PIC
port d'informació
científica

Dimensioning storage and computing clusters for efficient High Throughput Computing

G. Merino, on behalf of
PIC Services & Infrastructure Team

CHEP 2012, NYC, May 22nd 2012

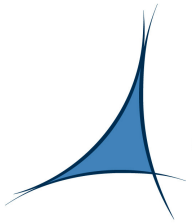
Generalitat de Catalunya
Departament d'Economia i Coneixement
Secretaria d'Universitats i Recerca

GOBIERNO DE ESPAÑA MINISTERIO DE CIENCIA E INNOVACIÓN
Ciemat
Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas

Institut de Física d'Altes Energies IFAE

UAB
Universitat Autònoma de Barcelona

UAB CEI
CAMPUS D'EXCEL·LÈNCIA INTERNACIONAL



PIC
port d'informació
científica

PIC Services & Infrastructure Teams:

E. Accion, V. Acin, A. Bria, R. Cruz,
G. Bernabeu, M. Caubet, X. Espinal,
F. Lopez, F. Martinez, E. Planas

Port d'Informació Científica

PIC
port d'informació
científica

PIC is a scientific-technological centre supporting research groups that need to analise large data sets in collaborative environments.

Spanish Tier1 centre for ATLAS, CMS and LHCb.

Data processing services for several disciplines:

- HEP
- Astrophysics & Cosmology
- Life Sciences

Support to multiple disciplines is in PIC's mandate.

Goal: technology transfer from the LHC Tier1 to other applications with similar needs.

LHC computing resources

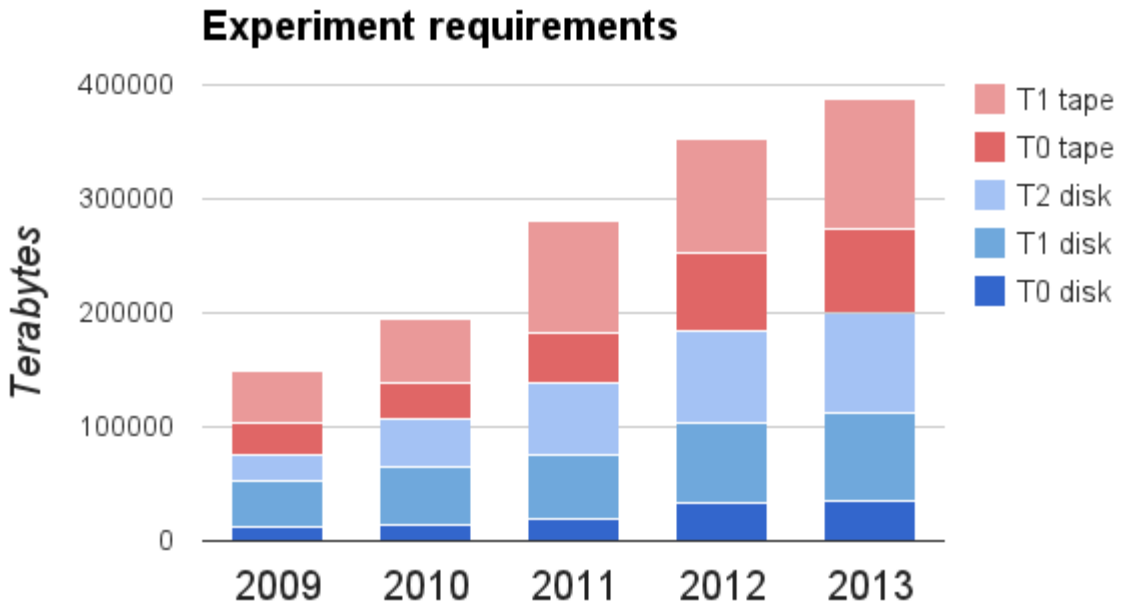
PIC
port d'informació
científica

Always quoted that LHC generates ~15 PB every year.

Looking at the total experiment requirements: it is much more.

In the process of analysing LHC data, several types of secondary data are generated.

Also, lots of Monte Carlo simulation is needed.



~ +60 PB every year on avg.

Tape

PIC
port d'informació
científica

Dealing with Petabytes of data (still) implies the need to use tapes: cost-effective.

