Noname manuscript No.
(will be inserted by the editor)

Monitoring New Efficiency Metrics: beyond PUE

A. Ruiz-Falcó · J. M. Martínez · C. Redondo · C. Sáenz de Pipaón

Received: date / Accepted: date

Abstract The Energy efficiency is one of the great challenges of the IT industry. This challenge is even greater for supercomputing centers, since they are very intensive in energy consumption. The FCSC is a center who was born in 2008 and its infrastructures were designed to achieve high energy efficiency. In addition, energy efficiency is the priority line of research in FCSCL, and so embarked on the MONICA project. The objective of this project is to develop a comprehensive monitoring and control system for the entire datacenter, which is regarded as an industrial plant. For that it is necessary to break the technological barriers and integrate under a single system all elements of the datacenter, from the input power to the last application, since all are related. Not only is a classic monitoring system that acquires real-time information, the goal is to perform intelligent dynamic control, as Monica should be able to make decisions on working parameters to improve efficiency and implement them in real time. But advances in the development of MONICA have shown that focused decisions only to the PUE can be very detrimental, as improve the PUE can mean worsen the efficiency as will be shown. It is therefore necessary to define new metrics that show the overall efficiency of a datacenter as a utility, and new units of measure will be proposed.

Keywords Metric · PUE · Energy Efficiency · Datacenter

A. Ruiz-Falcó · J. M. Martínez · C. Redondo
FCSCL, Ed CRAI/TIC, Campus de Vegazana s/n, E-24071 Leon (Spain)
Tel.: +34 987293160
E-mail: antonio.ruizfalco@fcsc.es
E-mail: jose.martinez@fcsc.es
E-mail: carlos.redondo@fcsc.es

A. Ruiz-Falcó · C. Sáenz de Pipaón
Caton S. L., C/ Parque de las Ciencias 1, E-18006 Granada (Spain)
Tel.: +34 958184332
E-mail: arf@caton.es
E-mail: csp@caton.es

1 Introduction to the issue of energy efficiency

The new computing technologies have brought about deep changes in the datacenter in recent years. In the mid 80's it was normal for every organization to have just a single computer, with a huge cost. The relationship between the costs of purchase of the computer (typically millions of dollars) and the energy cost was very unevenly because these computers just consumed a few kilowatts and moreover, were cheap. But the increased processing power and decrease in size, led, in the nineties, to a shift to client-server architectures, and many host servers were replaced by "pizza box" format ones. The reality is that in just a couple of decades, many data centers have grown from a single computer to hundreds or even thousands of servers.

A modest server now has 12 processing cores, and costs just 2,000€, depending on the configuration and consumes between 400w and 700w at full load. This means that the cost of electricity consumption in three years exceeds the purchase cost of the server.

This process of change has been especially significant in supercomputing centers and facilities for HPC in general: a Cray J916 of 1996 consumed less than 4Kw (the cabin of processors), and other peripherals 4Kw each cabin. However, nowadays usual large clusters are composed of thousands of nodes and tens of thousands of processing cores. Process cores operate at

very high frequencies and thus, the associated power consumption is enormous. All this brings the problem of density: since demand causes the increasing of HPC cluster nodes, the power density per square meter is also growing, and it is usual to have configurations with more than 100 servers and over 1000 processing cores per rack. In the case of Calendula, the supercomputer of the FCSCL, the configuration of MPI cluster nodes is 128 8-core servers per rack, which means 1024 cores / rack. These racks reach consumption $\geq$ 40Kw/rack therefore represents a significant engineering problem. It is necessary not only to refrigerate it, but to do it efficiently.

2 Energy efficiency metrics

In order to improve energy efficiency it is necessary a measurement parameter. In February 2007, The Green Grid consortium defined the PUE (Power Usage Effectiveness) as follows:

$$PUE = TFP/ITP \tag{1}$$

Where TFP is Total Facility Power, meaning, the total energy consumed divided by the ITP, which is IT Equipment Power: the total energy consumed by IT equipment.

