PIC
port d'informació
científica

21st International Conference on Computing in High Energy and Nuclear Physics **CHEP2015** Okinawa Japan: April 13 - 17, 2015

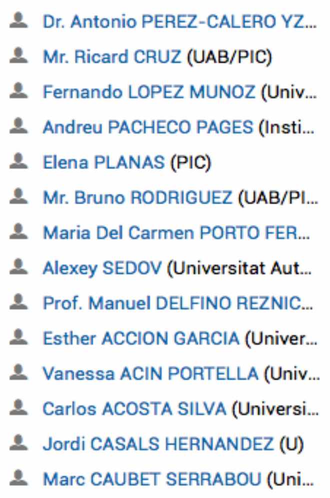
# Getting prepared for the LHC Run2: the PIC Tier-1 case

***J. Flix*** *

*On behalf of the PIC Tier1 team →*

** PIC Tier-1 project coordinator*
*WLCG Operations co-coordinator*
*CMS Resource Management Office co-coordinator*

*jflix@pic.es*     *@JosepFlixMolina*



- Dr. Antonio PEREZ-CALERO YZ...
- Mr. Ricard CRUZ (UAB/PIC)
- Fernando LOPEZ MUNOZ (Univ...
- Andreu PACHECO PAGES (Insti...
- Elena PLANAS (PIC)
- Mr. Bruno RODRIGUEZ (UAB/PI...
- Maria Del Carmen PORTO FER...
- Alexey SEDOV (Universitat Aut...
- Prof. Manuel DELFINO REZNIC...
- Esther ACCION GARCIA (Univer...
- Vanessa ACIN PORTELLA (Univ...
- Carlos ACOSTA SILVA (Universi...
- Jordi CASALS HERNANDEZ (U)
- Marc CAUBET SERRABOU (Uni...

EXCELENCIA
SEVERO
OCHOA

Institut de Física
d'Altes Energies

IFAE

Ciemat

Centro de Investigaciones
Energéticas, Medioambientales
y Tecnológicas

# PIC computing centre

**Port d'Informació Científica (PIC)** is the largest Grid centre in Spain, supporting **research** involving **analysis of massive sets of distributed data**

It provides computing services for many **applications**

- host the **Spanish WLCG Tier-1 centre** → **~85% of resources**

  *** Offer 5.1% of Tier1 computing resources for ATLAS, CMS and 6.5% for LHCb**

- host resources of the Spanish federated **ATLAS Tier-2**

- provides an **ATLAS Tier-3 facility**

**~5000 cores (~62 kHS06)**
**~6 PB disk**
**~12 PB magnetic tape**

Palau-solità i Plegamans
Terrassa
Parets del Vallès
Sabadell
Mollet del Vallès
PIC
Rubí
Martorell
Sant Cugat del Vallès
El Masnou
Montgat
Badalona
Sant Andreu de la Barca
Horta-Guinardó
Molins de Rei
Sant Martí
Sant Vicenç dels Horts
Barcelona
Torrelles de Llobregat
Sants - Montjuïc

~25 Kms from Barcelona
Autonomous University of Barcelona

# Tier-1 computing challenges for Run2

The **LHC experiments** will collect unprecedented data volumes in the next Physics run (Run2), with high pile-up collisions

## More data and more complex processing!

**Note1:** LHC experiments were asked to optimize the use of the available resources, in the midst of widespread funding restrictions, without penalizing Run2 physics objectives [Computing Model Update]

**Note2:** Most funding agencies asked (*forced*) their computing centers to operate with less money, without degrading performance

PIC
port d'informació
científica

# Tier-1 computing challenge~~s~~ for Run2

The real **challenge** during the last 2-3 years was to pave the road towards doing **MORE**, doing **BETTER…** with **LESS MONEY**!

# Tier-1 computing challenge~~s~~ for Run2

The real **challenge** during the last 2-3 years was to pave the road towards doing **MORE**, doing **BETTER…** with **LESS MONEY**!

# Tier-1 computing challenge~~s~~ for Run2

The real **challenge** during the last 2-3 years was to pave the road towards doing **MORE**, doing **BETTER…** with **LESS MONEY**!

