Energy-Aware Scheduling with Computing and Data Consolidation Balance in 3-tier Data Center

Manuel Combarro, Andrei Tchernykh
CICESE Research Center
Ensenada, México
cmanuel@cicese.edu.mx, chernykh@cicese.mx

Alexander Drozdov
Moscow Institute of Physics and Technology
Moscow, Russia
alexander.y.drozdov@gmail.com

Dzmitry Kliazovich
University of Luxembourg
Luxembourg, Luxembourg
dzmitry.kliazovich@uni.lu

Gleb Radchenko
South Ural State University
Chelyabinsk, Russia
gleb.radchenko@susu.ru

Abstract— **Energy consumption represents a large percentage of the operational expenses in data centers. Most of the existing solutions for energy-aware scheduling are focusing on job distribution and consolidation between computing servers, while network characteristics are not considered. In this paper, we propose a model of power and network-aware scheduling that can be tuned to achieve energy-savings, through job consolidation and traffic load balancing. We describe a methodology to find the best tuning of the Adjustable Scheduler.**

Keywords-data center, energy efficient, parallel machines, scheduling algorithms

I. INTRODUCTION

Data centers play a very important role in cloud computing hosting thousands of computing servers for providing a virtually unlimited computational and storage services [1]. They require a tremendous amount of energy to operate. The cost of the energy consumed by the servers may be greater than the cost of the equipment itself [1], [2].

In 2010, data centers consumed about the 1.5% of the electricity in the world [3]. In 2012, their energy consumption was about 15% of global ICT consumption and, it is expected, to be increased between 5 and 10% in 2017 [4]. By 2020, it is estimated that European data centers consume nearly 93 TWh [5]. Almost the 75% of the consumption is due to the IT equipment and cooling system; the rest is the energy distribution and data center operation lost. Annually, in terms of CO_2 emissions, the energy consumption is equivalent to more than 50 million of metric tons [6].

In this paper, we propose a scheduler, named Adjustable-Scheduler (AS), to minimize energy consumption in data centers. It is based on the configurable ACCURATE scheduler [7]. It can be configured for different types of workloads: computation-intensive, communication-intensive and balanced.

The rest of the paper is structured as follows: Section II presents background on data center and the motivation of the study. Section III reviews related works on energy-aware scheduling and network-aware scheduling. Section IV presents the problem definition and described AS. Section V proposes experimental setup and describes performance evaluation methodology. Section VI concludes the paper.

II. MOTIVATION AND BACKGROUND

In data centers, the scheduling problem is to allocate a finite set of resources to incoming Virtual Machines (VM), jobs or tasks [8], [9]. To assign resources (CPU cycles, RAM, storage, bandwidth) schedulers have to consider a set of constraints and requirements to be accomplished, based on the Service Level Agreement (SLA) [10], such as: minimum provided computation power, bandwidth, due date, storage capacity, etc.

Schedulers can have different objectives, such as: makespan, load balancing, Quality of Service (QoS), energy consumption or a combination of them [11], [12]. It has been given a particular interest to optimize energy consumption of computing resource, with the purpose of reduce data center energy consumption through software solutions.

The main focus of traditional approaches to minimize energy consumption is to consolidate VMs on a minimum number of physical resources. Another important aspect is the data communication. Computing consolidation and data balance are conflicting goals. In realistic scheduling systems they should be considered together [13]. Only few works take into account the data center network characteristics to developed energy-aware scheduling strategies [14], [6], [15], [7]. The state of the data center network could affect job response time, packet loss, deadlines, violations of SLA, reduced quality of service, etc.

It has been showed that workloads in clouds are highly heterogeneous [16], [17]. Data centers can receive a great range of jobs, from High-Performance Computing (HPC) to Data-Intensive.

Fig. 1 [7] shows the data center topology considered in this paper, the three-tier topology [18], which is the most popular data center topology [19].