The other unit of measure defined by the Green Grid consortium is DCIE (Data Center Infrastructure Efficiency). This parameter, defined as the percentage of efficiency is the inverse of PUE:

$$DCIE = 1/PUE = ITP/TFP \times 100\% \tag{2}$$

The Green Grid defines IT Equipment Power as the load associated with all the IT equipment such as computers, storage, and network equipment, along with supplemental equipment such as KVM switches, monitors, and workstations/laptops used to monitor or otherwise control the datacenter. And the definition of Total Facility Power 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 miscellaneous component loads such as datacenter lighting.

The Green Grid proposes four categories of measuring PUE:

1. PUE category 0: demand (peak) in a period of twelve months. IT load measured at the output of the UPS.
2. PUE category 1: total consumption (kWh) for twelve months. IT load measured at the output of the UPS.
3. PUE category 2: total consumption (kWh) for twelve months. IT load measured at the PDU's.
4. PUE category 3: measured load in all the samples.

The units of measure proposed by the Green Grid provide an easy way to identify important aspects of the datacenter:

- Opportunities to improve a datacenters operational efficiency.
- How a datacenter compares with competitive datacenters.
- If the datacenter operators are improving the designs and processes over time.
- Opportunities to repurpose energy for additional IT equipment.

3 Measurement Systems

The first problem to solve is to provide a measuring and monitoring system. This requires addressing two problems:

1. Organizational problem: in many organizations it is common that IT equipment and infrastructure that support them (datacenter, chillers, power, ups, security, etc...) have different hierarchical dependencies.
2. Technical problem: each computer (hardware, networking, UPS, cooling, generator, etc) is provided with its own system of measurement (if available), but it is unusual to have a system of monitoring, control and measurement unifying the entire system, and even more unusual that this system controls IT equipment and infrastructure simultaneously.

The difficulty in establishing measurement systems is large. Normally there is a measuring instrument for the total energy consumed. But it is not common to have measurement systems to obtain accurate consumption of IT load. In some cases it is possible to provide partial data, and manually (for example, there are installations in which it is possible to see the load at the output of the UPS, but it is necessary to perform a manual process of data collection).

This means that, in many organizations, it becomes common to perform a static measure of the PUE: taking manually a reading of total energy consumed and the load of the UPS output (or simply an estimate of IT load power ratings using nominal consumptions, which has a huge error rate, as discussed below).

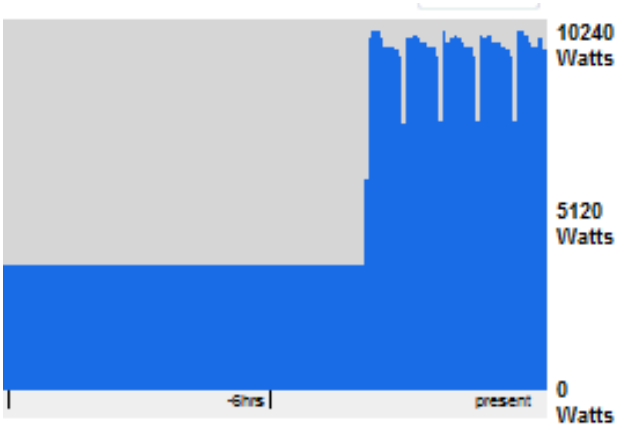

Fig. 1 Power Consumption with and without CPU load

One of the most common mistakes is to consider that the IT load is flat over time. The difference in consumption with load and without it is about three times more in a modern server. Figure 1 show the consumption of a C7000 chassis with 32 BL2x220c servers (E5450 Xeon processors with two each). On the left side of the figure without load (server on and operating system loaded), and on the right side running a Linpack test with np = 256, N = 245,000, NB = 160, P = 16, Q = 16 parameters. As shown, the consumption goes from 4kw to over 10Kw.

4 The Monica Project

In order to achieve efficiency in a facility like the FCSCL it is essential to have a system of monitoring and dynamic control of the facility. The aim of the MONICA project is to develop a monitoring technology that integrates the whole CPD, all-in-one, resolving the fact that existing technologies are just specific and not all integrated. This means that technology provides monitoring of the various elements of a CPD but these tools provide a partial view of its operation. They have tools to monitor the status of power supply systems, cooling systems, fire systems, network communications and computing infrastructure and services, but each of these tools provides an overview part of CPD. **The MONICA project has as one of its objectives to provide a vision of a CPD as a complete, integrated manufacturing plant, providing comprehensive information on all the parameters of its operation, and possible to establish metrics that include information collected from subsystems of different nature** (outside temperature, chiller consumption, consumption per PDU, state of charge of processors, etc.).

