First tests with Tier-3 facility for the ATLAS experiment at IFIC(Valencia)

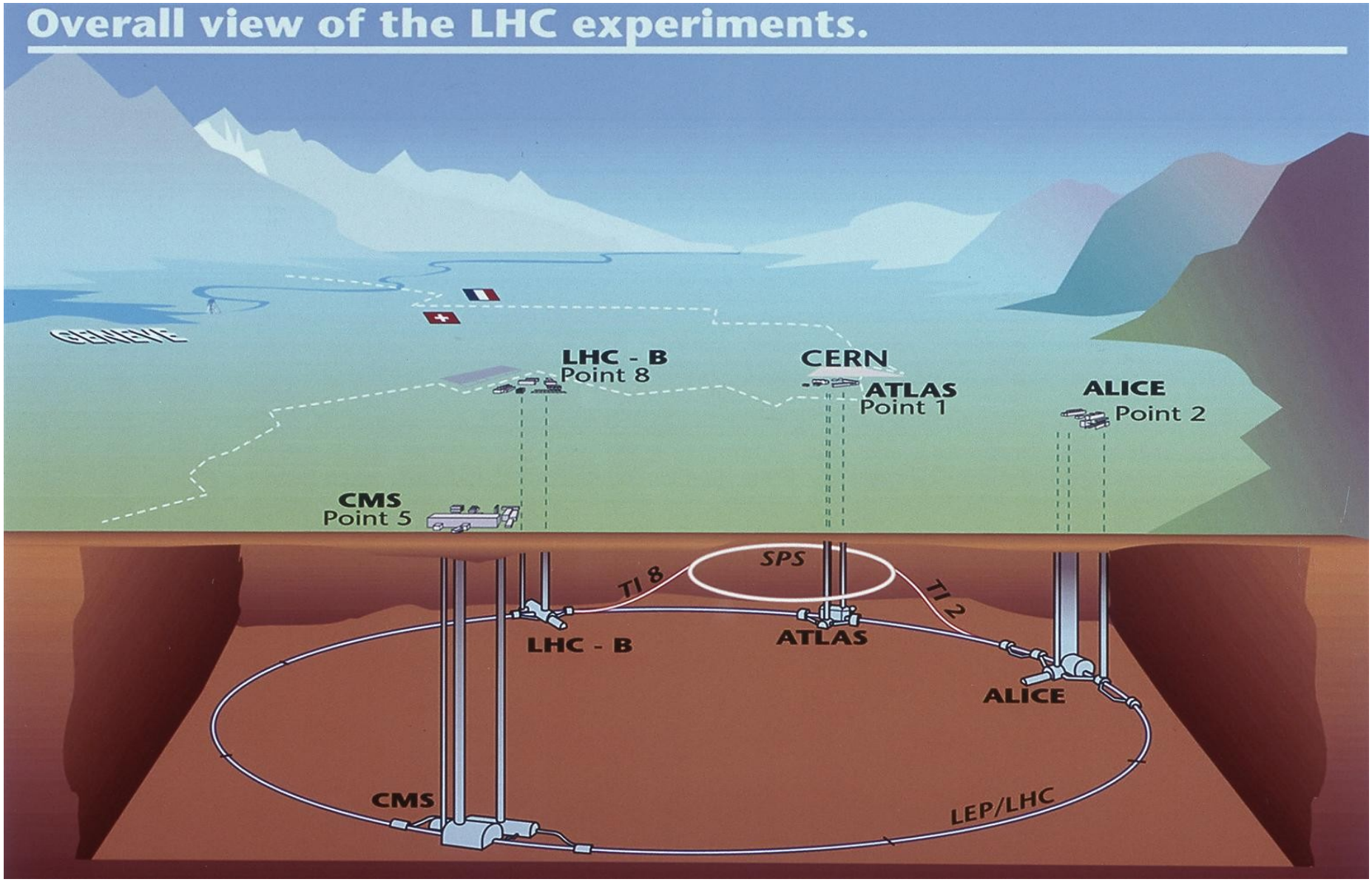
M.Villaplana , S.González de la Hoz, E.Oliver, J.Salt, J.Sánchez

IFIC - Instituto de Física Corpuscular
Valencia (Spain)

