The EPOC project: Energy Proportional and Opportunistic Computing system

Nicolas Beldiceanu[1], Bárbara Dumas Feris[2], Philippe Gravey[2], Sabbir Hasan[3], Claude Jard[4], Thomas Ledoux[1], Yunbo Li[1], Didier Lime[5], Gilles Madi-Wamba[1], Jean-Marc Menaud[1], Pascal Morel[6], Michel Morvan[2], Marie-Laure Moulinard[2], Anne-Cécile Orgerie[7], Jean-Louis Pazat[3], Olivier Roux[5] and Ammar Sharaiha[6]

[1]*Mines de Nantes, LINA, France* [2]*Telecom Bretagne, Brest, France* [3]*INSA de Rennes, IRISA, France*
[4]*Université de Nantes, LINA, France* [5]*Ecole Centrale de Nantes, IRCCyN, France*
[6]*ENIB, LabSTICC-ENIB, France* [7]*CNRS, IRISA, France*

Keywords: Data-Center, Energy-Efficiency, Virtualization, Task Placement, Optical Network

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 criteria 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 EPOC project which focuses on optimizing the energy consumption of mono-site DCs connected to the regular electrical grid and to renewable energy sources.

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.

However, the extension of hydroelectric power plants is severely limited by environmental issues, and other renewable energy sources provide intermittent electricity over time. In the EPOC 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 a local renewable energy sources (such as windmills or solar cells).

Pioneering solutions have recently 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.

On the generation side, it is estimated that 10% of electric energy produced by power plants is currently lost during transmission and distribution to the consumers, with 40% of these losses occurring on the distribution network (Feng et al., 2009). For instance in 2006, in the United-States, the total energy and distribution losses were about 1,638 billion and 655 billion kWh, respectively (Feng et al., 2009). Most

of the energy-efficient Cloud frameworks proposed in the literature do not consider electricity availability and renewable energy in their models. This is a major drawback since significant amounts of electricity are lost during transportation and storage.

In the EPOC project, the 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 to integrate 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. EPOC is an ongoing project, and the aim of this paper is to present the DC architecture designed in this context, from hardware layer to middleware layer.

2 EpoCloud principles

Our first goal is to design an energy-proportional-computing system (EPCS), 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 EPCS, we need to have the capability to turn on/off servers dynamically. Vary-on/vary-off (VOVO) policy reduces the aggregate-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 permits to keep using VOVO policy in the on-line application context. 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. 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 more concentrated on Workload-driven approach. EpoCloud adapts the power consumption of the DC depending on the application workload. Our second objective is more focused on Power-driven SLA. The Power-driven approach implies shifting or scheduling the postponable workloads to the time period when the electricity is 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 a 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 without SAN

Recent studies on companies' data-centers show that a VM consume 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. Consequently, 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. Turning off a server involves transferring 7.5 TB on average. Given this scenario, a high broadband network is required, but is a 100 Gb/s network card really exploitable with current server technologies? In this article, we present an innovative network architecture, detailed in section 3.1, a pre-study in section 3.2, and finally, we describe in section 3.3 architectural motives and principles for the integration of renewable energy.

3.1 Network Architecture

A classical interconnection architecture is based on a 3-Tier fat-tree topology as presented in (Kachris and Tomkos, 2012). Whereas the three main switching layers are: core, aggregation, and ToR (top-of-the-rack); each layer, based on Electrical Packet Switches (EPS). Servers accommodated into racks are connected through the ToR switches to the aggregation layer, and from there to the core layer using the aggregate switches. Finally, by means of the core switches, servers can be interconnected 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, 2013). 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) that does not meet the needs of a variable traffic over time. Some hybrid architectures, involving several hierarchy levels, could have the potential to connect millions of servers in giant DC (Gumaste and Bheri, 2013). As already noted, EPOC aims at focusing on small/medium size data centers.

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 TWIN (Time-domain Wavelength Interleaved Networks), originally proposed by Lucent (Sanjee and Widjaja, 2004), and POADM (Packet Optical Add and Drop Multiplexer) proposed by Alcatel-Lucent (Chiaroni, 2008).

In the EPOC project, we decided to investigate a third option, derived from TWIN, which was presented in (Indre et al., 2014) under the name of POPI (Passive Optical Pod Interconnect). The main motivation for this choice is that POPI uses a purely passive optical network, with power consumption concentrated at networks edge. This architecture is simpler than the classical EPS one (Kachris and Tomkos, 2012), in the sense that there is no ToR switch and the existence of racks will depend on the bandwidth assigned per server (see POPI scheme depicted at Figure 1). Therefore, there is no difference between inter- or intra-rack communications. Servers are independent and can connect to each other by means of a passive coupler. Each server i has a transmitter constituted by a tunable laser, and a receiver adjusted to wavelength i. The estimate total power consumption of POPI is around 25% of the classical EPS architecture power consumption (Indre et al., 2014).