Hence, **significant efforts** for experiments and sites were needed

- with the goal of providing a context compatible with flat funding

# Tier-1 data management upgrades

With better and increased network capabilities among centers, the Tier-1s become **data servers** to the whole Grid

- XRootD **fail-back** activated in PIC (WNs can read data from remote centers)
- ATLAS/CMS PIC data can be **XRootD** accessed **from remote centers** (~4 PB)
- LHCb data can be **HTTP** accessed **from remote centers** (~800 TB)

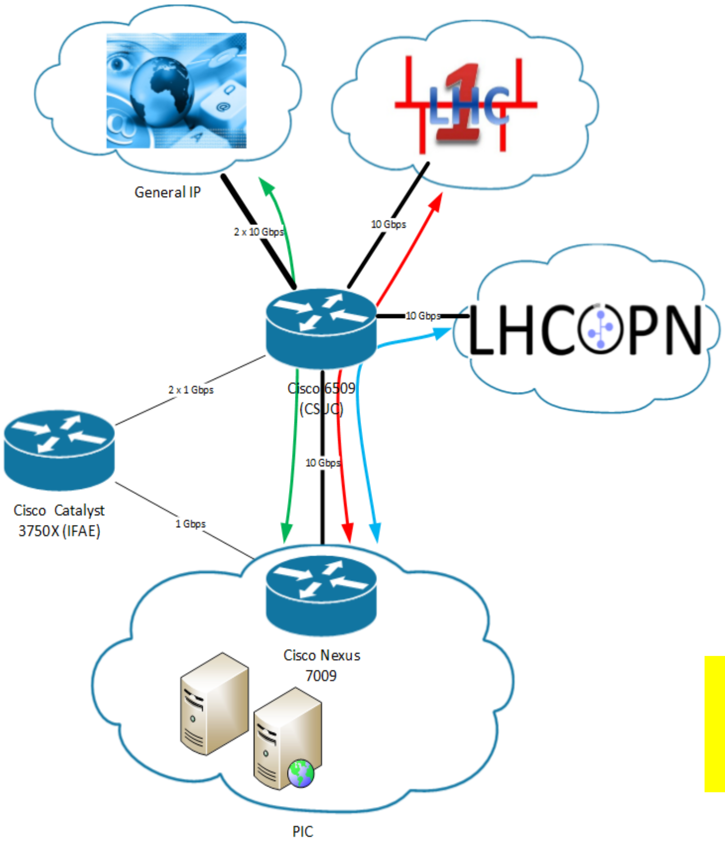
Joining the data federations requires(d) substantial **R&D and tuning**

- Deployment of compatible data management *software* (PIC: *dCache*)
- Creation of **disk-only pools**, to protect tape systems against uncontrolled *rw*
- Integration of dedicated experiment **monitoring plugins**
- Deployment of **site 'local' XRootD redirectors**
- Implementation of **protection mechanisms**

# Tier-1 data management upgrades

**Network access** to data allows for a valuable *cost optimisation*, as disk is the most expensive resource

But, this puts more load on the network and **network is not free!**



## Upgrades

- Careful planning
- Impact on LAN costs
  - * New switches and router upgrades
  - * The need for more powerful Firewalls (IPv6)
- Increase of WAN last-mile costs
- Deployment efforts

Not yet saturating, but WAN bandwidth increase is being drafted with involved parties

# WLCG multicore jobs @ PIC

Given the evolution of LHC running conditions at the restart of the data taking in 2015, experiments are developing **multicore applications**