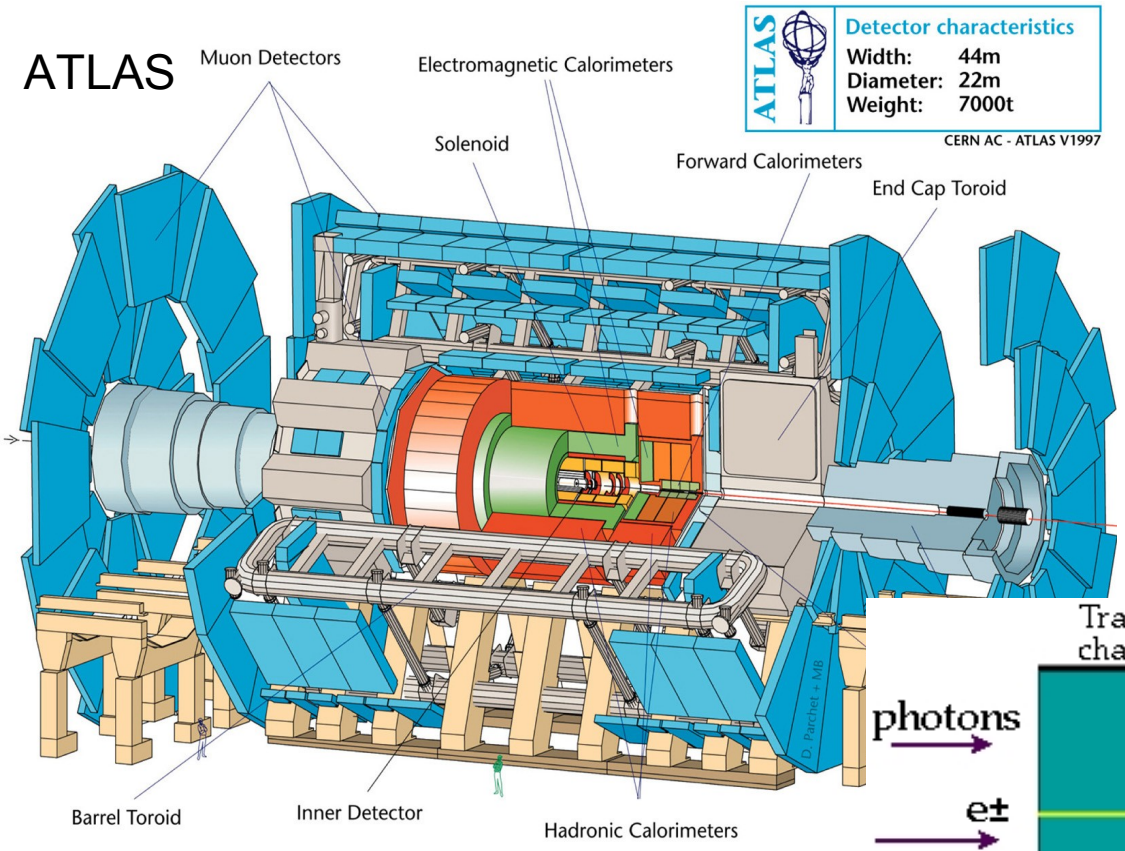
Outline:

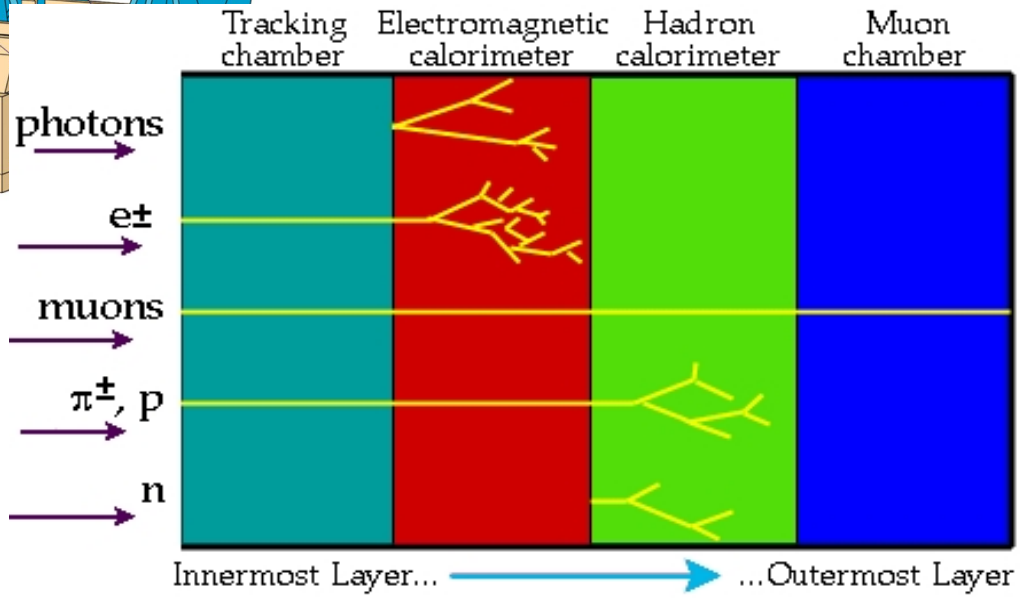
- Introduction
  - LHC, ATLAS
  - Event Data Model
  - Computing Model
- Tier-3
  - Atlas T3 Taskforce
  - T3 at IFIC
- Performance Tests

Overall view of the LHC experiments.

