Cluster Comput
DOI 10.1007/s10586-011-0177-4

DENS: data center energy-efficient network-aware scheduling

Dzmitry Kliazovich · Pascal Bouvry · Samee Ullah Khan

Received: 18 April 2011 / Accepted: 10 August 2011

Abstract In modern data centers, energy consumption accounts for a considerably large slice of operational expenses. The existing work in data center energy optimization is focusing only on job distribution between computing servers based on workload or thermal profiles. This paper underlines the role of communication fabric in data center energy consumption and presents a scheduling approach that combines energy efficiency and network awareness, named DENS. The DENS methodology balances the energy consumption of a data center, individual job performance, and traffic demands. The proposed approach optimizes the tradeoff between job consolidation (to minimize the amount of computing servers) and distribution of traffic patterns (to avoid hotspots in the data center network).

Keywords Network-aware scheduling · Energy-efficient · Data center · Cloud computing · Congestion

1 Introduction

Data centers are becoming increasingly popular for the provisioning of computing resources. The cost and operational expenses of data centers have skyrocketed with the increase in computing capacity [7].

Energy consumption is a growing concern for data centers operators. It is becoming one of the main entries on a data center operational expenses (OPEX) bill [11, 28]. The Gartner Group estimates energy consumptions to account for up to 10% of the current OPEX, and this estimate is projected to rise to 50% in the next few years [14].

The slice of roughly 40% is related to the energy consumed by information technology (IT) equipment [7], which includes energy consumed by the computing servers as well as data center network hardware used for interconnection. In fact, about one-third of the total IT energy is consumed by communication links, switching, and aggregation elements, while the remaining two-thirds are allocated to computing servers [29]. Other systems contributing to the data center energy consumption are cooling and power distribution systems that account for 45% and 15% of total energy consumption, respectively.

Early solutions implemented distributed algorithms for making data center hardware energy efficient [6]. There are two popular techniques for power savings in computing systems. The Dynamic Voltage and Frequency Scaling (DVFS) technology, adjusts hardware power consumption according to the applied computing load and the Dynamic Power Management (DPM), achieves most of energy savings by powering down devices at runtime. To make DPM scheme efficient, a scheduler must consolidate data center jobs on a minimum set of computing resources to maximize the amount of unloaded servers that can be powered down (or put to sleep) [22]. Because the average data center workload often stays around 30%, the portion of unloaded servers can be as high as 70% [23].

Most of the existing approaches for job scheduling in data centers focus exclusively on the job distribution between computing servers [30] targeting energy-efficient [3] or thermal-aware scheduling [32]. To the best of our knowledge, only a few approaches have considered data center

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

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

Published online: 22 September 2011

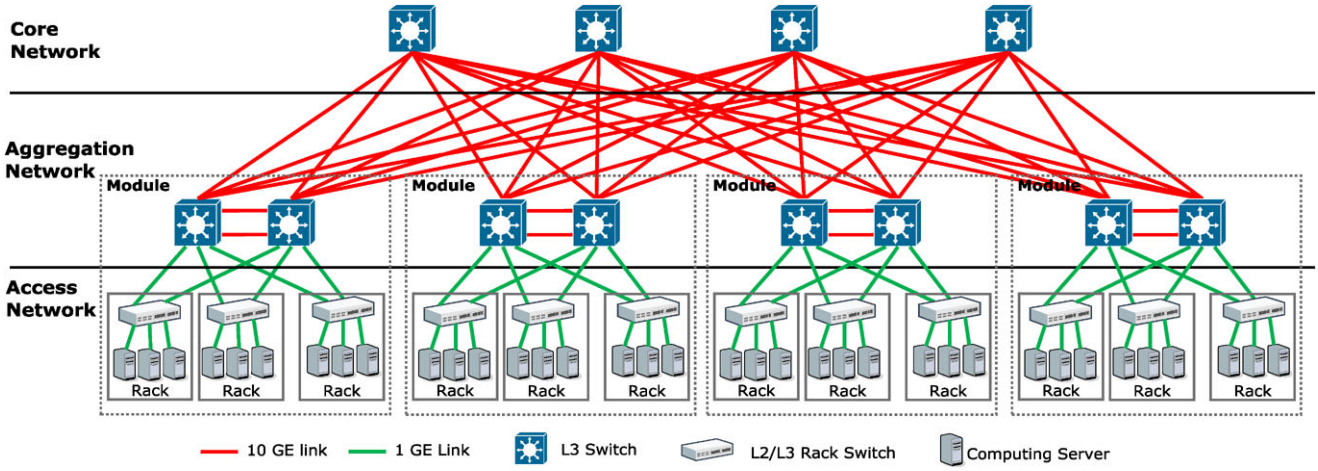

Fig. 1 Three-tier data center architecture

network and traffic characteristics for developing energy-efficient data center schedulers [1, 26, 31].

Reference [1] identifies the problem associated with existing multi-path routing protocols in typical fat tree network topologies. Two large traffic flows may be assigned to share the same path if their hash values collide leaving other paths under-loaded. The problem is solved with the introduction of a complex central scheduler that performs flow differentiation and analysis of flow traffic demands across the data center network. Traffic-aware virtual machine placement is proposed in [26]. Relying on the knowledge about network topology, virtual machines are placed to optimize traffic flows inside a data center network. The approach presented in [31], also allows job migration control during runtime with a specifically designed network-aware scheduler. The migration scheduler is aware of the migration delays and bandwidth resources required.

This paper presents a data center scheduling methodology that combines energy efficiency and network awareness. The methodology is named DENS, which is an acronym for data center energy-efficient network-aware scheduling. The network-awareness refers to the ability of DENS approach to receive and analyze a run-time feedback from the data center switches and links as well as take decisions and actions based on the network feedback. The DENS methodology aims to achieve the balance between individual job performances, job QoS requirements also defined in Service Level Agreement (SLA), traffic demands, and energy consumed by the data center. Data intensive jobs require low computational load, but produce heavy data streams directed out of the data center as well as to the neighboring nodes. Such data intensive jobs are typically produced by popular video sharing or geographical information services. The scheduling approach presented in this paper is designed to avoid hotspots within a data center while minimizing the number of computing servers required for job execution. In the proposed methodology, the network awareness is achieved with the introduction of feedback channels from the main network switches. Moreover, the proposed methodology reduces computational and memory overhead compared to previous approaches, such as flow differentiation, which makes the proposed methodology easy to implement and port to existing data center schedulers.

