PIC
port d'informació
científica

Oxford Physics HEPiX
HEPIX SPRING 2015 WORKSHOP

PIC Site Report

HEPiX Spring 2015 at Oxford University, UK
23-27 March 2015

*J. Flix**

On behalf of the PIC Tier1 team

** PIC Tier-1 project coordinator*

EXCELENCIA SEVERO OCHOA
Institut de Física d'Altes Energies
IFAE
Ciemat
Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas

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**

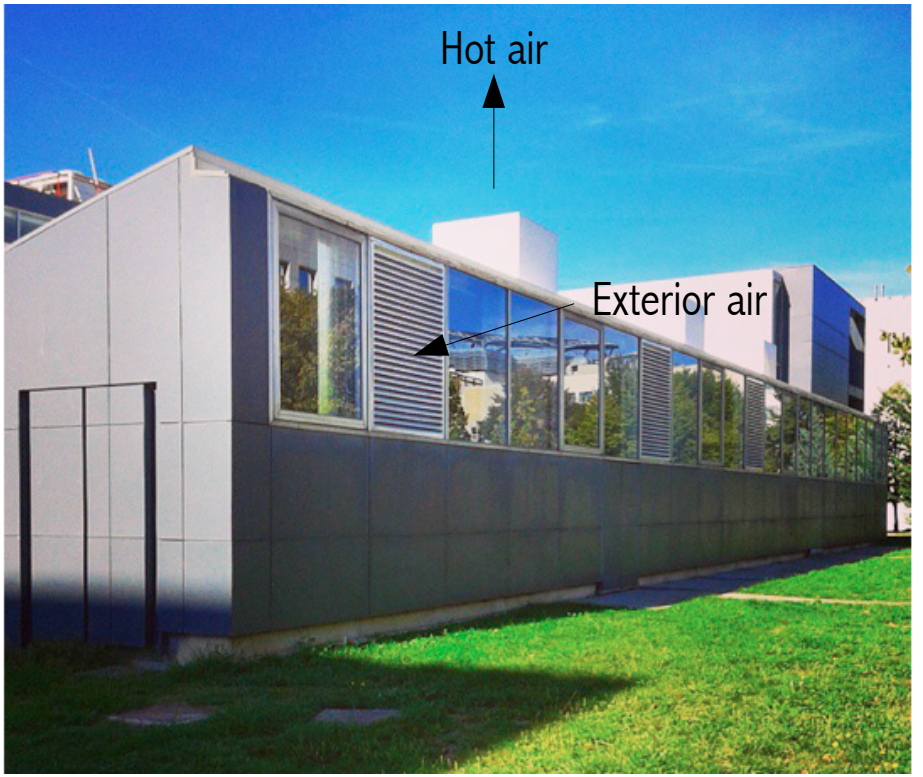
After:

- 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 is in the range **1.45-1.3**