978-1-5090-4553-2/16 $31.00 © 2016 IEEE
DOI 10.1109/EnT.2016.13

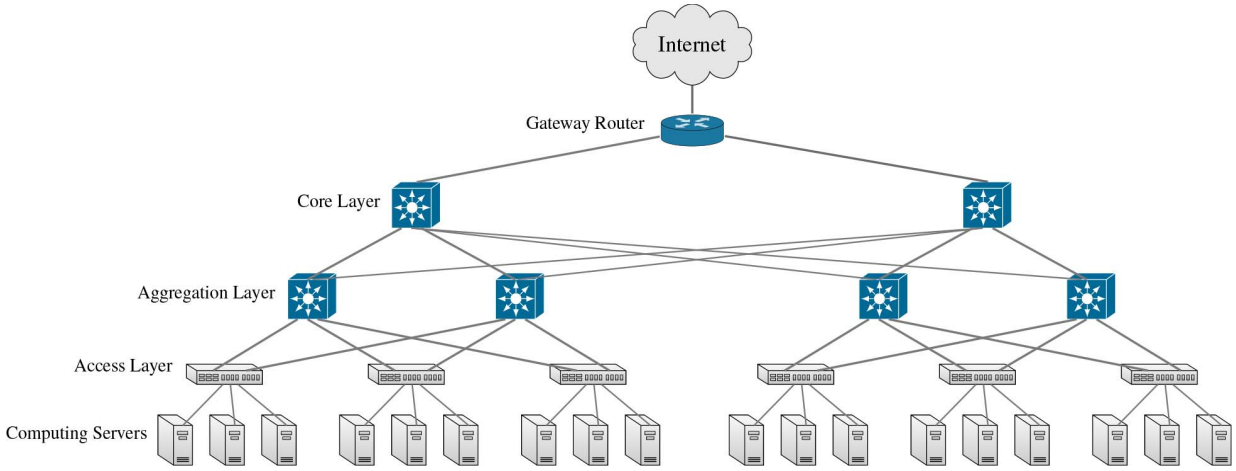

Figure 1. Three-tier data center architecture.

III. Related Work

There are several studies on energy efficient schedulers that take into account deadline constraints, but do not consider the conditions of the data center network.

ECTC and MaxUtil [20] use consolidation to reduce energy consumption without violation of deadline constraints.

In GCSM (Green Cloud Scheduling Model) [21], and in MESF (Most Efficient Server First) [22], a task is allocated to the most efficient node (computational resource) that can finish it before the deadline. These strategies save energy by putting into sleep modes idle servers. In GCSM, the model proposes that the more energy is consuming a node the more inefficient it is. Both heuristics set a threshold and if the consumption is above it the node will be discarded, the task will be only allocated on it if there is no other node that can complete the task before the deadline.

In [23], the objective is to minimize the energy consumption and SLA violations. The SLA establishes the minimum MIPS, RAM and bandwidth. The energy saving is achieved through VM consolidation. The idle computational resources are putted into the sleep mode. VM migrations are done if the host utilization is above an upper threshold, if it is below a lower threshold (in this case, all the host's VM are reallocated and the host is put into sleep mode), if a VM has an intensive communication with other VM located in a different host, and, lastly, if temperature exceeds a value. In [24], the authors present an improvement of this strategy, by taking into account the SLA violations related with MIPS.

Few energy-aware strategies consider network characteristics.

DENS – Data Center Energy-Efficient Network-Aware Scheduling, proposed in [14], optimizes the relation between jobs consolidation and traffic distribution (avoiding hot spots). It is used a lineal energy model and idle servers are putted into sleep mode to save energy.

e-STAB [7] is very similar to DENS. They differ in how the traffic is analyzed. In e-STAB, the traffic can be balanced and the criteria to select servers are little different. In both works, the server with highest communication capacity and with more utilization is chosen to execute the task.

HEROS – Energy-Efficient Load Balancing for Heterogeneous Data Centers, proposed in [15], is based on DENS and e-STAB. It is designed to work in a heterogeneous environment, at the contrary to DENS and e-STAB that work on a homogenous one. HEROS can be applied to topologies different to three-tier, which is not effective in DENS and e-STAB.

ACCURATE – Adaptive Computing and Communication Resource Allocation Scheme for Cloud Computing Data Centers proposed in [7] is similar to DENS and e-STAB but it can be configured and the score function takes into account another metric: energy proportionality factor. This factor represents how increase the power consumption with respect to load increase for any utilization level. The major contribution is to introduce a configurable scheduler that can be dynamically tuned for different types of workloads, and different interests: energy saving, SLA violations and data transmission distribution.

IV. Adjustable Scheduler

A. Infrastructure Model

Let us consider a set of n identical tmachines. Each machine m is described by the tuple $(l_m^{cp}(t), W_m^{cp}(t))$ where, $l_m^{cp}(t)$ is the CPU load at time t, and $W_m^{cp}(t) = F_m^{cp}(l_m^{cp}(t))$ is the power at time t. We suppose that power depends on CPU load at time t as CPU load is the most important factor to determine the energy consumption of a server [25]. $F_m^{cp}(l_m^{cp}(t))$ is a function that computes the power consumed in a time t based on CPU load.