The rest of the paper is organized as follows. Section 2 summarizes the background knowledge on a typical data center architecture, energy consumption models, and data center network congestion. Section 3 presents the core of the scheduling approach and defines the necessary components of the proposed methodology. In Sect. 4, we will present and discuss experimental results. Finally, Sect. 5 will conclude the paper and outline directions for future work on the topic.

2 Background

2.1 Data center topology

Three-tier trees of hosts and switches form the most widely used data center architecture [10]. It (see Fig. 1) consists of the core tier at the root of the tree, the aggregation tier that is responsible for routing, and the access tier that holds the pool of computing servers (or hosts). Early data centers used two-tier architectures with no aggregation tier. However, such data centers, depending on the type of switches used and per-host bandwidth requirements, could typically support not more than 5,000 hosts. Given the pool of servers in today's data centers that are of the order of 100,000 hosts [24] and the requirement to keep layer-2 switches in the access network, a three-tiered design becomes the most appropriate option.

Although 10 Gigabit Ethernet (GE) transceivers are commercially available, in a three-tiered architecture the computing servers (grouped in racks) are interconnected using

1 GE links. This is due to the fact that 10 GE transceivers: (a) are too expensive and (b) probably offer more capacity than needed for connecting computing servers. In current data centers, rack connectivity is achieved with inexpensive Top-of-Rack (ToR) switches. A typical ToR switch shares two 10 GE uplinks with 48 GE links that interconnect computing servers within a rack. The difference between the downlink and the uplink capacities of a switch defines its oversubscription ratio, which in the aforementioned case is equal to $48/20 = 2.4 : 1$. Therefore, under full load, only 416 Mb/s will remain available to each of the individual servers out of their 1 GE links.

At the higher layers of hierarchy, the racks are arranged in modules (see Fig. 1) with a pair of aggregation switches servicing the module connectivity. Typical oversubscription ratios for these aggregation switches are around 1.5:1, which further reduces the available bandwidth for the individual computing servers to 277 Mb/s.

The bandwidth between the core and aggregation networks is distributed using a multi-path routing technology, such as the Equal Cost Multi-Path (ECMP) routing [33]. The ECMP technique performs a per-flow load balancing, which differentiates the flows by computing a hash function on the incoming packet headers. For a three-tiered architecture, the maximum number of allowable ECMP paths bounds the total number of core switches to eight. Such a bound also limits the deliverable bandwidth to the aggregation switches. This limitation will be waved with the (commercial) availability of 100 GE links, standardized in June 2010 [18].

Designing data center topologies is an extremely important research topic. Fat-tree successors are constantly proposed for large-scale data centers [15, 16]. However, the fact that not even a single data center has been built (to this date) based on such proposals, we constrict the scope of this paper to the three-tiered architecture. Nevertheless, we must note that all of the findings of this research will remain valid for any or all types of data center topologies.

2.2 Energy models

Computing servers account for a major portion of data center energy consumption. The power consumption of a computing server is proportional to the CPU utilization. An idle server consumes around two-thirds of its peak-load consumption to keep memory, disks, and I/O resources running [8]. The remaining one-third changes almost linearly with the increase in the level of CPU load.

There are two main approaches for reducing energy consumption in computing servers: (a) DVFS [27] and (b) DPM [4]. The DVFS scheme adjusts the CPU power (consequently the performance level) according to the offered load. The aforementioned is based on the fact that power in a chip decreases proportionally to $V2 \cdot f$, where V is a voltage, and f is the operating frequency. The scope of the DVFS optimization is limited to CPUs. Therefore, computing server components, such as buses, memory, and disks remain functioning at the original operating frequency. On the contrary, the DPM scheme can power down computing servers (that includes all components), which makes such a technique very energy efficient. However, if there occurs a need to power up (powered down) the server, a considerable amount of energy must be consumed compared to the DVFS scheme.

Switches form the basis of the interconnection fabric that delivers job requests to the computing servers for execution. Energy consumption of a switch depends on the: (a) type of switch, (b) number of ports, (c) port transmission rates, and (d) employed cabling solutions. The energy consumed by a switch can be generalized by the following [25]:

$$P_{switch} = P_{chassis} + n_{linecards} \cdot P_{linecard} + \sum_{i=0}^{R} n_{ports} \cdot P_r \quad (1)$$

where $P_{chassis}$ is the power consumed by the switch base hardware, $P_{linecard}$ is the power consumed by an active linecard, and P_r corresponds to the power consumed by an active port (transmitter) running at the rate r. In (1), only the last component, P_r, scales with a switch's transmission rate. This fact limits the benefits of any rate adaptive scheme as the combined consumption of switch transceivers accounts for just 3–15% of switch's total energy consumption [25]. Both $P_{chassis}$ and $P_{linecard}$ do not scale with the transmission rate and can only be avoided when the switch hardware is powered down (given that there is no traffic to be handled by the switch).

Obviously, not all of the switches can dynamically be put to sleep. Each core switch consumes a considerable amount of energy to service large switching capacity. Because of their location within the communication fabric and proper ECMP forwarding functionality, it is advisable to keep the core network switches running continuously at their maximum transmission rates. On the contrary, the aggregation switches service modules, which can be powered down when the module racks are inactive. The fact that on average most of the data centers are utilized around 30% of their compute capacity [23], it makes perfect sense to power down unused aggregation switches. However, such an operation must be performed carefully by considering possible fluctuations in job arrival rates. Typically, it is enough to keep a few computing servers running idle on top of the necessary computing servers as a buffer to account for possible data center load fluctuation [8].