PIC
port d'informació
científica

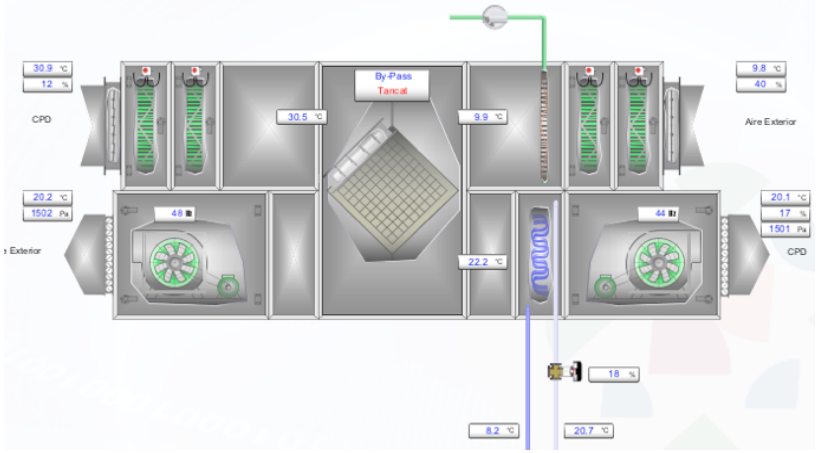
Free-Cooling at PIC

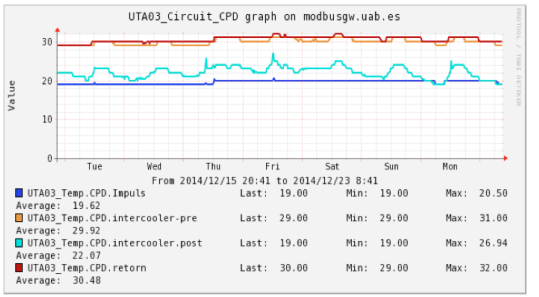
Installation of free-cooling units



New technical area

Free-cooling unit control/monitoring





rdd graphs

PIC
port d'informació
científica

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 years estimated at **~100k€/year**

More **work ahead**:

- Current UPS has loses of ~15% → **New UPS of 550 KVA** w/IGBT technology to be installed in two weeks, w/efficiency in the range of 97%-99%
- Compact module (2/3 of CPU) to be upgraded w/liquid cooling solutions (*immersion*)

PIC
port d'informació
científica

Service upgrades in PIC

Services virtualisation: Testbed with oVirt 3.5

- 4 Hypervisors (SunBlades – 2 VMs/2mgt) to test services and the new platform
  → Supervisors connected to a NetApp via 2x1GbE
  → ~60 test services
  * To replace the prod. **RedHat Enterprise Virt.** (RHEV 3.4.2), KVM-based

Migrated local repositories from **SVN to GIT** / ***gitlab*** for projects code depl.

Adapted all of the configuration management to **Puppet 3.6**

Pilot tests with **OpenStack** & **Docker** for Astro/Cosmological projects

PIC FTS2 stopped for WLCG. **FTS3** instance deployed & used by other VOs

Currently running **dCache 2.10.20**

Enstore4 was deployed in 2014 (+ new HW). Currently at **Enstore 4.2.2-3**

PIC Tier-1: Computing Farm

2014 WN purchases (~10 kHS06 [1] & ~15 kHS06 [2] – 2300 slots)

[1] Ivy Bridge: **E5-2650-v2** (dual processor 8 cores/proc) – 3 blades
[2] Haswell: **E5-2640-v3** (dual processor 8 cores/proc) – 3 blades
→ Equipped w/2x1TB disks (RAID0 for 2TB), 4 GB RAM/core, 10GbE/node

- 1 blade of **[1]** w/o HT: for ATLAS HighMemory (dynamic usage)
- The rest w/HT: optimized for Multicore jobs (24 cores/node)

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%**

Currently **~55 kHS06 – 4368 slots**

- ~16 kHS06 X5650 *off*, used to compensate in periods

Current farm running under Torque 2.5.13 + Maui 3.3-4

- **HTCondor** seen as plausible replacement – Testbed in place
- Collaboration w/other sites (CERN) – next: ARC CE / Grid tests

