European Association for the
Development of Renewable Energies, Environment
and Power Quality (EA4EPQ)

International Conference on Renewable Energies and Power Quality
(ICREPQ'12)
Santiago de Compostela (Spain), 28th to 30th March, 2012

Optimization of energy consumption in HPC centers: Energy Efficiency Project of Castile and Leon Supercomputing Center - *FCSCL*

C. Redondo Gil[1], A. Ruiz-Falcó[1, 2] and J. M. Martínez[1]

[1] Castile and León Technological Centre for Supercomputing (FCSC) · http://www.fcsc.es

Edificio CRAI-TIC · University of Leon
Campus of Vegazana s/n E24071 León (Spain)
e-mail: **carlos.redondo@fcsc.es**, **antonio.ruizfalco@fcsc.es**, **jose.martinez@fcsc.es**

[2] Catón, S. L.
C/ Parque de las Ciencias nº 1 Of. 2-A
18006 Granada (Spain)

Abstract.

Supercomputers are very intensive in energy consumption. Large clusters are composed of hundreds or thousands of processors working in parallel, as is easily seen, the processors are the system component that consumes the most. An indicative figure is that a standard server tripled their consumption when their processors work at 100% load.

In addition, we must say that the power consumption of computers is converted into heat at a high rate, which causes that the power consumption of the Supercomputing Center is a bigger problem than the general-purpose data centers.

These reasons force to design new systems (data centers and computer) with the most efficient energy balance, which means to have low PUE (Power Usage Efectiveness). This article shows the PUE monitoring and control system implanted in Calendula, the FCSCL Supercomputer.

Keywords

Supercomputer, Monitoring, Datacenter, PUE, Energy Efficiency

1. Introduction

More and more there is a greater concern about the problems of energy efficiency and IT industry is no stranger to it. In addition, it is estimated that the 2% of U.S. electricity consumption in 2007 was in datacenters. This implies that the IT sector is the second most polluting industry in the world only behind aviation, and that at current rates soon will be the first.

Advances in technology have brought up deep changes in data centers in the recent years. In the mid 80's it was normal for every organization to have a single computer. This represented a very high purchase cost, so all the costs associated with its operation (engineers, facilities, electricity, etc.) accounted just for a small fraction over the cost of acquisition or rental of the computer. It is evident that the performance of these systems were low and very energy inefficient.

But because of these limited benefits the power consumption was not excessive.

Since the early 90's, with the arrival of client/server technologies, many of these hosts were replaced by servers in a "pizza box" format. The fact is that in just a couple of decades, many datacenters have gone from one computer to hundreds or even thousands of servers. Nowadays, a modern low end server has 12 processing cores, costs around 2,000€ and, depending on the configuration, consumes between 400 and 700W at full load. This means that the cost of power consumption for three years exceeds the purchase cost of the machine itself.

2. The P.U.E.

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:

$$PUE = \frac{TFP}{ITP}$$

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:

$$DCIE = \frac{1}{PUE} = \frac{ITP}{TFP} \times 100\%$$

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:

- PUE category 0: demand (peak) in a period of twelve months. IT load measured at the output of the UPS.
- PUE category 1: total consumption (kWh) for twelve months. IT load measured at the output of the UPS.
- PUE category 2: total consumption (kWh) for twelve months. IT load measured at the PDU's.
- 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 datacenter's 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. Power Consumption in HPC

The following problems of efficiency are especially important in facilities for High Performance Computing (HPC). Today the usual way to solve large computational scientific and technical problems is through large parallel computing clusters. These consist of a number of conventional servers working in parallel.

There are two factors that differentiate an HPC datacenter to the conventional one:

- The number of servers. A low power HPC cluster has today hundreds of servers. Being in the top 100 in the Top500 list (the list of the 500 fastest supercomputers in the world) in November 2011 called for a cluster of more than 8000 processing cores, about 700 servers with two 6core processors. (Calendula, the FCSC supercomputer, was ranked in 2009 as number 53 in the Top500 list. Calendula has more than 300 servers and 3000 cores).
- But the main difference is the level of workload. In a conventional scenario it is common to see servers with a CPU load of about 5-10%. Obviously, in a system dedicated to calculation it is normal, if the programs are well done and the facility properly used, that processes are continuously at 100% load.

This last factor is especially important because the consumption of a processor grows cubically with the processor frequency (f^3). The parallel computing cluster of Calendula consists of Hewlett Packard BLx220c servers on a C7000 blade chassis. These servers are "double blade", which means that each blade has two dual-processor servers each. In its maximum configuration, the C7000 chassis supports 32 servers. In FCSCL configuration, each chassis has 32 BLx220C servers with two Intel Xeon E5450 each and 16GB of RAM. That is, the overall chassis has 32 servers, 256 processor cores and 512GB of RAM.