2.3 Data center tasks models

In cloud computing incoming requests are typically generated for such applications like web browsing, instant messaging, or various content delivery applications [9, 35]. The

tasks tend to be more independent, less computationally intensive, but have a strict completion deadline specified in SLA. The majority of such requests can be classified according to the amount of computing and communications they require into three categories:

- **Computationally Intensive Workloads (CIWs)** model High-Performance Computing (HPC) applications and aim at solving advanced computational problems. CIWs demand large amount of computing resources, but produce 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 severs 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.

2.4 Data center network congestion

Utilizing a communication fabric in data centers entails the concept of running multiple types of traffic (LAN, SAN, or IPC) on a single Ethernet-based medium [13]. On one side, the Ethernet technology is cheap, easy to deploy, and relatively simple to manage, on the other side, the Ethernet hardware is less powerful and provisions for small buffering capacity. A typical buffer size in an Ethernet network is in the order of 100 s of KB. However, a typical buffer size of an Internet router is in the order of 100 s of MB [2]. Small buffers and the mix of high-bandwidth traffic are the main reasons for network congestion.

Any of the data center switches may become congested either in the uplink direction or the downlink direction or both. In the downlink direction, the congestion occurs when individual ingress link capacities overcome individual egress link capacities. 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's aggregate uplink capacity.

Congestion (or hotspots) may severely affect the ability of a data center network to transport data. Currently, the Data Center Bridging Task Group (IEEE 802.1) [17] is specifying layer-2 solutions for congestion control, termed IEEE 802.1Qau specifications. The IEEE 802.1Qau specifications introduce a feedback loop between data center switches for signaling congestion. Such a feedback allows overloaded switches to hold off heavy senders from sending with the congestion notification signal. Such a technique may avoid congestion-related losses and keep the data center network utilization high. However, it does not address the root of the problem as it is much more efficient to assign data-intensive jobs to different computing servers in the way that jobs avoid sharing common communication paths. To benefit from such spatial separation in the three-tiered architecture (see Fig. 1), the jobs must be distributed among the computing servers in proportion to the job communication requirements. Data-intensive jobs, like ones generated by video sharing applications, produce a constant bit-stream directed to the end-user as well as communicate with other jobs running in the data center. However, such a methodology contradicts the objectives of energy-efficient scheduling, which tries to concentrate all of the active workloads on a minimum set of servers and involve minimum number of communication resources. This tradeoff between energy-efficiency, data center network congestion, and performance of individual jobs is resolved using a unified scheduling metric presented in the subsequent section.

3 The DENS methodology

The DENS methodology minimizes the total energy consumption of a data center by selecting the best-fit computing resources for job execution based on the load level and communication potential of data center components. The

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 [30] that model data centers as a homogeneous pool of computing servers, the DENS methodology develops a hierarchical model consistent with the state of the art data center topologies. For a three-tier data center, we define DENS metric M as a weighted combination of server-level f_s, rack-level f_r, and module-level f_m functions:

$$M = \alpha \cdot f_s + \beta \cdot f_r + \gamma \cdot f_m \tag{2}$$

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 racks. Higher β values will prioritize computationally loaded racks with low network traffic activity. Higher γ values favor selection of loaded modules. The γ parameter is an important design variable for job consolidation in data centers. Taking into account that $\alpha + \beta + \gamma$ must equal unity, the values of $\alpha = 0.7$, $\beta = 0.2$, and $\gamma = 0.1$ are selected experimentally (see Sect. 4 for details) to provide a good balance in the evaluated three-tier data center topology.

The factor related to the choice of computing servers combines the server load $L_s(l)$ and its communication potential $Q_r(q)$ that corresponds to the fair share of the uplink resources on the ToR switch. This relationship is given as:

$$f_s(l, q) = L_s(l) \cdot \frac{Q_r(q)^{\varphi}}{\delta_r} \tag{3}$$

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 $Q_r(q)$. Similar to the case of computing servers, which was encapsulated in (3), the factors affecting racks and modules can be formulated as:

$$f_r(l, q) = L_r(l) \cdot \frac{Q_m(q)^{\varphi}}{\delta_m} = \frac{Q_m(q)^{\varphi}}{\delta_m} \cdot \frac{1}{n} \sum_{i=1}^{n} L_s(l) \tag{4}$$

$$f_m(l) = L_m(l) = \frac{1}{k} \sum_{j=0}^{k} L_r(l) \tag{5}$$

where $L_r(l)$ is a rack load obtained as a normalized sum of all individual server loads in the rack, $L_m(l)$ is a module load obtained as a normalized sum of all of the rack loads in this module, n and k are the number of servers in a rack and the number of racks in a module respectively, $Q_m(q)$ is proportional to the traffic load at the module ingress switches, and δ_m stands for the bandwidth overprovisioning factor at the module switches. It should be noted that the module-level factor f_m includes only a load-related component l. This is due to the fact that all the modules are connected to the same core switches and share the same bandwidth using ECMP multi-path balancing technology.

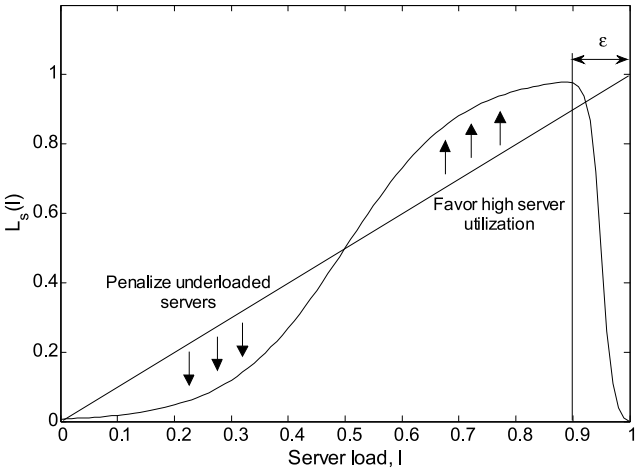

Fig. 2 Computing server selection by DENS metric

The fact that an idle server consumes energy that is almost two-thirds of its peak consumption [8], suggests that an energy-efficient scheduler must consolidate data center jobs on a minimum possible set of computing servers. On the other hand, keeping servers constantly running at peak loads may decrease hardware reliability and consequently affect the job execution deadlines [21]. To address the aforementioned issues, we define the DENS load factor as a sum of two sigmoid functions:

$$L_s(l) = \frac{1}{1 + e^{-10(l - \frac{1}{2})}} - \frac{1}{1 + e^{-\frac{10}{\varepsilon}(l - (1 - \frac{\varepsilon}{2}))}}. \tag{6}$$

The first component in (6) defines the shape of the main sigmoid, while the second component servers as a penalizing function aimed at the convergence towards the maximum server load value (see Fig. 2). The parameter ε defines the size and the incline of this falling slope. The server load l is within the range $[0, 1]$. For the tasks having deterministic computing load, l the server load can be computed as the sum of computing loads of all of the running tasks. Alternatively, for the tasks with predefined completion deadline, the server load l can be expressed as the minimum amount of computational resource required from the server to complete all the tasks right-in-time.