Data processing and analysis models need to make use of the "tiered storage" model.

Interaction of processing nodes (CPU) and storage needs to be tuned to minimise CPU inefficiency due to I/O waits.

- Online data (disk): needs to flow reliably
- Nearline data (tape): needs coordinated prestage on buffer disk

PIC HTC centre figures

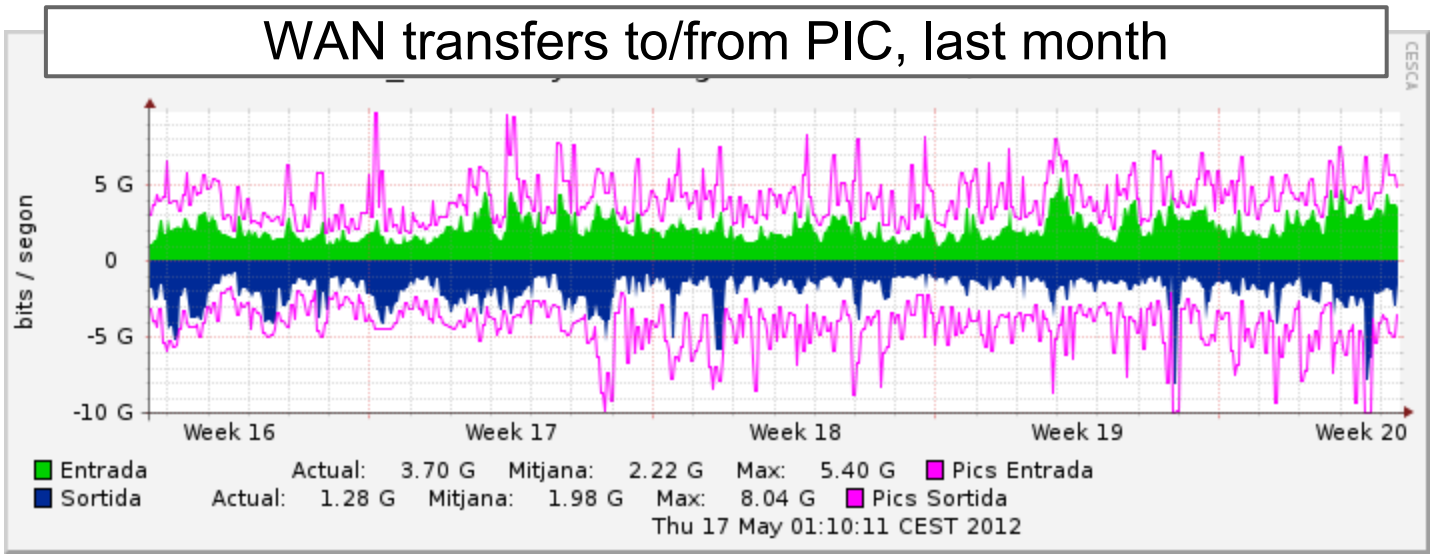
PIC
port d'informació
científica

~4000 cores CPU farm and ~4 PB disk (+4 PB tape)

LAN WN - Storage ~ 20 Gbps average

WAN imports/exports ~ 2 Gbps average

- Sometimes reaching saturation (10 Gbps) for short periods of time.
- Ready to upgrade to 2x10GE when need arises.



