Consiglio Nazionale delle Ricerche
Istituto di Calcolo e Reti ad Alte Prestazioni

Efficiency and Green Metrics for Distributed Data Centers

Carmine De Napoli[1], Agostino Forestiero [1,2], Demetrio Laganà[1], Giovanni Lupi[1], Carlo Mastroianni[1,2], Leonardo Spataro[1]

RT-ICAR-CS-16-04 **September 2016**

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 712363

1) Eco4Cloud srl, Piazza Vermicelli, Rende (CS), email: {denapoli,lagana,spataro}@eco4cloud.com
2) ICAR-CNR, Via P. Bucci 7/11 C, Rende (CS), email: {forestiero,mastroianni}@icar.cnr.it

Consiglio Nazionale delle Ricerche, Istituto di Calcolo e Reti ad Alte Prestazioni (ICAR)
– Sede di Cosenza, Via P. Bucci 7-11C, 87036 Rende, Italy, URL: *www.icar.cnr.it*
– Sezione di Napoli, Via P. Castellino 111, 80131 Napoli, URL: *www.icar.cnr.it*
– Sezione di Palermo, Via Ugo La Malfa, 153, 90146 Palermo, URL: *www.icar.cnr.it*

Table of contents

Introduction

This report has two objectives: (i) individuate and discuss the efficiency and green metrics that can be used to assess the performances of distributed data centers (ii) individuate the quantitative and qualitative targets related to such metrics. The content of this report derives from the activities of the EcoMultiCloud project, a project of the CNR spinoff Eco4Cloud srl. EcoMultiCloud is funded by the European Community under the “SME Instruments” tool of the Horizon 2020 program.

The report is organized as follows:

- Section 1 introduces and describes the performance metrics that can be used to assess the performance of a multi data center environment. Such metrics are classified into: metrics regarding the energy efficiency, metrics concerning the use of green energy, metrics on latency and communication performances, metrics on the cost of energy, metrics about the quality of service perceived by users and metrics on the recovery from situations of interruption or limitation of electricity;
- Section 2 explains why there is a clear lack of standard metrics about the efficiency of the computational components of data centers, and presents a new metric introduced by Eco4Cloud, the HUE (Host Usage Effectiveness) that intends to fill this gap. This metric will be used to assess the efficiency of the strategies and policies for the assignment and redistribution of the workload among the data centers;
- Section 3 focuses on the metrics related to the efficient usage of hardware resources in data centers. In particular, the section illustrated the CPU Ready Time and the Memory Ballooning indices that are used to assess the correct usage of CPU and RAM memory, respectively;
- Section 4 concludes the report.

1. Definition and Application of Metrics

The performance of EcoMultiCloud will be measured in accordance to different types of metrics, which will be discussed in the following paragraphs.

1.1. Energy Efficiency Metrics

Fig. 1 illustrates how the energy is used and distributed in a data center. This is the basis to understand the different sets of metrics that can be used to measure the energy efficiency of a data center.

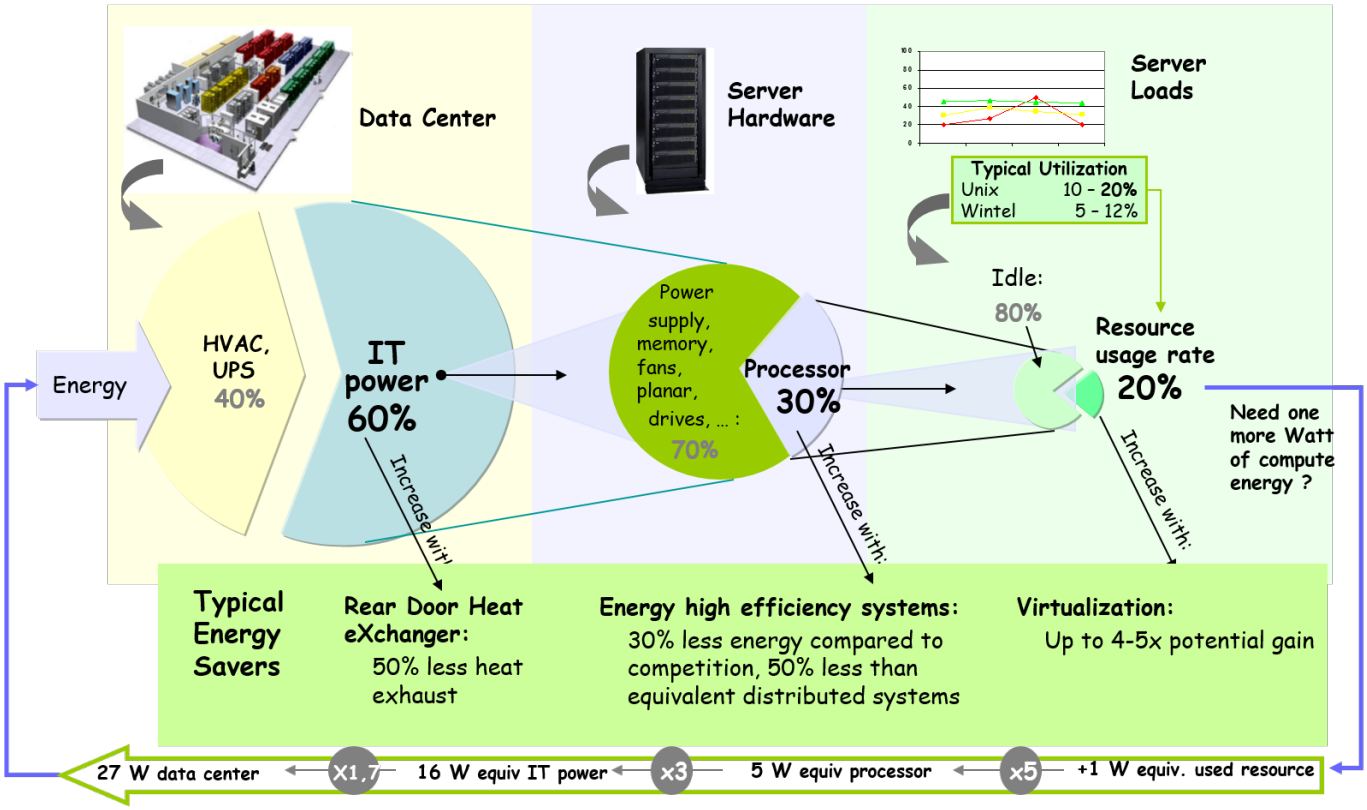

Fig. 1 Energy chain in a data center.

The figure shows that not all the energy entering the system is actually used in the IR component, i.e., the component that runs the applications and performs the useful work. In the example illustrated in the figure, 40% of the energy is used for components like HVAC (Heating, Ventilating and Air Conditioning) and UPS (Uninterruptible Power Supply). Then, the figure shows that also in the IT component, only a fraction of energy, 30% in this example, is strictly used for computation, while the rest is used for power supply, fans, drives, etc. Finally, the rightmost part of the figure shows that the computational facilities is often highly underutilized, i.e.,

the energy is used to run much less computation than the server could support. The virtualization approach certainly helps to increase the utilization, but Eco4Cloud has demonstrated the algorithms for consolidation re essential to efficiently redistributed the workload in a dynamic environment.