But MONICA has not only been designed as a monitoring and data acquisition. MONICA must process the information gathered, make decisions about the operating system installation and deploy them in real-time. MONICA has to make active decisions on the best possible configuration to improve efficiency (eg, shutdown inactive nodes, moving loads between servers, configuration changes in the cooling system according to the Calendula operating, outside weather, etc.,).

To develop MONICA, it has been necessary to provide the FCSCL with the installation of the hardware elements necessary (especially data acquisition interfaces and electrical infrastructure), develop data acquisition interfaces for heterogeneous devices (SNMP, Modbus, etc.) and interpolate data not directly existing in the equipment.

Monica has two major differences compared to traditional monitoring systems:

1. First of all **MONICA conceived the datacenter as an "industrial plant" where everything is integrated**. It is impossible to do the proper management of a server (for example, risk management as proposed by ISO 27001) if the monitoring system that takes no account of the elements on which it depends including the infrastructure. MONICA so not only monitors the IT hardware and software, but also controls all auxiliary infrastructures involved.
2. The data obtained from monitoring are not only for the management of events and alarms. **MONICA should be able to make decisions and implement them automatically in real time according to predefined business rules.**

One of these rules is the energy efficiency. MONICA must control all the parameters, calculating the PUE and decide the best configuration of the cooling system at a given time.

4.1 PUE is not constant

A common mistake is to think that the PUE is flat. Actually, due to increasing awareness, more organizations have been interested in knowing the PUE. The PUE, as defined, requires data collection for twelve months, and their optimization requires to work proactively in times when it is too high. That is, acting in the instant PUE to avoid the peaks that raise it during the period of twelve months integration.

The first thing MONICA has to do is to calculate the instant PUE. To do this, MONICA captures all the necessary information (the total energy consumed by the FCSCL (cooling consumption, consumption of IT load, etc.) and displays this data. Figures 2 and 3 shows

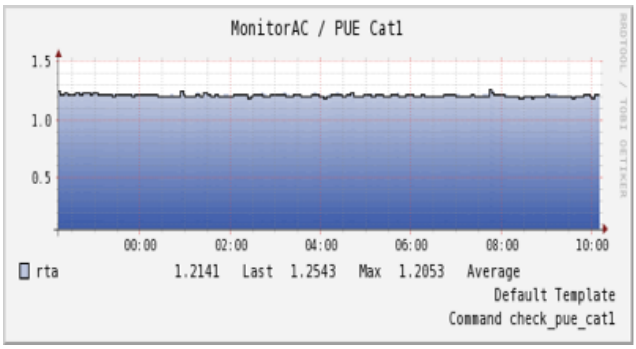

Fig. 2 PUE Evolution in a 12h period

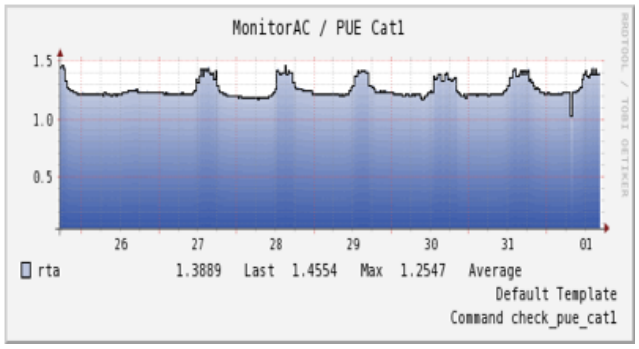

Fig. 3 PUE evolution un a weekly period

the evolution of the PUE in FCSCL for a period of 12 hours and in a week.

Although if it may be seem that the PUE remains unchanged, it is clear that it does in terms of operating parameters:

- State of system load: we have seen how the differences in consumption are between unloaded servers and fully loaded (100%) servers. That is, the denominator of the equation (ITP) can vary greatly depending on time.
- Weather: In an installation where there is free cooling is obvious that consumption will vary depending if it is working with free cooling or normal cooling.
- System Configuration Parameters: the configuration of the coolers in the computer room (in the case of FCSCL, 16 units APC InRow RC), water temperature setpoint, cold air temperature setpoint of entry into the racks, etc.