- PIC co-coordinates the WLCG Multicore deployment Task Force

Oral #333

The **challenge for sites** in this new scenario

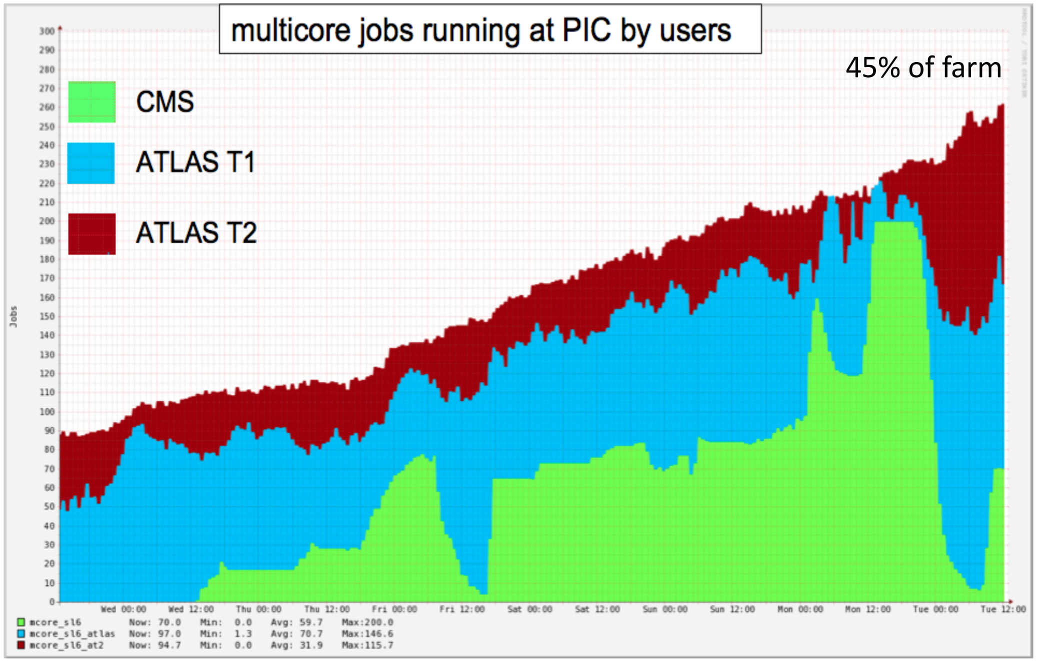
- Effective scheduling of both multicore and single-core jobs, that will still be used by all the VOs using shared sites
- Maximize CPU usage: minimize idle CPUs while there are jobs in queue
  → In particular **avoiding static splitting of resources**

In order to schedule multicore jobs, the n-core slots must be created

- Preventing single core jobs taking resources of ending jobs (***draining***)
  → **Backfilling** (using short running jobs while sufficient resources are being reserved to create a multicore slot) is not currently available/practical
- Therefore, draining represents a wastage, an **unavoidable price to be paid**
- Once the cost has been paid, **avoid multicore slot destruction**

