Towards energy-proportional Clouds partially powered by renewable energy

Nicolas Beldiceanu, Bárbara Dumas Feris, Philippe Gravey, Md Sabbir Hasan, Claude Jard, Thomas Ledoux, Yunbo Li, Didier Lime, Gilles Madi-Wamba, Jean-Marc Menaud, et al.

Computing

Towards energy-proportional Clouds partially powered by renewable energy

Nicolas Beldiceanu, Bárbara Dumas Feris, Philippe Gravey, Sabbir Hasan, Claude Jard, Thomas Ledoux, Yunbo Li, Didier Lime, Gilles Madi-Wamba, Jean-Marc Menaud, Pascal Morel, Michel Morvan, Marie-Laure Moulinard, Anne-Cécile Orgerie, Jean-Louis Pazat, Olivier Roux and Ammar Sharaiha

Received: December 2015 / Accepted: June 2016

Abstract With the emergence of the Future Internet and the dawning of new IT models such as cloud computing, the usage of data centers (DC), and consequently their power consumption, increase dramatically. Besides the ecological impact, the energy consumption is a predominant criterion for DC providers since it determines the daily cost of their infrastructure. As a consequence, power management becomes one of the main challenges for DC infrastructures and more generally for large-scale-distributed systems. In this paper, we present the EpoCloud prototype, from hardware to middleware layers. This prototype aims at optimizing the energy consumption of mono-site Cloud DCs connected to the regular electrical grid and to renewable-energy sources.

Keywords Data-Center · Energy-Efficiency · Virtualization · Task Placement · Optical Network · Prediction algorithms · SLA

N. Beldiceanu · T. Ledoux · Y. Li · G. Madi-Wamba · J.-M. Menaud
Mines de Nantes, LINA, France

B. Dumas Feris · P. Gravey · M. Morvan · M.-L. Moulinard
Telecom Bretagne, Lab-STICC, France

S. Hasan · J.-L. Pazat
INSA de Rennes, IRISA, France

C. Jard
Université de Nantes, LINA, France

D. Lime · O. Roux
Ecole Centrale de Nantes, IRCCyN, France

P. Morel · A. Sharaiha
ENIB, Lab-STICC, France

A.-C. Orgerie (✉)
CNRS, IRISA, France

1 Introduction

A data center (DC) is a facility used to house tens to thousands of computers and their associated components. These servers are used to host applications available in the Internet, from simple web server to multi-tier applications, but also some batch jobs. With the explosion of online services, particularly driven by the extension of cloud computing, DCs are consuming more and more energy. The growth of energy consumption by DCs is, at the same time, a technical, environmental and financial problem. Technically, in some areas (like Paris), the electrical grid has already saturated, thus preventing new DC installation or expansion of the existing ones. From an environmental point of view, the electricity production causes many CO_2 emissions, whereas financially the OPEX (Operational Expenditure) have exceeded CAPEX (Capital Expenditure). Although over the last few years, computer servers have become less expensive and highly energy efficient, the price of electricity has significantly increased even in countries known of having lower electricity price (e.g. France). To some extent, these operating costs are mainly related to the power consumption. Several actions are possible to reduce these impacts/costs. One of them consists in using a local power generation based on renewable energy, like Microsoft, Google, and Yahoo who have built new DCs close to large and cost-efficient-hydroelectric power sources for instance.

In the EPOC (Energy Proportional and Opportunistic Computing systems) project, we aim at focusing on energy-aware task execution from the hardware to the application's components in the context of a mono-site and small DC (all resources are in the same physical location), which is connected to the regular electric Grid and to local-renewable-energy sources (such as windmills or solar cells).

Pioneering solutions have been proposed to tackle the challenge of powering small-scale DC with only renewable energies Goiri et al (2014). In the context of EPOC, we are considering a hybrid approach relying on both the regular grid and a renewable-energy source, like sun or wind for instance. We also assume that renewable energy produced locally cannot be stored (i.e. no battery). Consequently, we do not consider energy loss due to battery inefficiency. Instead of storing renewable energy during its production, it is consumed by performing opportunistic-computing tasks.

In EPOC, our first challenge consists in developing a transparent (for users) energy proportional computing (EPC) distributed system, from system to service-oriented runtime, mainly based on hardware and virtualization capabilities. The second challenge addresses the energy issue through a strong synergy inside infrastructure-software stack and more precisely between applications and resource management systems designed to tackle the first challenge. This approach must allow adapting the Service-Level Agreement (SLA) by seeking the best trade-off between energy cost (from regular electric grid), its availability (from renewable energy), and service degradation (from application reconfiguration to jobs suspension). The third challenge embarks to set energy-efficient-optical networks as key enablers of future internet

and cloud-networking service deployment through the convergence of optical-infrastructure layer with the upper layers. Another strength of the EPOC project is the integration of all research results into a common prototype named EpoCloud. This approach allows the pooling of development efforts, and validates solutions on common and reproducible use-cases. In this paper, we present the EpoCloud DC architecture, from hardware layer to middleware layer, including the resource-management algorithms.

The paper is organized as follows. Section 2 sets the principles of EpoCloud, the hardware architecture is described in Section 3 and the EpoCloud manager is detailed in Section 4. Finally, Section 5 concludes and presents future work.

2 EpoCloud principles

Our first goal is to design an energy-proportional-computing system, which implies no energy consumption, whenever there is no activity. To date, dynamic power management has been widely used in embedded systems as an effective energy-saving method with a policy that attempts to adjust the power mode according to the workload variations Sridharan and Mahapatra (2010). Unfortunately, servers consume energy even when they are idle. For an efficient energy-proportional-computing system, we need to be abble to have the capability to turn on/off servers dynamically. Vary-on/vary-off (VOVO) policy reduces the aggregated-power consumption of a server cluster during periods of reduced workload. The VOVO policy turns off servers so that only the minimum number of servers that can support the workload are kept alive. However, much of the applications running in a data center must be online constantly. To solve this problem, dynamic placement using application live migration[1] permits to keep using VOVO policy in the on-line application context.

Currently, the most efficient system for live migration is the use of virtualization. Virtualization refers to the creation of a Virtual Machine (VM) that acts like a real computer with an operating system but software executed on these VMs is separated from the underlying-hardware resources. Virtualization also allows snapshots, fail-over and globally reduce the IT-energy consumption by consolidating VMs on a physical machine (i.e. increasing the server utilization and thus reducing the energy footprint). Furthermore, dynamic consolidation uses live migration for effective placement of VMs on the pool of DC servers to reduce energy, increase security, etc. But, live migration requires significant network resources.