The data center has a three-tier topology. It is modelled by a graph $G(V, E)$, where V is composed by the machines and switches, and E by the communication links. S is a set of switches. Every server can reach the gateway router through a set of paths P. A path $p_k \to G$ is a non-repeated sequence of nodes that connects the machine m_k with G. It is represented as follows $p_k \to G = (m_k, s_1, s_2, s_3, G)$, where s_1, s_2, s_3 are the switches in access, aggregation and core layers respectively. All the path has the following tuple $\left(l_p^{cm}(t), W_p^{cm}(t)\right)$, where $W_p^{cm}(t) = F_{ac}^{cm}\left(l_{ac}(t)\right) + F_{ag}^{cm}\left(l_{ag}(t)\right) + F_{co}^{cm}\left(l_{oc}(t)\right)$ is the power consumed at time t and is function of the load of s_1, s_2, s_3 at time t. $l_p^{cm}(t)$ is the path's load and it is computed as the rate of bits sent/received by the machine divided between the end-to-end transmission bandwidth of the path; the former is equal to the link with less effective bandwidth. The value of the path load is in the range from 0 to 1. Paths with low load are preferred for job allocation to maintain the network balanced.

B. Job Model

Every job is described by the tuple $(r_j, l_j^{cp}, l_j^{cm})$, where r_j is the release time, l_j^{cp} and l_j^{cm} are the computational (MIPS) and communicational (Mbps) requirements, respectively. The last two define the requirements of the QoS to satisfy the SLA. The job description is fully known only when the job has been submitted.

C. Energy Model

The energy consumed to operate the IT equipment is E^{IT} and is computed as:

$$E^{IT} = E^{cp} + E^{cm} \tag{1}$$

where E^{cp} and E^{cm} are the energy consumed by the servers and the switches respectively. Both values are computed for the time interval $[0, T_{max}]$.

IT devices have different power consumption profiles. To represent this diversity, we use the Energy Proportionality Coefficient (EPC) [26]. Each device i has an EPC_i and it is computed as:

$$\text{EPC}_i = \int_0^1 \sin 2\alpha_i(l)\, dl = \int_0^1 \frac{2\tan\alpha_i(l)}{1+\tan^2\alpha_i(l)}\, dl \tag{2}$$

where $\tan\alpha_i(l) = dW/dl$ represents the deviation of the power function from the ideal curve having considered the load l to be normalized in the range [0,1].

D. Scheduling Criteria

The objective is to minimize the energy consumption and SLA violations. This multiobjective problem is addressed using a weighted score function f. This function has three criteria: instantaneous load of computing servers, paths load and EPC. Job j is allocated to the suitable machine i with the highest score. The machine i is selected from M_a set, which is the set of servers that have the available MIPS and bandwidth (Mbps) required by job j. If no servers meet the requirements, then M_a is composed by the machine(s) with the highest available MIPS.

The score function is computed for each server i as follow:

$$f_i = \alpha f_i^{cp} + (1-\alpha) f_i^{cm}, \tag{3}$$

where: f_i^{cp} and f_i^{cm} are the computation and communication components. Both values are between 0 and 1.

- α is a balancing coefficient that gives more importance to f_i^{cp} or f_i^{cm}, $\alpha \in [0,1]$.

1) Computational equipment: The score due to computational equipment is computed as follows:

$$f_i^{cp} = \beta \bar{f}_i + (1-\beta) EPC_i^{cp}, \tag{4}$$

where:

- EPC_i^{cp} is the energy proportionality coefficient of the machine i
- β is a balancing coefficient that shows a relative importance between the two terms, $\beta \in [0,1]$.
- $\bar{f}_i$ is a function of the server load $l_i^{cp}(t)$. It is based on DENS strategy. It is calculated as follows:

$$\bar{f}_i\left(l_i^{cp}(t)\right) = \frac{1}{1+e^{-10\left(l_i^{cp}(t)-\frac{1}{2}\right)}} - \frac{1}{1+e^{-\frac{10}{k}\left(l_i^{cp}(t)-\left(1-\frac{k}{2}\right)\right)}}, \tag{5}$$

where $k \in [0,1]$. The first part of the equation is increased with the workload. The second term, instead, makes the score decreasing with the increase of load. As a result, it prevents selection of overloaded servers during jobs consolidation. The parameter k allows fine tuning of the maximum server load defining overloaded servers.