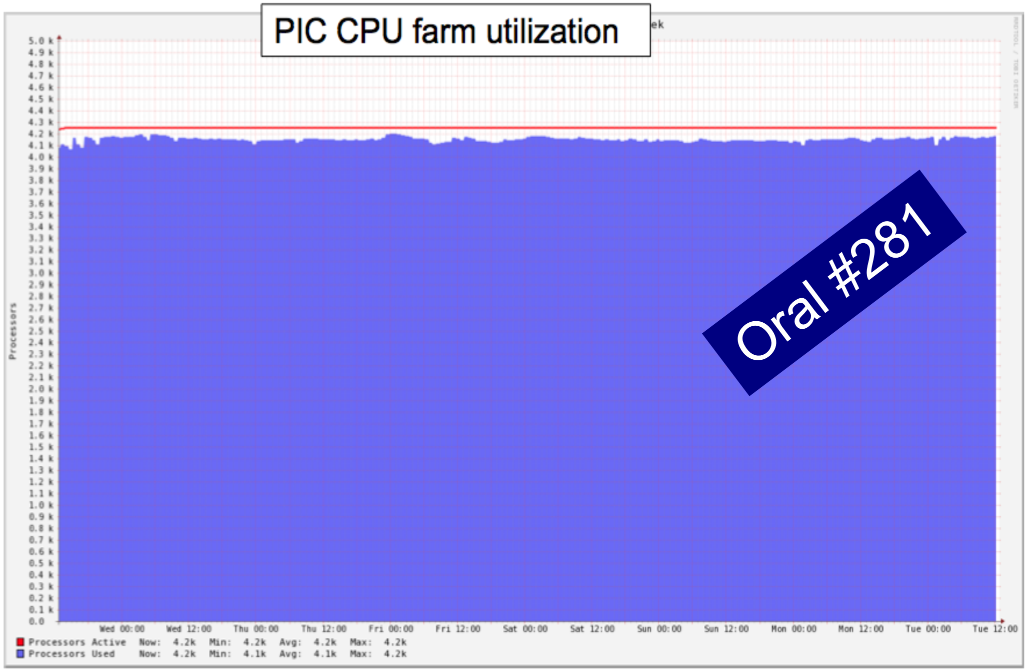
# WLCG multicore jobs @ PIC

Controlled draining and multicore slot conservation at PIC achieved with **dynamic partitioning** of site resources: implemented by **mcfloat** tool (NIKHEF) for Torque/Maui


multicore jobs running at PIC by users

45% of farm

| | | | | |
|---|---|---|---|---|
| 🟩 mcore_sl6 | Now: 70.0 | Min: 0.0 | Avg: 59.7 | Max:200.0 |
| 🟦 mcore_sl6_atlas | Now: 97.0 | Min: 1.3 | Avg: 70.7 | Max:146.6 |
| 🟥 mcore_sl6_at2 | Now: 94.7 | Min: 0.0 | Avg: 31.9 | Max:115.7 |


PIC CPU farm utilization

Oral #281

| | | | | |
|---|---|---|---|---|
| 🟥 Processors Active | Now: 4.2k | Min: 4.2k | Avg: 4.2k | Max: 4.2k |
| 🟦 Processors Used | Now: 4.2k | Min: 4.1k | Avg: 4.1k | Max: 4.2k |

*Controlled ramp up of multicore resources reduces draining impact on farm utilization*
*98% full farm while ramping up under combined pressure*

PIC
port d'informació
científica

# Free-Cooling at PIC

In 2014, PIC has improved the energy efficiency of its main computing room

→ 15 weeks of work, without any downtime, interruption and/or negative impact in Ops

## Before:

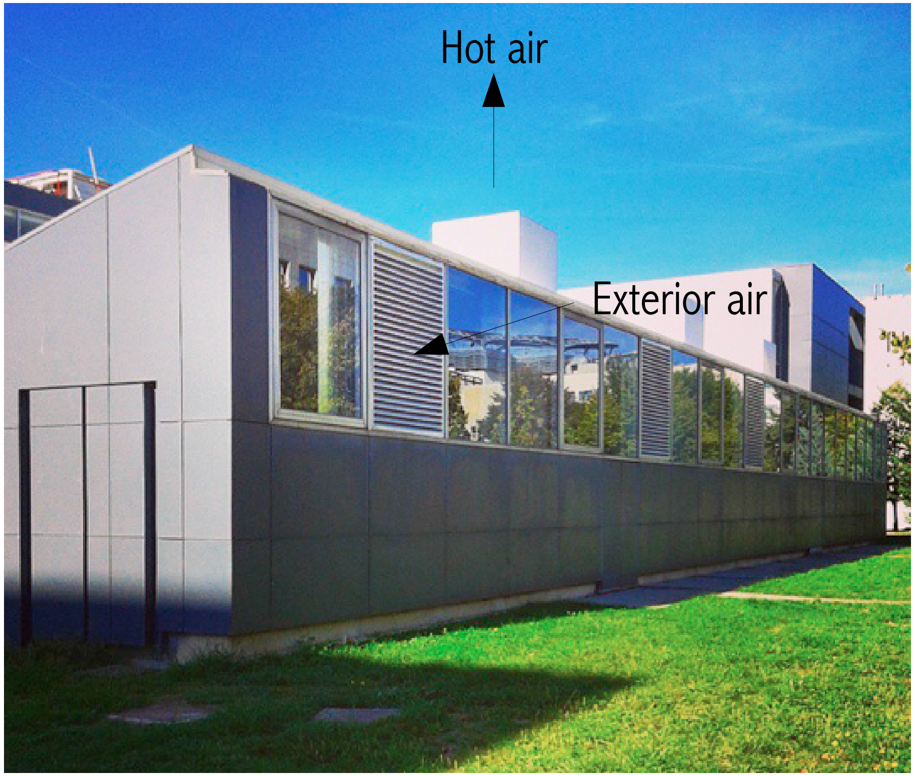
- No separation of cold/hot air in the room
- Several CRAH's (Computer Room Air Handler) managing the air through a cold water battery, injecting air at 14º C to get a room temperature of 22-23º C (*inefficient*)
- PUE (Power Usage Effectiveness) was about **1.8**