The Figure 3 shows that most of the time, the instant PUE is $\simeq$ 1.2 (the graph indicates that the 1 week average PUE is 1.2547), while in a reduced percentage of the time the PUE goes up, but peaks $\leq$ 1.5. These peaks are due to low system load: less CPU load implies less IT load, so the denominator of the equation decreases. In other words, the efficiency is better with higher densities. The data acquired at the FCSCL shows very high efficiency when the load is

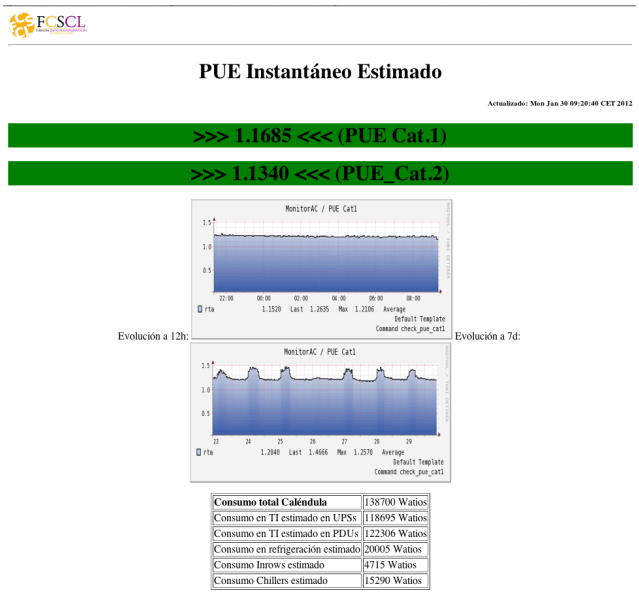

Fig. 4 MONICA main's web page

$\geq 30Kw/rack$.The FCSCL has a modern and energy efficient installation. But thanks to MONICA, the FCSCL has reduced the PUE implementing active policies of control.

The first requisite is to show the acquired information. Figure 4 shows the web page with instantaneous PUE (measured in the UPS -PUE cat 1- and in the PDU's, -PUE cat 2-). If PUE is $\leq$ 1.3 is shown in green, if 1.3 $\leq$ PUE $\leq$ 1.5 the band is Orange and red if PUE $\geq$ 1.5.

As noted, data are measured at the instant PUE UPS (cat 1) and the PDU's (cat 2), and the fundamental parameters of the equation: total load, IT loads, and so on.

The next information that MONICA shows in real time is the installation schema with its parameters in the given instant (Figure 5). As shown in this schema, are contained not only the parameters of electric load (TFP 138.7Kw, ITP 118Kw, etc.), but also covers environmental and functioning parameters: impelling water temperatures (9.6°C) and return (13.8°C) from the cooling water circuit, outlet temperature in the location of the chiller (-1.0°C), working regime of the chiller (Free cooling to 42% power, etc.). MONICA control does not end here. The datacenter of FCSCL consists of a closed cube with hot aisle with two rows of racks and sandwiched between them APC InRow RC coolers (eight racks and seven InRows per row. There are eight racks more and two InRows in the facilities room where are the UPSs). The next web page of real time information is the outline of racks that can be seen in Figure 6. This diagram shows important operating data