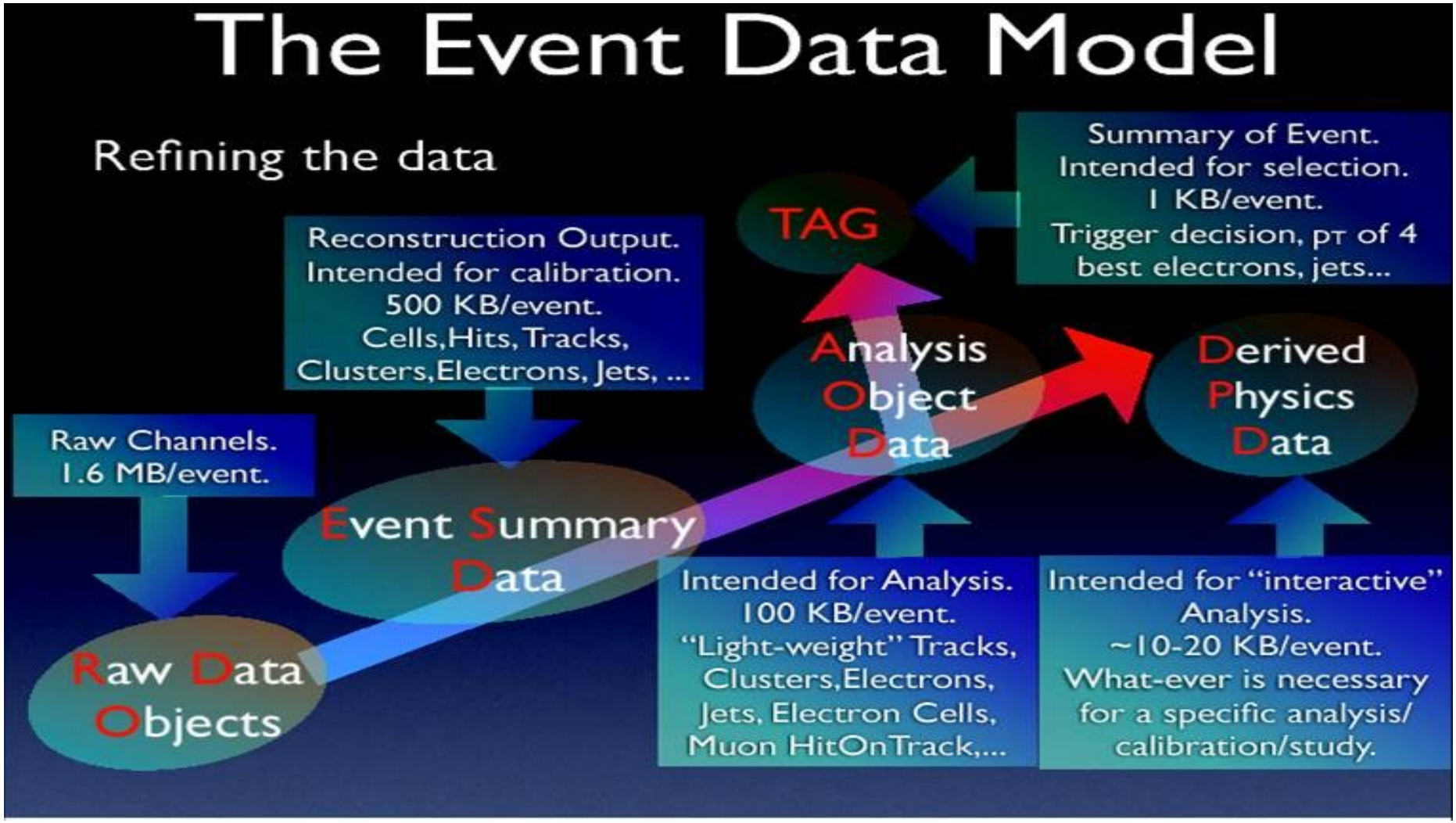


ATLAS


Detector characteristics
Width: 44m
Diameter: 22m
Weight: 7000t
CERN AC - ATLAS V1997
Muon Detectors
Electromagnetic Calorimeters
Solenoid
Forward Calorimeters
End Cap Toroid
Barrel Toroid
Inner Detector
Hadronic Calorimeters


Tracking chamber
Electromagnetic calorimeter
Hadron calorimeter
Muon chamber
photons
e±
muons
π±, p
n
Innermost Layer...
...Outermost Layer

- **Event rate 40MHz, interesting events 100-1000Hz**
- **Raw data will be transferred to the Tier-0 input buffer at 320 MB/s (average)**



The Computing Model

- Resources Spread Around the GRID

•Interactive Analysis
•Plots, Fits, Toy MC, Studies, ...

Tier 3

•Reprocessing of full data with improved calibrations 2 months after data taking.
•Managed Tape Access: RAW, ESD
•Disk Access: AOD, fraction of ESD

DPD

30 Sites Worldwide
Tier 2

•Derive 1st pass calibrations within 24 hours.
•Reconstruct rest of the data keeping up with data taking.

AOD

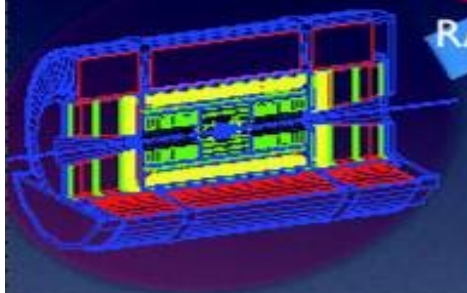
Tier 1

10 Sites Worldwide

RAW/ AOD/ ESD

Tier 0

RAW



CERN Analysis Facility

•Production of simulated events.
•User Analysis: 12 CPU/ Analyzer
•Disk Store: AOD

•Primary purpose: calibrations
•Small subset of collaboration will have access to full ESD.
•Limited Access to RAW Data.