Worker Nodes

PIC
port d'informació
científica

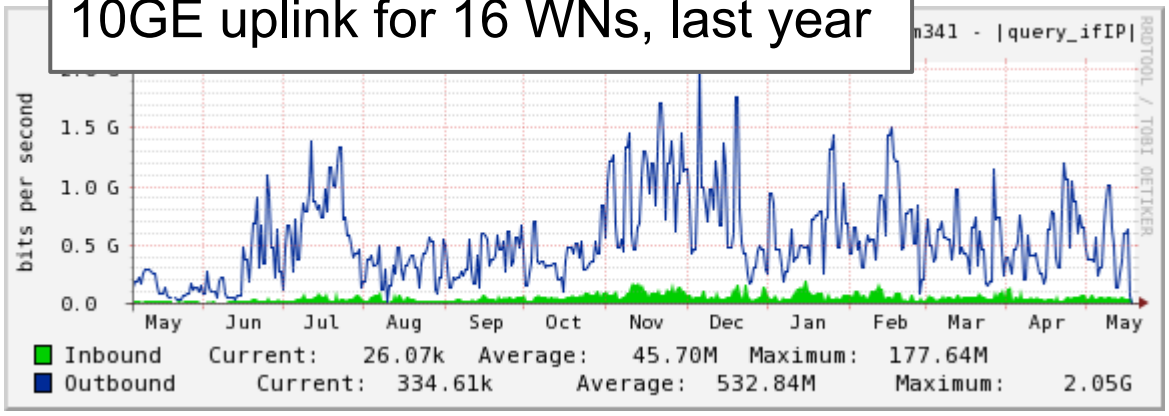
Dell PE C6100

- 2x1GE at the WN
- 1x10GE uplink/16 WNs

HP BL260 and BL460 blades

- 2x1GE at the WN, 2x10GE(1GE) uplinks/16 WNs
- 1x10GE at the WN, 2x10GE uplinks/16 WNs

10GE uplink for 16 WNs, last year



Worker Nodes

PIC
port d'informació
científica

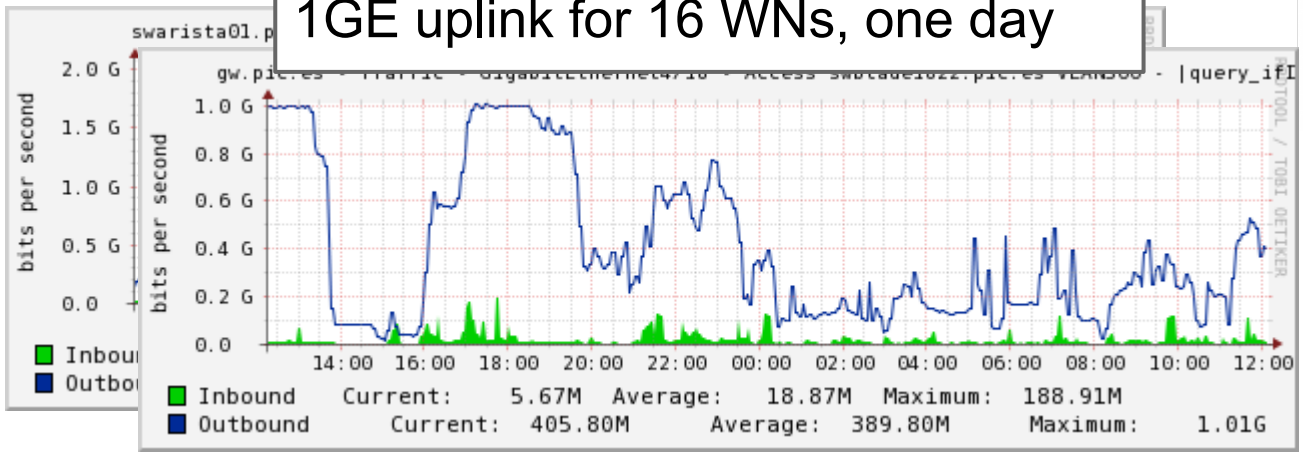
Dell PE C6100

- 2x1GE at the WN
- 1x10GE uplink/16 WNs

HP BL260 and BL460 blades

- 2x1GE at the WN, 2x10GE(1GE) uplinks/16 WNs
- 1x10GE at the WN, 2x10GE uplinks/16 WNs