It is therefore essential to understand the different metrics are adopted to measure the use of energy at the different stages of the energy chain and at the different components of a data center.

The general formulation of energy efficiency can be expressed with the following expression:

$$Efficiency = \frac{Computation}{Total\ Energy(P_{IN})} = \left(\frac{1}{PUE}\right) \times \left(\frac{1}{SPUE}\right) \times \left(\frac{Computation}{P_{IT}}\right)$$

In the expression, the different terms are:

- **PUE** captures the inefficiencies due to power delivery and cooling of the datacenter
- **SPUE** captures the inefficiencies due to power delivery and cooling of the server. These can be server’s power supply, voltage regulator modules and cooling fans. SPUE is defined as the ratio of the total server input power over the useful server power, i.e the power consumed by motherboards, CPU, DRAM, I/O cards, etc. The combination of PUE and SPUE measures the total losses associated to non critical components that exist in the data center’s NCPI (Network-Critical Physical Infrastructure) and IT equipments.
- **Computation** is the useful work
- $\mathbf{P_{IT}}$ is the power delivered to IT equipments

The rest of this section is devoted to the illustration of the metrics regarding the physical efficiency. The computational will be illustrated in Section 2, which also introduces a new metric, the Host Usage Effectiveness.

Energy efficiency metrics (PUE, DCiE)

The Green Grid [1] is an association of IT professionals that, with a series of proposals, aim to increase the energy efficiency of data centers. The Green Grid proposed the use of Power Usage Effectiveness (PUE) and its reciprocal Datacenter Infrastructure efficiency metrics, (DCiE), to enable the rapid assessment of energy efficiency of a data center and compare the results with other data centers and eventually evaluating appropriate corrections to improve the situation.

The **PUE** is defined as:

$$PUE = \frac{\text{Total Facility Power}}{\text{IT Equipment Power}}$$

and its reciprocal, the **DCiE** is defined as: $$\frac{1}{PUE} = \frac{IT\ Equipment\ Power}{Total\ Facility\ Power} \times 100\%$$

In this expression, **Total Facility Power** is defined as the power provided to the datacenter, while **IT Equipment Power** is defined as the equipment that is used to manage, process, store, or route data within the data center.

The components for the loads in the metrics can be described as follows:

- IT EQUIPMENT POWER. This includes the load associated with all of the IT equipment, such as compute, storage, and network equipment, along with supplemental equipment such as KVM switches, monitors, and workstations/laptops used to monitor or otherwise control the datacenter
- TOTAL FACILITY POWER. This includes everything that supports the IT equipment load such as:
  - Power delivery components such as UPS, switch gear, generators, PDUs, batteries, and distribution losses external to the IT equipment.
  - Cooling system components such as chillers, computer room air conditioning units (CRACs), direct expansion air handler (DX) units, pumps, and cooling towers.
  - Compute, network, and storage nodes.
  - Other component loads such as datacenter lighting.

IT Equipment Power would be measured after all power conversion, switching, and conditioning is completed and before the IT equipment itself. The most likely measurement point would be at the output of the computer room power distribution units (PDUs). This measurement should represent the total power delivered to the compute equipment racks in the datacenter.

The PUE can be used to measure if the energy allocation in the datacenter is efficient. The PUE can range from 1.0 to infinity. For example, if a PUE is 2.0, this indicates that the datacenter demand is two times greater than the energy necessary to power the IT equipment. In addition, the PUE can be used as a multiplier for calculating the real impact of the system’s power demands. For example, if a server demands 300 watts and the PUE for the datacenter is 2.0, then the power from the utility grid needed to deliver 300 watts to the server is 600 watts.

Ideally, a PUE value approaching 1.0 would indicate 100% efficiency, as all the power is used by IT equipment only. Some researches indicate that PUE values of 1.6 are today achievable with proper design in any data center.

The DCiE index is useful as well. A DCiE value of 33% (equivalent to a PUE of 3.0) suggests that the IT equipment consumes 33% of the power in the datacenter.

A number of components influence the total facility load. The cooling infrastructure may consume 40% of the incoming electricity as in the example above. For this reason, a user may want to specifically measure and trend consumption in the central plant. In Table 1, Total facility Power is the sum of all other value and the PUE is calculated as

PUE = 300 / 170 = 1,76

Now, let's calculate DCIE:

DCIE = 1 / 1,76 = 170 / 300 x 100% = 56,66% (about 57%)

PUE and DCiE Calculation	
Total IT Power	170kW
Cooling Infrastructure	96kW
Power System power	30kW
Lighting power	4kW
Total Facility power	300kW
PUE	**1.76**
DCiE	**56%**

Table 1. Example of PUE AND DCiE Calculation

The Green Grid propose a table of PUE and DCiE reference values - so, our example is for almost efficient data center:

PUE	DCiE	Level of Efficiency
3.0	33%	Very Inefficient
2.5	40%	Inefficient
2.0	50%	Average
1.5	67%	Efficient
1.2	83%	Very Efficient

Table 2. PUE Reference values

Some achievements regarding the PUE index

In October 2008, Google's Data center was noted to have a ratio of 1.21 PUE across all 6 of its centers, which at the time was considered as close to perfect as possible. Right behind Google, was Microsoft, which had another notable PUE ratio of 1.22 [2].

Through proprietary innovations in liquid cooling systems, French hosting company OVH has managed to attain a PUE ratio of 1.09 in its data centers in Europe and North America [3].

In October 2015, Allied Control has a claimed PUE ratio of 1.02 [4] through the use of 3M Novec 7100 fluid.

As of the end of Q2 2015, Facebook's Prineville data center had a power usage effectiveness (PUE) of 1.078 and its Forest City data center had a PUE of 1.082 [5].

In January 2016, the *Green IT Cube* in Darmstadt was dedicated with a 1.07 PUE [6]. It uses cold water cooling through the rack doors.

1.2. Green Metrics

This section discusses the most utilized metrics involving the environmental sustainability of a data center. They are: CUE and WUE.

Carbon emissions (CUE)

The most important metric related to environmental sustainability of a data center, defined by The Green Grid consortium [7], is the CUE, or Carbon Usage Effectiveness. The CUE considers the carbon footprint of a given datacenter, measuring the greenhouse gas emissions in relation to IT energy consumption.

The impact of operational carbon usage is emerging as extremely important in the design, location, and operation of current and future data centers. When used in combination with the power usage effectiveness (PUE) metric, data center operators can quickly assess the sustainability of their data centers, compare the results, and determine if any energy efficiency and/or sustainability improvements need to be made. CUE is the second metric (after PUE) in the family of *x*UE metrics designed to help the data center community better manage the energy, environmental, societal, and sustainability-compliance parameters associated with building, commissioning, operating, and de-commissioning data centers.

CUE is defined as follows:

$$CUE = \frac{Total\ CO_2\ emissions\ caused\ by\ the\ Total\ Data\ Center\ Energy}{IT\ Equipment\ Energy}$$

Total Data Center Energy is the same value as the numerator of the PUE metric. The numerator in this CUE metric is the total carbon emissions caused by the use of the energy in the PUE metric. The units of the CUE metric are kilograms of carbon dioxide (kgCO2eq) per kilowatt-hour (kWh).