PIC
port d'informació
científica

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

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**

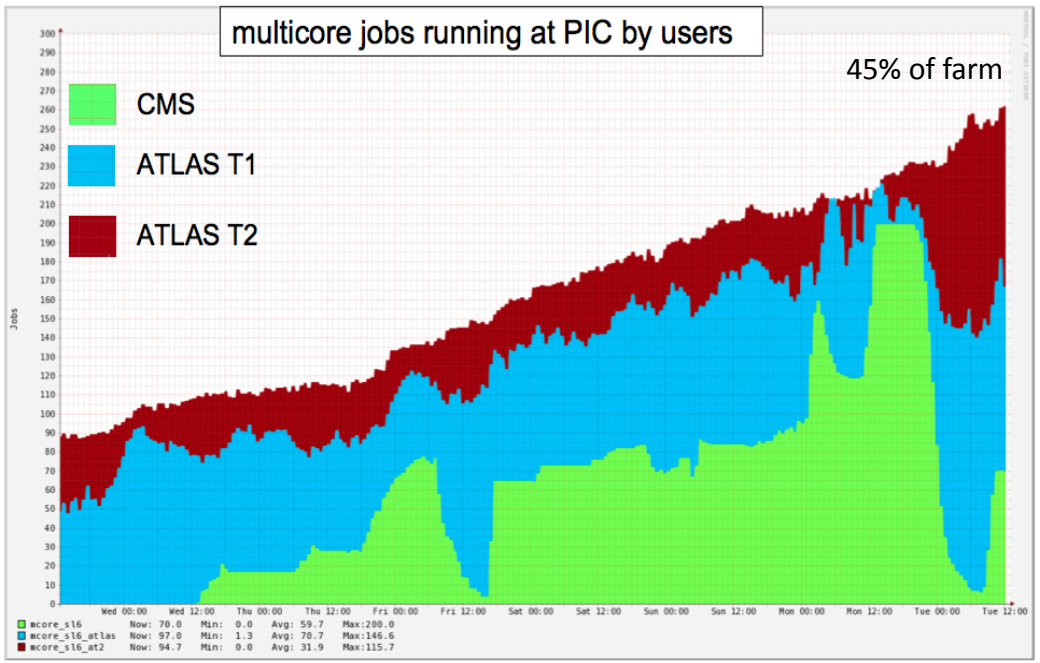
In order to schedule multicore jobs, the n-core slots must be created

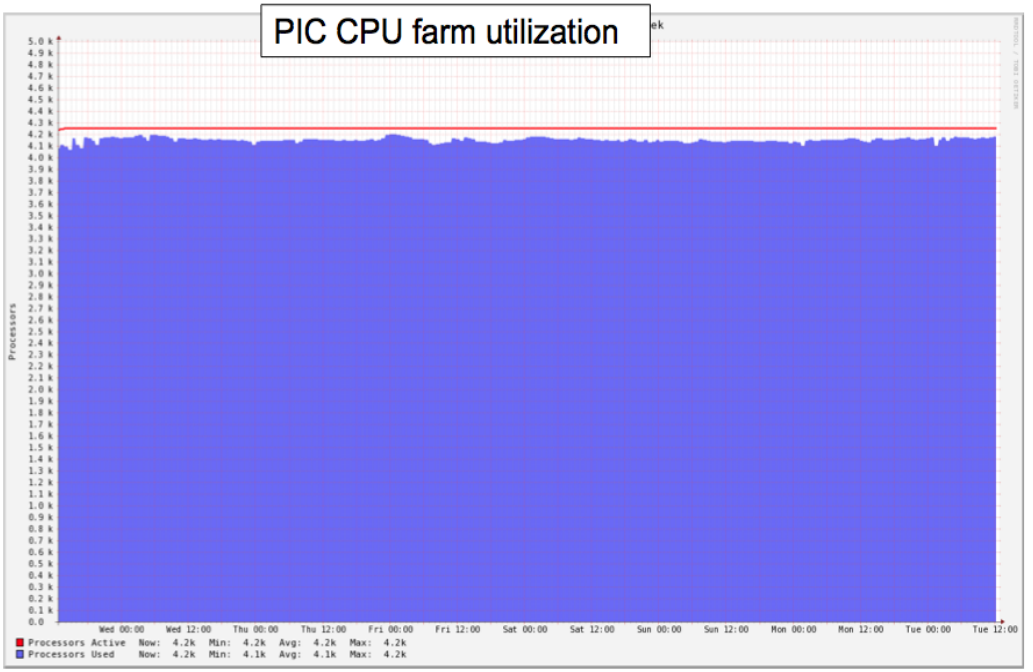
- Preventing single core jobs taking resources of ending jobs (***draining***)
  → **Backfilling** (using short running jobs while sufficient resources are being reserved to create a multicore slot) is not currently available/practical
- Therefore, draining represents a wastage, an **unavoidable price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

WLCG multicore jobs @ PIC

Multicore slot conservation can be achieved with **dynamic partitioning** of site resources: implemented by **mcfloat** tool (NIKHEF) for Torque/Maui

- Moving WNs between separated pools for single and multicore jobs
- **Controlled draining**: only a small percentage of the total number of cores in a site is drained simultaneously – multicore slots preserved for a while when jobs end