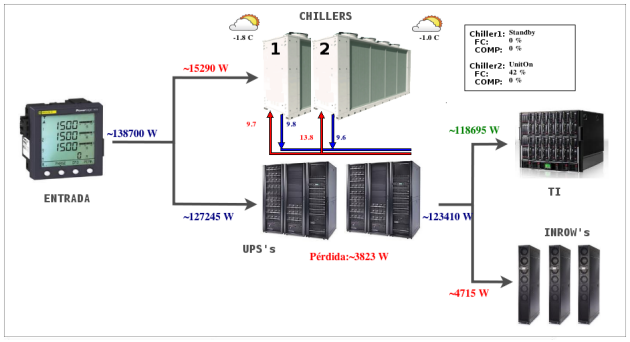

Fig. 5 MONICA main's web page

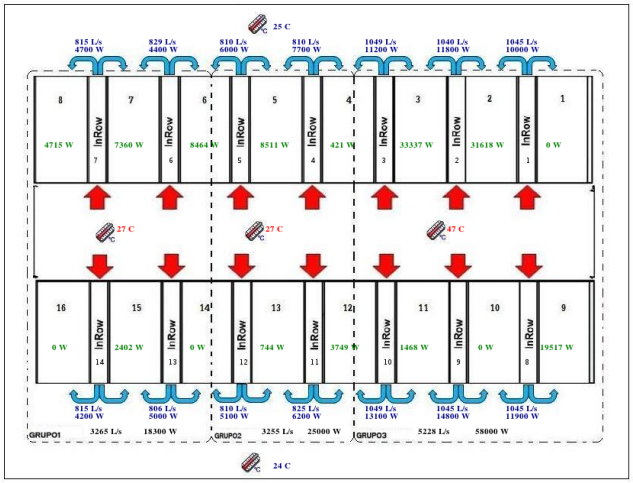

Fig. 6 MONICA installation's schema and parameters

such as power consumption per rack (note that at the time of image capture consumption of racks 2, 3 and 9 it was of 31.6, 33.3 and 19.5Kw respectively), the air flow from the InRows, cooling capacity provided by the InRow groups, temperatures at different points of the cold and hot aisles, etc.

MONICA captures more than 3,000 parameters of all kind from Calendula every five minutes (state of the hardware, applications, etc.). From these parameters, it uses 150 for optimizing the PUE (power load at PDUs and UPS, main load, power load and working mode of the chillers, outlet temperature, water temperature, InRow, parameters, etc. With this data MONICA calculates the optimum operating model and how to set the setpoints of the key elements. The problem that exists today is that there are things that can be implemented automatically (for example, turn off or on servers) but others are not possible because there is no suitable transducer, so that in this case MONICA is unable to deploy it automatically (the most important example is the setpoint temperature water drive). It is necessary to implement these transducers to achieve full automation, which will result in a better optimization

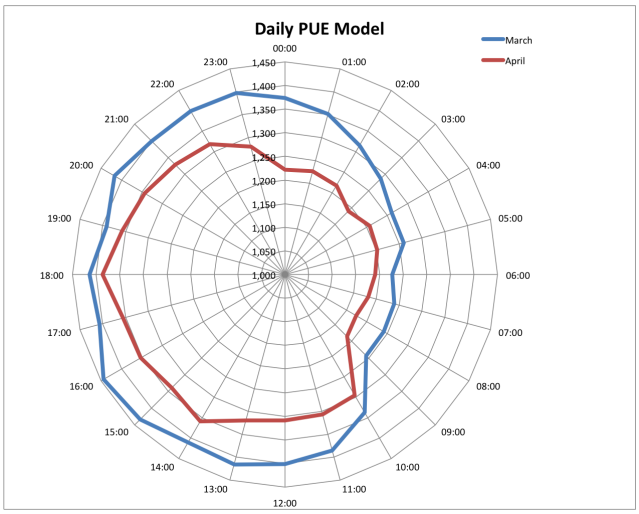

Fig. 7 Daily PUE model at the FCSCL

(feedback will be much faster by not needing human intervention in any case).

5 Results at the FCSCL

The FCSCL data center is the MONICA testing laboratory. Due to space constraints, the FCSCL has a very small datacenter and has therefore been necessary to achieve high density. The system is composed by:

- MPI Thin Cluster Nodes: 400 nodes between production and testing.
- Fat Cluster Nodes: 8 nodes (256GB RAM) + 8 nodes (128Gb RAM).
- Virtualization Farm and auxiliary systems: 20 servers.
- Storage: SAN & Tape Robot.
- Networking: 10GbE and GbE switches, InfiniBand switches, routers, etc..

Due to the high system density (128 servers/rack, in the case of the thin nodes cluster, the cooling system must be prepared for a system with this high density. As shown in Figure 6 the installation has intercooler water/air APC InRows every two racks and the chillers are equipped with free cooling.

The FCSCL is not a classic supercomputing center in which the system is permanently overloaded. The nature of the projects that run on it and the SLA's associated cause that the load levels have a significant variance in time.