Total CO2 Emissions include the CO2 emissions from local and energy grid–based energy sources. Ideally, the CO2 emissions will be determined for the actual mix of energy delivered to the site (e.g., the electricity may have been generated from varying CO2-intensive plants—coal or gas generate more CO2 than hydro or wind. The mix also must include other energy sources such as natural gas, diesel fuel, etc.). The total CO2 emissions value will include all GHGs, such as CO2 and methane (CH4). All emissions will need to be converted to “CO2 equivalents.”

Alternatively CUE can be calculated as **PUE x CEF**. CEF is the carbon dioxide emission factor, calculated and available from the EPA website or

$$CEF = \frac{CO_2\ emmited\ (kg)}{unitEnergy\ (kWh)}$$

Like PUE, CUE uses the familiar value of total IT energy as the denominator but, unlike PUE, CUE has dimensions. Another important difference is the range of values. PUE has an ideal value of 1.0, as said in the previous

paragraph, implying that all the energy used at the site goes to the IT equipment. The ideal value of CUE is 0.0, i.e., no carbon use is associated with the data center’s operations. Like PUE, CUE has no theoretical upper boundary.

Both CUE and PUE simply cover the operations of the data center. They do not cover the full environmental burden of the life-cycle of the data center and IT equipment. For example, attempting to determine the carbon generated in the manufacturing of the IT equipment and its subsequent shipping to the data center would make the metric far too difficult to measure, calculate, or use. The Green Grid considers the full life-cycle to be important to the overall sustainability of the industry but, for practical considerations, they excluded it from this metric.

Water emissions (WUE)

A great amount of electricity is used by Data centers. Electricity produced by thermoelectric and nuclear power consumes a huge amount of water in the power plant through steam condensation (i.e., water evaporates from cooling towers into the environment). Most data centers are located in places where installation of advanced cooling systems is not feasible or economical. It was estimated that a 15MW data center could consume 360,000 gallons of cooling water each day. The problem of cooling water wasting is very strong even in regions in which this element is abundant. Indeed, reducing water by 10-25% is a prerequisite for green certifications (e.g. LEED program [8]) which provide tax/zoning benefits and are being actively pursued by 77% of large data.

The Green Grid [9] developed a metric for measuring data center water efficiency as Water Usage Effectiveness (WUE), which is defined as the ratio of total water consumption to the IT energy usage. Water consumption (e.g., evaporation into the air) is more accurate than water withdrawal because it is a more accurate indicator of how much water does not return to source (i.e., “lost”).

Like PUE and CUE, the WUE metric uses the value of **IT Equipment Energy** as its denominator. So the same value determined for PUE or CUE, should be used as the denominator for this new metric as well. Unlike PUE, WUE and CUE have dimensions, while PUE is unit-less. Another important difference is the range of values. PUE has an ideal value of 1.0, implying that all energy used at the site goes to the IT equipment. There is no theoretical upper boundary for PUE. WUE has an ideal value of 0.0, indicating that no water use is associated with the data center’s operations. WUE has no theoretical upper boundary, like PUE. The metric for water usage in the data center is defined at a high level as:

$$WUE = \frac{\text{Annual Water Usage}}{\text{IT Equipment Energy}}$$

The units of WUE are liters/kilowatt-hour (L/kWh).

With WUE, the issue of a “source-based” versus “site-based” metric must be considered. The main issue is that water use or changes to a site’s water use strategy generally affects other site use parameters and also can affect the supply chain for different utilities. A reduction in water use on-site can be accomplished in a number of ways. The most attractive way is to employ optimal design, then increase operational efficiencies and tune the existing systems. Recommissioning a facility can accomplish this.

1.3. Latency and Communication Metrics

Cloud computing systems provide on-demand access to the pool of shared computing resources over the Internet. Therefore, communication processes, not computing, often define the efficiency of the cloud. In this section, we present a set of metrics which capture performance and describe energy efficiency of data center communication systems.

Network Latency: Cloud applications are found to be extremely sensitive to communication delays [10], [11]. Therefore, an ability to monitor and control network latency is especially important to guarantee Quality of Service (QoS) and Service Level Agreements (SLAs). The Uplink/Downlink Communication Latency (UDCL), or Uplink/Downlink Hop Distance (UDHD), if expressed in the number of hops, measures the time (in seconds) needed for a request incoming to the data center to reach a computing server (downlink), or the time it takes for a computing result to leave the data center network (uplink) and be on the way to the end user. UDCL is added on top of the task execution time for every user request. Network topologies hosting computing servers closer to the data center gateway have smaller UDCL and can provide faster response times.

The network delay of a single packet is composed of the transmission delay Dt and link propagation delay Dp. Dt is expressed as a ratio between packet size S and link rate R, while Dp is defined as the link length L over the signal propagation speed P. P defines the physical characteristic of the medium. In copper, the propagation speed is two thirds of the light speed.

$$delay\ of\ a\ single\ packet = \frac{transmission\ delay\ (Dt)}{link\ propagation\ delay\ (Dp)}$$

$$Dp = \frac{Link\ Lenght}{signal\ propagation\ speed}$$

$$Dt = \frac{packet\ size}{link\ rate}$$

Network Losses: The packets travelling in a data center network may be lost and fail to reach destination due to link errors. Packet loss is typically caused by network congestion. Packet loss is measured as a percentage of packets lost with respect to packets sent. These errors may cause significant communication delays, as retransmissions are performed at the transport layer using TCP protocol. Therefore, measuring error rates is important to assure network performance and to help detecting hardware faults. Packet loss can reduce throughput for a given sender, whether unintentionally due to network malfunction, or intentionally as a means to balance available bandwidth between multiple senders when a given router or network link reaches or approximates its maximum capacity (see https://en.wikipedia.org/wiki/Packet_loss - cite_note-4)

When reliable delivery is necessary, packet loss increases latency due to additional time needed for retransmission. Assuming no retransmission, packets experiencing the worst delays might be preferentially dropped (depending on the queuing discipline used) resulting in lower latency overall at the price of data loss

In addition, it is possible to diversify resource allocation strategies taking into account the sensitivity of cloud applications to transmission errors. In data centers, interconnection links are not identical. The amount of packet loss that is acceptable depends on the type of data being sent. For example, for Voice over IP, missing one or two packets every now and then will not affect the quality of the conversation. Only losses between 5% and 10% of the total packet stream will affect the quality significantly. Less than 1% packet loss can be considered "good" for streaming audio or video, while 1-2.5% is "acceptable". On the other hand, when transmitting a text document or web page, a single dropped packet could result in losing part of the file, which is why a reliable delivery protocol must be used for this purpose (which means retransmission of dropped packets).