Being assigned into racks, the servers share the ToR switch uplink channels for their communication demands. However, defining a portion of this bandwidth used by a

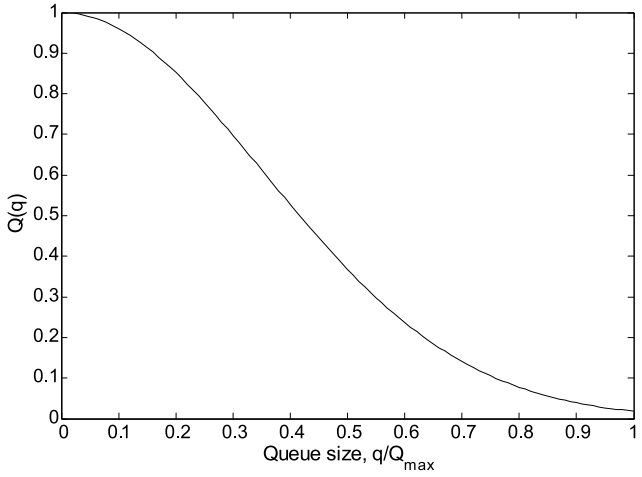

Fig. 3 Queue selection by DENS metric

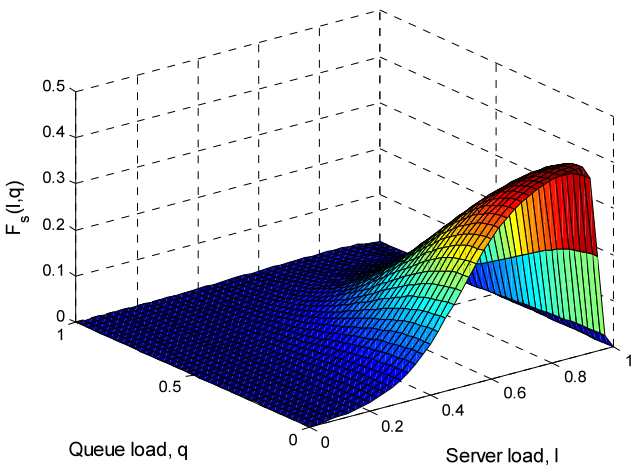

Fig. 4 Server selection by DENS metric according to its load and communicational potential

given server or a flow at the gigabit speeds during runtime is a computationally expensive task. To circumvent the aforementioned undesirable characteristic, both (3) and (4) include a component, which is dependent on the occupancy level of the outgoing queue $Q(q)$ at the switch and scales with the bandwidth over provisioning factor δ.

Instead of relying on the absolute size of the queue, the occupancy level q is scaled with the total size of the queue Q_{max} within the range [0, 1]. The range corresponds to none and full buffer occupancy. By relying on buffer occupancy, the DENS metric reacts to the growing congestion in racks or modules rather than transmission rate variations. To satisfy the aforementioned behavior, $Q(q)$ is defined using inverse Weibull cumulative distribution function:

$$Q(q) = e^{-(\frac{2q}{Q_{max}})^2}. \tag{7}$$

The obtained function, illustrated in Fig. 3, favors empty queues and penalizes fully loaded queues. Being scaled with the bandwidth over provisioning factor δ in (3) and (4) it favors the symmetry in the combined uplink and downlink bandwidth capacities for switches when congestion level is low. However, as congestion grows and buffers overflow, the bandwidth mismatch becomes irrelevant and immeasurable. The (7) is inspired by the Random Early Detection (RED) [12] and Backward Congestion Notification (BCN) [5] technologies.

Figure 4 presents the combined $f_s(l, q)$ as defined in (3). The obtained bell-shaped function favors selection of servers with the load level above average located in racks with the minimum or no congestion.

The following algorithm is used to compute the DENS metric during runtime:

DENS Algorithm

Initialization
- **set** weighted coefficient $\alpha = 0.7$, $\beta = 0.2$, $\gamma = 0.1$
- **set** proportional coefficient $\varphi = 2$
- **get** server load l
- **get** queue size at access and aggregate switches q

Server selection
- **FOR** all servers **DO**
  - **compute** server load $L_s(l)$, rack load $L_r(l)$, and module load $L_m(l)$
  - **compute** communications potentials of rack $Q_r(q)$ and module $Q_m(q)$
  - **compute** metric factors related to servers $f_s(l, q)$, racks $f_r(l, q)$, and modules $f_m(l)$
  - **compute** DENS metric as a weighted sum of $f_s(l, q)$, $f_r(l, q)$, and $f_m(l)$
- **ENDFOR**
- Select server with highest DENS metric

4 Performance evaluation

4.1 GreenCloud simulator

For performance evaluation purposes, the proposed DENS methodology was implemented in the GreenCloud simulator [19, 20]. GreenCloud is a cloud computing simulator developed by us to capture data center communication processes at the packet level. It is developed as an extension of network simulator Ns2 [34] allowing it to exploit realistic TCP/IP processes in a large variety of network scenarios. GreenCloud offers users a detailed fine-grained modeling of