The following chart shows the energy consumption of the 32 servers on, and the operating system loaded, but without running any applications:

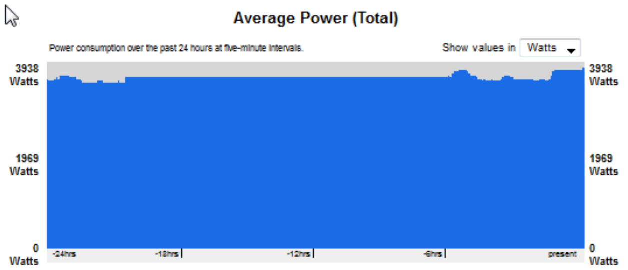


Fig. 1. C7000 Chassis Power consumption without load

The data is extracted directly from the Onboard Administrator (OA) of the C7000 chassis. As shown, the chassis consumption in this case is about 4Kw. This consumption is the sum of the energy required by the electronics of administration (OA and iLO<), switches (4 gigabit switches + 4Infiniband switches), the six fans and 32 servers without any load.

The following chart shows now that consumption increased significantly when these systems are 100% load:

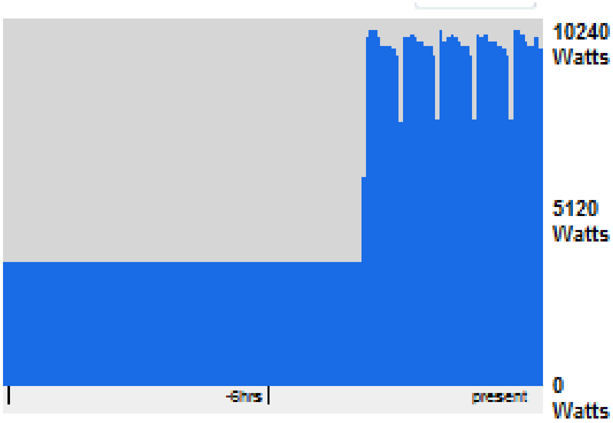


Fig. 2. C7000 Chassis Power consumption with load

That is, consumption is almost multiplied by 2.5 or more when executing a compute-intensive program. In this case we have released several jobs that each one involves two runs of LINPACK test (solving a linear equations system) simultaneously. Each test with parameters N=175,000, NB=160, P=8 and Q=16. The energy consumed by the calculation programs can vary according to specific parameters of each run. In fact, the graph shows the consumption drop due to the completion of a job and starts the next. In any case, the fully loaded chassis consumption will vary between 10Kw and 11.5Kw.

4. High Density Datacenters

The immediate consequence of consumption shown in the previous paragraph is that the HPC data centers have very high densities.

In the case of the FCSCL, the MPI parallel computing cluster uses the HP servers mentioned above. That is, C7000 chassis with 32 servers each. There have installed four C7000 chassis per rack, so the result is a very high density:

- 4 Chassis per rack.
- 128 servers per rack.
- 1024 CPU cores per rack.
- 2TeraBytes RAM per rack

With the data shown above, this represents a minimum consumption of 14kW per rack (with 128 servers on and no load). This data, at any other datacenter would be considered high density. But doing the multiplication is easy to see that each rack, fully loaded, **is a consumption exceeding 40kW**.

This high density is a double problem:

1. Traditional datacenter technologies are not ready for such densities.
2. Not only must make the installation work, but does so efficiently.

5. New DataCenter cooling technology

The high consumption of HPC clusters is converted into heat, and it is necessary to dissipate it. As traditional datacenter technology was not enough, it was necessary to design new datacenter technology, from the traditional concept of cool room ("room cooling systems") to "dynamic smart cooling".

Specifically, the most significant are:

- **Elimination of hot spots**. With the traditional concept of cooling the room, the higher consumption units condition the temperature and humidity. That is, in a conventional datacenter that has a single point of consumption as representing the frame shown in this article, need to cool the entire room to its needs.
- **Hot and cold aisles**: this technique is to prevent outgoing warm air from a machine to mix with cold air to cool the others, lowering the air conditioning system performance.
- In **“InRow” cooling and “Inrack”** cooling systems, the heat removal is as close as possible to where it is produced (as opposed to traditional systems). This makes it possible to adapt the cooling needs of each specific rack, without having to cool the entire room. Closed systems cooling rack cabinets consist of fully enclosed, which makes heat exchange inside.
- **Free Cooling**: When the ambient temperature is lower than in the datacenter, it is possible to use systems that take advantage of the cold environment, by introducing air directly from outside (Direct Free Cooling) and using the environment to cool the cooling water circuit heat In rows of the enclosures (Indirect Free Cooling). In either cases avoids using conventional chillers.
- **Datacenter Modular design** is becoming more common to establish separate compartments in the data center so that an operation does not affect the other. This also helps avoid the inefficiencies due to the over dimension. Within this data center should be considered standard containers 20 and 40 feet.