Tier-3s are non-ATLAS funded or controlled centers

It is up to the different institutions to propose possible Tier-3 configurations

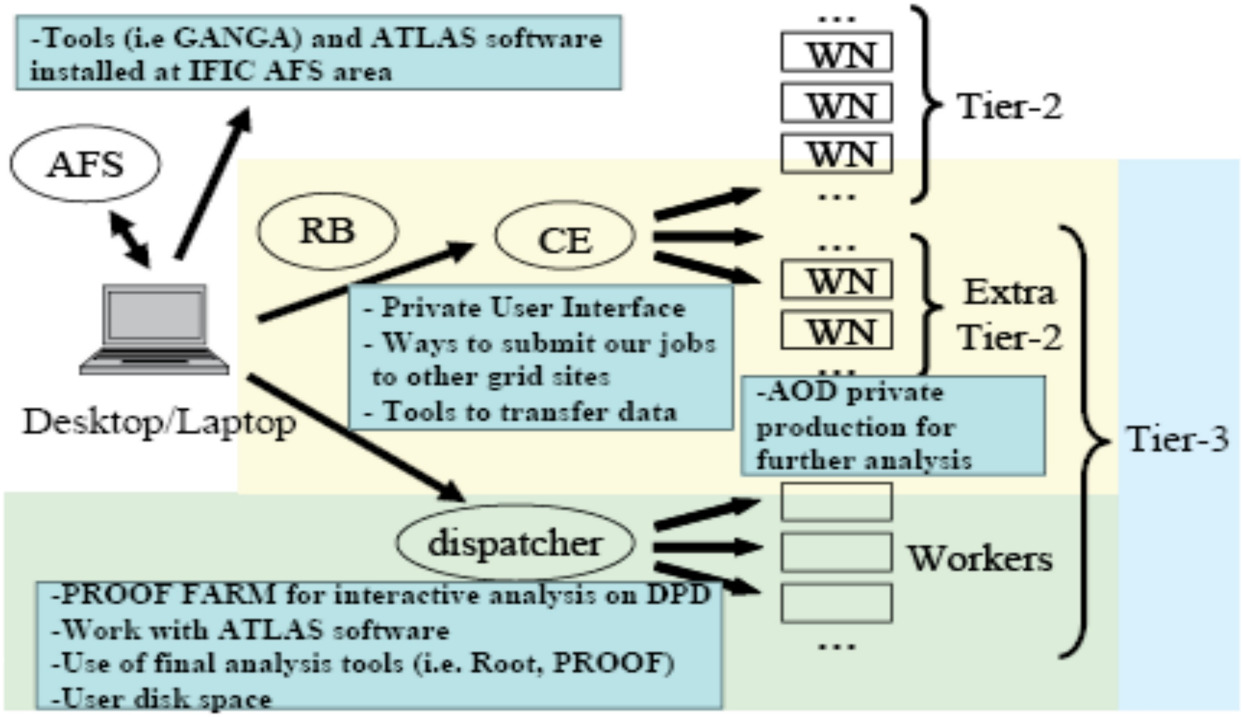
ATLAS Tier-3 Taskforce
https://twiki.cern.ch/twiki/bin/view/Atlas/AtlasTier3

→ **Try to converge the various existing Tier-3 prototypes on a small number of models**

- **Document the current usage in Atlas Tier-2 and Tier-3 sites**
- **Determine and make available best practices guidelines**
- **Develop suggestion for deployment at all Tier-3 sites**
- **Propose test metrics for the considered design and tabulate the results**

→ **Main Goal: Provide a document + some twiki pages with installation recommendations in a Tier-3**

The Tier-3 at IFIC



- IFIC's Tier-3 is attached to a Tier-2 that has 50% of the Spanish Federated Tier-2 resources

- Tier-3 resources are split into two parts
  - Some resources are coupled to IFIC Tier-2 in a GRID environment
  - A computer farm to perform interactive analysis outside the GRID framework

Resources coupled to Tier-2

Tier-2 Resources

- Storage:
  - SUN X4500 and X4540 → 316 TB
- Connectivity:
  - Switch Cisco 6509
  - 10 Gbit to backbone
  - 1 Gbit to worker nodes and disk servers
- Lustre v1.8 (in hardware with iSCSI + HA)
- One metadata server (MDS) Lustre server with redundancy RAID1.
- Disk servers aggegated using linux (RHEL5) + Lustre + RAID5(software)
- The 48 disks are distributed into 6 OSTs
- Every OST has 8 disks but one that has 6 (2 disks for the System OS(RAID1))

Tier-3 Resources

- Around 100 TB → 60 TB under DDM control + 40 TB under IFIC control
- Space token dedicated to Tier-3 → ATLASLOCALGROUPDISK
  - To manage local users' data.
  - It has an area on a SE but points to non-pledged space

Interactive analysis: PROOF Farm

The ROOT system provides a set of OO frameworks with all the functionality needed to handle and analyze large amounts of data.