1GE uplink for 16 WNs, one day



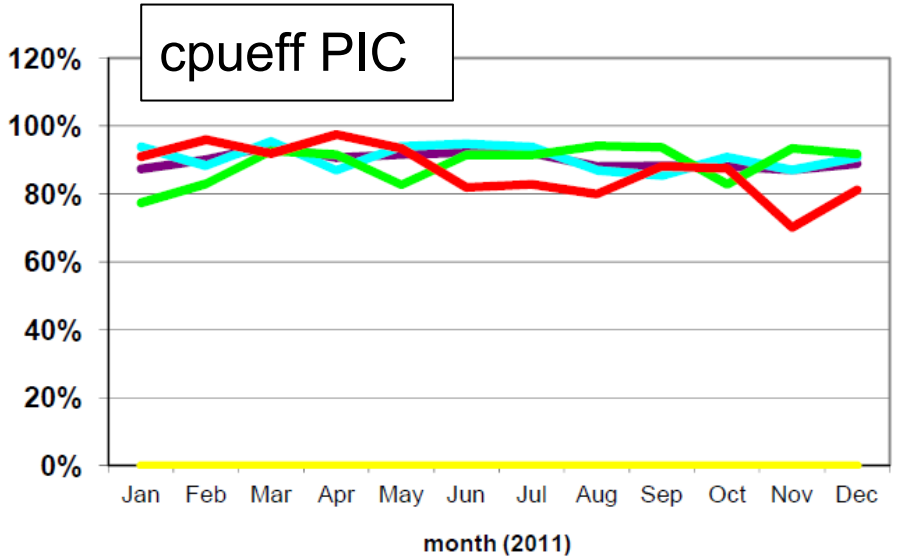
CPU efficiency

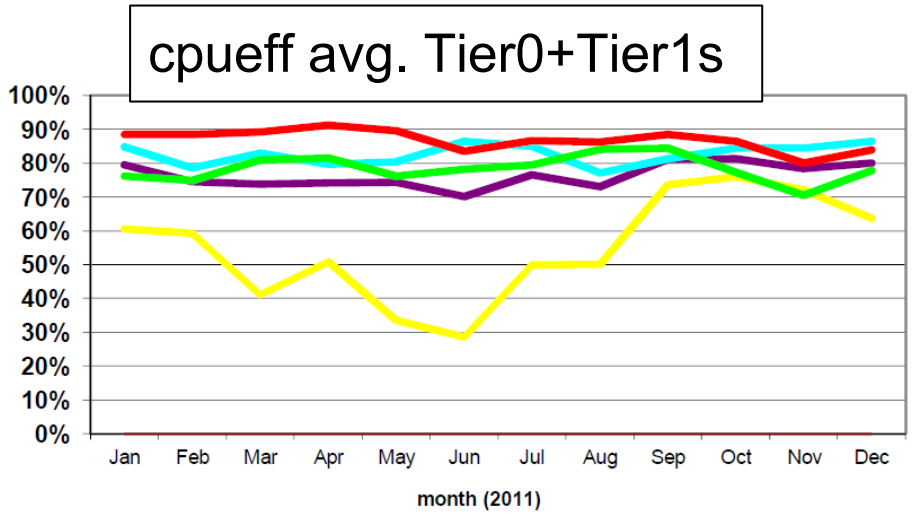
PIC
port d'informació
científica

In the old generation equipment (1GE uplinks) sometimes observe short periods of network saturation.

No visible impact in the cpu efficiency of the overall service.

The figure of merit is throughput. The ~1h spikes get diluted.





Disk servers: SAN

PIC
port d'informació
científica

2x 600 HD (2 TB per HD)

2 controllers (RAID6 of 8+2 disks)

FC connected to 2x 8 servers

- each server 2x 10GE (active-passive)

=> 8.3 MB/s/TBraw, per server

(~24h to migrate a full server)

Disk servers: DAS

PIC
port d'informació
científica

w/o hw RAID: Sun x4500

- 17 servers, 48 HD/server, 1 TB/HD
- 4x1GE per server
- => 10,4 MB/s/TBraw

with hw RAID: SuperMicro, SGI, Flytech

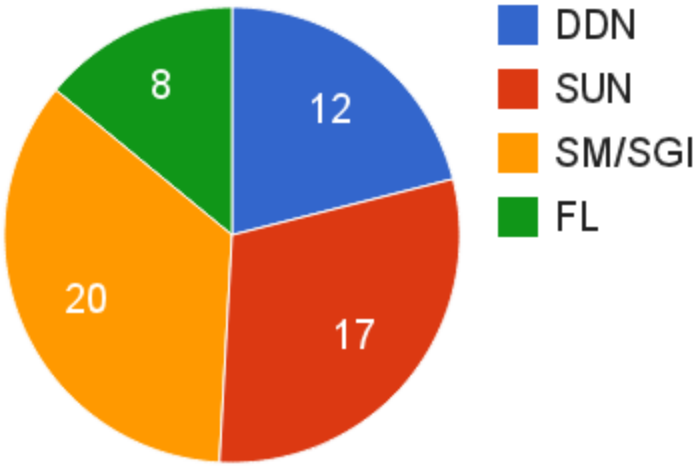
- 2 TB/HD: 20 servers of 36 HD/server
  - 1x 10GE per server => 17,3 MB/s/TBraw