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

PIC Tier-1: Disk Storage

5 FlyTech SC847 disk servers were acquired in 2014

- Each server has ~350 TB net space:
  → 6 TB disks, 4K native
  → 2x10GE network
- Retired 1.2 PB (old SuperMicro 80TB/server)

Currently, **~6 PB Disk Storage** managed by dCache

NFS 4.1 enabled in PIC

- particularly for non-LHC projects for which Grid access is inconvenient

Enabled **HTTP/WebDAV** access:

→ ATLAS namespace renaming & deletions
→ Joined LHCb HTTP federation

Deployed **XRootD** and joined AAA and FAX federations

In 2014, we installed a small **Ceph** cluster to test its functionalities

PIC Tier-1: Tape Storage

All data now goes to **StorageTek** (STK) SL8500 robot

- IBM TS3500 library was decommissioned: 490 LTO3 & 500 LTO4 data → STK

In 2014, **2 new T10KC drives** were added in the STK robot

- Currently: 10 drives LT04; 3 drives LT05; 8 drives T10KC

Old LHC data was migrated from old technology to T10KC cartridges

- Around 900 T10KC cartridges were purchased in 2014
- All of the MAGIC Telescope's data is written as of today to T10KC
- No T10KD is considered: sufficient slots / amortize investment in LTO4/O5 & T10KC

Currently, **~12 PB Tape Storage** at PIC

- 3107 LT04 tapes / 1416 LT05 tapes / 1378 T10KC tapes

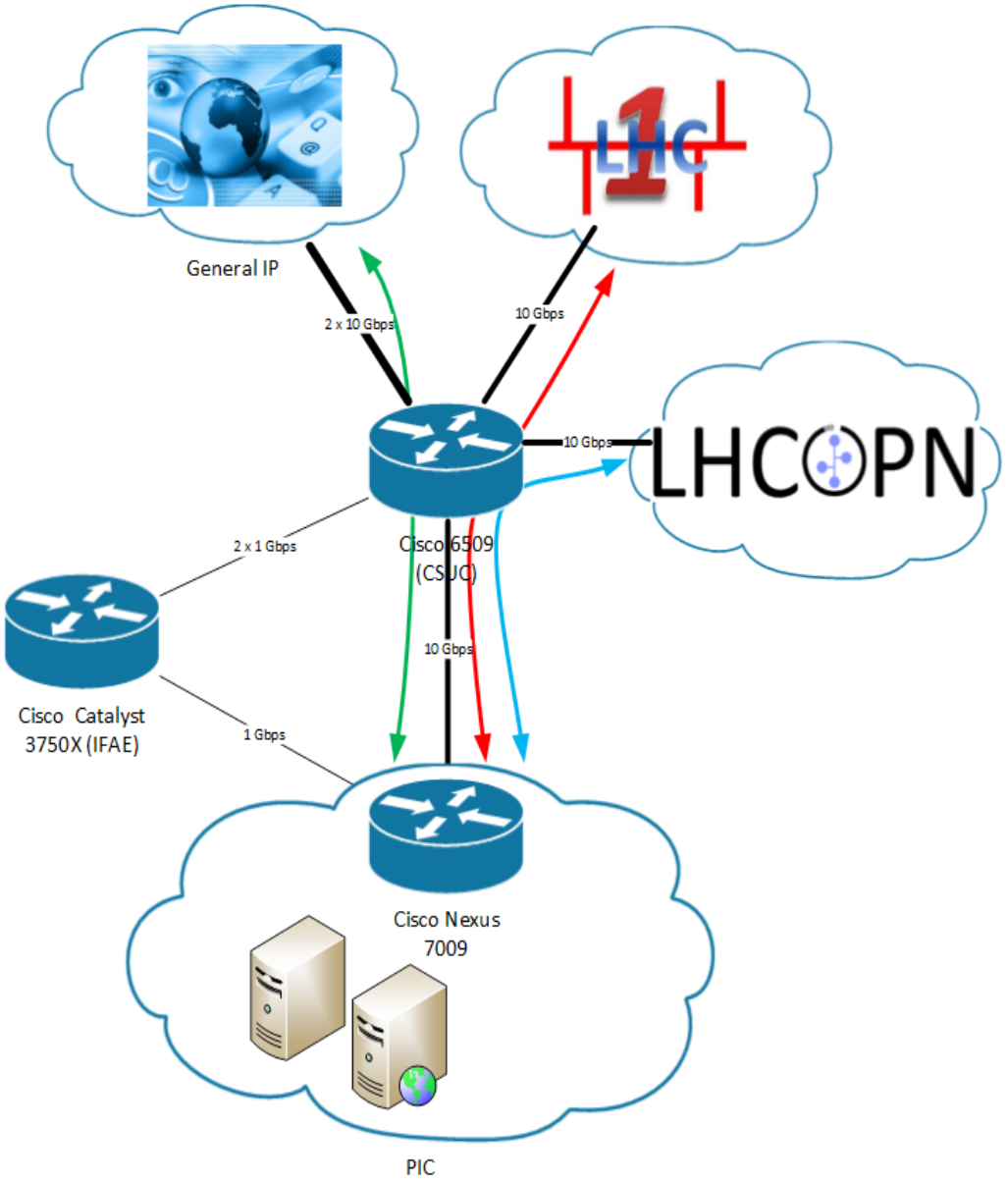
Since Q3/2013, **we add 10% resources** above WLCG pledges to account for operational inefficiencies (data deleted subject to repack & recycling)