Migration time and downtime

Live migration allows to move a running virtual machine (VM) to a new physical host with minimal service interruption. This is a very attractive tool for various scenarios in dependable computing. Currently the predominant use of live migration is in data centers or compute clouds where VMs can be moved across physical hosts for load balancing, server consolidation or maintenance. In all these cases knowing the downtime involved by moving the VM is essential when service availability guarantees have to be fulfilled: the time of service interruption must not exceed the clients retry intervals. Within the different existing virtualization frameworks with live migration support, the basic principle is that the virtualization cluster management actively moves a virtualized system while it is still executing and is still changing the hardware’s and software’s state. Today’s products realize this by a delta-copying approach where modified memory regions are incrementally transferred until a lower threshold for data to be moved is reached. In the subsequent phase in which the VM is stopped, the remaining resources are copied and reconfigured and the VM is resumed on the destination host. This leads to the two metrics:

- **migration time**: it is the time from start of the live migration process until the virtualization framework declares the physical source host to be no longer relevant for the execution of the migrated virtual machine. The maximum tolerable migration time is determined by internal dependability assumptions at the provider side, e.g., maintenance intervals.
- **downtime or blackout time**: it is the phase during live migration when there is a temporary (potentially user perceptible) service unavailability, caused by the virtual machine suspending execution for the finalization of the movement. From a dependability perspective, blackout time is a crucial quantity when a virtualized service (e.g. server application) needs to fulfill reliability guarantees. Blackout time limits are therefore driven by dependability contracts between the service provider and client.

The most difficult procedure in live migration is the transfer of main memory state. As live migration environments typically share storage within the migration cluster, swapped out memory pages do not have to be considered. Both metrics are actually only determinable through direct experimentation.
The overall duration of live migrations and the short downtime during this process are essential properties when implementing service availability agreements.

We intend to perform a set of migrations of VMs having different profile (e.g., CPU-intensive or disk-intensive VMs), in order to understand how the migration time is related to the characteristics of VMs. More details about the planned experiments are given in Section 4.2.

1.4. Energy Cost

The cost of the energy consumed in a single data center can be obtained with the expression:

$C = P * PUE * C_{comp}$

In this expression, Ccomp is the energy consumed by the IT component of the data center, and P is the price of electricity. PUE accounts for the cost of the energy consumed by the physical components: power distribution, cooling, etc. The overall cost can be obtained by summing up the cost related to the different data centers.

In a geo-distributed environment, with several data centers located in different regions or countries, it is then possible to exploit the “follow the moon” paradigm, i.e., dynamically move portions of the workload where the price of electricity and/or the PUE is lower. Indeed, the cost of electricity is generally different from site to site and also varies with time, even on a hour-to-hour basis in some countries.

On the other hand, also the PUE varies both spatially and temporally. In a multi data center scenario, some data centers can be more efficient than others, for example due to more efficient chillers or simply because of the different weather conditions. Moreover, the same data center can have lower values of PUE in night hours, because cooling needs are less severe.

Some examples of these variations are given in Section 4.2, where we introduce the scenario of some experiments that we are performing to assess the advantages related to the “follow the moon” paradigm.

1.5. Quality of Service

The metrics defined in Section 1.3 are useful to assess the quality of the service offered to users, in particular the downtime and the latency experienced during migrations. Other metrics are related to the quality of service: among them, the balance of load among the data centers, the CPU Ready Time and the Memory Ballooning. The load balancing is discussed in the following, while the CPU Ready Time and the Memory Ballooning are discussed in Section 3, where we also describe an innovative approach defined by Eco4Cloud to efficiently handle the memory: the *Smart Ballooning*.

Load balancing

The capacity of a data center is related to the usage of the bottleneck resource, which typically is the RAM memory. Indeed, in most data centers on which Eco4Cloud has operated, the percentage of applications and VMs that are memory intensive is generally higher than the percentage of CPU- and disk-intensive applications and VMs.

Therefore, the capacity of a data center can be defined as the percentage of RAM used with respect to the overall amount of memory available on all the hosts of the data center.

The load balancing measures the balance of the capacities measured on the different data centers. Indeed, the quality of service is deteriorated when one or more data centers are overloaded with respect to the others. The overloading of a data center can lead to the failed adherence to the Service Level Agreements, for example in terms of responsiveness and reliability of applications. We measure the load balancing using the HUE metric introduced in Section 2: the objective is to have similar values of the HUE metric in the different data centers.

1.6. Adaptation and Recovery

The benefits of Cloud computing, such as the ability to execute long-running, computationally and data intensive scientific experiments, the achievement of high availability of applications and so on, are hindered by significant problems in the underlying computing and networking infrastructures. For example, many Cloud service providers sometimes suffer from failures in the power supply networks and Internet connectivity issues. For Cloud-based services, dynamic planning for infrastructure disruptions and subsequent orchestration and management of disaster recovery are major issues. Although such infrastructure-related problems are particularly severe in developing countries, they certainly exist also in the developed parts of the world too, and hinder the adoption of the Cloud computing technology.

The EcoMultiCloud solution provides the ability to tailor resilience, availability and other QoS parameters to meet the needs of a wide range of potential scenarios, where the main requirements could be any one of:

- **maximize service availability** to all customers;
- **minimize the impact of any failure** on any given customer. For example, the service should degrade gracefully, which means that non-critical components of the service may fail but critical functions still work;
- **minimize the number of customers** affected by any failure;
- **minimize the number of minutes** that a customer (or customers) cannot use the service in its entirety (e.g. by using migration where necessary in order to resume normal service);
- **maintain service performance** during periods of increased load, in a cost-effective way;
- **maximize business continuity.** Focus on how the organization and the service respond when a failure does occur - including the implementation of disaster recovery strategies.

In case of energy interruption, it is needed to reduce the energy consumption for a given interval of time, which depends on the expected duration of the interruption and the amount of energy provided by the UPS. To reduce the energy consumption, it is necessary to lower the load of the data center to a predetermined amount, called “emergency target”. The emergency target can be defined as the maximum overall CPU utilization, since the energy consumption on the data center servers mainly depends on the CPU utilization.

The achievement of the requirements mentioned above is strictly related to the reduction of CPU utilization so as to reach the emergency target. Therefore, we intend to measure the time needed to recover the data center

affected by a power interruption, i.e., the time needed to reduce the data center utilization to the emergency target. This is strictly related to the time needed to migrate VMs to more stable data centers.

The policy adopted for the assignment of VMs on the different data centers can affect the recovery time, since the migration time highly depends on the specific characteristics of the VMs running on the critical data centers. For example, a CPU-intensive VM can be migrated more rapidly than a disk-intensive VM, and with better effect on the reduction of power consumption.

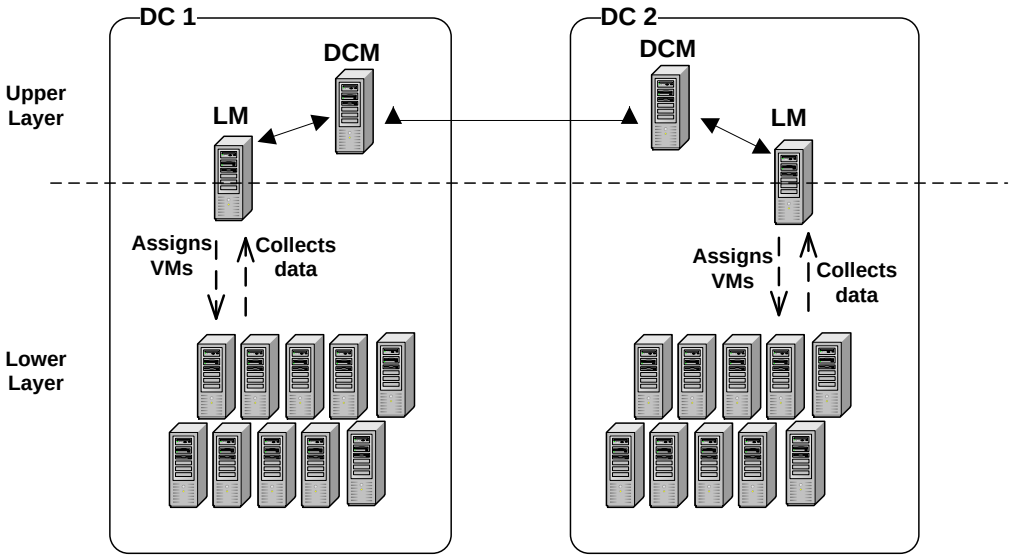

Fig. 2. EcoMultiCloud hierarchical architecture

The reference scenario is depicted in Fig. 2, which shows the upper and lower layer for two interconnected data centers, as well as the main involved components. At each data center, a data center manager (DCM) runs the algorithms of the upper layer, while the local manager (LM) performs the functionalities of the lower layer.