- 3 TB/HD: 8 servers of 81 HD/server
  - 2x 10GE per server => 10,3 MB/s/TBraw

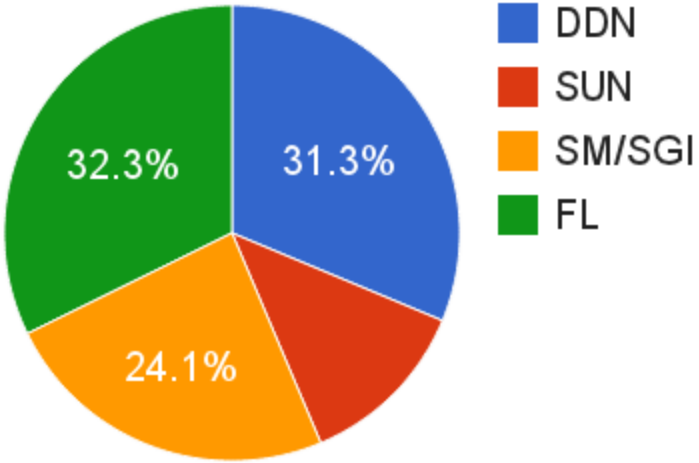
Disk servers

PIC
port d'informació
científica

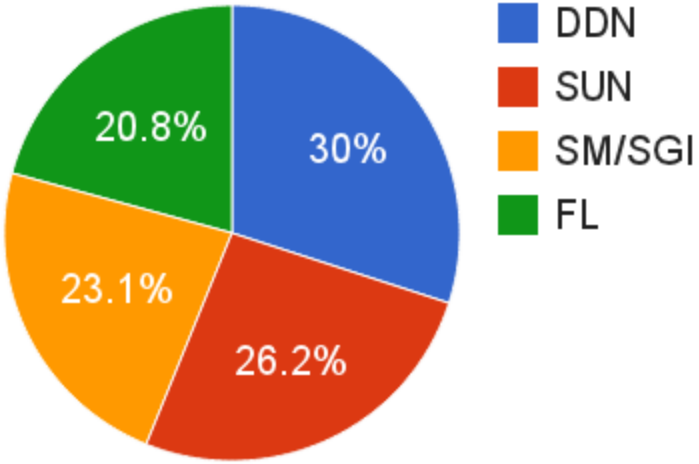
Number of servers



Terabytes



Number of spindles



57 Disk servers
4742 Terabytes
3119 Spindles

RAID configurations

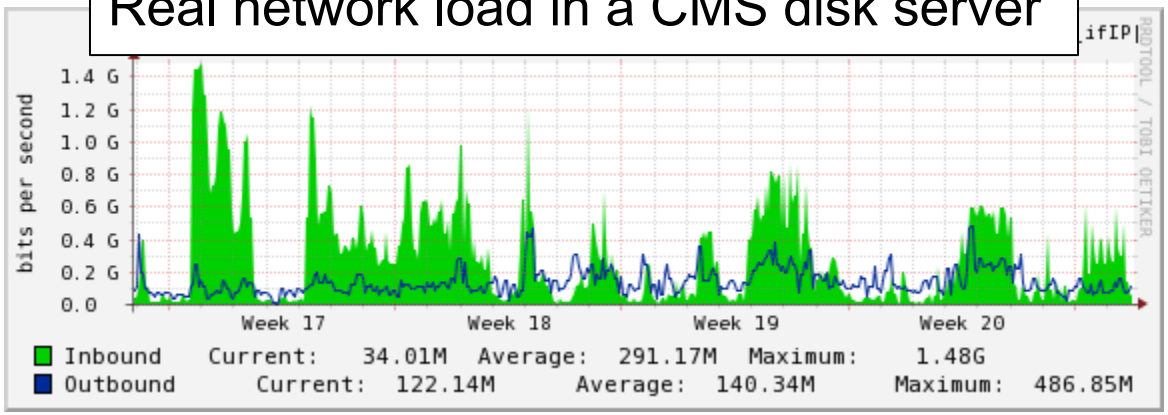
PIC
port d'informació
científica

RAID 60: 3x (RAID6 10+2) => Controller CPU saturation.