2) Communication component: For each server i we have:

$$f_i^{cm} = \delta\left(1 - \frac{1}{1+e^{-10 l_i^{\rho}}}\right) + (1-\delta) EPC_i^{cm}, \tag{6}$$

where:

- l_i^{ρ} is the load of the path $p_i \to G$.
- EPC_i^{cm} is the overall energy proportionality of the path $p_i \to G$ and is computed as:

$$EPC_i^{cm} = \frac{1}{n}\sum_{k=1}^{n} EPC_{s_k},$$

where: EPC_{s_i} is the EPC value of the switch $s_k \in p_i \to G$. n is number of switches in the path.

The first term of f_i^{cm} is a function of the load of the path from server *i* to the data center gateway and is based on [14]. The second term is the overall energy proportionality on that path.

Fig. 2 [7] illustrates a possible example of scoring function f_i with $\alpha = 0.5$.

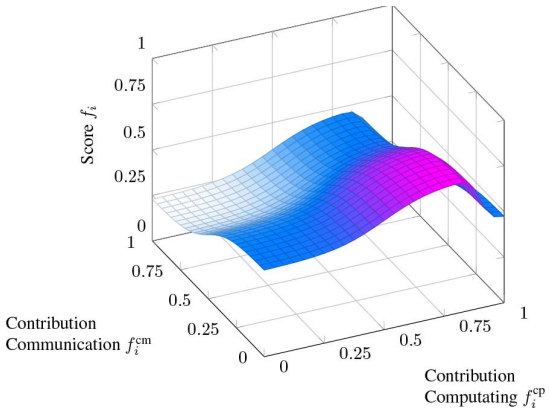

Figure 2. Example of score function f for $\alpha = 0.5$.

V. Performance Evaluation

A. GreenCloud Simulator

We evaluate the performance of strategies using the cloud computing data center simulator: GreenCloud [27]. This simulator is energy-aware and network-aware. It offers a detail model of the energy consumed by data center devices such as: computing servers, switches and communication links. It supports three-tier topology and implements DVFS [28] and DPM [29] to minimize the energy consumption of servers and switches.

GreenCloud is an extension of the well-known packet-level network simulator Ns2 [30], which considers TCP/IP protocols for packet transmission. Hence, communications are simulated in a more realistic way and this is what makes this simulator different from the others.

B. Workload

The types of workloads are characterized mainly by the computational and communicational components of jobs. The computational component defines the amount of computing that must be executed per second and it is given in MIPS. The communicational component indicates the size of data transfers that goes from data center gateway to the server (input communication) and from server to data center gateway (output communication).

The three types of workloads that we use are computational-intensive, communication-intensive, and balanced. The first one requires a high amount of MIPS but almost no communication, the second one requires a low amount of MIPS but high degree of communications. The last one requires computation and communication in the same proportion.

C. Performance evaluation methodology

The scheduling problem we face is bi-objective, we want to minimize the energy consumption and SLA violations. In this section, we describe how to compare the strategies taking into account both criteria.

First, we calculate the degradation in performance of each strategy for each metric. This is done relatively to the best performing strategy for the metric, as follows: $\frac{strategy\ criterion\ value}{best\ found\ criterion\ value} - 1$. Next, we calculate the approximation of the Pareto front.

When we have the approximated Pareto front for each strategy, the fronts are compared using the set cover metric [31] that calculates the proportion of solutions in B, which are dominated by solutions in A.

$$SC(A,B) = \frac{|\{b \in B | \exists a \in A : a \preccurlyeq b\}|}{|B|} \quad (7)$$

A metric value $SC(A,B) = 1$ means that all solutions of B are dominated by at least one solution of A, and $SC(A,B) = 0$ indicates that any solution of B is dominated by any solution of A. This way, the larger $SC(A,B)$, the better the Pareto front A with respect to B. Since the dominance operator is not symmetric, $SC(B,A)$ is not necessarily equal to $1 - SC(A,B)$, and both $SC(A,B)$ and $SC(B,A)$ have to be computed for understanding how many solutions of A are covered by B and vice versa.