2. HUE – Host Usage Effectiveness

The objective if this section is to introduce a new metric, defined by Eco4Cloud, which aims to fill the gap regarding the current lack of standard metrics for the assessment of the computational efficiency of data centers. In modern data centers physical efficiency is generally high, but computational efficiency is far from being acceptable. Yet most standard performance indices are still focused on the first aspect.

In the Fig. 3 below we see that the IT efficiency is on average 30% on the data centers owned by big companies, and this average reduces to only 15% on the data center of a generic company. This is clearly unacceptable.

Reliable Datacenter PLC		Colossalcloud inc	
Servers/IT equipment • Branded servers • Bought by customer?	Full price (100%)	Servers/IT equipment • ODM or custom built • Large volume discount	50%
IT efficiency (utilization) • Few VMs per server • Storage utilization low	15%	IT efficiency (utilization) • Many floating VMs per server • Storage managed	30%
Throughput per watt • Moore's Law • Servers swapped out slowly	2x every 2-6 years	Throughput per watt • Moore's Law drives performance • Servers swapped out rigorously	2 x every 2 years or less
Energy price • Must pay local rates	12 cents per KWh	Energy price • Site selection secures low price	3 cents per KWh
Energy overhead (PUE) • Higher tier limits energy savings • Site, customers limit innovation	x 1.7	Energy overhead (PUE) • Engineers drive down energy • Business model allows low tiers	x 1.3
Facility build costs • Higher tiers cost more • City sites cost more, limit space	€12m per MW	Facility build costs • Reduced infrastructure costs less • Remote sites cost less	€6m per MW
Financing costs • Capital may need to be raised • VAT, rates, income tax can be high	3-8% capital & taxes	Financing costs • Cash on balance sheet or cheap • VAT, rates, income tax negotatied	Very low

Fig. 3 Technical and business characteristics of regular and very big data centers (Source: Andy Lawrence, 451 Research, "Datacenters in a cloud storm")

In Section 1 we have seen that there are several well accepted indices for physical efficiency, among which:

- PUE (Power Usage Effectiveness): total energy of the data center divided by the IT energy consumed
- CUE (Carbon Usage Effectiveness): greenhouse gas emissions in relation to IT energy consumption
- WUE (Water Usage Effectiveness): water used on-site for data center operations

- DCeP (Datacenter Energy Productivity): complicated formula to express the useful work divided by total energy consumed to produce this work

On the other hand, “there has yet to be designed a dynamic metric for IT efficiency that is scalable and adjustable for various types of operations”, as it is reported in *[451 Reseach] Datacenter environmental initiatives - Government lags industry*

There are several reasons why physical efficiency indices have been standardized while computational indices are not. In our opinion, the main two are: (i) it easier to improve the physical efficiency, because this can be done without inspecting the actual workload which is carried by data centers, so companies tend to measure and advertise improvements that they have actually achieved; (ii) it is easier to *define* the physical efficiency, while computational efficiency often depends from the type of data center, the type of carried workload, the metrics used to measure the computation itself, etc.

An index for computational efficiency must measure how effective is the management of computational resources, and how much workload can be supported for a given amount of resources.

The basic requirements of a computational efficiency index are the following:

1) It must be easy to understand and measure
2) Its definition should be similar to other established indices (PUE, WUE etc.): a ratio, with the optimal value set be 1, as for the PUE
3) It should clearly indicate the margin for improvement, and what will happen if better strategies/algorithms/mechanisms for workload management are adopted
4) It should be able to compare the effectiveness of different data centers

On the other hand, as said before, no standard definition of a computational efficiency index has been accepted so far. The main obstacles for the definition of a universally accepted standard are:

1) The definition of workload is uncertain and may depend on the specific scenario/application
2) Even if a definition of workload is accepted, it is difficult to measure, due to the vast heterogeneity of applications (e.g., CPU-intensive, memory-intensive, transactional, batch, centralized, parallel/distributed etc.)
3) Relevant computational resources may be different in different cases: CPU, memory, bandwidth etc.
4) Even if the relevant resource is correctly identified, it is difficult to quantify its usage, due to the vast heterogeneity and typology of servers

Eco4Cloud has defined an index that overcomes all the mentioned obstacles and matches all the requirements listed before. Consequently, Eco4Cloud uses this definition to measure the improvements it brings both to single data centers and to multi data center environments.

This index is called HUE: Host Usage Effectiveness. This index is first defined for each subset of homogenous servers, then for an entire data center or multi data center. Indeed, it is normal practice that the servers of a data

centers are acquired at different times, and that the servers acquired together are equal or very similar, i.e., they have the same characteristics in terms of operating system, virtualization environment, CPU, memory, etc.

In the following, we introduce the definition of HUE for a homogeneous set of servers, then we extend and generalize the definition for a heterogeneous set of servers.

2.1. HUE for a homogeneous set of servers

The definition of the HUE index derives from the well accepted consideration that data center resources are under-utilized. First let us first consider the index definition for a set of homogeneous servers. The steps are:

1) Select the bottleneck resource, **R**. It is the hardware resource (CPU, RAM, disk, bandwidth…) with the highest relative utilization. In many cases, it is the RAM memory
2) Compute/retrieve the avg. utilization of the bottleneck resource on the servers, **U**
3) Set the utilization threshold, **T**, for the bottleneck resource in a single server, i.e., the maximum value of utilization that still meets the QoS requirements. For example, if the bottleneck resource is the CPU and the CPU utilization threshold is 80%, **T** =0.8

If **N** is the number of homogenous servers, **Ui** is the utilization of server *i*, and **U** is the average utilization, the total occupancy (load) of the bottleneck resource is:

$$\mathbf{L = \sum_{i=1}^{N} U_i = N * U}$$

The same load can ideally be supported by a lower number of servers in which the utilization of the bottleneck resource is the maximum allowed, **T**

$$\mathbf{L = N_{eq} * T}$$

In this expression, $\mathbf{N_{eq}}$ is referred to as the number of *equivalent servers*

Now, we can give the definition of the HUE for a homogeneous set of servers:

$$\mathbf{HUE = \frac{N}{Neq} = \frac{L/U}{L/T} = \frac{T}{U}}$$

The meaning of the HUE index is the following: the HUE index is the ratio of the number of servers actually used to the *minimum number of servers* that could be used to support the same load if they were used at full capacity.