## After:

- CRAH's replaced by 3 free-cooling units: indirect heat exchangers with outside air and equipped with adiabatic cooling humidifiers
- Implemented separation of hot and cold flows in the room
- Hot aisle containment and confinement + installation of ceiling to contain the hot air
- Increase of inlet temperature according to the ASHRAE recommendations
- Installation of dedicated monitors for the most important climate parameters
- PUE expected in the range **1.45-1.3**

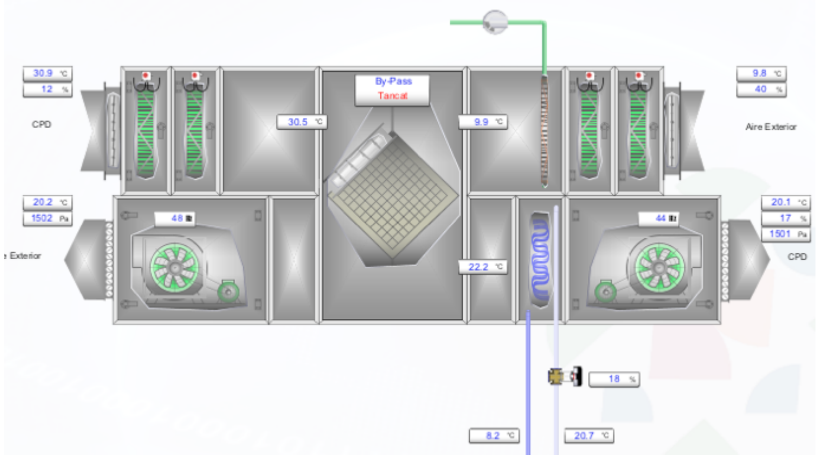
# Free-Cooling at PIC

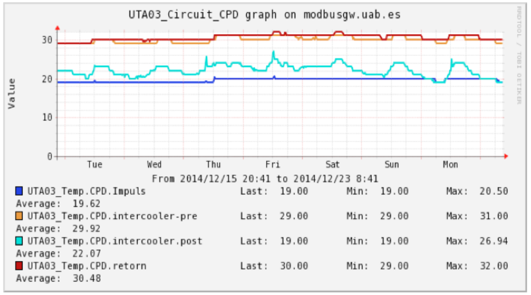
Installation of free-cooling units



New technical area