Based on $SC(A,B)$ and $SC(B,A)$, we calculate two ranks. In the first rank, all the strategies are ordered by their dominance with respect to the others Pareto fronts. In the second rank, all the strategies are ordered by how much they are dominated by the others Pareto fronts. As higher the first one and lower the second one the better is the strategy. This two ranks are combined to obtain the best strategy.

D. Setup Parameters

The considered data center architecture is a three-tier topology, with 1536 servers grouped in 64 racks, 16 aggregation switches and 8 core switches. Links connecting servers to top of rack switches are 1 Gbps, and links between access network and aggregation network, and between aggregation network and core network are 10 Gbps. The data center load will be 50% of its capacity.

1) Scheduler configurations: Each combination of α, β, δ is a configuration of the AS.

The variations of α, β, δ values are as follow:

- $\alpha \rightarrow 0.2, 0.4, 0.6, 0.8$
- $\beta \rightarrow 0, 0.25, 0.5, 0.75, 1$
- $\delta \rightarrow 0, 0.25, 0.5, 0.75, 1$

When $\alpha = 1$, the values of δ do not affect the score function. The same happens with β, when $\alpha = 0$. So for $\alpha = 1$ and $\alpha = 0$, we have:

- $\alpha = 1, \beta \rightarrow 0, 0.25, 0.5, 0.75, 1$
- $\alpha = 0, \delta \rightarrow 0, 0.25, 0.5, 0.75, 1$

This gives a total of $4 \times 25 + 10 = 110$ scheduling strategies that has to be compared. Due the execution cost of one solution and the amount of executions necessary to obtain valid statistical values, we limit the search to 110 combinations. We run each strategy 30 times for computation-intensive and communication-intensive workloads. This gives 6600 executions ($110 \cdot 30 \cdot 2 = 6600$).

2) Performance Comparisson: We also compare it with Green scheduler, Round-Robin scheduler [14], and DENS. The first one is a greedy heuristics that allocates jobs to the most loaded server, it is an energy-efficient strategy. The

second one allocates jobs using round-robin method. It is a network-balanced strategy.

VI. Conclusions

In this paper, we introduce a model of adjustable scheduling strategy that can be configured to minimize energy consumption through job consolidation and, at the same time, to balance traffic loads of the data center. We describe the procedure of tuning AS for three types of workloads: computation-intensive, communication-intensive and balanced. We present its performance evaluation methodology.

Acknowledgment

The work is partially supported by Ministry of Education and Science of Russian Federation under contracts RFMEFI58214X0003 and 02.G25.31.0061/ 12/02/2013 (Government Regulation No 218 from 09/04/2010), CONACYT (Consejo Nacional de Ciencia y Tecnología, México), grant no. 178415.

References

[1] A. Vasan, A. Sivasubramaniam, V. Shimpi, T. Sivabalan, and R. Subbiah, "Worth their watts? - An Empirical Study of Datacenter Servers," in The Sixteenth International Symposium on High-Performance Computer Architecture, 2010, pp. 1–10.

[2] J. Scaramella, "Worldwide Server Power and Cooling Expense 2006-2010 Forecast," Market Analysis, 2006.

[3] J. G. Koomey, "GROWTH IN DATA CENTER ELECTRICITY USE 2005 TO 2010," Analytics Press, 2011.

[4] A. Andrae and P. M. Corcoran, "Emerging Trends in Electricity Consumption for Consumer ICT." 2013.

[5] G. Sauls, "Measurement of data centre power consumption." Falcon Electronics Pty LTD, 2009.

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

[7] F. Giordano, C. Fiandrino, D. Kliazovich, A. Tchernykh, P. Giaccone, M. Guzek, and P. Bouvry, "ACCURATE : Adaptive Computing and Communication Resource Allocation Scheme for Cloud Computing Data Centers," unpublished.

[8] M. Guzek, P. Bouvry, and E.-G. Talbi, "A Survey of Evolutionary Computation for Resource Management of Processing in Cloud Computing [Review Article]," IEEE Computational Intelligence Magazine, vol. 10, no. 2, pp. 53–67, May 2015.

[9] D. I. G. Amalarethinam and T. L. A. Beena, "Cloud Scheduling - A Survey," International Journal of Computer Applications, vol. 97, no. 13, pp. 27–31, 2014.

[10] A. Tchernykh, L. Lozano, U. Schwiegelshohn, P. Bouvry, J. E. Pecero, S. Nesmachnow, and A. Y. Drozdov, "Online Bi-Objective Scheduling for IaaS Clouds Ensuring Quality of Service," Journal of Grid Computing, vol. 14, no. 1, pp. 5–22, Mar. 2016.