Our first main objective is focused on a workload-driven approach. EpoCloud adapts the power consumption of the DC depending on the application workload, and predicts this workload to be more reactive. Our second objective is focused on power-driven SLA. The power-driven approach implies shifting or scheduling the postponable workloads to the time period when the electricity is

[1] Live migration moves a running application between different physical machines without disconnecting the client or application. Memory, storage, and network connectivity are transferred from the original host machine to the destination.

available (from the renewable-energy sources) or at the best price. For on-line application, power-driven approach implies a degradation of services when energy is at an insufficient level, while maintaining SLAs. In addition, EpoCloud takes advantage of the available energy to perform some tasks. Some of them allow limitations on application degradation. We describe our EpoCloud architecture and EpoCloud manager in section 3 and 4 respectively.

3 High throughput optical networks for VM migration

Recent studies on data-centers companies show that a VM consumes an average of 4 GB of memory and 128 GB of storage. Thus, it will take a minimum of 17.5 minutes (resp. 1.75 minutes) with a 1 Gb/s (resp. 10 Gb/s) network to realize a complete VM migration. Moreover, a classical consolidation ratio in virtualized data centers is 50 VMs per server. According to the approach that we are considering in EPOC (VOVO Policy), our data center needs to be able to migrate all the VMs running on a server (7.5 TB), whenever the hypervisor requests to turn this server off in order to save power. Having one optical port per rack means that its bandwidth might be shared by the servers located in this rack. Then, is this bandwidth enough to migrate all the VMs in one server? Using 10 Gb/s this operation takes around 2 hours. However, if we consider an example, 32 servers per rack, the same operation would take about 53 hours, since now the bandwidth is being shared by the 32 servers. So, increasing the bit rate of the interconnection network becomes a must.

To overcome the aforementioned problem, classical dynamic consolidation system uses live migration with a Storage Area Network (SAN). In this case, the VM storage is shared between all servers and live migration is limited to transfer VMs memory. Nevertheless, adding a SAN impacts on the global-DC-energy consumption. EpoCloud proposes to suppress the SAN, which is a dedicated network providing access to consolidated data storage. Among various components of a data center, storage is one of the biggest consumers of energy. An industry report Inc (2002) shows that storage devices account for almost 27% of the total energy consumed by a DC. By suppressing the SAN we optimize the energy consumption but we introduce a strong hypothesis on the technical architecture: for accessing data of applications and systems, we can only use local disk servers and the regular network linking the servers. In this article, we present an innovative network architecture, detailed in section 3.1, and we describe in section 3.2 architectural motives and principles for the integration of renewable energy.

3.1 Network Architecture

The topology of most current-DC networks is based on a 3-Tier fat-tree topology (see Figure 1(a)). Each of the three main switching layers: core, aggregation and ToR (top-of-the rack), uses Electrical Packet Switches (EPS).

Servers accommodated into racks are connected through ToR switches to the aggregation layer, and from there to the core layer using aggregation switches. Core switches provide interconnection to the internet (or outside the DC).

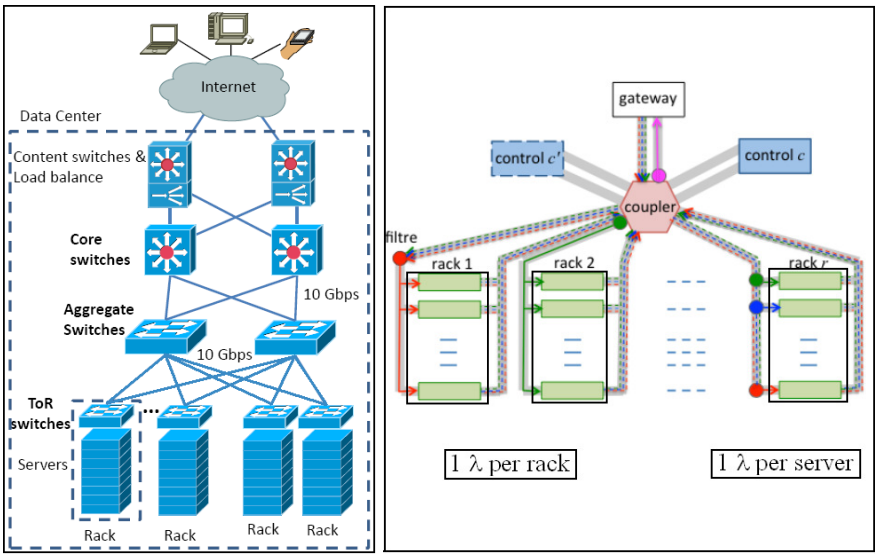
The introduction of optical communications seems to be crucial, because it can achieve very high data rates, low latency and low power consumption Kachris and Tomkos (2012). This has recently become a hot research topic inside the optical-networking community. Some authors propose a direct migration to all-optical architectures, most of them based on Optical Circuit Switching (OCS) Singla et al (2010)-Liboiron-Ladouceur et al (2008) that does not meet the needs of a variable traffic over time. Other authors propose to introduce optical networking gradually by the use of hybrid (electrical-optical) solutions Farrington et al (2013). OCS could be used for transferring large blocks of data between ToR switches, while EPS would serve for background traffic. More sophisticated hybrid architectures, involving several hierarchy levels, could have the potential to connect millions of servers in giant DCs.

As already noted, EPOC aims at focusing on small/medium size DCs. For transferring 7.5 TB, implementing a full optical interconnection architecture could be an attractive option, in terms of latency, power consumption and control complexity. This implies using Optical Packet Switching (OPS) technology, whose maturity is still highly questionable, in spite of several decades of investigation for telecom network applications Yoo (2006). Nevertheless, several techniques, relying on fast wavelength-tunable-optical-emitters, have recently gained a renewed attention, in particular for metropolitan-area-network applications. These techniques include Time-domain Wavelength Interleaved Networks (TWIN), originally proposed by Lucent Sanjee and Widjaja (2004), and Packet Optical Add and Drop Multiplexer (POADM), proposed by Alcatel-Lucent Chiaroni et al (2010).

In the EPOC project, we decided to investigate a third option, derived from TWIN Indre et al (2014), under the name of Passive Optical Pod Interconnect (POPI). Mainly, because POPI uses a purely passive optical network with power consumption concentrated at networks edge and its passive nature provides a high reliability.

3.1.1 POPI architecture

POPI is simpler than the classical EPS architecture Kachris and Tomkos (2012), since it does not use neither ToR nor aggregation switches (see Figure 1(b)). Servers, two controllers (for redundancy) and one gateway are interconnected by one incoming and one outgoing fiber in a tree topology routed on a passive star coupler. Servers and gateway are equipped with fast-tunable transmitters and fixed receivers enabling them to send optical bursts to any destination by selecting the appropriate wavelength. To avoid collisions, the controller receives reports from servers and gateways and issues grants in order to allocate them suitable transmission windows. A connection towards the Internet is provided by the gateway.