6. MONICA monitoring and control system

Like any other datacenter, the FCSCL needs to monitor the installation. This was the reason for the MONICA project: **MON**itorización **I**ntegral de **CA**léndula, Calendula Integrated Monitoring System. The project, which is strategic for the FCSCL, is a research project funded by the Plan Avanza2 call from the Spanish Ministry of Industry, Tourism and Trade (MYTIC). The project was developed in collaboration between the FCSCL and Catón S. L.

Monica has two major differences compared to traditional monitoring systems:

1. The traditional view has been that datacenter IT equipment depends on the IT department (and it is monitored by that department), while the ancillary infrastructure (electricity, UPS,

generators, cooling, security, etc.) depends on the infrastructure and maintenance department. However, 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.

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 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.

To do this, 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. The following chart shows the evolution of the PUE in FCSCL for a period of 12 hours:

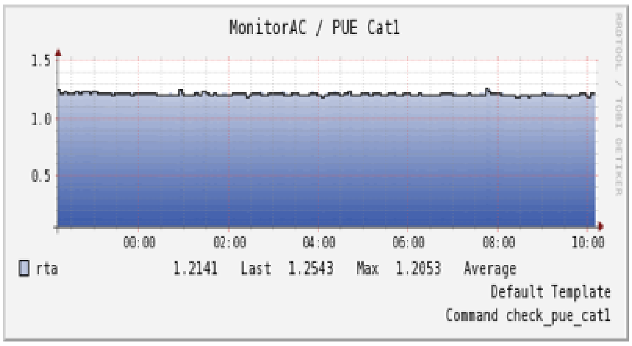


Fig. 3. PUE Evolution in a 12h period

Although if it may 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 graph below shows an evolution of the PUE in a week:

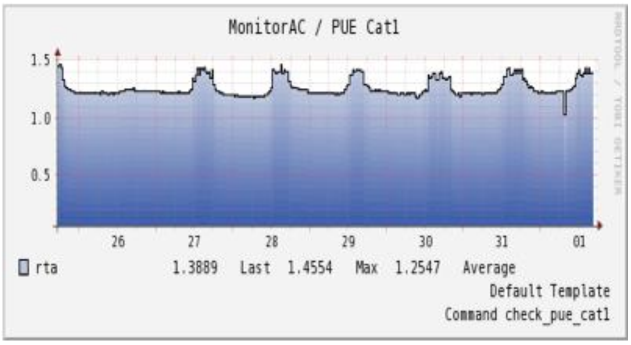


Fig. 4. PUE Evolution in a week

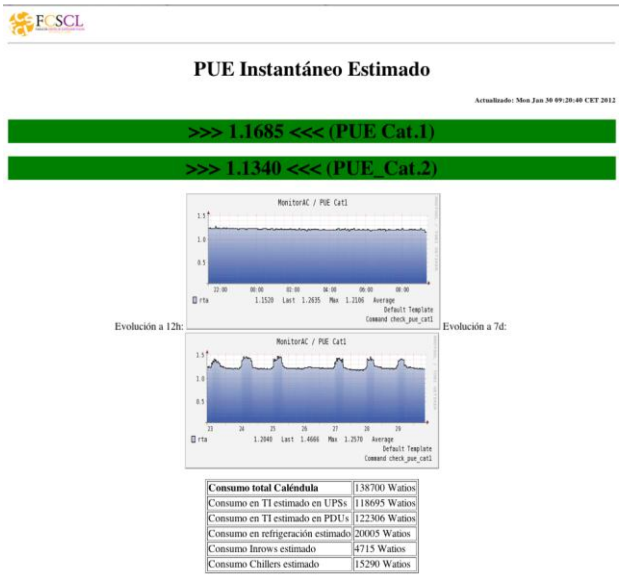
This graph shows that most of the time, the instant PUE is ≈ 1.2 (the graph indicates that de 1 week average PUE is 1.2547), while in a reduced percentage of the time the PUE goes up, but peaks not over ≤ 1.5. These peaks are due to low system load: less CPU load implies less IT load, so the denominator 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 >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 best way to understand this is an example: the FCSCL chillers are two Emerson SuperChillers SBH030 units with a cooling capacity of 330kW each. Each chiller has four compressors, each one starts under the thermal demand, and the individual consumption per compressor is ≈25Kw. In some cases if it exceeds the threshold narrowly that goes from free cooling to compressor or start of a new compressor, this can be avoided by changing the cooling system parameters.