[11] S. Abrishami and M. Naghibzadeh, "Deadline-constrained workflow scheduling in software as a service Cloud," Scientia Iranica, vol. 19, no. 3, pp. 680–689, 2012.

[12] Y. Kessaci, N. Melab, and E.-G. Talbi, "Multi-level and Multi-objective Survey on Cloud Scheduling," in IEEE International Parallel & Distributed Processing Symposium Workshops, 2014, pp. 480–488.

[13] D. Kliazovich, J. E. Pecero, A. Tchernykh, P. Bouvry, S. U. Khan, and A. Y. Zomaya, "CA-DAG: Modeling Communication-Aware Applications for Scheduling in Cloud Computing," Journal of Grid Computing, vol. 14, no. 1, pp. 23–39, Mar. 2016.

[14] D. Kliazovich, P. Bouvry, and S. U. Khan, "DENS: data center energy-efficient network-aware scheduling," Cluster Computing, vol. 16, no. 1, pp. 65–75, Mar. 2013.

[15] M. Guzek, D. Kliazovich, and P. Bouvry, "HEROS: Energy-Efficient Load Balancing for Heterogeneous Data Centers," in IEEE 8th International Conference on Cloud Computing, 2015, pp. 742–749.

[16] C. Reiss, A. Tumanov, G. R. Ganger, R. H. Katz, and M. A. Kozuch, "Heterogeneity and dynamicity of clouds at scale: Google Trace Analysis," in Proceedings of the Third ACM Symposium on Cloud Computing, 2012, pp. 1–13.

[17] Q. Qi Zhang, M. F. Zhani, R. Boutaba, and J. L. Hellerstein, "Dynamic Heterogeneity-Aware Resource Provisioning in the Cloud," IEEE Transactions on Cloud Computing, vol. 2, no. 1, pp. 14–28, Jan. 2014.

[18] Cisco, "Cisco Data Center Infrastructure 2.5 Design Guide." 2011.

[19] B. Wang, Z. Qi, R. Ma, H. Guan, and A. V. Vasilakos, "A survey on data center networking for cloud computing," Computer Networks, vol. 91, pp. 528–547, 2015.

[20] Y. C. Lee and A. Y. Zomaya, "Energy efficient utilization of resources in cloud computing systems," The Journal of Supercomputing, vol. 60, no. 2, pp. 268–280, May 2012.

[21] T. Kaur and I. Chana, "Energy aware scheduling of deadline-constrained tasks in cloud computing," Cluster Computing, vol. 19, no. 2, pp. 679–698, Jun. 2016.

[22] Z. Dong, N. Liu, and R. Rojas-Cessa, "Greedy scheduling of tasks with time constraints for energy-efficient cloud-computing data centers," Journal of Cloud Computing: Advances, Systems and Applications, vol. 4, no. 1, p. 14, Dec. 2015.

[23] A. Beloglazov and R. Buyya, "Energy Efficient Resource Management in Virtualized Cloud Data Centers," in 2010 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing, 2010, pp. 826–831.

[24] A. Beloglazov, J. Abawajy, and R. Buyya, "Energy-aware resource allocation heuristics for efficient management of data centers for Cloud computing," Future Generation Computer Systems, vol. 28, no. 5, pp. 755–768, May 2012.

[25] C. Mobius, W. Dargie, and A. Schill, "Power Consumption Estimation Models for Processors, Virtual Machines, and Servers," IEEE Transactions on Parallel and Distributed Systems, vol. 25, no. 6, pp. 1600–1614, Jun. 2014.

[26] C. Fiandrino, D. Kliazovich, P. Bouvry, and A. Y. Zomaya, "Performance Metrics for Data Center Communication Systems," in IEEE 8th International Conference on Cloud Computing, 2015.

[27] D. Kliazovich, P. Bouvry, and S. U. Khan, "GreenCloud: a packet-level simulator of energy-aware cloud computing data centers," The Journal of Supercomputing, vol. 62, no. 3, pp. 1263–1283, Dec. 2012.

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

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

[30] "ns-2." [Online]. Available: http://www.isi.edu/nsnam/ns/.

[31] E. Zitzler, "Evolutionary Algorithms for Multiobjective Optimization: Methods and Applications," 1999