(a) (b)

Fig. 1: (a) Classical EPS data center architecture (from Kachris and Tomkos (2012)) (b) POPI architecture (from Indre et al (2014)).

3.1.2 Power consumption

We estimate the power consumption of the classical EPS architecture and the one of POPI's. For a fair comparison we assume that in both cases each server is equipped with a 10 Gb/s transceiver (using Non Return to Zero (NRZ) format). The components required by each architecture, considering 32 racks and 1000 servers, are listed in Table 1. Hence, the total power consumption of POPI is around 4 kW, i.e. about 20 % of the classical-EPS-architecture's.

3.1.3 Capacity issue

The maximum number of connected entities by POPI (considering 1 λ per server), N, is limited on one hand, by the insertion losses and on the other hand, by the number of channels that can be packed into the band of the fast-tunable emitters.

First of all, we quantify the maximum intrinsic power loss allowed for the coupler (L_C) for a Symbol Error Rate (SER) of 10^{-3}, according to:

$$L_C = P_{Tx} - P_{Rx} - L_{EAM} - L_f - L_{connectors} - L_{BPF} - M - L_{C,excess} \quad (1)$$

where P_{Tx} represents the transmission optical power; P_{Rx}, the required received power for a SER of 10^{-3}. L_{EAM}, L_f , $L_{connectors}$ and L_{BPF} are the losses due to the modulator (3.3 dB), the optical fiber (0.2 dB), connectors (1 dB) and the optical band-pass filter (5 dB), respectively. M represents a

Classical EPS architecture Kachris and Tomkos (2013), Singla et al (2010)			
Component	Amount	Power consumption per unity	Subtotal
Transceivers	1000	0.3 W	300 W
ToR switches	32	600 W	19,200 W
Aggregation switches	2	600 W	1,200 W
Total			≈ 21 kW
POPI architecture Indre et al (2014)			
Component	Amount	Power consumption per unity	Subtotal
Transceivers	1000	2 W	2,000 W
Optical filters	32	1 W	32 W
Controllers	2	20 W	40 W
Control channels	2	1 W/server	2,000 W
Total			≈ 4 kW

Table 1: Power consumption of classical EPS and POPI architectures.

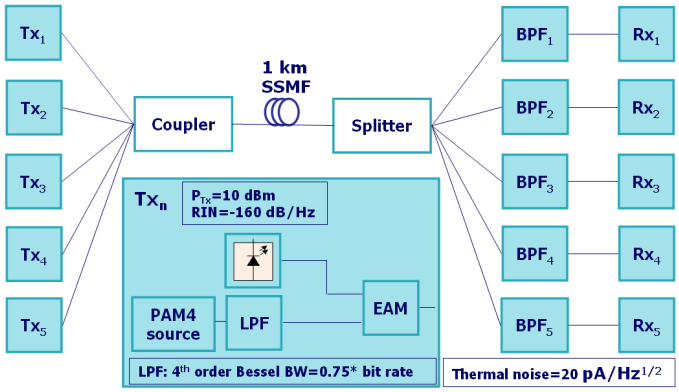


Fig. 2: Simulation setup. Tx: Transmitter, LPF: Low-Pass Filter, EAM: Electro-Absortion Modulator, BPF: Band-Pass Filter, Rx: Receiver.

margin of 1.5 dB. We shall estimate the coupler-excess loss ($L_{C,excess}$), by using the following empirical relation, derived by analyzing several commercial datasheets: $L_{C,excess} = (L_C + 3)/6$. Then, the maximum number of connected entities due to the insertion losses $N_{max,IL} = 10^{L_C/10}$.

Therefore, we simulate many possible scenarios by the use of $VPItransmissionMaker^{TM}9.3$ software, taking into account the transmission of five adjacent channels (we include crosstalk effects). The simulation setup is depicted in Figure 2. We consider NRZ and Pulse Amplitude Modulation 4 (PAM4) formats, for different bit rates: 14, 28 and 56 Gb/s, using two different receivers: PIN-photodiode and Avalanche Photodiode (APD). Note that each study case presents different optical bandwidths, which will further limit channel spacing. Simulation results of SER versus the received optical power, considering a P_{Tx} of 10 dBm, for PIN-photodiode and APD receivers, are shown in Figure 3.

From these figures we get P_{Rx}, necessary for calculating L_C and $N_{max,IL}$. As expected, P_{Rx} increases with the bit rate and of course when passing from

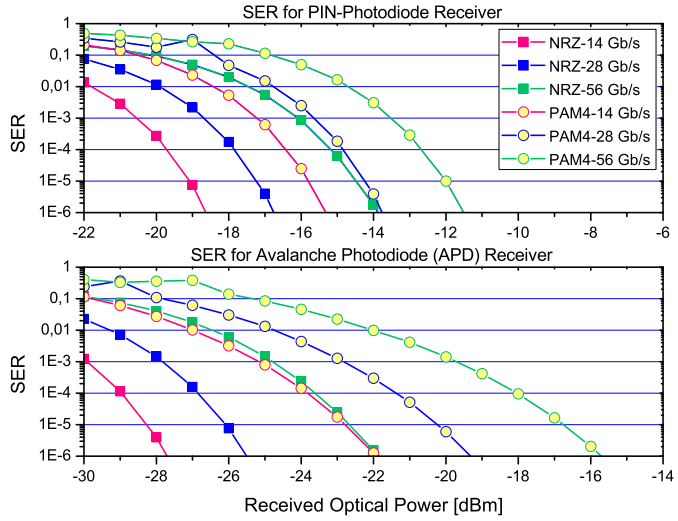


Fig. 3: Simulated SER vs received optical power for PIN-photodiode (top) and APD (bottom) receivers.

NRZ to PAM4 (since the decision thresholds get closer). It can be seen that these results can be improved until around a factor of 10, by use of an APD receiver; for instance, the 14-Gb/s-NRZ case passes from -20.5 to -29.9 dBm.

Regarding spectral issues, we consider a fast-tunable laser ($\leq$ 50 ns) providing 64 50-GHz-spaced ITU channels in a 3200-GHz band Simsarian and Zhang (2004). Additionally, the PAM-signal bandwidth must fit in the chosen grid. Usually the ITU grid fixed at 50 or 100 GHz is taken into account, however in this paper we consider fixed grids coming from 25 GHz to 100 GHz with a 12.5-GHz step in order to maximize the number of channels for each case. Then, the maximum number of connected entities using a X-GHz grid, $N_{max,XGHz}$, will depend on the laser band according to the following expression: $3200GHz/X$. It is worth to mention that wavelength tuning between 25-GHZ-spaced channels has not been presented yet, maybe because of stability issues in the laser electronics, but this might be demonstrated soon.