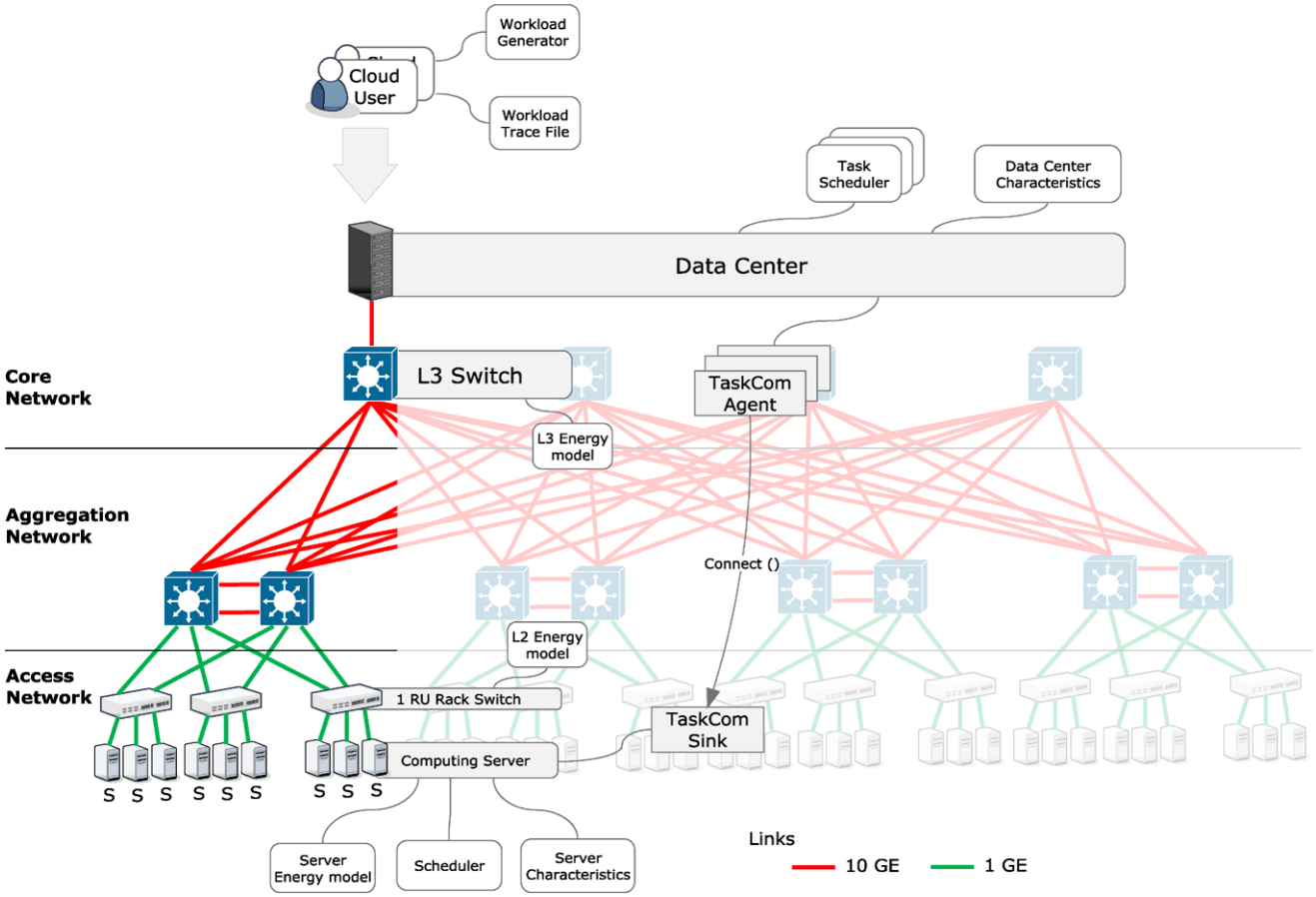

Fig. 5 Structure of the GreenCloud simulator

the energy consumed by the elements of a data center, such as servers, switches, and communication links.

Moreover, GreenCloud offers a thorough investigation of workload distributions. Furthermore, a specific focus is devoted to 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. Implemented energy-efficient optimization techniques include DVFS [27] and DPM [4] approaches. In addition, a set of energy monitoring tools operating in data center servers, switches, and other components is included.

Figure 5 presents the structure of the GreenCloud simulator mapped onto the simulated three-tier data center architecture. Computing servers implement single core processing model that have a preset on a processing power in MIPS (million instructions per second) or FLOPS (floating point operations per second), have their own memory and disk storage resources, and can follow different task scheduling policies. Network switches and links provide communication fabric for workloads distribution. Their characteristics may vary depending on the technology used for interconnection. For example, for data rates of up to 1 Gb/s energy profiles of network links and switches' transceivers are driven by twisted pair technology while for greater rates of 10 Gb/s optical multimode transmitters are used.

For workload execution GreenCloud employs deadline-based model, i.e. each task should be able to perform a specified amount of computations and transmit a given amount of data before a specified deadline for successful completion. The deadline aims at introducing Quality of Service (QoS) parameters. On the communicational side each workload is characterized by the size of the workload which should be transmitted to the pool of servers for the workload execution as well as size of internal and size of external to the data center transfers.

4.2 Simulation scenario

A three-tier tree data center topology comprised of 1536 servers arranged into 32 racks each holding 48 servers, served by 4 core and 8 aggregation switches (see Fig. 1), was used in all simulation experiments. We used 1 GE links for interconnecting servers in the inside racks while 10 GE links were used to form a fat-tree topology interconnecting

access, aggregation, and core switches. The propagation delay on all of the links was set to 10 ns.

The workload generation events are exponentially distributed in time to mimic typical process of user arrival. As soon as a scheduling decision is taken for a newly arrived workload it is sent over the data center network to the selected server for execution. The size of the workload is equal to 15 KB. Being fragmented, it occupies 10 Ethernet packets. During execution, the workloads produce a constant bitrate stream of 1 Mb/s directed out of the data center. Such a stream is designed to mimic the behavior of the most common video sharing applications. To add uncertainties, during the execution, each workload communicates with another randomly chosen workload by sending a 75 KB message internally. The message of the same size is also sent out of the data center at the moment of task completion as an external communication.