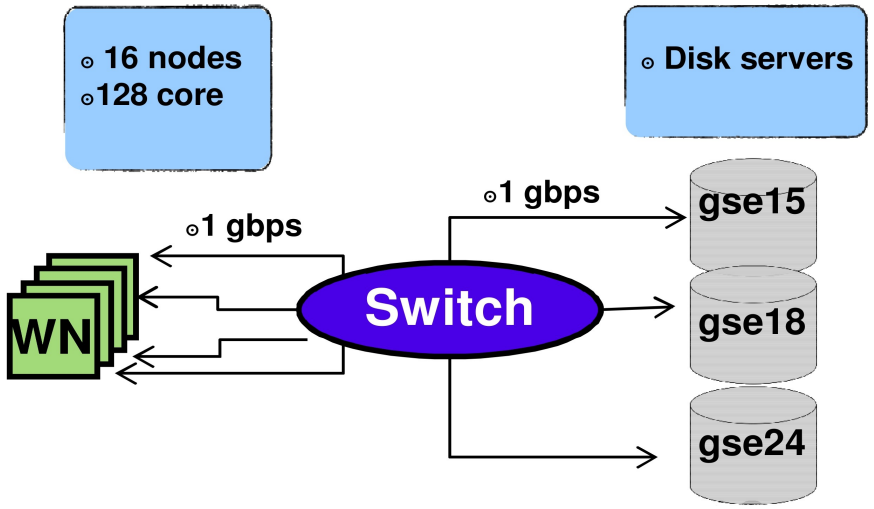
http://root.cern.ch/drupal/

The Parallel ROOT Facility, PROOF enables interactive analysis of large sets of ROOT files in parallel on clusters of computers or many-core machines.

PROOF Farm at IFIC

- 3 disk servers dedicated exclusively to Tier-3 to avoid overlap with Tier-2

- The only shared resource between Tier-2 and Tier-3 is the Lustre metadirectory server (MDS)

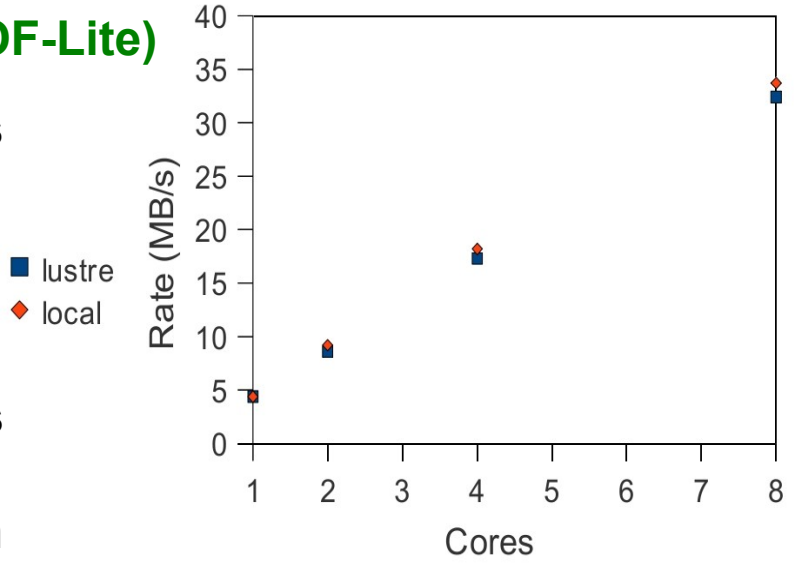


- 128 cores (16 nodes):
  - 16 x HP BL460c, 8 cores, 2 x Intel Xeon E5420@2.5 Ghz
  - 16 GB RAM
  - 2 HD SAS 146 GB (15000 rpm)

PROOF Farm: Performance tests

a) Test using one machine with 8 cores (PROOF-Lite)

- 3684500 events (7675.24 MB), 372 files (22MB per file)
- Data stored locally and on Lustre file system

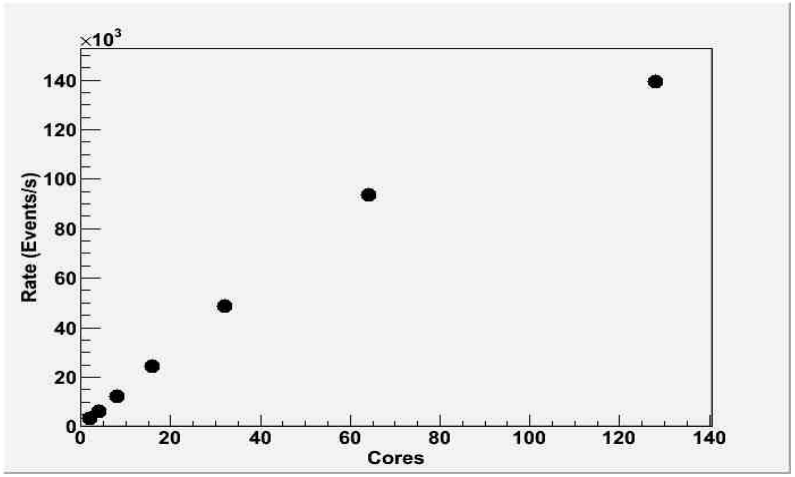


- CPU more important than i/o
- Lustre had a nearly equivalent behavior as local storage
- Only when 8 cores were busy reading from Lustre started to slightly deviate from linearity

