the C_1–C_2 and C_2–C_3 links in the 3Ths architecture is ten times of that in 3T and corresponds to 100 GE and 10 GE, respectively. The availability of 100 GE links allows keeping the number of core switches as well as the number of paths in the ECMP routing limited to 2 serving the same amount switches in the access. The propagation delay of all the links is set to 10 ns.

The workload generation events and the size of the workloads are exponentially distributed. The average size of the workload and its computing requirement depends on the type of task. For CIW workloads, the relation between computing and data transfer parts is chosen to be $1/10$, meaning that with a maximum load of the data center its servers will be occupied for 100% while the communication network will be loaded for 10% of its maximum capacity. For DIW workloads the relation is reverse. Under the maximum load, the communication network is loaded for 100% while computing servers for only 10%. Balanced workloads load computing servers and data center network proportionally.

The workloads arrived to the data center are scheduled for execution using energy-aware "green" scheduler. This "green" scheduler tends to group the workloads on a minimum possible amount of computing servers. In order to account for DIW workloads, the scheduler continuously tracks buffer occupancy of network switches on the path. In case of congestion, the scheduler avoids using congested routes even if they lead to the servers able to satisfy computational requirement of the workloads.

The servers left idle are put into sleep mode (DNS scheme) while on the underloaded servers the supply voltage is reduced (DVFS scheme). The time required to change the power state in either mode is set to 100 ms.

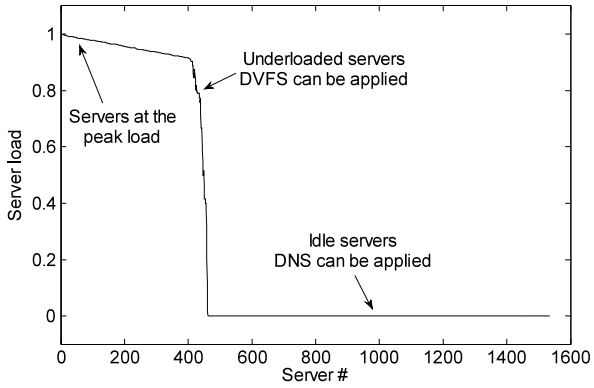

Fig. 6 Server workload distribution with a "green" scheduler

Table 3 Power consumption of data center components

Parameter	Power consumption (W)		
	Servers		
Server peak	301		
Server CPU peak	130		
Server other (memory, peripherial, mother board, fan, PSU losses)	171		
Server idle	198		
	Switches		
	Top-of-Rack (C_3)	Core (C_1)	Aggregation (C_2)
Chassis	146 (1 G)	1.5 K (10 G)	15 K (100 G)
Linecard	–	1 K (10 G)	12 K (100 G)
Port transceiver	0.42 (1 G)	0.3 K (10 G)	1.6 K (100 G)

Figure 6 presents a workload distribution among servers. The whole load of the data center (around 30% of its total capacity) is mapped onto approximately one third of the servers maintaining load at a peak rate (left part of the chart). This way, the remaining two thirds of the servers can be shut down using DNS technique. A tiny portion of the approximately 50 out of 1536 servers which load represents a falling slope of the chart are under-utilized on average, and DVFS technique can be applied on them.

Table 3 presents the power consumption of data center components. The server peak energy consumption of 301 W is composed of 130 W (43%) allocated for a peak CPU consumption [17] and 171 W (56%) consumed by other devices like memory, disks, peripheral slots, mother board, fan, and power supply unit [7]. As the only component which scales with the load is the CPU power, the minimum consumption of an idle server is bounded and corresponds to 198 W (66%) where also a portion of CPU power consumption of 27 W required to keep the operating system running is included.

Table 4 The distribution of data center power consumption

Parameter	Power consumption (kW h)		
	Two-tier (2T)	Three-Tier (3T)	Three-tier high-speed (3Ths)
Data center	477.8	503.4	508.6
Servers	351	351	351
Switches	126.8	152.4	157.6
Core (C_1)	51.2	25.6	56.8
Aggregation (C_2)	–	51.2	25.2
Access (C_3)	75.6	75.6	75.6

The switches' consumption is almost constant for different transmission rates as most of the power (85–97%) is consumed by their chassis and line cards and only a small portion (3–15%) is consumed by their port transceivers. Switch power consumption values are derived from [9, 21] with a twisted pair cable connection considered for the rack switch (C3) and optical multimode fiber for the core (C1) and aggregation (C2) switches. Depending on the employed data center topology, the core and aggregation switches will consume differently. For the 3T topology where the fastest links are 10 G the core and aggregation switches consume a few kilowatts, while in the 3Ths topology where links are of 10 G speed faster switches are needed which consume tens of kilowatts.