The average load of the data center is kept at 30% that is distributed among the servers using one of the three evaluated schedulers: (a) DENS scheduler proposed in Sect. 3 of this paper, (b) Green scheduler performing the best-effort workload consolidation on a minimum set of servers, and (c) a round-robin scheduler which distributes the workloads equally.

The servers left by the schedulers idle are powered down using DPM technique to reduce power consumption. A similar technique is applied to the unused network switches in aggregation and access networks. The core network switches remain always operational at the full rate due to their crucial importance in communications.

4.3 Simulation results

Figure 6 presents the server load distribution for all three of the evaluated schedulers. Figure 7 reports a combined uplink load at the corresponding rack switches. The Green scheduler consolidates the workload leaving the most (around 1016 on average) servers idle in the evaluated data center. These servers are then powered down. However, the load of the loaded servers (left part of the chart) is kept close to the maximum and no consideration of network congestion levels and communication delays is performed. As a consequence, a number of workloads scheduled by the Green scheduler produces a combined load exceeding ToR switch forwarding capacity and causes network congestion. The round-robin scheduler follows a completely opposite policy. It distributes computing and communicational loads equally among servers and switches; thereby the network traffic is balanced and no server is overloaded. However, the drawback is that no server or network switch is left idle for powering down, making the round-robin scheduler as the least energy-efficient.

The DENS methodology achieves the workload consolidation for power efficiency while preventing computing

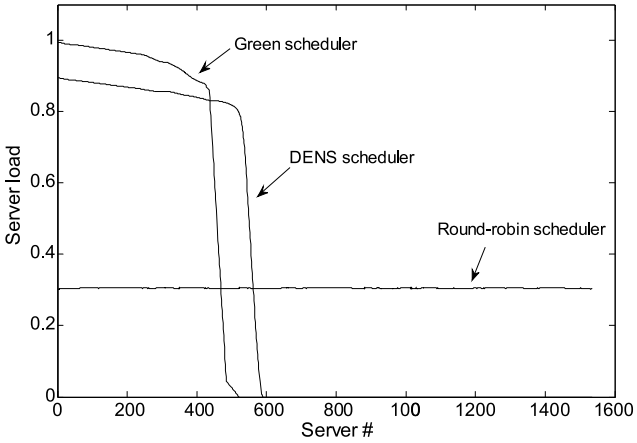

Fig. 6 Server workload distribution performed by DENS, green, and round-robin schedulers

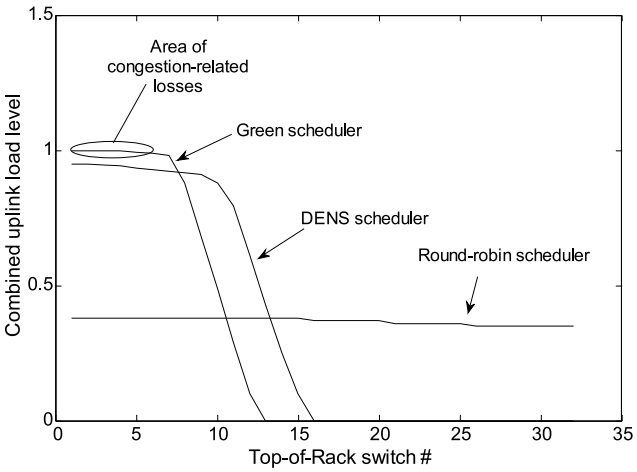

Fig. 7 Combined uplink traffic load at the rack switches

servers and network switches from overloading. In fact, the average load of an operating server is around 0.9 and the average load of the rack switch uplink is around 0.95. Such load levels ensure that no additional delays in job communications are caused by network congestion. However, this advantage comes at a price of a slight increase in the number of running servers. On average, DENS scheduler left 956 servers as opposed to 1016 servers left idle by the Green scheduler.

To explore the uplink load in detail, we measured the traffic statistics at the most loaded switch ToR switch (the leftmost in Fig. 7). Figure 8 presents a combined ToR switch uplink load evolution, while Fig. 9 presents the uplink queue evolution at the same switch for the first 15 seconds of simulation time. Under the Green scheduler, the link is constantly overloaded and the queue remains almost constantly full, which causes multiple congestion losses. All queues were limited to 1000 Ethernet packets in our simulations. Under the DNS scheduler, the buffer occupancy is mostly below

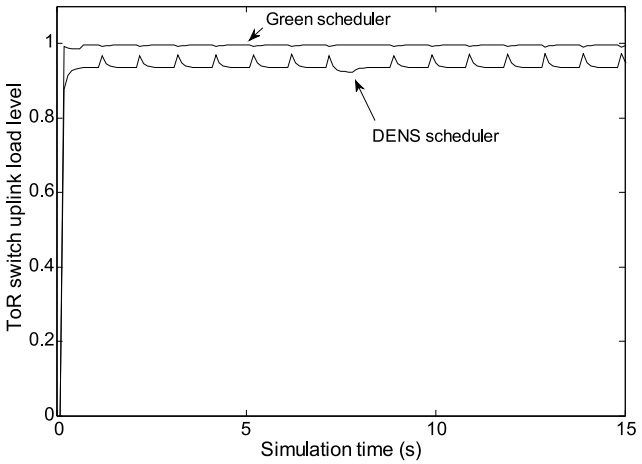

Fig. 8 ToR switch uplink load

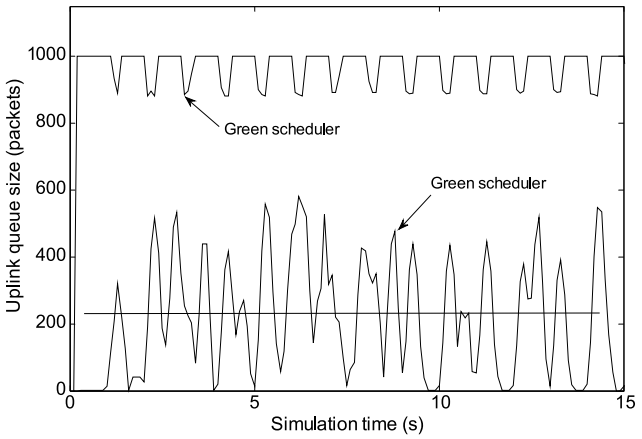

Fig. 9 ToR switch uplink buffer occupancy

the half of its size with an average of 213 packets, displayed with a dashed line in Fig. 9. At certain instances of time the queue even remains empty having no packets to send. This fact results in a slightly reduced uplink utilization level of 0.95.