But, before selecting a grid, we check whether optical BPF suited to it is actually feasible. In order to do so, we consider trapezoid BPFs and we limit their slopes to less than 800 dB/nm Pincemin et al (2014); where 0.75 of the first lobe of the PAM-signal bandwidth and 20 dB of filter attenuation (to ensure a reasonable crosstalk), are taken into account.

Finally, Table 2 presents the maximum number of connected entities for each study case, for both the PIN-photodiode and the APD receivers. Results are shown in different colors in order to visualize the strongest limitation. Red represents $N_{max,IL}$; and blue, $N_{max,X-GHz}$.

On one hand, for the PIN-photodiode only 14 Gb/s under the NRZ format can be limited by the channel-spacing constraint (blue). For the other cases

Modulation format	NRZ			PAM4		
Bit rate [Gb/s]	14	28	56	14	28	56
Optical bandwidth [GHz]	21	42	84	10.5	21	42
PIN-photodiode receiver						
P_{Rx} for $SER = 10^{-3}$ [dBm]	-20.5	-18.7	-16.1	-17.3	-15.7	-13.5
L_C [dB]	16.3	14.7	12.5	13.5	12.2	10.3
$N_{max,IL}$	42	29	17	22	16	10
N for 25-GHz-spaced-channel grid	NA	NA	NA	22	NA	NA
N for 37.5-GHz-spaced-channel grid	42	NA	NA	22	16	NA
N for 50-GHz-spaced-channel grid	42	29	NA	22	16	10
N for 62.5-GHz-spaced-channel grid	42	29	NA	22	16	10
N for 75-GHz-spaced-channel grid	39	29	NA	22	16	10
N for 87.5-GHz-spaced-channel grid	33	29	NA	22	16	10
N for 100-GHz-spaced-channel grid	29	29	17	22	16	10
APD receiver						
P_{Rx} for $SER = 10^{-3}$ [dBm]	-29.9	-27.9	-24.7	-25.2	-22.8	-19.8
L_C [dB]	24.3	22.6	19.9	20.3	18.3	15.7
$N_{max,IL}$	271	183	97	107	66	37
N for 25-GHz-spaced-channel grid	NA	NA	NA	107	NA	NA
N for 37.5-GHz-spaced-channel grid	82	NA	NA	82	66	NA
N for 50-GHz-spaced-channel grid	61	61	NA	61	61	37
N for 62.5-GHz-spaced-channel grid	48	48	NA	48	48	37
N for 75-GHz-spaced-channel grid	39	39	NA	39	39	37
N for 87.5-GHz-spaced-channel grid	33	33	NA	33	33	33
N for 100-GHz-spaced-channel grid	29	29	29	29	29	29

Table 2: Estimation of the maximum number of connected entities by POPI, for PIN-photodiode and APD receivers. Colors red and blue represent $N_{max,IL}$ and $N_{max,X-GHz}$, respectively. NA: Not Applicable.

this limitation is much smaller than the insertion-losses's one, since the receiver sensitivity gets easily higher (and L_C smaller) when increasing the bit rate or when passing from NRZ to PAM4. On the other hand, for the APD receiver $N_{max,IL}$ is larger than $N_{max,X-GHz}$ for most of cases (except under PAM4 at 28 and 56 Gb/s, for 37.5- and until 75-GHz grids, respectively), leading the channel-spacing constraint to become the biggest limitation. Finally, note that the NRZ format uses twice the optical bandwidth of the used by PAM4. Here the channel spacing gets an important role. Specially at 56 Gb/s using the APD receiver, since the former can only fit in the 100-GHz grid, limiting the maximum number of connected entities to 29, compared to 37, for PAM4 using a 50-GHz grid.

As detailed in Beldiceanu et al (2015b), current servers using Solid-State Drives (SSD) can reach 100 Gb/s throughput capacity and thus, it can make efficient use of the optical network designed in the EPOC project.

3.2 Integrating Local Renewable Energy

Although several research efforts have been made to reduce energy consumption in infrastructure layer, still the goal for alleviating carbon footprint is

being underachieved. Given the circumstances, explicit or implicit integration of renewable energy to the DC can be the only way to reduce carbon footprint at an acceptable level. Besides, the demand for green services is ever increasing, thus integrating renewable sources to the data center leave no choice. Few green-cloud providers, e.g., greenQloud GreenQloud (2010), green House Data Green House Data (2007) and academic researchers Goiri et al (2014) integrated renewable sources to the data center explicitly, which offers green computing services with partial SLA fulfillment.

In contrast, renewable sources are known to be very intermittent in nature, thus providing green services or running Servers and VMs only by on-site-renewable energy becomes very unrealistic. Moreover, some research efforts have also explored how to incorporate off-site renewable energy to a data center, as the best location for producing renewable energy does not always have the best potential to build a DC. Transporting the off-site energy is arduous since wheeling charge imposed by the Grid might be more than the expectation and power losses through transmission line are inevitable. Besides the above explicit involvement, some implicit options for reducing carbon footprint also exist through renewable energy certificate (REC) and power purchasing agreement (PPA) Haque et al (2013). REC, is a tradable commodity proving that electricity was generated using renewable sources. Therefore, purchasing of a green certificate equals to purchasing a claim that the certificate owner consumed energy from the renewable portion of the whole energy grid Hasan et al (2014). Due to the intermittency, predicting the amount of renewable energy production ahead of real time might demonstrate greater error statistics in DC power management. Nonetheless, excessive production of renewable energy can go to waste if it is not consumed. One way to overcome the challenge is to use energy storage or battery to store this superfluous green energy which can be discharged later for peak shaving of DC power demand or for fulfillment of a planned energy consumption target, when renewable energy is needed but not available. Energy storage incurs additional costs to DCs CAPEX and OPEX, and energy losses due to battery's efficiency and finite capacity. Therefore it is not an attractive solution for small-scale data centers. Moreover, storages have finite capacities to recharge energy and their lifetime is a decreasing function of DOD and charge/discharge cycles Ren et al (2012). Therefore, if the production of renewable energy is above the capacity of storage, remaining energy goes to waste. Even the state of the art batteries have 80-85% efficiency on charging and discharging capabilities, which implies 28-36% loss of energy.