Table 4 presents simulation results obtained for three evaluated data center topologies with no energy saving management involved for an average load of the data center of 30%. The obtained numbers aim to estimate the scope of the energy-related spending components in modern data centers and define where the energy management schemes would be the most efficient.

On average, the data center consumption is around 432 kW h during an hour of the runtime. On the yearly basis, it corresponds to 4409 MW h or $441 thousand with an average price of 10 c per kW h.

The processing servers share around 70% of total data center energy consumption, while the communicational links and switches account for the rest 30%. Furthermore, the consumption of switches breaks with 17% allocated for core switches, 34% for aggregation switches, and 50% for the access switches. It means that after computing servers lowering the power consumption of access switches will have the highest impact. The core and aggregation switches together account for 15% of total energy consumption. However, taking into account the requirements for network performance, load balancing, and communication robustness, the obvious choice is to keep core and aggregation switches constantly running possibly applying communication rate reduction in a distributed manner.

The data center network accounts for the differences between power consumption levels of different data center architectures. With the respect to the 2T architecture, the 3T architecture adds around 25 kW for aggregation layer which enables the data center scale beyond 10 000 nodes. The 3Ths architecture contains fewer core and aggregation switches. However, the availability of 100 G links comes at a price of the

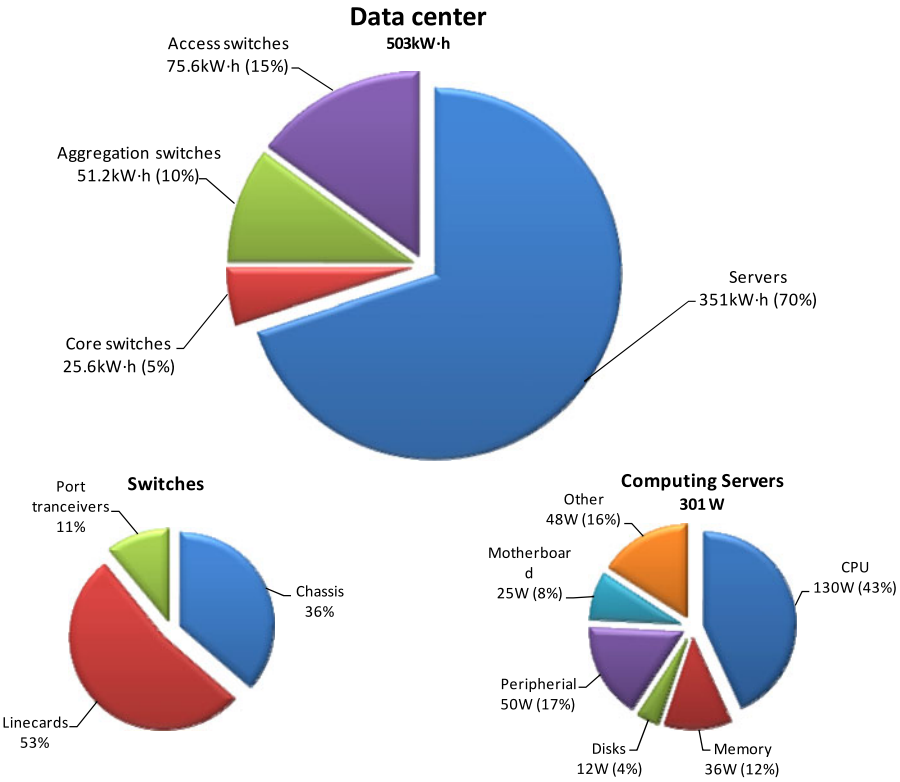

Fig. 7 Distribution of energy consumption in a data center

increase per-switch energy consumption. As a result, a 3Ths network consumes more than a 3T network. Figure 7 reports an average distribution of energy consumption in a 3T data center.

Table 5 compares the impact on energy consumption of DVFS, DNS, and DVFS with DNS schemes applied on both computing several and networking equipment. The results are obtained for balanced tasks loading both computing servers and interconnection network equally for an average system load of 30%.