The main characteristics of the HUE are the following:

1. It is very easy to compute
2. It clearly indicates the computational efficiency. For example:
    a) HUE=1 means that the computational efficiency is optimal;
    b) HUE=2 means that the number of utilized servers is twice the number of servers strictly necessary to support the same workload.

3. It is flexible, and adapts to different types of applications/scenarios: the definition is not tied to a specified resource, but to the bottleneck resource for the specific environment (CPU, RAM, etc.)
4. It can be easily used to compare the computational efficiency of different data centers, or of different portions of a data center
5. It can be used to measure the effectiveness of a strategy aimed at increasing the computational efficiency. For example, assume that in a data center with 200 servers, the number of equivalent servers is 100. Then the HUE is equal to 2. Lowering the HUE from 2 to 1.6 means one of two things:
    a. the same 200 servers *are supporting a higher load*, i.e., a larger number of equivalent servers, in this case 125 (N_{eq}=N/HUE=200/1.6=125)
    b. the same load as before (the same number of equivalent servers, 100) is supported by a *lower number of physical* servers, in this case 160 (N=N_{eq}*HUE=100*1.6=160)

2.2. HUE for a heterogeneous set of servers

The servers in a data center environment are not all homogeneous, but fortunately they are not all different from each other. Usually, a data center contains sets of homogeneous servers, belonging to different classes and acquired at different times. The same consideration applies when we consider multiple data centers belonging to the same environment.

The HUE for a data center environment is computed as the weighed average of the HUEs computed at the different sets of homogeneous servers, where the weighs are the total loads supported by the respective sets of servers

$$\mathbf{HUE} = \frac{\sum_i \mathbf{Li} * \mathbf{HUEi}}{\sum_i \mathbf{Li}}$$

In this expression, L_i = load on set of servers i; HUE_i = HUE index computed at set i

Now, we examine which is the specific data that we need to compute the HUE index . Such data is specified for the two phases:

1) first phase: computation of local HUE on sets of homogeneous servers;
2) second phase: computation of the global HUE on an entire data center environment.

The data needed for the first phase is:

- *Specification of the bottleneck resource (most commonly it is the RAM memory)*
- *The maximum desired utilization of the bottleneck resource on a single server (e.g., 80%)*
- *The average utilization of the bottleneck resource*

All this data is easily made available by any virtualization platform.

The data needed for the second phase is:

- *The bottleneck resource capacities of the different sets of homogeneous servers (used to compute the normalization coefficients)*

This data is also easily made available by any virtualization platform.

3. CPU Ready Time and Memory Ballooning

As mentioned in Section 1.5, the CPU Ready Time and the Memory Ballooning are important metrics that can be used to assess the quality of the service offered to the users. Indeed the value of these metrics helps to understand whether hardware resources, specifically CPU and RAM memory, are managed properly and with a proper degree of *overcommitment*.

CPU Ready Time and Ballooned Memory are symptoms of contention on CPU and RAM, respectively. These metrics represent, in IT literature, the universally recognized two most significant indicators of the fact that virtual machines are experiencing bad performance.

Eco4Cloud computes the ideal placement of VMs among physical hosts, in order to decrease both CPU Ready Time and Memory Ballooning, enabling higher performance and VMs density.

In the following of this section, we first illustrate the concept of overcommitment and the related concept of resource contention, then we separately describe the CPU Ready Time and the Memory Ballooning, which measure, respectively, the degree of CPU and memory contention. Finally, we present an innovative technique introduced by Eco4Cloud, referred to as “Smart Ballooning”, to exploit the available memory more efficiently, without affecting the degree of memory contention.

3.1. Overcommitment and resource contention

Though EcoMultiCloud will be platform-independent, the topic of overcommitment is here described, for simplicity, for the case of the most popular virtualization platform, i.e., VMware. VMware® ESX™ is a hypervisor designed to efficiently manage hardware resources including CPU, memory, storage and network among multiple concurrent virtual machines [12]. ESX uses high-level resource management policies to compute a target memory allocation for each virtual machine (VM), based on the current system load and parameter settings for the virtual machine (shares, reservation, and limit [13]).

The computed target allocation is used to guide the dynamic adjustment of the memory allocation for each virtual machine; in case host memory is overcommitted, the target allocations are achieved by invoking several lower-level mechanisms to reclaim memory from virtual machines.

VMware ESX allows running VMs with total configured resources that exceed the amount available on the physical machine: this is called *overcommitment*.

Overcommitment raises the consolidation ratio, increases operational efficiency and lowers the total cost of operating virtual machines. However, if out of control, overcommitment leads to *resource contention,* a typical situation where several VMs are competing over the same resources, waiting for the VMware scheduler to assign them. This is the main reason for performance issues in virtualized environment and, as such, it is the very first key performance indicator to be monitored in a virtual farm.

Contention on CPU and memory is measured, respectively, via *CPU Ready Time* and *Memory Ballooning*.

3.2. CPU Ready Time

CPU Ready Time is the period of time in which a VM waits in a “ready-to-run” state (meaning it has work to do) before being scheduled by the hypervisor on one or more physical CPUs. Therefore, CPU Ready Time is a metric showing how much time a virtual CPU is ready to be scheduled on a given physical host. In general terms, it is normal for VMs to have small values of CPU Ready Time. For VMs with multiple virtual CPUs (vCPUs), the amount of ready time will generally be higher than for VMs with fewer vCPUs, since they require more resources to schedule/co-schedule and each CPU accumulates the time separately.

Even in best designed environments there will be some CPU contention. The generally accepted industry best practice based on VMware’s guidelines is that CPU Ready Time values up to 5% (per vCPU) fall within acceptable parameters. Once the percentage is in between 5% and 10%, attention should be paid on adding more virtual machines and/or CPU cores to the virtual machines. This can be considered as the warning area. When the percentage is higher than 10%, we are in the dangerous area, and most probably bad performance will impact the virtual machines. The hosts will experience strong CPU contention in the environment, thus affecting the overall performance.

CPU contention is one of the hidden issues in a virtualized environment. Some tools that can be used to check any CPU contention in your environment are: ESXTOP from inside the service console of the host, RESXTOP from the vMA appliance, or third-party tools, like the Eco4Cloud solution. The best defense against CPU contention is knowledge and comprehension of scheduler interactions with multi-processor virtual machines.

There are two common scenarios in which high values of CPU Ready Time can occur. The first is host over-subscription, where too many vCPUs have been allocated per pCPU (physical CPU). Typically performance problems arise when a host is in the 2-2.5X over-subscription range. The second most common scenario that leads to a high CPU Ready Time is when a large VM, for example a 4-8 vCPUs, runs on a host having a lot of smaller VMs with 1-2 vCPUs for application servers. Depending on the number of physical processors and on the total number of vCPUs allocated on the host, a larger resource allocation for the VM results in longer waiting time, because the hypervisor has to preempt the necessary physical CPUs to schedule/co-schedule the workload. When this issue occurs, the software vendor increases vCPUs requirements, due to performance problems for the VM. Unfortunately, if CPU Ready Time is the root cause, increasing vCPUs number actually does not improve performance, on the contrary things get worse.