In order to avoid using storage or batteries in small-scale DC, we could virtualize Hasan et al (2015) the green energy. The energy can be virtually green for a specific period of time if abundance of green energy is available aperiodically in shorter time interval along with the deficit of green energy in rest of the time frame. Therefore, the virtualization concept can increase the greenness of energy, rather increasing the amount of green energy. Concretely, when the availability of green energy is more than demand, we use the whole portion of available green energy but characterize the interval as surplus interval. When green energy is insufficient to meet the demand, we

nullify the degraded interval with the surplus interval. We use the term virtualization because we nullify a degraded interval (lack of green energy) with a surplus interval (excessive green energy than demand), but from the clients or SaaS providers perspective, they realize both the interval as ideal interval (when supply meet the demand), though the green energy was not present instantaneously rather present virtually.

In this way, energy storage is not needed and neither of the portion of renewable energy is wasted. Because, the idea behind proposing the virtualization of green energy is to use every watts of energy when it is available. Furthermore, total expenditure of energy purchasing can be reduced since no green energy goes to waste and additional cost for using storage is not needed. Even energy aware SLA between IaaS and SaaS providers can be fulfilled. For instance: assuming a IaaS provider has established a SLA to have some portion of renewable energy available for each time slot e.g., T=30/60 minutes to run applications, by using the virtualization concept of green energy maximum availability of green energy could be ensured by taking the average over time period.

4 EpoCloud Manager

Architectural principles for small data centers were defined in the previous sections. They rely on innovative infrastructure where a limited number of servers (without SAN) are connected by a high speed optical network and supplied by local sources of renewable energy, composed of a limited number of server (without SAN) connected by a high speed network. To take advantage of this architecture, the EPOC project develops an innovative task management system: the EpoCloud Manager including a smart task scheduler (Section 4.1), an energy-aware SLA oriented management system (Section 4.3), and a workload prediction module (Section 4.2).

4.1 Opportunistic energy-aware resource allocation

In the EPOC project, we propose to design a disruptive approach to Cloud's resource management which takes advantage of renewable energy availability to perform opportunistic tasks. Let's recall that, the considered EpoCloud is mono-site (i.e. all resources are in the same physical location) and performs tasks (like web hosting or MapReduce tasks) running in virtual machines. The EpoCloud receives a fixed amount of power from the regular electrical grid. This power allows it to run usual tasks. In addition, the EpoCloud is also connected to renewable energy sources (such as windmills or solar cells) and when these sources produce electricity, the EpoCloud uses it to run more, less urgent, tasks.

The proposed resource management system integrates a prediction model to be able to forecast these extra-power periods of time in order to schedule

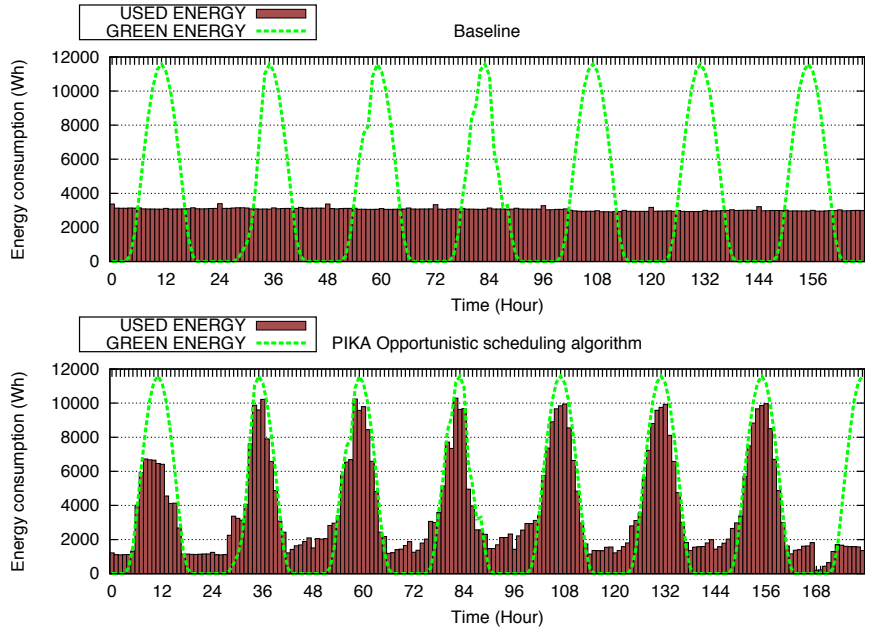


Fig. 4: Energy consumption of EpoCloud manager vs. baseline

more work during these periods. Given a reliable prediction model, it is possible to design a scheduling heuristic that aims at optimizing resource utilization and energy usage, problem known to be NP-hard. So, the proposed heuristics will schedule tasks spatially (on the appropriate servers) and temporally (over time, with tasks that can be planed in the future).

In order to achieve this energy-aware resource allocation, we distinguish two kinds of jobs to be scheduled: the web jobs which represent jobs requiring to run continuously (like web server) and the batch jobs which represent jobs that can be delayed and interrupted, but with a deadline constraint. The second type of jobs are the natural candidates of the opportunistic scheduling algorithm. Each job, batch or web, is executed in a dedicated VM.

Additionally for reducing further energy consumption in the EpoCloud, we are taking advantage of consolidation algorithms, on/off mechanisms on physical servers and over-commitment of RAM and CPU in order to optimize the number of powered-on resources. This energy-efficient manager is called PIKA (oPportunistic schedulIng broKer infrAstructure). First, the consolidation algorithm also relies on VM suspend/resume mechanisms for the batch jobs and live migration mechanisms of VMs for the web jobs. However, such mechanisms have a cost in terms of both time and energy, cost which is taken into account in PIKA to optimize the overall energy utilization. Secondly, for unused physical servers, PIKA applies VOVO policies, thus switching off idle servers to keep the energy-proportionality principle of EPOC. As it is costly to switch on again servers, PIKA is using workload prediction algorithm to

optimize the number of servers to keep on. The workload prediction algorithm is described in Section 4.2. Finally, the over-commitment policies are used to limit physical resource under-utilization (and consequently resource and energy waste) due to VM over-provisioning, which is typical in Cloud context Li et al (2015). Hence, over-commitment is exploited in PIKA to increase the server usage and to reduce the number of turned on servers if the VMs are not fully loaded.

We have evaluated the EpoCloud manager by using real-world workload traces from small-scale DC (55 servers), power values from real measurements on our experimental test-bed, and our own simulator Li et al (2015). Figure 4 presents the energy consumption of baseline algorithm versus PIKA which effectively manages to launch batch jobs when the renewable energy is available. The baseline algorithm is also using over-commitment techniques to improve resource utilization (classic approach in such a context), but has no energy-aware mechanisms.

From these simulation-based results using real workload traces, we show that PIKA is able to reduce non renewable energy consumption by 45% and to double the renewable energy utilization compared to the baseline algorithm.