Table 1 compares the impact of different scheduling policies on the level of data center energy consumption. The data is collected for an average data center load of 30%. The most energy inefficient is a round-robin scheduler. It does not allow any of the servers or network switches to be powered down for the whole duration of data center operation.

The Green scheduler is the most efficient. It releases around two-thirds of servers and network switches, which considerably reduces the energy consumption levels. With the Green scheduler, the data center energy consumption is slashed in half compared to when a round-robin scheduler is utilized. The DENS methodology when compared to the Green scheduler adds around: (a) 4% to the total data center consumption, (b) 3% in servers' energy consumption, and (c) 1% in switches' energy consumption. This slight

Table 1 Data center energy consumption

Parameter	Power consumption (kW·h)		
	Round Robin scheduler	Green scheduler	DENS scheduler
Data center	417.5 K	203.3 K (48%)	212.1 K (50%)
Servers	353.7 K	161.8 K (45%)	168.2 K (47%)
Network switches	63.8 K	41.5 K (65%)	43.9 K (68%)

increase in energy consumption is justified by the need of additional computing and communicational resources, detected by DENS methodology, and required for keeping the quality of job execution at the desired level. In contrast to the Green scheduler, DENS methodology uses network awareness to detect congestion hotspots in the data center network and adjust its job consolidation methodology accordingly. It becomes particularly relevant for data intensive jobs which are constrained more by the availability of communication resources rather than by the available computing capacities.

5 Conclusions

This paper underlines the role of communication fabric in data center energy consumption and presents a methodology, termed DENS, that combines energy-efficient scheduling with network awareness. The DENS methodology balances the energy consumption of a data center, individual job performance, and traffic demands. The proposed approach optimizes the tradeoff between job consolidation (to minimize the amount of computing servers) and distribution of traffic patterns (to avoid hotspots in the data center network). DENS methodology is particularly relevant in data centers running data-intensive jobs which require low computational load, but produce heavy data streams directed to the end-users.

The simulation results obtained for a three-tier data center architecture underline DENS operation details and its ability to maintain the required level of QoS for the end-user at the expense of the minor increase in energy consumption. Future work will focus on the implementation and testing of DENS methodology in realistic setups using testbeds. The design and specification of DENS metric is tight to the underlining data center architecture. In this paper it was the most widely used nowadays three-tier architecture. However, the adaptation of DENS approach to other existing and upcoming data center architectures is already on-going.

Acknowledgements The authors would like to acknowledge the funding from National Research Fund, Luxembourg in the framework of GreenIT project (C09/IS/05) and Marie Curie Actions of the European Commission (FP7-COFUND), the funding from VTT

Technical Research Centre of Finland in the framework of Intelligent Telecommunication Systems with Enhanced Cognitive Processing (ITSE) project as well as the research fellowship provided by the European Research Consortium for Informatics and Mathematics (ERCIM).