The DVFS scheme alone reduces power consumption to only 96% from the nominal level. Most of the power saving in servers comes from downshifting CPU voltage on the under-loaded servers. However, CPU accounts for 43% of server consumption only. On the other hand, DVFS shows itself ineffective for the switches as only 3% to 15% of the switch's energy is sensitive to the transmission rate variation.

The most effective results are obtained by DNS scheme. It is equally effective for both servers and switches as the most of their energy consumed shows no dependency on the operating frequency. However, in order to utilize DNS scheme effectively, its design should be coupled with the data center scheduler positioned to unload the maximum number of the servers.

Table 5 Comparison of energy-efficient schemes

Parameter	Power consumption (kW h)			
	No energy-saving	DVFS	DNS	DVFS+DNS
Data center	503.4	486.1 (96%)	186.7 (37%)	179.4 (35%)
Servers	351	340.5 (97%)	138.4 (39%)	132.4 (37%)
Switches	152.4	145.6 (95%)	48.3 (32%)	47 (31%)
Energy cost/year	$441k	$435k	$163.5k	$157k

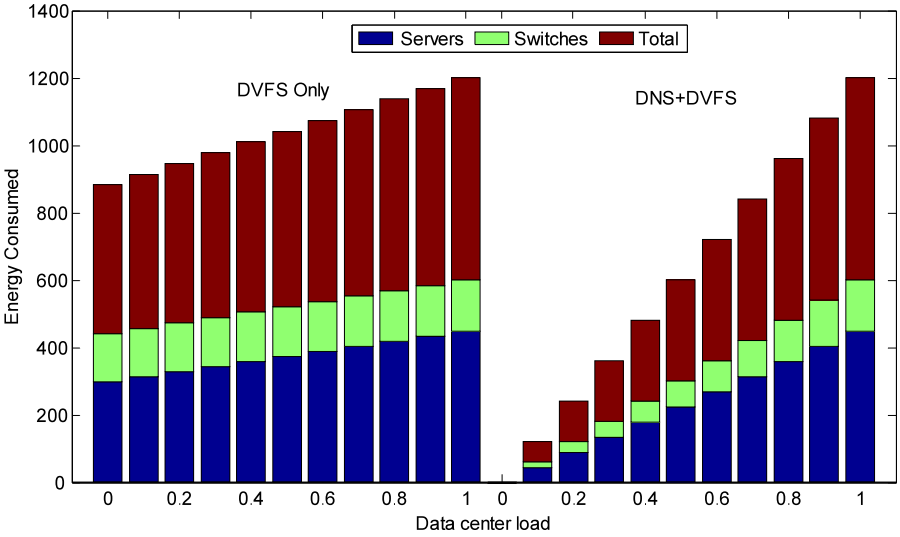

Fig. 8 Data center energy consumption comparison under variable load for DVFS only and DNS+DVFS power management schemes

The bottom of the table provides estimates of the data center energy cost on a yearly basis. Initial energy spending of $441 thousand can be reduced down to almost a third, $157 thousand, by a combination of DVFS and DNS schemes.

Figure 8 presents data center energy consumption under variable load conditions for DVFS only and DNS+DVFS power management schemes. The curves are presented for balanced type of the workloads and correspond to the total data center consumption as well as the energy consumed by the servers and switches. The DVFS scheme shows itself little sensitive to the input load of the servers and almost insensitive to the load of network switches. On the contrary, the DNS scheme appears to capture load variation precisely adjusting power consumptions of both servers and switches accordingly. The results reported are averaged over 20 runs with the random seed controlling random number generator. The introduced uncertainty affected mainly the way the workloads arrive to the data center slightly impacting the number of servers and network switches required to be powered. The maximum variance of 95% confidence intervals (not reported in Fig. 8 and Fig. 9) from the mean value