The first use has been to understand the operation of the facility and its cycles. For example, it has made possible to characterize the role of the PUE depending on the time of the day. Figure 7 shows the graphs of the daily behavior pattern of the PUE in March and April:

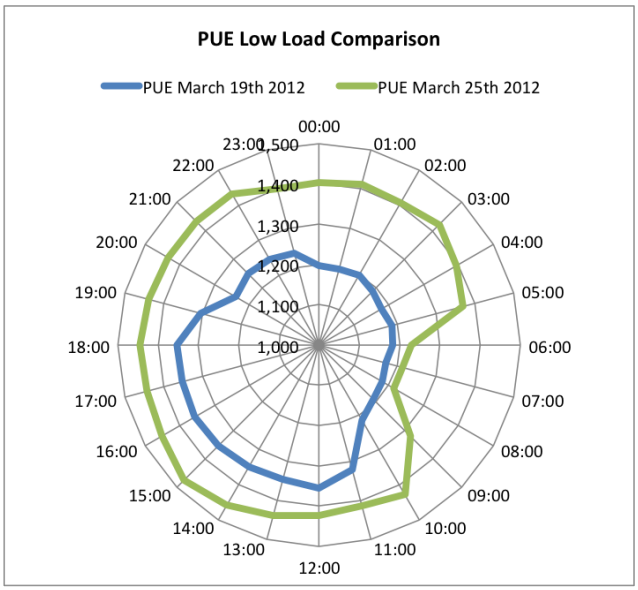

Fig. 8 PUE Low load Comparison

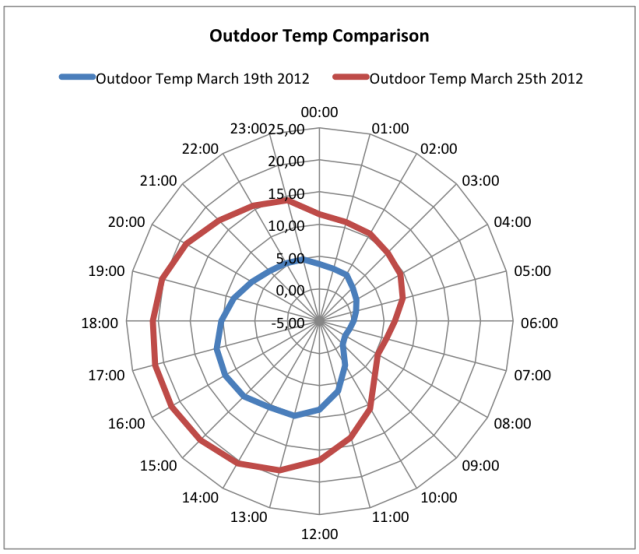

Fig. 9 OutDoor Temp Comparison March 19th and March 25th

6 Energy Efficiency versus IT Efficiency

The installation of the FCSCL is very modern and has been designed with energy efficiency criteria. Since the beginning, the FCSCL has had a low PUE, and Monica has served as a better understanding of the installation and its optimization. But it has also served to locate the problem of efficiency in its true dimension. To understand this, have been selected six days according to the following criteria: two days with very low computing load, two days with medium computing load and two days with high computing load. And, for each couple of days, one with low outdoor temp ($\leq 8°C$) and the other with high outdoor temp ($\geq 12°C$).

Let's see the first comparison (Figures 8 and 9). On March 19th, the medium outdoor temperature was 6.84°C, whereas on March 25th the medium temperature was 12.86°C. In both cases the load of Calndula was very low (below 1.000 CPU hours of a maximum of 62.000 in the part of the cluster used for the testing). That means, in order to perform the data collection necessary to show the idea presented in this article, days without almost any charge have been chosen. So, there have been elected the days when the workflow and SLA's allowed to delay executions, and let Calendula just with a minimum load (auxiliary systems, core processes, etc...). This minimum load (true bias of the system) is nearly constant in time. The rest of the systems and nodes of the clusters were on, operating system loaded and idle state.

As expected, the PUE is better on March 19th due to the lower outdoor temp. The data shows that on March 19th there where a lot of hours of free cooling.