PIC is the local organizer for the **Large Tape Users Group conference**

- LTUG will take place in Barcelona in October 2015

PIC
port d'informació
científica

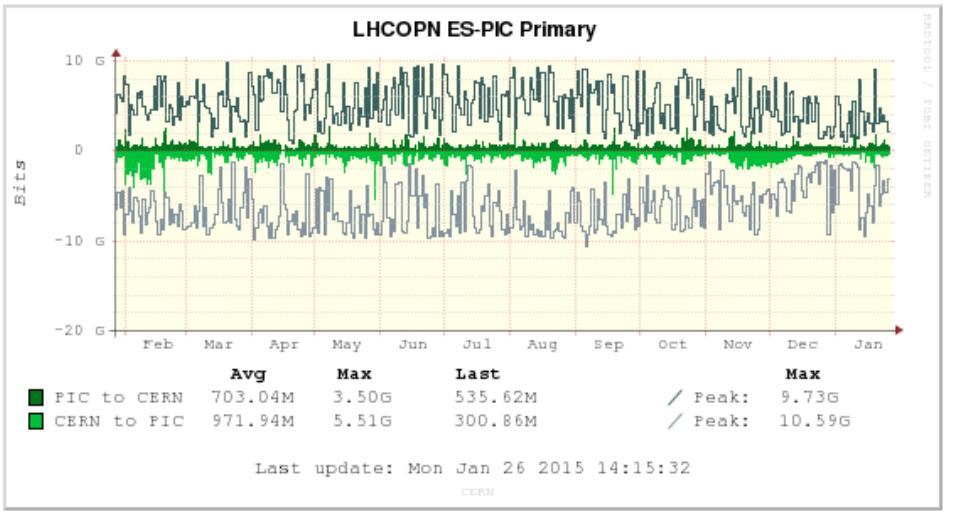
Network Upgrades



- **IPv6** testing of services (HEPiX IPv6 WG)
  - dCache + FTS3 + PhEDEx tests
  - **PerfSONAR-PS** is running in dual-stack
  - 2 new Firewalls (Fortinet) acquired
- New **48x10G** SFP+ card for Nexus 7009
- **2 new Nexus 2000** w/48x1G ports
- Traffic **saturation** is not (yet) significant

  → We monitor actively our 10 Gbps line

  → We might be upgrading to 20 Gbps



PIC team

Follow Us On twitter @pic_es

20 FTEs: Researchers/Engineers/Administration

Dedicated team with expertise in high-throughput mass storage and computing

LHC-specific interface team deals with specific applications of ATLAS, CMS, LHCb