4.2 Predicting workload

In order to design a scheduling heuristic that aims at optimizing resource utilization, we need to know in advance the eventual distribution of the resource usage in time. To do so, we propose a workload prediction model based on historical traces. The model is constructed as follows. Given a set of real workload traces modelled as time series, we first clusterize them into p clusters. This is motivated by the fact that the workload of a data center may both depend on the day (e.g. week, weekend) as well as of the type of services it provides at specific time periods. Second, we extract from each cluster cl_i, some key properties that correspond to typical features of the time series (e.g. its highest peak, number of peaks). Using these key properties, we build a model $m(cl_i)$ for each cluster cl_i. This offline prepossessing is the *learning step*. Finally the *prediction model* is given by $\cup_{i=0}^{p-1} m(cl_i)$ the union of all the models $m(cl_i)$ of each cluster cl_i. At any time t, the prediction model should be able to foretell in real time how will evolve the workload at time $t + \epsilon$, where ϵ is a time lapse to determine. We now recall some background on time series.

Background on time series

Beldiceanu et al (2015a) describe a large family of constraints for structural time series. A time series is characterized by the following concepts:

– Signature of a time series. The *signature* of a time series is a sequence of comparison operators taking values in the set $\{<, =, >\}$. Each element of the signature is obtained by comparing two adjacent input values.

– Pattern. A *pattern* is a regular expression over the alphabet $\{<, =, >\}$. To find a pattern occurrence in a time series, its signature has to be computed first. A pattern occurrence is a maximal occurrence of a sequence of characters from the signature that matches the regular expression of the pattern. Table 3 gives examples of patterns that may occur in a time series.

pattern	**regular expression**
`increasing`	$<$
`increasing_sequence`	$< (< \mid =)^* < \mid <$
`increasing_terrace`	$<=^+<$
`summit`	$(< \mid (< (= \mid <)^* <))(> \mid (> (= \mid >)^* >))$
`plateau`	$<=^*>$
`proper_plateau`	$<=^+>$
`strictly_increasing_sequence`	$<^+$
`peak`	$< (= \mid <)^* (> \mid =)^* >$
`inflexion`	$< (< \mid =)^* > \mid > (> \mid =)^* <$
`steady`	$=$
`steady_sequence`	$=^+$
`zigzag`	$(<>)^+(< \mid <>) \mid (><)^+(> \mid ><)$

Table 3: Examples of patterns and their corresponding regular expressions.

– Feature. Given a pattern occurrence, a *feature* is a quantifiable property of the pattern.
– Aggregation. Given one or more occurrence of a pattern p and a feature f of p, an *aggregation* is a function applied to the different feature values of each occurrence of pattern p.

– Footprint. Given a time series ts of length n and a pattern p, the *footprint* $fp_p(ts)$ of pattern p is a sequence of n values that identifies all occurrences of pattern p in the time series ts.

Offline step: Learning Step

Using constraint programming techniques, we analyze input workload traces to extract key properties. For each cluster cl_i a model $m(cl_i)$ is built after a 3-step analysis of cl_i. For each pattern p of interest and for each input time series ts belonging to the cluster cl_i we proceed as follows:

(1) The number of occurrence of pattern p in the time series ts is computed.
(2) The footprint $fp_p(ts)$ of the pattern p is computed.
(3) Different values for aggregation of features of pattern p are computed.

At the end, all the results for all the patterns of interest are put all together. They are analyzed to extract different ranges of values for each characteristic. Among these ranges, we select a subset of pertinent ones. Those pertinent ranges constitute the model $m(cl_i)$ of the cluster cl_i.

Prediction model

We introduce the following notion of prefix time series: Given an index $t < n$ (where n is the length of a time series) and a server s, the *prefix time series* induced by t ($pref(t)$) is the time series $x_0x_1 \dots x_t$ which represents the workload of server s from time 0 to time t. At time $t < n$, each value x_i ($i \in [0, t]$) is already known.

The prediction is done in three steps. Given a prefix time series ($pref(t)$), we first determine in which clusters it can belong. The second step gives an interval for the potential values of the prefix time series ($pref(t)$) at time $t+\epsilon$. And finally, we refine the model to make the prediction more precise.

(1) The first step checks the compatibility of the time series ts with each cluster model $m(cl)$. The time series ts is compatible with cluster cl_i if the value for each characteristic of ts (number of occurrence of each pattern, footprint of each pattern, aggregation values ...) falls in the corresponding range of the model $m(cl_i)$ of cluster cl_i.
(2) For each compatible cluster cl, we check the values $ts_j^{cl}(t+\epsilon)$ of each time series ts_j belonging to cl. From these values we build a probabilistic model for the potential value $ts(t+\epsilon)$ of the time series ts at time $t+\epsilon$. That probabilistic model is the smallest interval $I(t+\epsilon)$ containing all the values $ts_j^{cl}(t+\epsilon)$ for each compatible cluster cl and each time series $ts_j \in cl$.
(3) At this step, we reduce the size of the interval $I(t+\epsilon)$. To do this, we consider the center time series of each compatible cluster cl_i and compute the footprint of the patterns *strictly_increasing_sequence* and *strictly_decreasing_sequence* to identify location on time series were there should occur or not. We use it to discard some values from $I(t+\epsilon)$.

Evaluation of the workload prediction model

This section presents a preliminary evaluation of the prediction model. We clustered 95 real workload traces times series into 6 clusters $cl_1, cl_2, \dots, cl_6$ of cardinality $20, 9, 13, 13, 16, 24$ respectively. The learning was made using 70% of the time series of each cluster, and the prediction model were builded from the 30% remaining time series, that we call the test time series.

Each series is of length 24 and is associated to a server. For each prefix of length k (k from 1 to 23) of each test time series we performed 5 benches on the prediction of the $(t+1)^{th}$ value of the prefix:

(1) The first bench presented in Figure 5 evaluates the percentage of case were there is at least one cluster compatible with the prefix.
(2) The second bench presented in Figure 6 evaluates the average number of cluster compatible with the prefix.

Benches 1 and 2 (Figures 5 and 6) show that there are cases were the tested prefixes are compatible with no cluster. The major reason for this is that the number of times series on witch the learning was made is not enough. The three other benches are however more interesting:

(3) Accuracy of the cluster compatibility part of the prediction. We check whether the cluster to which the test time series belongs is included in the list of compatible clusters. The results are given in Figure 7. We observe that, for all prefixes of length bigger than or equal to 10, the correct cluster is included in the list of compatible cluster the cluster.

(4) The fourth bench evaluates the percentage of cases were the actual $(k+\epsilon)^{th}$ value of a test time series belongs to the interval $I(t+1)$ predicted. We did this evaluation twice: In one hand without refining the interval $I(t+1)$ (step (3) of the prediction) and in the other hand after refining the interval $I(t+1)$. The results are presented in Figure 8.