Free-cooling unit control/monitoring

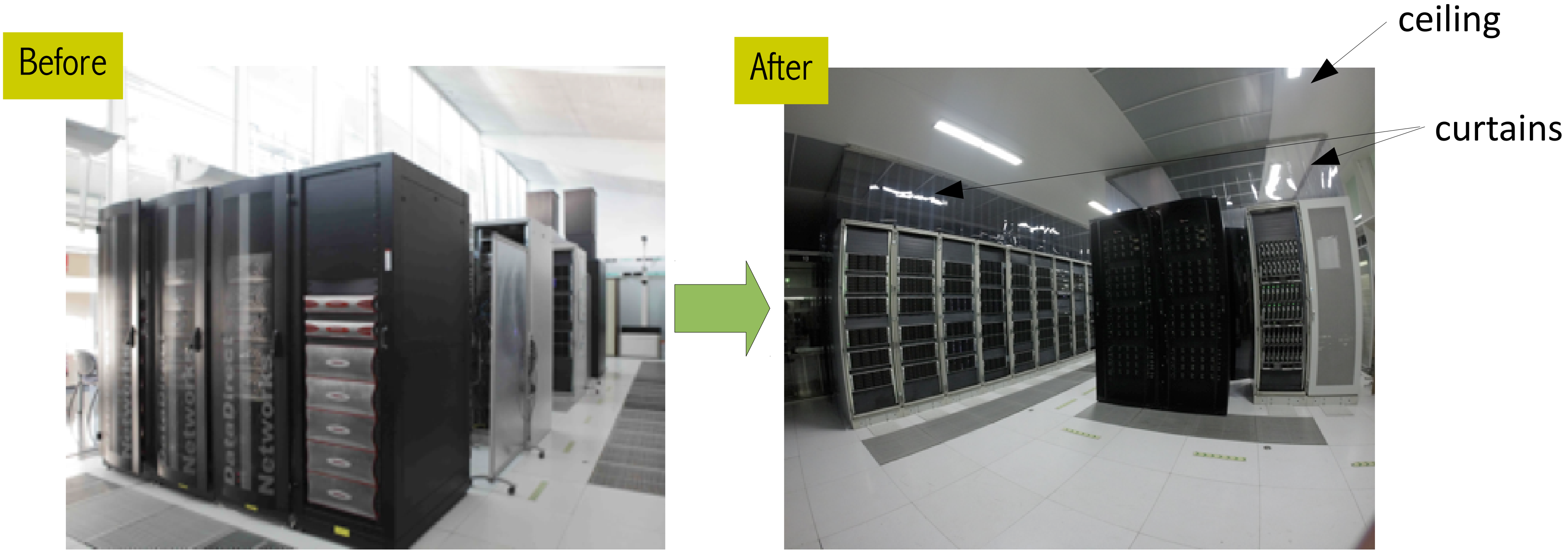




```
                    From 2014/12/15 20:41 to 2014/12/23 8:41
■ UTA03_Temp.CPD.Impuls              Last:  19.00     Min:  19.00     Max:  20.50
Average:  19.62
■ UTA03_Temp.CPD.intercooler-pre     Last:  29.00     Min:  29.00     Max:  31.00
Average:  29.92
■ UTA03_Temp.CPD.intercooler.post    Last:  19.00     Min:  19.00     Max:  26.94
Average:  22.07
■ UTA03_Temp.CPD.retorn              Last:  30.00     Min:  29.00     Max:  32.00
Average:  30.48
```

rdd graphs

# Free-Cooling at PIC



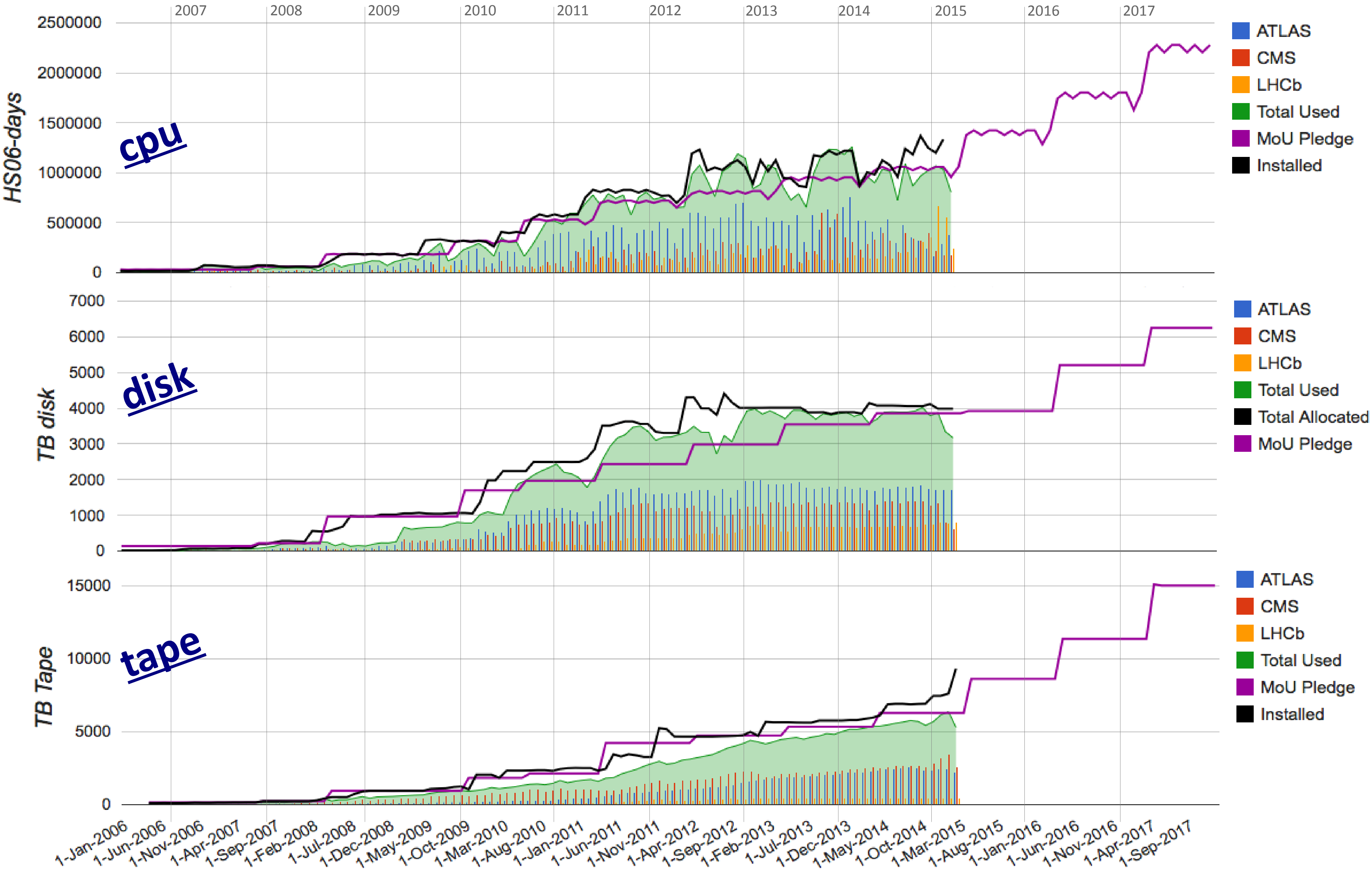
The work was completed in September 2014