3.3. Memory Ballooning

The resource scheduler in the kernel manages the assignment of resources to the virtual machines, for example determining how much machine memory is to be allocated to a given virtual machine. If the scheduler determines that the amount of memory allocated to the virtual machine must be increased, it can reclaim memory, using various methods. During reclamation by ballooning, application pages may have been swapped out by the guest OS to disk. Memory ballooning, executing as an application in the VM, can detect a request for increasing balloon memory, for decreasing balloon memory, or for resetting the balloon application. For example, if a request to increase balloon memory is detected, the balloon application requests memory from the guest OS.

VMware ballooning is a memory reclamation technique used when an ESXi host is running low on memory. This allows the physical host system to retrieve unused memory from certain guest virtual machines and share it with others [14].

Ballooning makes the guest operating system aware of the low memory status of the host. In ESX, a balloon driver is loaded into the guest operating system as a pseudo-device driver. It has no external interface to the guest operating system and communicates with the hypervisor through a private channel. The balloon driver polls the hypervisor to obtain a target balloon size. If the hypervisor needs to reclaim virtual machine memory, it sets a proper target balloon size for the balloon driver, making it “inflate” by allocating guest physical pages within the virtual machine.

Ballooned memory is a symptom of RAM memory contention. If the free memory of a host drops towards the 4% threshold, the hypervisor starts to reclaim memory, using ballooning. VM memory ballooning can create performance degradation. Ballooning is a CPU intensive process, and can eventually lead to memory swapping, when a balloon driver inflates to the point where the VM no longer has enough memory to run its processes. This will slow down the VMs, depending upon the amount of memory to recoup and/or the quality of the storage IOPS delivered to it.

Memory Ballooning is the first technique the hypervisor uses to reclaim memory. Absence or very low levels of ballooning is a sign of excellent/good health for a virtual farm.

3.4. EcoMultiCloud’s Smart Ballooning

Virtualization comes with some key benefits. One of them is virtual machines isolation, which is very useful for security and risk management. A drawback of virtual machines isolation is that the guest operating system is not aware that it is running inside a virtual machine and is not aware of the states of other virtual machines on the same host. When the hypervisor runs multiple virtual machines and the total amount of the free host memory gets low, none of the virtual machines will release guest physical memory since the guest operating system cannot detect the host’s memory shortage.

Smart Ballooning is a solution designed by Eco4Cloud for virtual machines memory management for the VMware virtualization platform. Smart Ballooning enables the release of memory which is actually not used by VMs and make it available to ESX, which will possibly allocate it to other VMs in demand. The solution has been developed leveraging the native memory ballooning mechanism, and consists of a series of tasks that – under specific conditions – force the platform to activate the memory ballooning selectively, i.e., only on the virtual machines where this memory reclamation technique is most productive, and with no impact on performance.

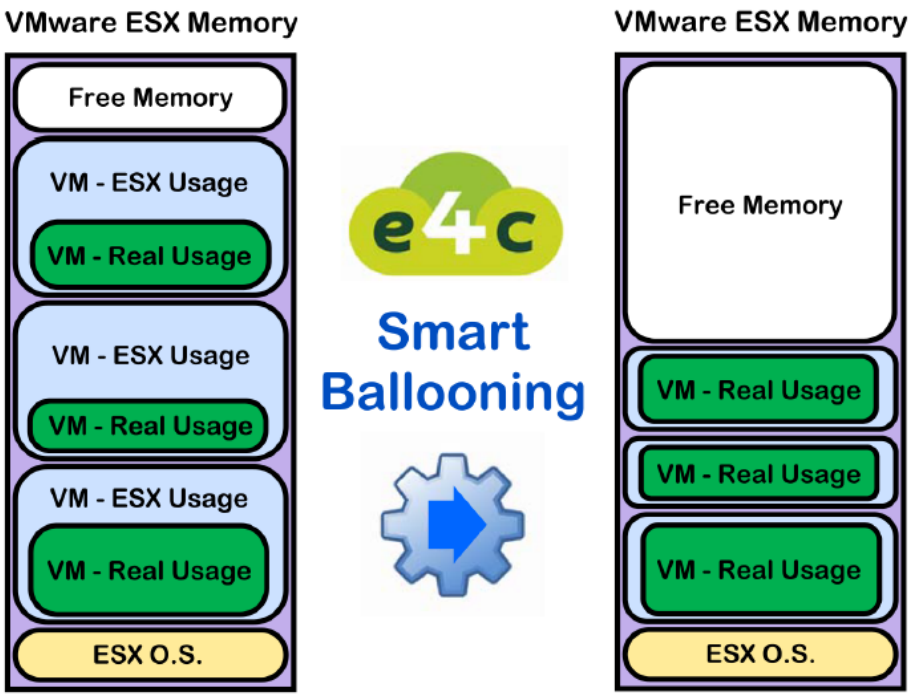

Fig. 4: Eco4Cloud Smart Ballooning

Eco4Cloud Smart Ballooning (Fig. 4) allows to automatically release memory from the virtual machines meeting particular conditions, releasing the unused RAM memory, and making it available for further virtual machines. Field results show that over 15% of RAM memory can be gained back, with full transparency for the virtual machines. Since memory is the most typical bottleneck resource in virtualized environment, this can lead to significant CapEx avoidance/deferral. Smart Ballooning prevents performance degradation due to memory ballooning hitting an entire ESX host, by proactively ballooning only overcommitted VMs.

Algorithms details

As described in Section 3.3, virtual-machine monitors of virtualization systems, such as for example VMware vSphere, possess a ballooning mechanism for freeing memory that is “unused” by the virtual machines and making it available for other VMs.

The Smart Ballooning solution can be used under the following conditions of applicability, which comprise both static and dynamic constraints:

The main static constraints or conditions are:

- the balloon driver must be installed;
- a limit for the consumed memory must be higher than the value referred to as “reservation” i.e., the minimum amount of memory for each virtual machine, indicated in the virtualization system;

- the option “Reserve all guest memory” (All Locked) must not be selected; in fact, this option maintains the memory reserve equal to the memory reserve of the virtual machines;
- the active memory (AM), i.e., the amount of guest memory that is currently being used by the guest operating system and by the applications, must be less than a given value, in particular 5%, of the memory configured on the virtual machine (CNF);
- the consumed memory (MC) minus the target size of the memory balloon must be greater than 40% of the configured memory CNF;
- There must be no other ballooning procedure in progress.

The main dynamic constraints are:

- the given virtual machine, i.e., the one selected, must be active, i.e., “On”;
- a current swapped memory (SW) must always be less than the initial swapped memory (if an amount of initial swapped memory was indeed present);
- a current compressed memory CM must always be less than the initial compressed memory (if an amount of initial compressed memory was indeed present);
- the active memory AM can increase only within given limits, control of these given limits is performed via linear regression of the point values;
- the rate of writing on the disk must not exceed a given value, preferably 50 Kbps.

The method according to the algorithm recovers memory automatically on all the virtual machines that present the static conditions CS listed above, until one of the dynamic conditions CD just described is violated.

