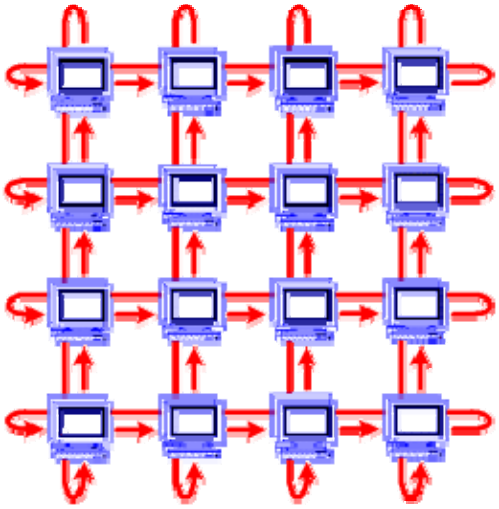
Status of the ALICE CERN Analysis Facility



Marco MEONI – CERN/ALICE
Jan Fiete GROSSE-OETRINGHAUS - CERN /ALICE
CHEP 2009 - Prague

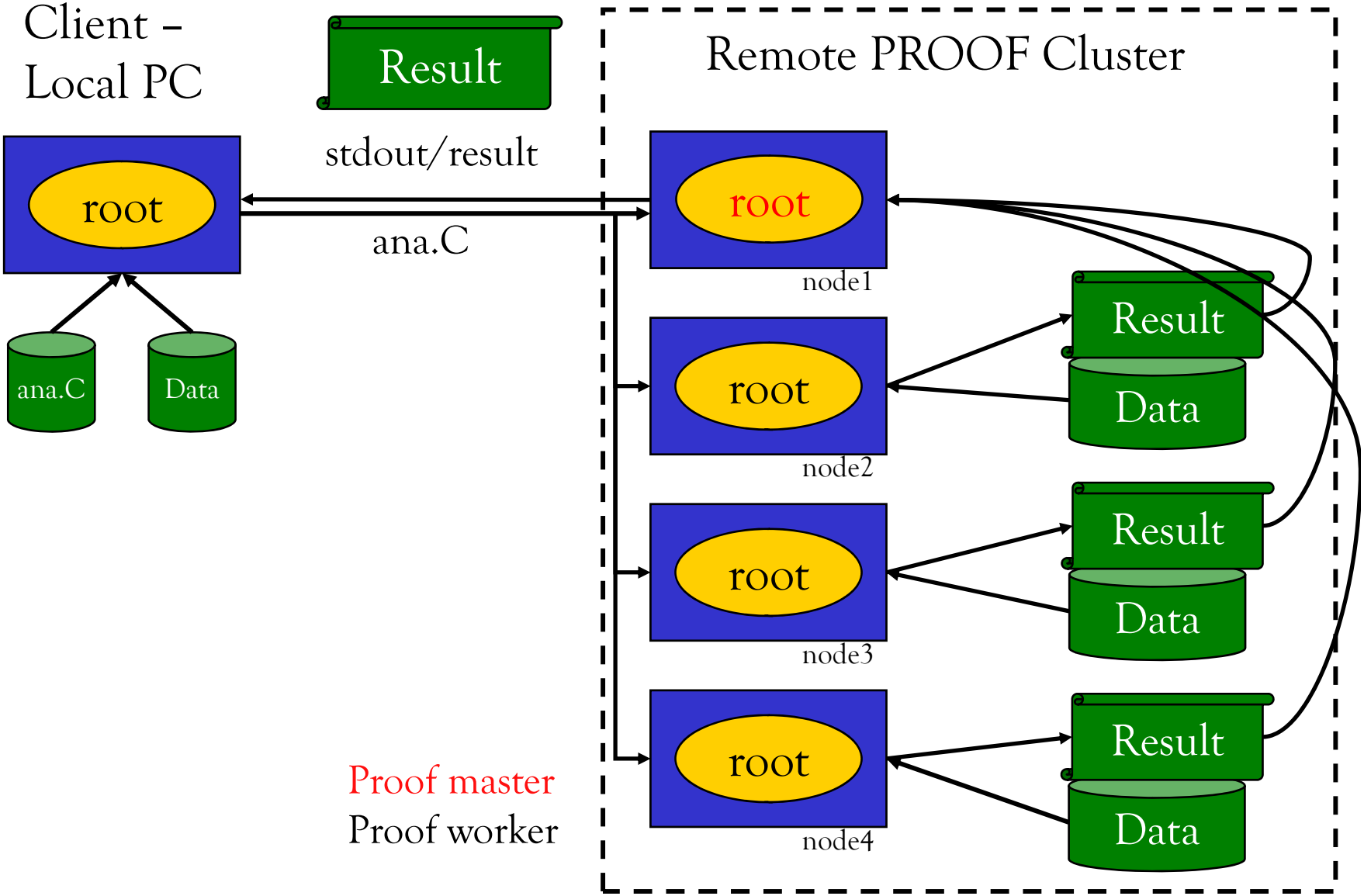
Introduction

- The ALICE experiment offers to its users a cluster for quick interactive parallel data processing
  - Prompt and pilot analysis
  - Calibration/Alignment
  - Fast Simulation and Reconstruction
- The cluster is called CERN Analysis Facility (CAF)
- The software in use is PROOF (Parallel ROOT Facility)
- CAF is operational since May 2006

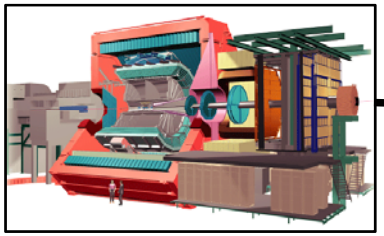
Outline

- PROOF
  - Schema
  - PROOF@CAF
- CAF Usage
  - Users and Groups
  - CPU Fairshare
  - File Staging and Disk Quota
  - Resource Monitoring
- Ongoing Development
  - PROOF on the Grid
- Outlook and Conclusions

PROOF Schema



CERN Analysis Facility (CAF)



ALICE

Grid SEs

CASTOR MSS

Direct Access

Staging

HW since Sep '08	
Arch.	AMD 64
CPU	15 x 8-core (IT standard)
Disk	15 x 2.33 TB RAID5
Workers	2/node
Mperf	8570

CAF

GSI

LDAP

Authentication

Proof master, xrootd redirector

MonALISA
MONitoring Agents using a Large Integrated Services Architecture

Proof worker
xrootd disk server

local disk

...