LVM grouping of 3x (RAID6 10+2)

- Multiple streams:
  - writes ~ 400 MB/s
  - reads ~ 1 GB/s
- Acceptable. Writes can be balanced into different disk servers.

Real network load in a CMS disk server

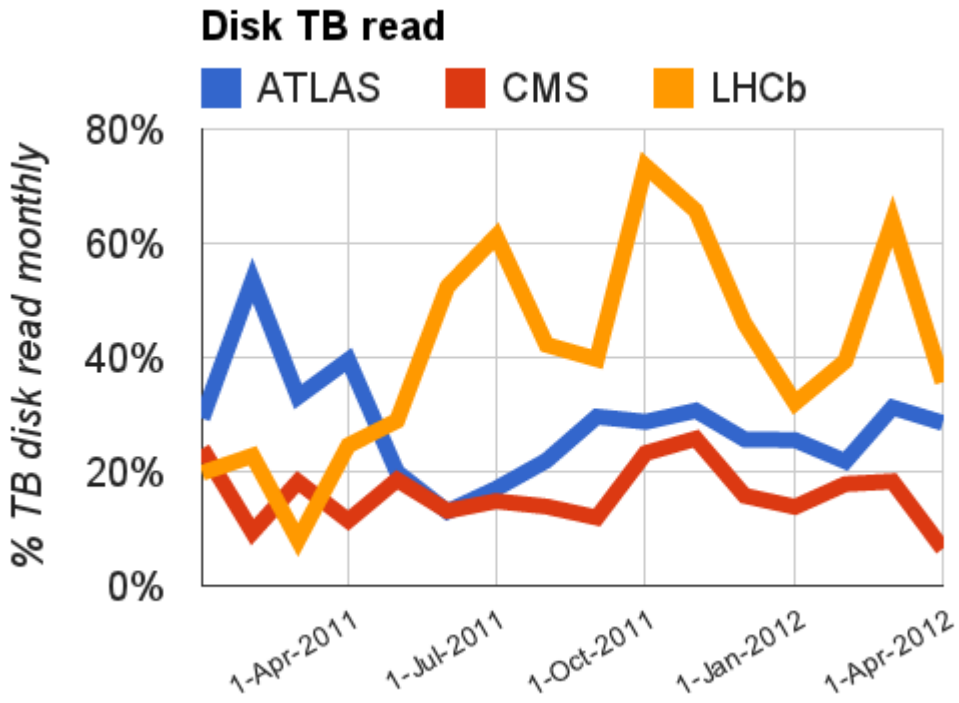


~< 10% of network max capacity

Data on disk usage

PIC
port d'informació
científica

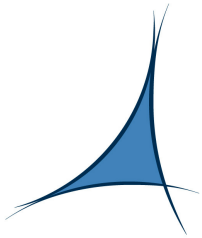
Looking at the fraction of the data on disk (different files) which is read every month, one sees that it is often quite low.



Given that LHCb represents ~10% of the resources

=> every month, **75 - 80%** of all the bytes on disk at PIC are not read.

dCache configuration



PIC
port d'informació
científica

All the pools are set up to accept any kind of workload or file by default:

- WAN, LAN or tape streams
- read or write
- any Space Token

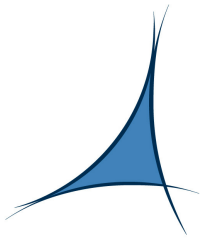
Pros:

- Maximise the nr. of spindles for any kind of workload.
- All available space at a given time, usable as buffer in front of tape.
- The impact of a disk server failure is diluted.

Cons:

- Potential interference between loads, contention effects.

dCache Issues



PIC
port d'informació
científica

The few incidents with PIC dCache system have been mostly caused by Issues associated to the pool cost calculation and pool-to-pool copies.