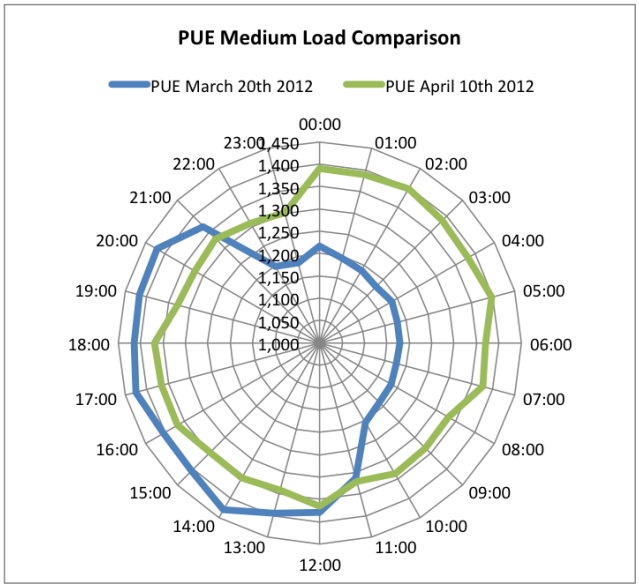

Fig. 10 PUE Medium Load Comparison

The next step (Figure 10) is to compare two days with medium load (less than 20.000 CPU hours in a day). And finally, the Figure 11 shows the comparison between two days with a workload greather than 30.000 CPU hours.

The data shows that on March 19th there were twelve hours with temperatures below six degrees, so there were many operating hours of free cooling. From these data anybody could think that March 19th has a better PUE than April 30th (1.250 vs. 1.256). With this information, anyone can deduce that the efficiency was better on March 19th than April 30th. The answer is not, because there is another difference between these two days: Day 19th is a day without charge. On the secon

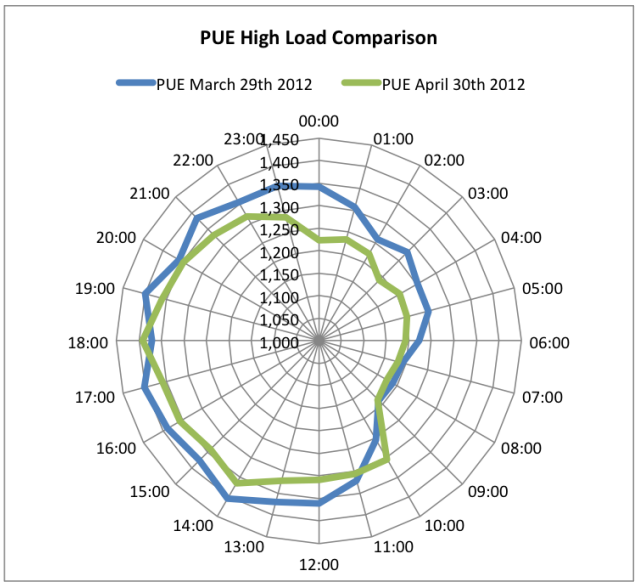

Fig. 11 PUE Medium Load Comparison

Table 1 CPU Load, Outdoor Temperature and PUE

Date	Outdoor Temp	CPU Hours	PUE
March 19	6.84	907	1.250
March 20	5.47	19.959	1.256
March 25	12.26	576	1.350
March 29	12.87	33.324	1.328
April 10	12.25	17.724	1.335
April 30	7.6	31.092	1.256

hand, the MPI cluster partition on which tests have been executed used over 33,000 hours of CPU. That is, the PUE was a little bit better (0.006) on the 19th than on day 29th, but the system was more efficient on day 29th since it executed a much higher workload (50 times more) than on day 19th.

Moreover, on day 19th, it was also energy inefficient, since the data shows that for higher density, higher energy efficiency. Consider the following table in which data is displayed, total kW, kW total consumption, IT consumption Kw, No IT Kw, PUE, temperature, hours of CPU and PUE:

The data acquiered at the FCSCL shows that PUE correlates with temp, but also with CPU load. The greater CPU load, then better PUE.it is easy to understand:if the weather allows operation with free cooling, cooling consumption is constant and independent of IT load. Cooling consumption is the sum of the consumption of the fans of free-cooling battery, the water circulation pumps and inrows consumption of the room. In this case, if it was appropriate to turn off unused servers (obviously that is a measure of efficiency because it saves Kw) the IT load would decrease. But the No IT load would remain the same, so the PUE would get worse.