Proof worker
xrootd disk server

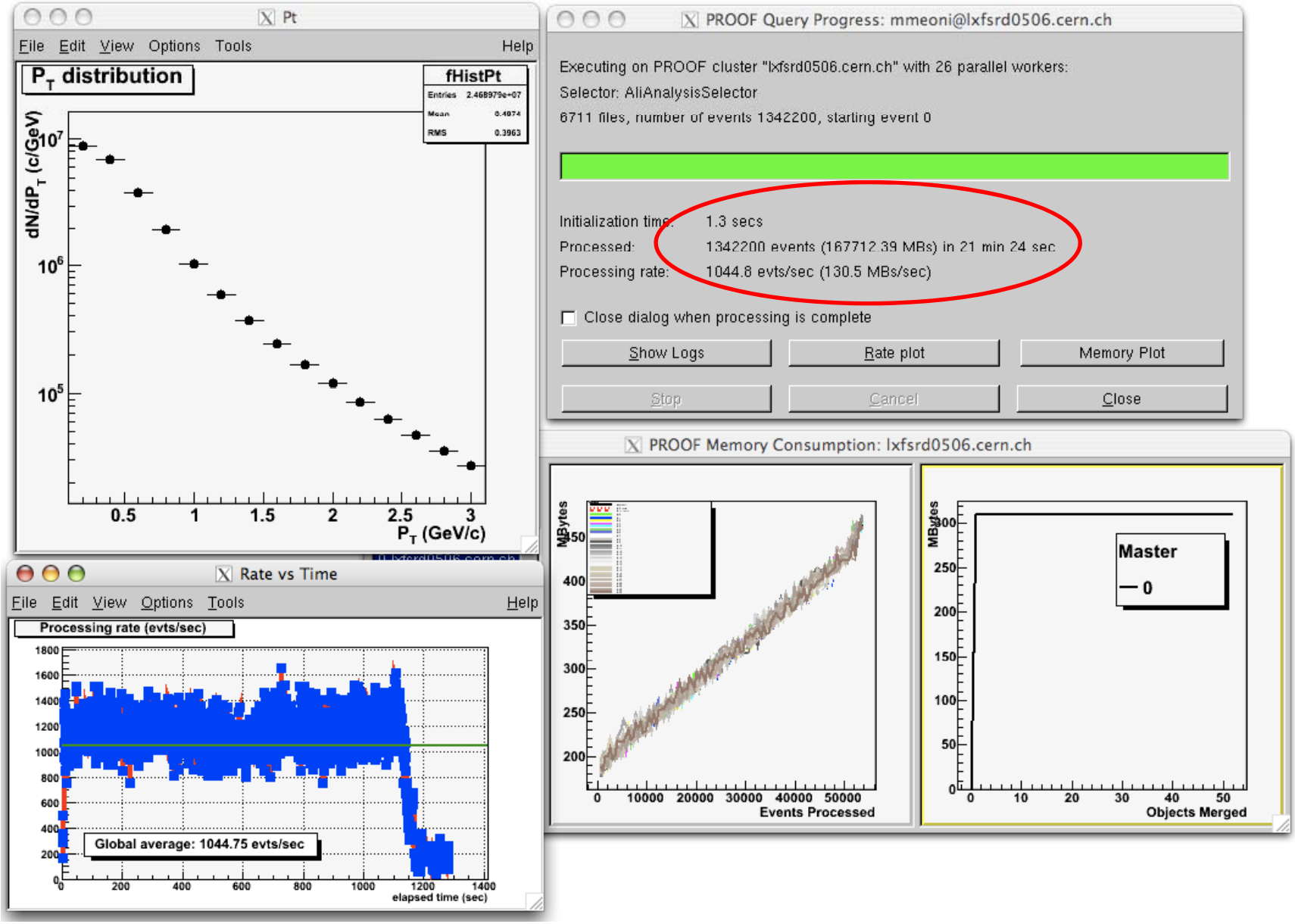
local disk

Access to local disks → Advantage of processing local data

CAF SW Components

- ROOT, Scalla sw Suite (Xrootd), MonALISA
- NEW! Transition from olbd to cmsd
  - Cluster Management Service Daemon
  - Provides dynamic load balancing of files and data name-space
  - ALICE file stager plugged into cmsd
- NEW! GSI (Globus Security Infrastructure) authentication
  - Uses X509 certificates and LDAP based configuration management
  - Same mean of authentication for Grid and CAF
  - Grid files can be directly accessed
    - Fast parallel reconstruction of raw data (see talk #457 from C. Cheshkov)

An Example


Pt
File Edit View Options Tools Help
P_T distribution
fHistPt
Entries 2.468979e+07
Mean 0.4974
RMS 0.3963
dN/dP_T (c/GeV)
P_T (GeV/c)
PROOF Query Progress: mmeoni@lxfsrd0506.cern.ch
Executing on PROOF cluster "lxfsrd0506.cern.ch" with 26 parallel workers:
Selector: AliAnalysisSelector
6711 files, number of events 1342200, starting event 0
Initialization time: 1.3 secs
Processed: 1342200 events (167712.39 MBs) in 21 min 24 sec
Processing rate: 1044.8 evts/sec (130.5 MBs/sec)
Close dialog when processing is complete
Show Logs
Rate plot
Memory Plot
Stop
Cancel
Close
PROOF Memory Consumption: lxfsrd0506.cern.ch
MBytes
Events Processed
Master
0
Objects Merged
Rate vs Time
File Edit View Options Tools Help
Processing rate (evts/sec)
Global average: 1044.75 evts/sec
elapsed time (sec)