The first requisite is to obtain information. With this purpose, FCSCL has installed the necessary transducers for the collection of data, which are displayed on a website. The first figure shown is the web page with instantaneous PUE (measured in the UPS –PUE cat 1- and in the PDU's, -PUE cat 2-). The PUE is shown in green if ≤1.3, if 1.3 <PUE≤1.5 the band is Orange and red if PUE> 1.5.

The following figure shows the main MONICA interface with the fundamental data.



FCSCL

PUE Instantáneo Estimado

Actualizado: Mon Jan 30 09:20:40 CET 2012

>>> 1.1685 <<< (PUE Cat.1)

>>> 1.1340 <<< (PUE_Cat.2)

Evolución a 12h:

Evolución a 7d:

Consumo total Caléndula	138700 Watios
Consumo en TI estimado en UPSs	118695 Watios
Consumo en TI estimado en PDUs	122306 Watios
Consumo en refrigeración estimado	20005 Watios
Consumo Inrows estimado	4715 Watios
Consumo Chillers estimado	15290 Watios

Fig. 5. MONICA main's web page

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 first thing MONICA does is verification that the data are consistent. It checks whether a measure is consistent with the difference from the others.

The following diagram shows in real time the installation with its parameters in the given instant:

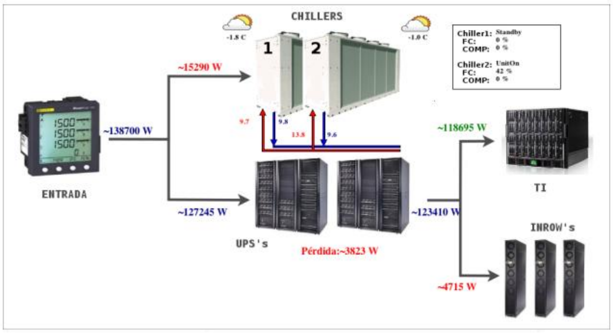


Fig. 6. MONICA installation's schema and parameters

As shown in this scheme, 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 to 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 below:

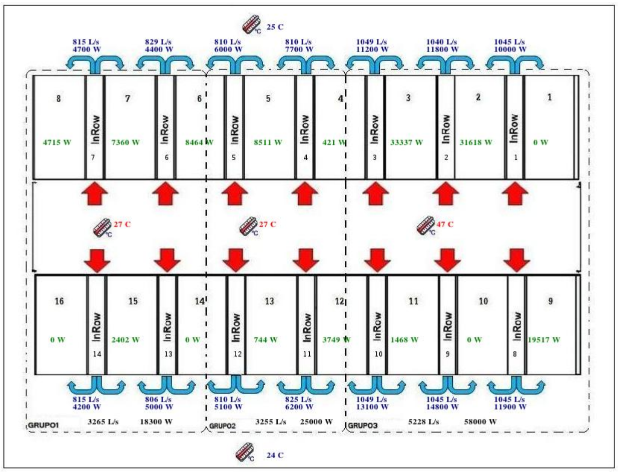


Fig. 7. MONICA racks schema

This diagram shows important operating data 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 19.5 kW 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 100 for optimizing the PUE (power load at PDU's 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 (feedback will be much faster by not needing human intervention in any case).

7. Conclusions

The most important conclusions for the Optimization of Calendula Energy Consumption are:

- Energy efficiency is one of the biggest challenges for the IT industry.
- To measure progress in energy efficiency is essential to establish measurement systems. The proposal by the Green Grid PUE/DCIE is correct.
- Supercomputers are very intensive installations in energy consumption, so in addition to

efficient facilities from an energy perspective, must make a dynamic PUE control.

- HPC facilities are very dense. But this problem also extends to general-purpose datacenters that use virtualization farms and that are also achieving higher server utilization rates and therefore high rates of use.
- Modern data center reaches a high-density level to which traditional technologies are unable to serve. It is necessary the adoption of new techniques (Free Cooling, Hot Aisle Containment System-HACS-InRow cooling, etc.).
- Modern datacenters must abandon their traditional view and adopt the model "industrial plant" with a comprehensive control and automated functioning.
- To achieve efficiency, data center must monitor PUE all the time.

Acknowledgement

The authors thank MYTIC the funding for the development of the MONICA project.

References

[1] Ruiz Falcó, A, “Design of an Ultra Dense HPC Datacenter with very high Energy Efficiency” Boletín RedIris nº 90, Madrid (2011), pp. 26-31.

[2] Belady, C, Rawson R, Pfleuger, J, Cader, T, “Green Grid Datacenter Power Efficiency Metrics: PUE and DCIE”, The Green Grid, pp. 1-9

[3] “Recommendations form Measuring and Reporting Overall Data Center Efficiency Version 2”, several councils including Green Grid, ASHRAE, Uptime Institute.