(5) Accuracy of the prediction. We computed the average length of the predicted interval for each prefix. We did this evaluation in one hand without refining the interval $I(t+1)$ and in the other hand after refining the interval $I(t+1)$. The results are presented in Figure 9.

Benches 4 and 5 (Figures 8 and 9) show that, refining the $I(t+1)$ increases the accuracy of the prediction, but slightly increase the number of cases were the actual $(k+\epsilon)^{th}$ value of a test time series do not belong to the interval $I(t+1)$ predicted.

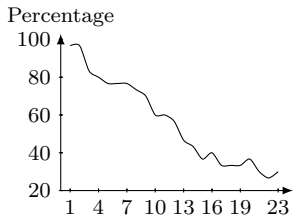


Fig. 5: Percentage of cases were there is at least one cluster that is compatible with the prefix.

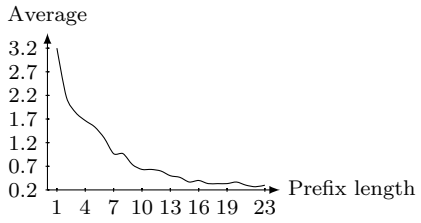


Fig. 6: Average number of cluster compatible with the prefix.

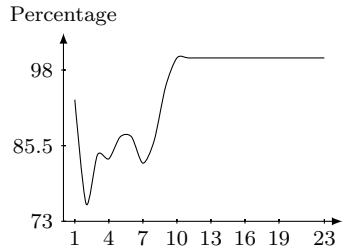


Fig. 7: Percentage of cases where the cluster of the prefix belongs to the list of compatible clusters

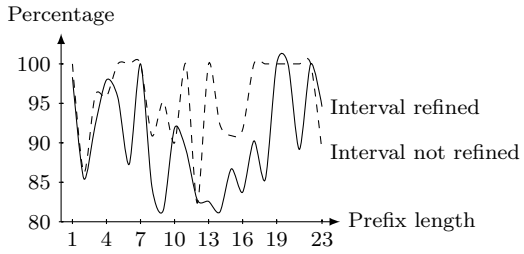


Fig. 8: Percentage of cases were the actual $(k+\epsilon)^{th}$ value of a test time series belongs to the interval $I(t+1)$ predicted.

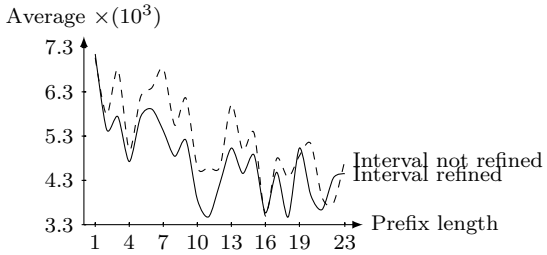


Fig. 9: Accuracy of the prediction.

The overall accuracy of the prediction is related to the accuracy of the first step of the prediction. The ideal would be that, for any prefix there exists at least one compatible cluster. But such a property can not be guarantied. Nevertheless, we can improve the accuracy of the cluster compatibility part of the prediction (Figures 5 and 6) by learning on a bigger data set.

4.3 Enforcing green SLA

While the proliferation of Cloud services have greatly impacted our society, how green are these services is yet to be answered. Usually, the traditional Cloud services are offered to clients having a Service Level Agreement (SLA), which includes availability and response time. Since, the demands for green products, services as well as social awareness for being green is ever increasing, its high time for service providers to consider offering green services using green energy. Therefore we propose a new paradigm of Service level objective (SLO) for SaaS provider, where service can be provided using proportional green energy with above mentioned classical objectives (i.e., Availability, Response time) in Figure 10.

To date, the problem for offering green services based on green energy has been undermined since green energy sources are very intermittent in nature and constantly providing the same amount of green energy is ungovernable. Therefore, reducing the energy consumption in application level leaves no choice if certain percentage of green energy requirement has to be respected while green energy is unavailable or scarce. Furthermore, service providers can propose greenness property as an extended scope to eco-friendly clients who are willing to accept degradation of traditional performance metrics in the absence of green energy; response time and availability. Traditionally, data center hosts heterogeneous applications; interactive and batch applications/jobs. Unlike batch jobs, interactive jobs can not be interrupted, hence need to be adapted with green energy period. Therefore, selecting algorithms depending on time, space complexity and choosing most energy compliant components in the software is necessary. Furthermore, by reducing energy consumption in applications, the percentage of green energy in data center can be increased in proportion to brown or traditional energy when green energy availability is

scarce in amount. Thus, proposed *green SLA based on green energy percentage* can be validated by SaaS providers.

To enable green energy awareness in an application, it must be adaptive with green energy production through dynamic reconfiguration capabilities in run-time. This reconfiguration can be realized by changing its business logic by replacing a component by another one or by reducing one or more loosely coupled components. Usually, modern application posses several components which is always activated throughout their life-cycle. We propose to tag some resource hungry application component as green and can only be activated in the presence of green energy since interactive/web jobs cannot be interrupted. Since, an ideal green data center should posses green energy most of the time, we mark the state of the application as standard case where all components are activated. In the absence of green energy, we deactivate the green-tagged component that reduces energy consumption and mark the state of the application as energy compliant. Figure 11 shows our architecture of standard and energy compliant state of a SaaS like application on top of RUBiS (auction site like ebay.com) prototype. So, *green SLA based on content* can be proposed to eco-friendly client who are conscious about environment and willing to help SaaS provider to reduce energy consumption during absence of green energy. Therefore, an efficient SLA-driven management is an interesting way to reduce resource consumption and indirectly energy footprint.