Example: One pool gets a "spike" of requests

-> load increases

-> dCache triggers p2p copies to try and balance intto other pools

-> overload gets worse

-> p2p copies hang and accumulate

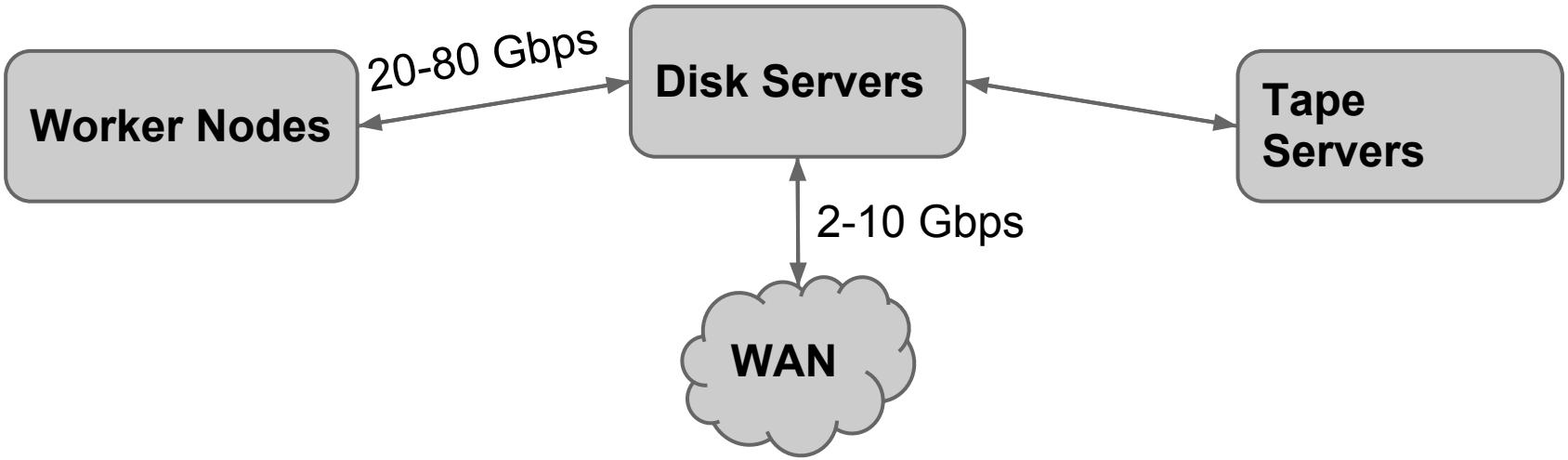
Looking forward to the new WASS (*Weighted Available Space Selection*) PoolManager cost system for a better file write distribution and expecting some improvements in the read cost balancing soon.

Network

PIC
port d'informació
científica

Star-like topology:

- Originally Cisco 6509 as central component.
- Limiting backplane speed => decided to setup a 10GE backbone with 2x Arista 7148SX (92x 10GE ports wirespeed)
  - Currently driving all the WN - Disk - Tape LAN traffic.
- Now running out of 10GE slots => New Nexus 7009 (336 10GE ports wirespeed)



Network tuning

PIC
port d'informació
científica

Jumbo frames: improvements of 30-40% in throughput seen after setting MTU to 9000 bytes.

Kernel parameters: It has been found that the default values do not fulfil HTC environment requirements.

- TCP max buffer size: default of 16 KB increased to 8 MB.
- Max backlog (# of unprocessed packets before kernel drops): default is 1000, increased to 25000.
- Transmit queue length: increased to 20000.

Network congestion is not rare in HTC environments. Recently switched from BIC to HTCP. Improvement seems significant (not quantified yet).

Summary

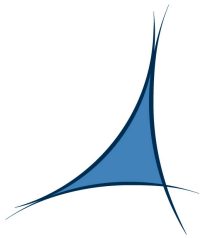
PIC
port d'informació
científica

Some of the main characteristics of the PIC Tier1 High Throughput Computing centre have been presented.

Tuning the balance between the Computing and Disk Storage components, as well as the network fabric in between, is key.

Running the systems with the highest possible efficiency is one of the main goals.

- For the CPU service, accounting figures are routinely at >~ 90%
- The Disk systems seen to run at lower load vs. full capacity.
  - Future work to try and diagnose this in detail and find if there is room for optimisation.