The algorithm is composed of the following steps:

- read the memory parameters of a given virtual machine; these memory parameters comprise, for example, the active memory and the configured memory;
- verify the occurrence of the static conditions;
- in the case of positive outcome, read the memory configured limit (Lconf) set by the virtual-machine monitor; if Lconf exceeds the threshold value L (beyond which memory allocation is to be performed), the algorithms proceeds, otherwise it is ended;
- the threshold value L is shifted to a new value L1, equal to a value of consumed memory MC minus a given difference Δ, preferably 5% of MC;
- enter a wait state until a start of ballooning operations by the virtual machine monitor is detected;
- check whether required dynamic conditions CD are verified by the host system; in the case of negative outcome, the method is ended;

- determine when the ballooning operation started by the virtual-machine monitor of the virtualization system is completed;
- at the end of ballooning (or when the dynamic conditions CD cease to arise), verify whether the consumed memory MC has dropped below the value of consumed memory MC minus a percentage PD of the given difference Δ, this percentage PD preferably being 90%;
- in the case of positive outcome, set the threshold L again at the configured limit value Lconf;
- put the given virtual machine to which the ballooning procedure has been applied in a blacklist, where it resides for a pre-set time TB

Hence, the method according to the algorithm substantially represents a method for memory management in virtual machines that comprises managing execution of ballooning operations in a smart way.

Fig. 5 shows a flowchart of operations of the method according to the algorithm.

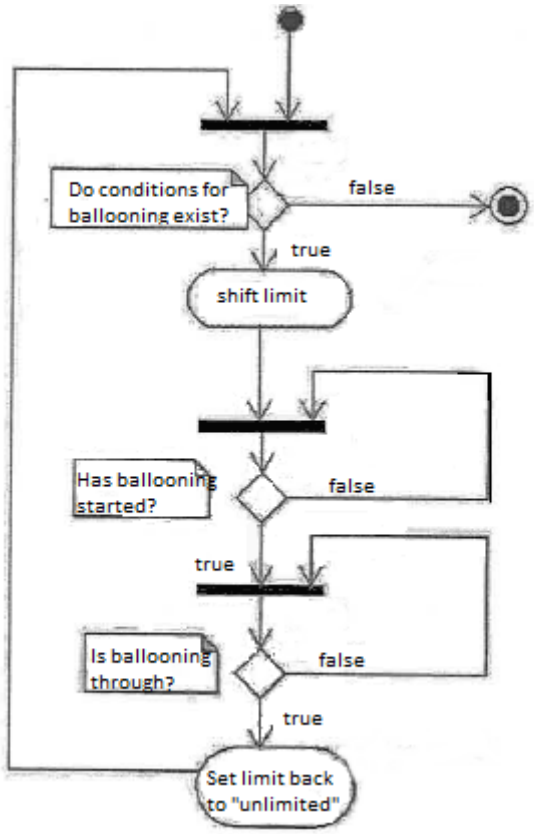

Fig. 5. Flowchart of operations

4. Conclusions

The objective of this report is to illustrate the efficiency and green metrics that can be used to assess the performances of distributed data centers.

In Section 1 we have describes a large set of business and technical metrics, most corresponding to international standards, which can be used to evaluate the performance of a management solution for a geo-distributed data centers. Metrics pertain to the energy efficiency, the use of green energy, the latency and communication performances, the cost of energy, the quality of service perceived by users and the ability to recovery from situations of interruption or limitation of electricity.

We have individuated the lack of standard metrics regarding the computation efficiency of data centers, i.e., the efficient use of the IT component, therefore in Section 2 we have introduced, motivated, and illustrated a new metric, the HUE (Host Usage Effectiveness) that is able to fill this significant gap. We will use HUE to assess the Eco4Cloud’s capacity to improve the computation effectiveness.

In Section 3 we focused on the very important issue regarding the efficient management of hardware resources, in particular CPU and memory, and have illustrated a new solution, referred to as “Smart Ballooning”, to improve the management of RAM memory in a data center.

References

[1] The Green Grid, "The Green Grid Data Center Power Efficiency Metrics: PUE and DCiE - See more at: http://www.thegreengrid.org/Global/Content/white-papers/The-Green-Grid-Data-Center-Power-Efficiency-Metrics-PUE-and-DCiE#sthash.s7Vss6y2.dpuf," 2007. [Online]. Available: http://www.thegreengrid.org/Global/Content/white-papers/The-Green-Grid-Data-Center-Power-Efficiency-Metrics-PUE-and-DCiE.

[2] S. Tuf, "Power Usage Effectiveness.," 17 Nov 2014. [Online]. Available: http://it.toolbox.com.

[3] OVH, "OVH, a green hosting provider," [Online]. Available: https://www.ovh.com/ca/en/about-us/green-it.xml.

[4] "Two-Phase Immersion Cooling," [Online]. Available: http://multimedia.3m.com/mws/media/1127920O/2-phase-immersion-coolinga-revolution-in-data-center-efficiency.pdf.

[5] "Energy Efficiency. Open Compute Project. Open Compute Project.," Mar 2016. [Online]. Available: http://www.opencompute.org/learn/energy-efficiency/.

[6] "Green IT Cube: Hocheffizientes Supercomputer-Domizil eingeweiht," 2016. [Online]. Available: http://www.heise.de/newsticker/meldung/Green-IT-Cube-Hocheffizientes-Supercomputer-Domizil-eingeweiht-3082605.html.

[7] The Green Grid, "Carbon Usage Effectiveness (CUE): A Green Grid Data Center Sustainability Metric," 2010. [Online]. Available: http://www.thegreengrid.org/en/Global/Content/white-papers/Carbon_Usage_Effectiveness_White_Paper.

[8] U.S. Green Building Council., "U.S. Green Building Council. Leadership in energy & environmental design,," [Online]. Available: http://www.usgbc.org/leed.

[9] The Green Grid., "Water usage effectiveness (WUE): A green grid," 2011. [Online]. Available: http://www.thegreengrid.org/en/Global/Content/white-papers/WUE.

[10] D. Kliazovich, J. E. Pecero, A. Tchernykh, P. Bouv, S. U. Khan and A. Y. Zomaya, "CA-DAG: Modeling Communication-Aware Applications for Scheduling in Cloud Computing," *Journal of Grid Computing,* 2015.

[11] P. Fan, J. Wang, Z. Zheng and M. Lyu, "Toward Optimal Deployment of Communication-Intensive Cloud Applications," in *IEEE International Conference on Cloud Computing*, 2011.

[12] C. A. Waldspurger, "Memory Resource Management in VMware ESX Server," in *Proceeding of the fifth Symposium on Operating System Design and Implementation*, Boston, 2002.

[13] Vmware, "vSphere Resource Management Guide," [Online]. Available: http://www.vmware.com/pdf/vsphere4/r40/vsp_40_upgrade_guide.pdf.

[14] VMware, "Understanding Memory Resource Management in VMware® ESX™ Server," [Online]. Available: http://www.vmware.com/files/pdf/perf-vsphere-memory_management.pdf.

[15] A. Forestiero, C. Mastroianni, M. Meo and G. Papuzzo, "Hierarchical approach for green workload management in distributed data centers," in *20th International European Conference on Parallel and Distributed Computing*, Porto, Portugal, 2014.

[16] A. Khosravi, S. Garg and R. Buyya, "Energy and carbon-efficient placement of virtual machines in distributed cloud data centers," in *Euro-Par 2013 Parallel Processing*, 2013.