5 Conclusion

In the EPOC project, we aim at focusing on energy-aware task execution from the hardware to application's components in the context of a mono-site data center (all resources are in the same physical location) which is connected to the regular electric Grid and to renewable energy sources (such as windmills or solar cells). In this paper, we have presented the EpoCloud prin-

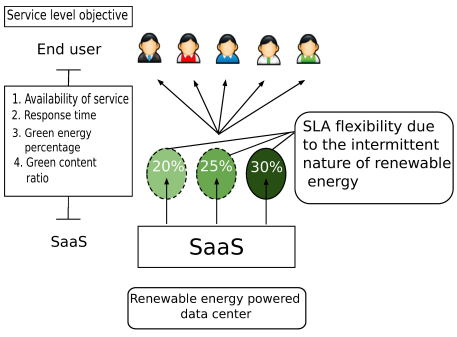


Fig. 10: Example of green SLA

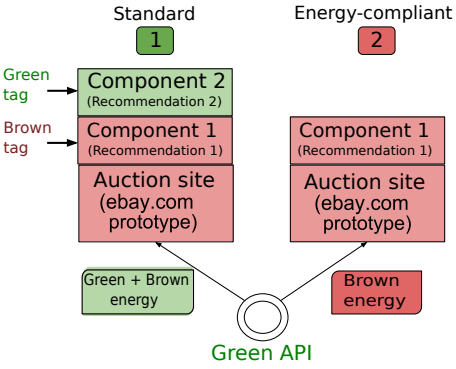


Fig. 11: Energy aware service differentiation

ciples, architecture and middleware components. EpoCloud is our prototype, which tackles three major challenges: 1) To optimize the energy consumption of distributed infrastructures and service compositions in the presence of ever more dynamic service applications and ever more stringent availability requirements for services; 2) To design a clever cloud's resource management, which takes advantage of renewable energy availability to perform opportunistic tasks, then exploring the trade-off between energy saving and performance aspects in large-scale distributed system; 3) To investigate energy-aware optical ultra high-speed interconnection networks to exchange large volumes of data (VM memory and storage) over very short periods of time.

In order to achieve these ambitious goals, we propose: 1) energy-aware SLA management policies considering energy as a first class resource and relying on the concept of virtual green energy to better utilize renewable energy; 2) energy-aware task scheduling algorithms based on the distinction of two kinds of tasks (web tasks and batch tasks) and leveraging renewable energy availability to perform opportunistic tasks without hampering performance thanks to prediction algorithms; 3) a specific OPS-based interconnection architecture to support the exchange of large data volumes (about 7.5 TB for the migration of all VMs hosted by a single server while allowing background traffic exchange between servers).

Acknowledgements This work has received a French state support granted to the CominLabs excellence laboratory and managed by the National Research Agency in the "Investing for the Future" program under reference Nb. ANR-10-LABX-07-01.

References

Beldiceanu N, Carlsson M, Douence R, Simonis H (2015a) Using finite transducers for describing and synthesising structural time-series constraints. Constraints pp 1–19

Beldiceanu N, Dumas Feris B, Gravey P, Hasan MS, Jard C, Ledoux T, Li Y, Lime D, Madi-Wamba G, Menaud JM, Morel P, Morvan M, Moulinard ML, Orgerie AC, Pazat JL, Roux O, Sharaiha A (2015b) The EPOC project: Energy Proportional and Opportunistic Computing system. In: International Conference on Smart Cities and Green ICT Systems (SMARTGREENS)

Chiaroni D, Santamaria GB, Simonneau C, Etienne S, Antona JC, Bigo S, Simsarian J (2010) Packet oadms for the next generation of ring networks. Bell Labs Technical Journal 14(4):265–283

Farrington N, Porter G, Forencich A, Ford J, Fainman Y, Vahdat A, Papen G (2013) Optical/electrical hybrid switching for datacenter communications. In: OptoElectronics and Communications Conference (OECC/PS), pp 1–2

Goiri I, Katsak W, Le K, Nguyen T, Bianchini R (2014) Designing and managing data centers powered by renewable energy. IEEE Micro 34(3):8–16

Green House Data (2007) Green house data. `http://www.greenhousedata.com/green-data-centers`

GreenQloud (2010) Greenqloud. https://www.greenqloud.com
Haque ME, Le K, Goiri Í, Bianchini R, Nguyen TD (2013) Providing green SLAs in high performance computing clouds. In: International Green Computing Conference (IGCC), pp 1–11
Hasan MS, Kouki Y, Ledoux T, Pazat JL (2014) Cloud Energy Broker: Towards SLA-driven Green Energy Planning for IaaS Providers. In: IEEE Internatonal Conference on High Performance Computing and Communications (HPCC)
Hasan MS, Kouki Y, Ledoux T, Pazat JL (2015) Exploiting Renewable sources: when Green SLA becomes a possible reality in Cloud computing. IEEE Transactions on Cloud Computing
Inc MI (2002) Power, heat, and sledgehammer. Tech. rep., University of Zurich, Department of Informatics
Indre RM, Pesic J, Roberts J (2014) Popi: A passive optical pod interconnect for high performance data centers. In: Optical Network Design and Modeling, 2014 International Conference on, IEEE, pp 84–89
Kachris C, Tomkos I (2012) A survey on optical interconnects for data centers. IEEE Communications Surveys & Tutorials 14(4):1021–1036
Kachris C, Tomkos I (2013) Power consumption evaluation of all-optical data center networks. Cluster Computing 16(3):611–623
Li Y, Orgerie AC, Menaud JM (2015) Opportunistic Scheduling in Clouds Partially Powered by Green Energy. In: IEEE International Conference on Green Computing and Communications (GreenCom), Sydney, Australia
Liboiron-Ladouceur O, Shacham A, Small BA, Lee BG, Wang H, Lai CP, Biberman A, Bergman K (2008) The data vortex optical packet switched interconnection network. Journal of Lightwave Technology 26(13):1777–1789
Pincemin E, Song M, Karaki J, Guillossou T, Thouenon G, Clavier R, Van Der Keur M, Le Bidan R, Gravey P, Froc G, Moulinard ML, Dumas Feris BP, Morvan M, Le Gall T, Jaouen Y, Poudoulec A, Betoule C, Grot D, Zia-Chahabi O (2014) Multi-band OFDM transmission at 100 Gbps with sub-band optical switching. Journal of lightwave technology 32(12):2202–2219, DOI 10.1109/JLT.2014.2322517
Ren C, Wang D, Urgaonkar B, Sivasubramaniam A (2012) Carbon-aware energy capacity planning for datacenters. In: IEEE International Symposium on Modeling, Analysis & Simulation of Computer and Telecommunication Systems (MASCOTS), pp 391–400
Sanjee I, Widjaja I (2004) A new optical network architecture that exploits joint time and wavelength interleaving. In: Optical Fiber Communication Conference, 2004. OFC 2004, vol 1, pp 446–448
Simsarian J, Zhang L (2004) Wavelength locking a fast-switching tunable laser. IEEE Photonics Technology Letters 16(7):1745–1747
Singla A, Singh A, Ramachandran K, Xu L, Zhang Y (2010) Proteus: a topology malleable data center network. In: Proceedings of the 9th ACM SIGCOMM Workshop on Hot Topics in Networks, ACM, p 8
Sridharan R, Mahapatra R (2010) Reliability aware power management for dual-processor real-time embedded systems. In: ACM Design Automation

Conference (DAC), pp 819–824
Yoo SJB (2006) Optical Packet and Burst Switching Technologies for the Future Photonic Internet. Lightwave Technology, Journal of 24(12):4468–4492