- one-year period ahead to study/adjust the system: reach maximum energy efficiency
- In December 2014, we already reached PUE of 1.3!
- Electricity costs savings in the next <4 years amortized investment

Poster #253

PIC
port d'informació
científica

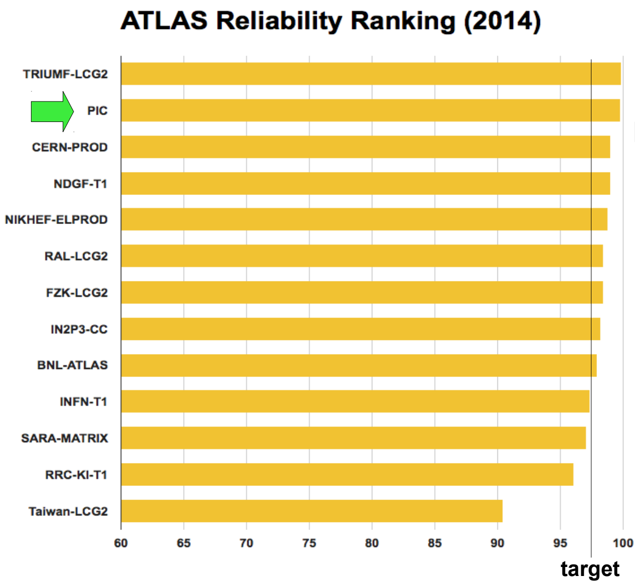
# PIC Tier-1 CPU and storage capacity growth...