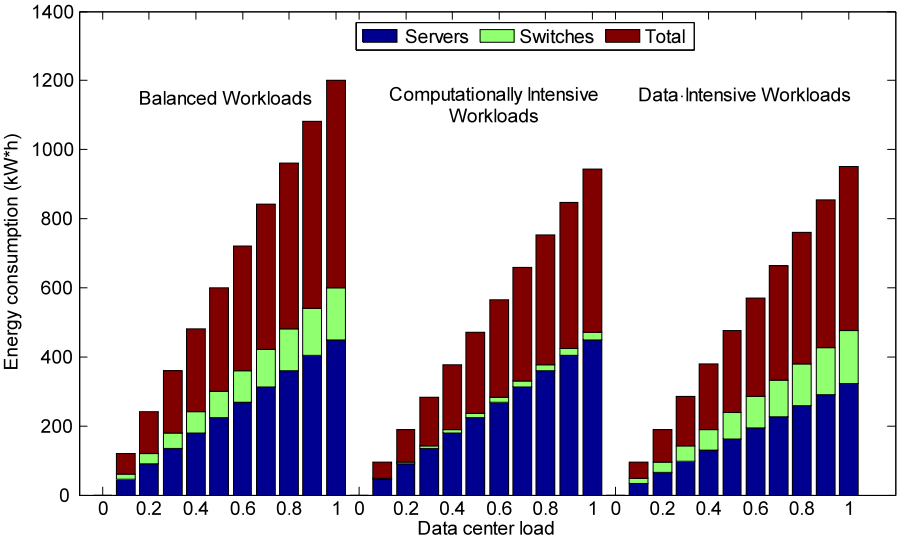

Fig. 9 Data center energy consumption for different types of workloads

accounted for less than 0.2% for the energy consumed by the servers and less than 0.1% for the energy consumed by the network switches.

Figure 9 presents data center energy consumption comparison for different types of user workloads: balanced, computationally intensive, and data-intensive workloads. Balanced workloads consume the most as the consumptions of both servers and switches become proportional to the offered load of the system. CIWs stress the computing servers and leave data center network almost unutilized. On the contrary, execution of DIWs creates a heavy traffic load at the switches and links leaving the servers mostly idle.

The process of scheduling for DIWs requires performing load balancing for redistributing the traffic from congested links. As a result, these workloads cannot be fully grouped at the minimum amount of the servers due to the limitations of the data center topology. This way, in real data centers with the mixed nature of workloads the scheduler may attempt a grouped allocation of CIWs and DIWs as optimal allocation policy.

5 Conclusions

In this paper, we presented a simulation environment for energy-aware cloud computing data centers. GreenCloud is designed to capture details of the energy consumed by data center components as well as packet-level communication patterns between them.

The simulation results obtained for two-tier, three-tier, and three-tier high-speed data center architectures demonstrate applicability and impact from the application of different power management schemes like voltage scaling or dynamic shutdown applied on the computing as well as on the networking components.

The future work will focus on the simulator extension adding storage area network techniques and further refinement of energy models used in the simulated components. On the algorithmic part, the research will be focused on the development of different workload consolidation and traffic aggregation techniques.

Acknowledgements The authors would like to acknowledge the support of Luxembourg FNR in the framework of GreenIT project (C09/IS/05) and the European Research Consortium for Informatics and Mathematics (ERCIM) for providing a research fellowship.