Energy efficiency is very important, and getting a low PUE is a good starting point. But do not forget that it is a measure of data center energy balance, and gives a precise idea of the quality of the engineering design of it. We can neither forget that computers were designed to execute instructions. It is pointless for the data center to have a very low PUE if the computers that host do not run programs. The simile is very easy: the consumption figures of a car give an idea of its performance, but not of its efficiency because, what is more efficient, a car that consumes 4.5l of fuel to transport one person at a distance of 100km in an hour, or one that consumes 6.5 l carrying four people at a distance of 100km in 50 minutes?

7 New efficiency metrics: beyond the PUE

It becomes necessary to use new metrics to give a general idea of the efficiency of a datacenter. We propose to keep the PUE and DCIE as measures of energy balance of the installation.

And we propose to define analogue units to the execution of instructions. The **first unit is the CUE (CPU Usage Efficiency)**, wich is defined as follows:

$$CUE = \frac{TotalGHzAvailableinaPeriod}{TotalGHzUsedinthePeriod} \quad (3)$$

The next unit is **the DCPE (Data Center Processor Efficiency)**, wich is:

$$DCPE = \frac{TotalGHzUsedinaPeriod}{TotalGHzAvailableinthePeriod} \times 100\% \quad (4)$$

And the measure that relates them is:

$$DCUE = DCIE \times DCPE \quad (5)$$

$$GUE = \frac{1}{DCUE} \quad (6)$$

Obviously, the measurement period must be the same in both cases. The Green Grid proposes a year to the correct measure of the PUE, and so must be done in the case of the proposed measures. But, as mentioned above, systems such as Monica can do an ongoing analysis of data in search of local maxima and minima to improve efficiency. Therefore, within the design of MONICA is the decision-making (and implementing them in real time) to improve the efficiency according to predefined rules.

The great benefit of establishing measurement systems as proposed is that they can establish good practice ranges depending on the type of work performed in the datacenter. For example, in supercomputer centers

with classic scheme of operation and high loading rate (CPU $\geq$ 75%) and a PUE of 1.5 would have a GUE = 2 and a DCUE = 50%. That means, a level of good practice for HPC centers can be GUE $\leq$ 2.5.

However, if we were talking about a critical service data center in virtualized systems, a GUE $\leq$ 2 should not be a good practice, because in addition of a high energy efficiency rates it would also imply CPU utilization that would seriously risk redundancy.

8 Conclusions

- Energy efficiency is not just a fad, is the great challenge of the IT industry today.
- It is essential to standardize metrics to not only measure, but to also compare between different systems.
- It is essential to establish in the data center, monitoring and measuring systems that contemplate it as an industrial plant acquiring all the necessary information, both IT and infrastructure components.
- This will allow the adoption of dynamic control systems that make decisions based on the operating parameters and implement them in real time.
- These decisions may be taken by energy efficiency criteria or other rules predefined by the user: risk management, prevention of failures, etc
- Sometimes, doing certain actions that improve energy efficiency (saving Kw consumption), exacerbate the PUE.
- The PUE is a measure of the energy balance of a datacenter, but not of the efficiency of its usefulness.
- It is necessary to establish performance metrics for the IT system similar to those established for energy efficiency.
- Establishing metrics that address global energy issues and implementing programs to measure the efficiency of a facility allow defining codes of practice depending on the installation target.

Acknowledgements – The MONICA project which has developed this work has been supported by the Avanza TIC Verdes call of the Spanish Ministry of Industry, Tourism and Trade, with the code TSI-080500-2011-79
– The authors thank Rosa M. Castellanos her help in correcting grammatically this article.

References

1. C. Redondo Gil, A. Ruiz-Falcó and J. M. Martínez, Optimizacion of energy consumption in HPC centers: Energy Efficiency Project of Castile and Leon Supercomputing Center FCSCL, International Conference on Renewable energies and Power Quality, (2012)
2. D. Gaffney, The Business Value of PUE and Beyond, The Green Grid Forum 2012, (2012)
3. J. Moore et al., Making Scheduling Cool: Temperature-Aware Workload Placement in Data Centers, Usenix 05, (2005)
4. T. Harvey et al., Updated air-side free cooling maps: the impact of ASHRAE 2011 allowable ranges, The Green Grid, EEUU (2012)
5. The Green Grid, Breaking New Ground on Data Center Efficiency, The Green Grid, EEUU (2012)
6. Cray Research, Preparing for a Cray J916 System Installation, Cray Research, EEUU (1997)