References

1. Al-Fares, M., Radhakrishnan, S., Raghavan, B., Huang, N., Vahdat, A.: 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
2. Alizadeh, M., Atikoglu, B., Kabbani, A., Lakshmikantha, A., Pan, R., Prabhakar, B., Seaman, M.: Data center transport mechanisms: congestion control theory and IEEE standardization. In: Annual Allerton Conference on Communication, Control, and Computing, September 2008
3. Beloglazov, A., Buyya, R.: Energy efficient resource management in virtualized cloud data centers. In: IEEE/ACM International Conference on Cluster, Cloud and Grid Computing (CCGrid), pp. 826–831, May 2010
4. Benini, L., Bogliolo, A., De Micheli, G.: A survey of design techniques for system-level dynamic power management. IEEE Trans. Very Large Scale Integr. **8**(3), 299–316 (2000)
5. Bergamasco, D., Baldini, A., Alaria, V., Bonomi, F., Pan, R.: Methods and devices for backward congestion notification. US Patent 2007/0081454
6. Berl, A., Gelenbe, E., Di Girolamo, M., Giuliani, G., De Meer, H., Dang, M.Q., Pentikousis, K.: Energy-efficient cloud computing. The Computer Journal **53**(7), 1045–1051 (2009)
7. Brown, R., et al.: Report to congress on server and data center energy efficiency: public law 109-431. Lawrence Berkeley National Laboratory, Berkeley, 2008
8. Chen, G., He, W., Liu, J., Nath, S., Rigas, L., Xiao, L., Zhao, F.: 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, 2008
9. Chen, M., Zhang, H., Su, Y., Wang, X., Jiang, G., Yoshihira, K.: EffectiveVM sizing in Virtualized Data Centers. In: 12th IFIP/IEEE International Symposium on Integrated Network Management (IM), Dublin, Ireland, May 2011
10. Cisco Data Center Infrastructure 2.5 Design Guide. Cisco Press, March 2010
11. Fan, X., Weber, W.-D., Barroso, L.A.: Power provisioning for a warehouse-sized computer. In: Proceedings of the ACM International Symposium on Computer Architecture, San Diego, CA, June 2007
12. Floyd, S., Jacobson, V.: Random early detection gateways for congestion avoidance. IEEE/ACM Trans. Netw. **1**(4), 397–413 (1993)
13. Garrison, S., Oliva, V., Lee, G., Hays, R.: Ethernet alliance: data center bridging, November 2008
14. Gartner Group: available at: http://www.gartner.com/ (2011)
15. Guo, C., Wu, H., Tan, K., Shiy, L., Zhang, Y., Luz, S.: DCell: a scalable and fault-tolerant network structure for data centers. In: ACM SIGCOMM, Seattle, Washington, USA 2008
16. Guo, C., Lu, G., Li, D., Wu, H., Zhang, X., Shi, Y., Tian, C., Zhang, Y., Lu, S.: BCube: A high performance, server-centric network architecture for modular data centers. In: ACM SIGCOMM, Barcelona, Spain 2009
17. IEEE 802.1 Data Center Bridging Task Group, available at: http://www.ieee802.org/1/pages/dcbridges.html (2011)
18. IEEE std 802.3ba-2010, Media access control parameters, physical layers and management parameters for 40 Gb/s and 100 Gb/s Operation, June 2010
19. Kliazovich, D., Bouvry, P., Audzevich, Y., Khan, S.U.: GreenCloud: a packet-level simulator of energy-aware cloud computing data centers. In: IEEE Global Communications Conference (GLOBECOM), Miami, FL, USA, December 2010
20. Kliazovich, D., Bouvry, P., Khan, S.U.: GreenCloud: a packet-level simulator of energy-aware cloud computing data centers. J. Supercomp., special issue on Green Networks, 2011
21. Kopparapu, C.: Load Balancing Servers, Firewalls, and Caches Wiley, New York (2002)
22. Li, B., Li, J., Huai, J., Tianyu, W., Li, Q., Zhong, L.: EnaCloud: an energy-saving application live placement approach for cloud computing environments. In: IEEE International Conference on Cloud Computing, Bangalore, India, 2009
23. Liu, J., Zhao, F., Liu, X., He, W.: 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
24. Mahadevan, P., Sharma, P., Banerjee, S., Ranganathan, P.: Energy aware network operations. In: IEEE INFOCOM Workshops, pp. 1–6 (2009)
25. Mahadevan, P., Sharma, P., Banerjee, S., Ranganathan, P.: A power benchmarking framework for network devices. In: Proceedings of the 8th International IFIP-TC 6 Networking Conference, Aachen, Germany, 11-15 May 2009
26. Meng, X., Pappas, V., Zhang, L.: Improving the scalability of data center networks with traffic-aware virtual machine placement. In: IEEE INFOCOM, San Diego, California, March 2010
27. Pouwelse, J., Langendoen, K., Sips, H.: Energy priority scheduling for variable voltage processors. In: International Symposium on Low Power Electronics and Design, pp. 28–33 (2001)
28. Raghavendra, R., Ranganathan, P., Talwar, V., Wang, Z., Zhu, X.: No "power" struggles: Coordinated multi-level power management for the data center. In: APLOS (2008)
29. Shang, L., Peh, L.-S., Jha, K.N.: 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 (2003)
30. Song, Y., Wang, H., Li, Y., Feng, B., Sun, Y.: Multi-tiered on-demand resource scheduling for VM-based data center. In: IEEE/ACM International Symposium on Cluster Computing and the Grid (CCGRID), pp. 148–155, May 2009
31. Stage, A., Setzer, T.: 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, May 2009
32. Tang, Q., Gupta, S.K.S., Varsamopoulos, G.: Energy-efficient thermal-aware task scheduling for homogeneous high-performance computing data centers: a cyber-physical approach. IEEE Trans. Parallel Distrib. Syst. **19**(11), 1458–1472 (2008)
33. Thaler, D., Hopps, C.: Multipath issues in unicast and multicast nexthop selection. In: Internet Engineering Task Force Request for Comments 2991, November 2000
34. The Network Simulator Ns2: available at: http://www.isi.edu/nsnam/ns/ (2011)
35. Verma, A., Dasgupta, G., Nayak, T., De, P., Kothari, R.: Server workload analysis for power minimization using consolidation. In: USENIX Annual Technical Conference (USENIX'09) (2009)

Dzmitry Kliazovich is an AFR Research Fellow at the Faculty of Science, Technology, and Communication of the University of Luxembourg. He holds an award-winning Ph.D. in Information and Telecommunication Technologies from the University of Trento, Italy. Prior to joining the University of Luxembourg he was an ERCIM Research Fellow at the VTT Technical Research Centre of Finland and a Scientific Advisor for Wireless Communications at the Create-Net Research Centre, Italy.
In 2005 he was a Visiting Researcher at the Computer Science Department of the University of California at Los Angeles (UCLA). A year later he joined Nokia Siemens Networks with the responsibility of starting up a research direction focusing on 3G Long-Term Evolution (LTE).
Dr. Kliazovich is a holder of several scientific awards including fellowship grants provided by the European Research Consortium for Informatics and Mathematics (ERCIM), the IEEE Communication Society, Italian Ministry of Education, and the University of Trento. His work on energy-efficient scheduling in cloud computing environments received Best Paper Award at the IEEE/ACM International Conference on Green Computing and Communications (GreenCom) in 2010.
Dr. Kliazovich is the author of more than 50 research papers, Editorial Board Member of the IEEE Communications Surveys and Tutorials, Features Editor of the ICaST magazine, contributing member of the IEEE ComSoc Technical Committee on Communication Systems Integration and Modeling (CSIM), and the organizer of the CSIM flagship workshop CAMAD in 2006, 2010, and 2011. His main research activities are in the field of energy efficient communications and next-generation networking.

Pascal Bouvry earned his undergraduate degree in Economical & Social Sciences and his Master degree in Computer Science with distinction (‘91) from the University of Namur, Belgium. He went on to obtain his Ph.D. degree (’94) in Computer Science with great distinction at the University of Grenoble (INPG), France. His research at the IMAG laboratory focussed on Mapping and scheduling task graphs onto Distributed Memory Parallel Computers. Dr. Bouvry is currently heading the Computer Science and Communications (CSC) research unit of the Faculty of Sciences, Technology and Communications of Luxembourg University, and serving as Professor. Pascal Bouvry is also treasurer & member of the administration board of CRP-Tudor, and member of various scientific committees and technical workgroups.

Samee Ullah Khan is Assistant Professor of Electrical and Computer Engineering at the North Dakota State University, Fargo, ND, USA. Prof. Khan has extensively worked on the general topic of resource allocation in autonomous heterogeneous distributed computing systems. As of recent, he has been actively conducting cutting-edge research on energy-efficient computations and communications. A total of 110 (journal: 40, conference: 50, book chapter: 12, editorial: 5, technical report: 3) publications are attributed to his name. For more information, please visit: http://sameekhan.org/.