b) Test on a cluster of machines

- 1440 files (32GB) running on 128 cores
- With 128 cores we start loosing linearity

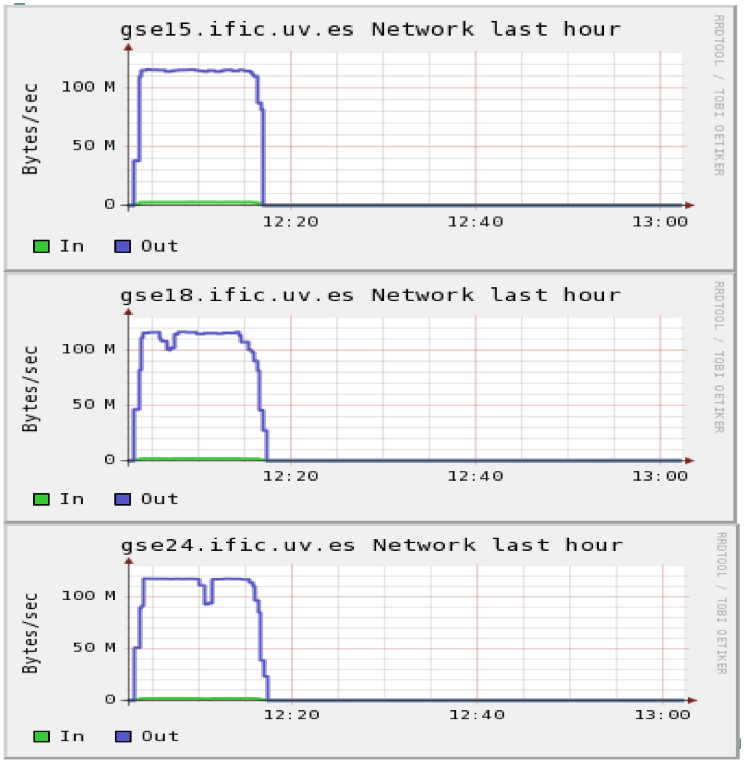
- The following tests showed that we are limited by our disk server interface



PROOF Farm: Performance tests

c) Sequential read test

- Made each of the 128 cores read 100 random files with dd (bs=32k).
- A total of 10995 files (225 GB) were used.
- Test showed a Bandwidth = 357 MB/s and that the disk server interfaces were saturated.
- Bandwidth values obtained from the switch CISCO X6509 counters (5 minute intervals) using CACTI



PROOF Farm: Performance tests

d) Test using 4 simultaneous PROOF sessions

1 PROOF session: 3684500 events (372 files, 7 GB)

N	Init(s)	Elapsed(s)	Rate(evts/s)	Rate(MB/s)
128	2.5	36	101634.4	228.3

4 simultaneous PROOF sessions:

Each PROOF session run the same analysis but reading from a diferent copy

N	Init(s)	Elapsed(s)	Rate(evts/s)	Rate(MB/s)
128	6.0	2:38	23234.3	53.8
128	8.1	2:39	23133.0	53.8
128	8.1	2:36	23530.9	54.4
128	7.3	2:37	23362.0	54.7
Total			93260.2	216.7

Conclusions

- The Tier-3 at IFIC-Valencia is no longer a prototype but a real working facility with around 20 users

- The design might change in the future acording to users needs.

- As for now, performance tests have shown good PROOF behaviour. The farm has shown correct scalability and concurrent use is possible without added degradation

- In adition to this, Lustre performance is adequate and no sensible degradation has been observed while concurrent access is made. Even though Lustre performance is limited by disk server's ethernet interface there is room still open to improvement aggregating a second interface (channel bonding)

Thank you