References

1. Buyya R, Ranjan R, Calheiros RN (2009) Modeling and simulation of scalable cloud computing environments and the CloudSim toolkit: challenges and opportunities. In: Proceedings of the 7th high performance computing and simulation conference, Leipzig, Germany, June
2. Chen Y, Das A, Qin W, Sivasubramaniam A, Wang Q, Gautam N (2005) Managing server energy and operational costs in hosting centers. In: Proceedings of the ACM SIGMETRICS international conference on measurement and modeling of computer systems. ACM, New York, pp 303–314
3. Chen G, He W, Liu J, Nath S, Rigas L, Xiao L, Zhao F (2008) Energy-aware server provisioning and load dispatching for connection-intensive internet services. In: The 5th USENIX symposium on networked systems design and implementation, Berkeley, CA, USA
4. Cisco Systems (2008) Cisco Data Center Infrastructure 2.5 Design Guide, Inc., May
5. CSIM Development Toolkit for Simulation and Modeling (2010) Available at http://www.mesquite.com/
6. Fan X, Weber W-D, Barroso LA (2007) Power provisioning for a warehouse-sized computer. In: Proceedings of the ACM international symposium on computer architecture, San Diego, CA, June
7. Fan X, Weber W-D, Barroso LA (2007) Power provisioning for a warehouse-sized computer. In: Proceedings of the 34th annual international symposium on computer architecture (ISCA). ACM, New York, pp 13–23
8. Fan X, Weber W-D, Barroso LA (2007) Power provisioning for a warehouse-sized computer. In: Proceedings of the 34th annual international symposium on computer architecture (ISCA '07). ACM, New York, pp 13–23
9. Farrington N, Rubow E, Vahdat A (2009) Data center switch architecture in the age of merchant silicon. In: Proceedings of the 17th IEEE symposium on high performance interconnects (HOTI '09). IEEE Computer Society, Washington, pp 93–102
10. Gartner Group (2010) Available at: http://www.gartner.com/
11. Greenberg A, Lahiri P, Maltz DA, Patel P, Sengupta S (2008) Towards a next generation data center architecture: scalability and commoditization. In: Proceedings of the ACM workshop on programmable routers for extensible services of tomorrow, Seattle, WA, USA, August 22–22
12. Guo C, Wu H, Tan K, Shiy L, Zhang Y, Luz S (2008) DCell: a scalable and fault-tolerant network structure for data centers. In: ACM SIGCOMM, Seattle, Washington, USA
13. Guo C, Lu G, Li D, Wu H, Zhang X, Shi Y, Tian C, Zhang Y, Lu S (2009) BCube: a high performance, server-centric network architecture for modular data centers. In: ACM SIGCOMM, Barcelona, Spain
14. Horvath T, Abdelzaher T, Skadron K, Liu X. (2007) Dynamic voltage scaling in multitier web servers with end-to-end delay control. IEEE Trans Comput 56(4):444–458
15. IEEE Std. 802.3ad-2000 (2000) Amendment to carrier sense multiple access with collision detection (CSMA/CD) access method and physical layer specifications-aggregation of multiple link segments. IEEE Press, New York
16. IEEE std 802.3ba-2010 (2010) Media access control parameters, physical layers and management parameters for 40 Gb/s and 100 Gb/s operation. June
17. Intel Inc. (2010) Intel® Xeon® Processor 5000 Sequence. Available at: http://www.intel.com/p/en_US/products/server/processor/xeon5000
18. Li B, Li J, Huai J, Wo T, Li Q, Zhong L (2009) EnaCloud: an energy-saving application live placement approach for cloud computing environments. In: IEEE international conference on cloud computing, Bangalore, India
19. Lim S-H, Sharma B, Nam G, Kim EK, Das CR (2009) MDCSim: a multi-tier data center simulation, platform. In: IEEE international conference on cluster computing and workshops (CLUSTER)

20. Liu J, Zhao F, Liu X, He W (2009) Challenges Towards Elastic Power Management in Internet Data Centers. In: Proceedings of the 2nd international workshop on cyber-physical systems (WCPS), in conjunction with ICDCS 2009, Montreal, Quebec, Canada, June
21. Mahadevan P, Sharma P, Banerjee S, Ranganathan P (2009) Energy aware network operations. In: IEEE INFOCOM workshops, pp 1–6
22. Mahadevan P, Sharma P, Banerjee S, Ranganathan P (2009) A power benchmarking framework for network devices. In: Proceedings of the 8th international IFIP-TC 6 networking conference, Aachen, Germany, May 11–15
23. Moore J, Chase J, Ranganathan P, Sharma R (2005) Making scheduling "cool": temperature-aware workload placement in data centers. In: USENIX annual technical conference
24. Postel J (1981) Internet control message protocol. Internet engineering task force request for comments 792, September
25. Raghavendra R, Ranganathan P, Talwar V, Wang Z, Zhu X (2008) No "power" struggles: coordinated multi-level power management for the data center. In: APLOS
26. Rasmussen N (2010) Calculating total cooling requirements for data centers. White paper, APC Legendary Reliability. Available at: http://www.ptsdcs.com/whitepapers/23.pdf
27. Rawson A, Pfleuger J, Cader T (2008) Green grid data center power efficiency metrics: PUE and DCIE. The Green Grid White Paper #6
28. Rimal BP, Choi E, Lumb I (2009) A taxonomy and survey of cloud computing systems. In: The fifth international joint conference on INC, IMS and IDC, pp 44–51
29. Shang L, Peh L-S, Jha NK (2003) Dynamic voltage scaling with links for power optimization of interconnection networks. In: Proceedings of the 9th international symposium on high-performance computer architecture table of contents
30. Thaler D, Hopps C (2000) Multipath issues in unicast and multicast nexthop selection. Internet engineering task force request for comments 2991, November
31. The Network Simulator Ns2 (2010) Available at: http://www.isi.edu/nsnam/ns/
32. Wickremasinghe B, Calheiros RN, Buyya R (2008) CloudAnalyst: a CloudSim-based visual modeller for analysing cloud computing environments and applications. In: International conference on advanced information networking and applications (AINA 2010), Perth, Australia, April 20–23