PIC
port d'informació
científica

Thank you

PIC is maintained through a collaboration between the Generalitat de Catalunya, CIEMAT, IFAE and the Universitat Autònoma de Barcelona.
This work was supported in part by grant FPA2007-66152-C02-01/02 and FPA2010-21816-C02-01/02
from the Ministerio de Educación y Ciencia, Spain.
We would like to specially thank the dCache team in Fermilab and Desy, and the Enstore team in Fermilab for their hard work, support and co-operation.

Tape system: ENSTORE

PIC
port d'informació
científica

Provided by FNAL. Excellent support.

2 types of libraries:

- IBM TS3500 (2015 slots)
- Oracle/STK SL8500 (6632 slots)

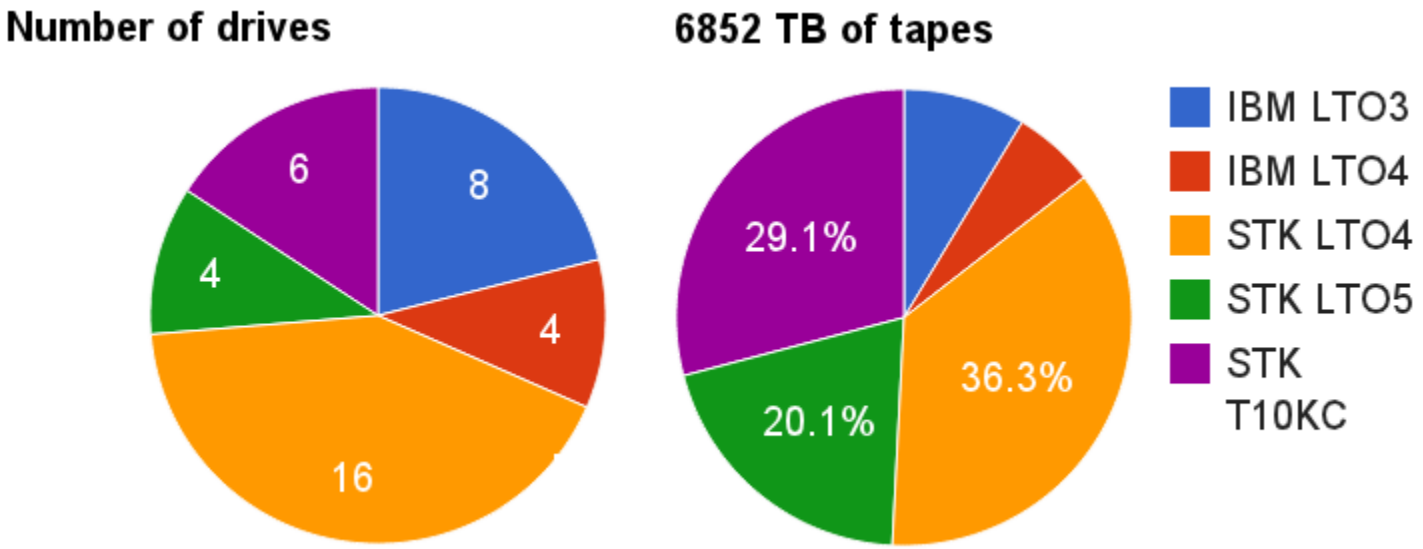
Tape servers: 8x Dell R710 with 2x FC HBA = 4Gbps FC ports per server

- 4 drives per server (power footprint reduced)
- Each server has different tape technologies: minimise the impact of one server failure + load balancing
- 1x 10GE network per server (T10KC drives, up to 240 MB/s)
- 32 GB RAM (12 GB system + 5 GB per mover process)

Tapes and drives

PIC
port d'informació
científica

6872 TB of tape capacity in LTO3, LTO4, LTO5 and T10KC technologies.



ENSTORE features

PIC
port d'informació
científica

Data distribution on tapes (*file_family_width*): max. number of streams used to migrate some given data to tape.

- LTO: used to have FFW = 3
- T10KC: FFW = 1 (higher capacity tapes and faster drives)

Disk servers contention on disk-to-tape copies (*discipline*): limits number of concurrent accesses to tape drives from a disk server (=1)

Minimise tape mount/unmount overhead:

- Tape will stay mounted for few minutes if there are no new requests.
- Ordering of requests according to tape.
- (dCache) min. amount of data to trigger a migration (=20 GB)