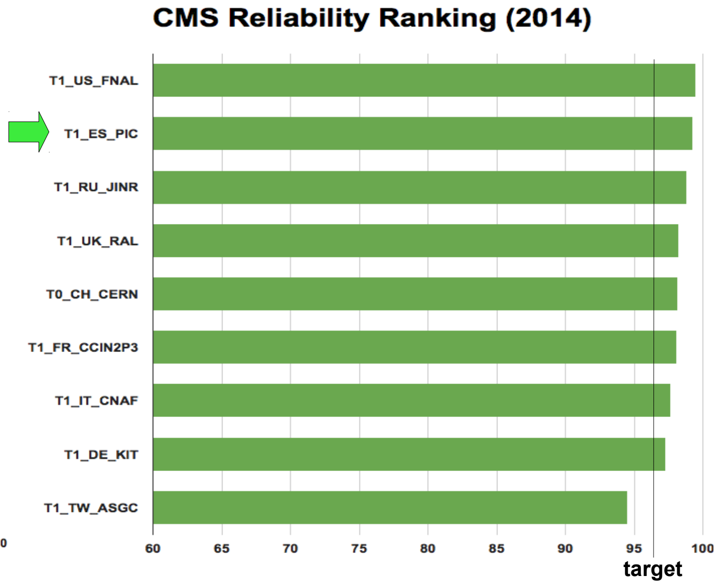


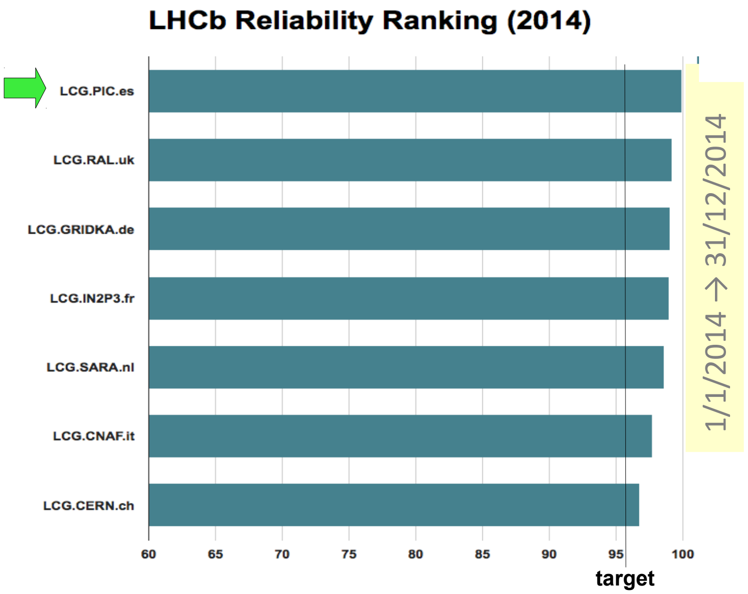
✔ PIC Tier1 delivering in terms of deploying pledged capacity

# … with excellent reliability and efficiency

From Jan. 2014, WLCG measures the reliability using more detailed **experiment probes**


ATLAS Reliability Ranking (2014)


CMS Reliability Ranking (2014)


LHCb Reliability Ranking (2014)

1/1/2014 → 31/12/2014

PIC Tier1 is at the **top of Reliability Rankings** (99.9% ATLAS, 99.4% CMS, 99.9% LHCb)

YES, being smaller makes it a bit easier to be reliable

BUT, being a multi-experiment site makes it harder

✔ PIC Tier1 delivering in terms of service quality

# Let's save more money: costs/efforts

Deployed a **RedHat Enterprise Virtualization system** (RHEV 3.4.2), KVM-based

**7 Hypervisors, each**: 16 cores / 96GB RAM (HP Proliant BL460c) with 2x10GbE
NetApp FAS3220 (2 TB, Thin Provisioning - QCOW2) is FC-connected to the HPBlade Box

→ This reduces the number of physical machines by a factor 10, without impact on the reliability and services performance – at ⅓ costs!

→ Testing **Ovirt 3.5** at scale – to save license costs

Constant efforts to improve **configuration management** and **automation**

A **new powerful** (Insulated Gate Bipolar Transistor) **UPS of 550 KVA** was recently installed, w/efficiency in the range of 97%-99% (small loses)

We **adjust** the PIC farm power to electricity cost, since beg. 2013

→ Less CPU during high cost periods, and vice-versa, keeping annual pledges OK

→ Reduction of electricity bill is **~10%**

PIC Tier-1 is operated with less personnel as compared to average Tier-1 values

…

# Conclusions

PIC Tier-1 **compliant** with the new WLCG requirements for Run2

The needed **resources** are in place

Computing center **infrastructure** has been improved to reduce costs

**Operational** and **maintenance** costs have been as well reduced, without compromising any of the objectives

The implementations done in PIC are **flexible** enough to rapidly evolve following changing technologies

**Thanks!**
**Questions?**

FOLLOW US ON twitter @pic_es