The maximum capacity of POPI in terms of number of servers (we consider one wavelength per server and no rack) is related to the limitations imposed by every component of the architecture. The tunable laser could present one of these limitations. In this paper we have taken into account the laser presented in (Chiaroni et al., 2010), with STM64 50 GHz-spaced channels and a tuning speed of 5 ns. We reserve three wavelengths: two for the controllers and one for the gateway. Thus, if we consider one wave-

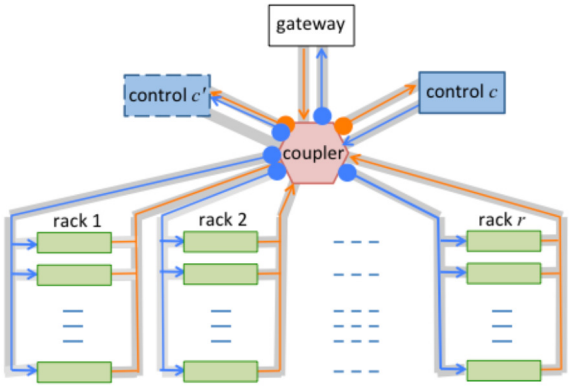

Figure 1: POPI architecture (from (Indre et al., 2014)).

length per server, the maximum number of servers shall be 61.

3.2 Server throughput capacity

In EPOC project, the DC consolidates the VMs periodically by using the live migration technology. Each server owns a 100 Gb/s transmission capacity and the live migration migrates the whole VM including the storage. It requires huge network resource to maintain the performance level therefore less degradation is employed during the process. In order to achieve a read/write speed of 100 Gb/s, we prefer SSD (*Solid-State Drive*) over HDD (*Hard Disk Drive*), since SSD is faster and less energy consuming than HDD. A single SSD of 800 GB capacity with a of PCI-E 2.1 x8 (*Peripheral Component Interconnect Express*) interface, can achieve 2 GB/s reading speed and 1 GB/s writing speed. This implies that we still need several SSDs in RAID (*Redundant Array of Inexpensive Disks*) technology in order to attain the 100 Gb/s-data rate. In the near future, a PCI-E 3.0 x16 shall offer a 15.75 GB/s network data rate, so this will be achieved by a single SSD.

Despite of higher speed, energy consumption for SSD is largely reduced to about 2 W compared to 6 W for HDD. Consequently, SSD generates less heat than HDD; making SSD more suitable to our project purposes.

3.3 Integrating Local Renewable Energy

Although several research efforts have been made to reduce energy consumption by designing/implementing server consolidation, hardware with better power/performance trade-offs, workload migration and software techniques for energy aware scheduling, 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 that, the demand for green services is ever increasing, thus integrating renewable sources to the data center left 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.

As renewable power sources are very intermittent in nature, hence 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 imbalance the Grid as renewable energy is connected to the Grid via grid-tie device, which combines electricity produced from renewable sources and Grid. 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 energy aware SLA between Infrastructure-as-a-Service (IaaS) and Software-as-a-Service (SaaS) provider when renewable energy 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.

In order to avoid using storage or batteries in small-scale DC, in the EPOC project, we propose to virtualize the green energy. Virtualization of energy implies nullifying the degraded interval (lack of green energy) with the surplus interval (excessive green energy than demand), whenever the availability of green energy is over the demand. From clients or SaaS providers perspective, they realize both the interval as ideal interval (when supply meets the demand), though the green energy was not present instantaneously rather present virtually as shown in Figure 2. So, whenever the green/renewable energy is present, we use the whole portion of the available green energy and draw the other portion from the grid. In this way energy storage is not needed and neither of the portion of renewable energy is wasted. Furthermore, total expenditure of energy purchasing can be reduced since no 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 if there is any.

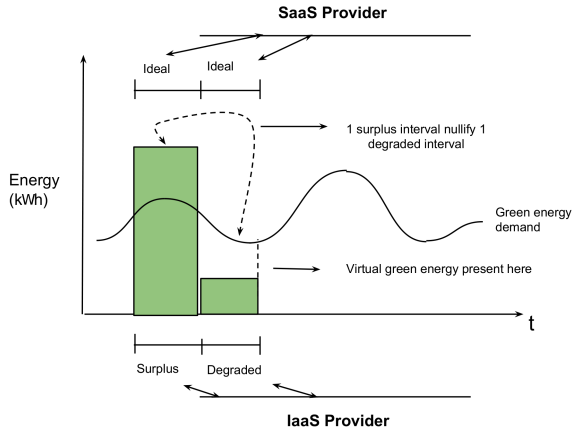

Figure 2: Concept of Virtualized Green Energy

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), and an energy-aware SLA oriented management system (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 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 on the data center: 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. Additionally for reducing further energy consumption in the EpoCloud, we are taking advantage of consolidation algorithms and on/off mechanisms to optimize the number of powered-on resources. These consolidation algorithms 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, and so, the algorithms take these costs into account to optimize the overall energy utilization.

4.2 Energy-aware SLA

In cloud systems, applications are embedded in VMs. A VM must be executed on a single server. So for an application, a transparent elastic system, also called vertical scaling (or scale up), consists in adapting resources to the VM needs, typically involving the addition of CPUs or memory. Vertical scaling relies on cloud computing models and virtualization techniques to scale up/down applications based on their performance metrics. Although those proposals can reduce the energy footprint of applications and by transitivity of cloud infrastructures, they do not consider the internal characteristics of applications to finely define a trade-off between the application's Quality of Service (QoS) and their energy footprint.

Contrary to the previous approach, Horizontal scaling (or scale out) implies to add more resources for an application, such as adding a new VM to a distributed software application. Nonetheless, one needs to change the application configuration. For example an Apache Web application is scaled out by adding/removing VM apache workers and by modifying the mod_proxy_balancer file configuration.

One of the main challenges is to consider both the application internals and the global system to reduce the energy footprint in our cloud infrastructure. More precisely, we focus on adding the usual scaling up/down by considering all application's internal to

be able to use several and different application configurations (corresponding to different quality of service (QoS) level). Each application is equipped with one autonomic loop in charge of determining the minimum amount of resources required to provide the best QoS possible while an additional loop manages the physical resources at the infrastructure level. The autonomic loop may switch from one configuration to another according to the incoming charge, the QoS expectations and the infrastructure constraints.

5 Conclusion

In this paper, we have presented the EpoCloud principles, architecture and middleware components. EpoCloud is our prototype, which will tackle 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) To determine 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) To evaluate 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; 3) To assess the ability of 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).

ACKNOWLEDGMENTS

The author acknowledges the support of the CominLabs Labex.

REFERENCES

Chiaroni, D. (2008). Optical packet add/drop multiplexers for packet ring networks. In *Optical Communication, 2008. ECOC 2008. European Conference on*, pages 1–4.

Chiaroni, D., Neilson, D., Simonneau, C., and Antona, J. C. (2010). Novel Optical Packet Switching Nodes for Metro and Backbone Networks. In *Optical Network Design and Modeling (ONDM), International Conference on*.

Feng, X., Peterson, W., Yang, F., Wickramasekara, G., and Finney, J. (2009). Smarter grids are more efficient. ABB review.

Goiri, I., Katsak, W., Le, K., Nguyen, T., and Bianchini, R. (2014). Designing and managing data centers powered by renewable energy. *Micro, IEEE*, 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`.

Gumaste, A. and Bheri, B. (2013). On the architectural considerations of the FISSION (Flexible Interconnection of Scalable Systems Integrated using Optical Networks) framework for data-centers. In *Optical Network Design and Modeling (ONDM), International Conference on*, pages 23–28.

Inc, M. I. (2002). Power, heat, and sledgehammer. Technical report, University of Zurich, Department of Informatics.

Indre, R.-M., Pesic, J., and Roberts, J. (2014). POPI: A Passive Optical Pod Interconnect for high performance data centers. In *Optical Network Design and Modeling, 2014 International Conference on*, pages 84–89.

Kachris, C. and Tomkos, I. (2012). A survey on optical interconnects for data centers. *Communications Surveys Tutorials, IEEE*, 14(4):1021–1036.

Kachris, C. and Tomkos, I. (2013). Power consumption evaluation of all-optical data center networks. *Cluster Computing*, 16(3):611–623.

Sanjee, I. and Widjaja, I. (2004). A new optical network architecture that exploits joint time and wavelength interleaving. In *Optical Fiber Communication Conference, 2004. OFC 2004*, volume 1, pages 446–448.

Singla, A., Singh, A., Ramachandran, K., Xu, L., and Zhang, Y. (2010). Proteus: A Topology Malleable Data Center Network. In *ACM SIGCOMM Workshop on Hot Topics in Networks (Hotnets)*, pages 8:1–8:6.

Sridharan, R. and Mahapatra, R. (2010). Reliability aware power management for dual-processor real-time embedded systems. In *Proceedings of the 47th Design Automation Conference*, DAC '10, pages 819–824, New York, NY, USA. ACM.

Yoo, S. J. B. (2006). Optical Packet and Burst Switching Technologies for the Future Photonic Internet. *Lightwave Technology, Journal of*, 24(12):4468–4492.