CAF Users

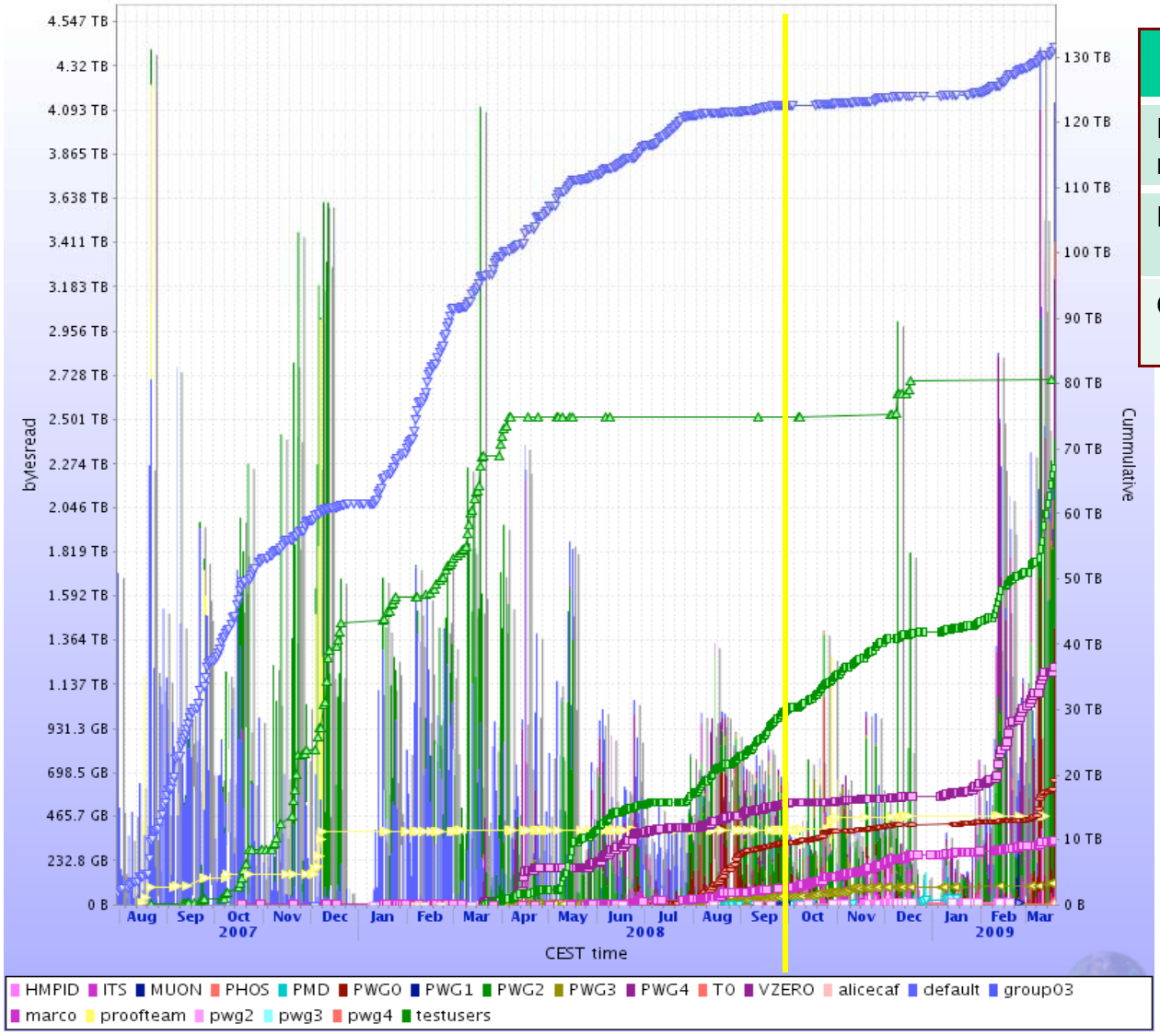
- Available disks and CPUs must be fairly used
- Users
  - are grouped into sub-detectors and physics working groups (PWG)
  - can belong to several groups
- Groups
  - have a disk space (quota) which is used to stage datasets from the Grid
  - have a CPU fairshare target (priority) to regulate concurrent queries

Groups	#Users
PWG0	22
PWG1	3
PWG2	39
PWG3	19
PWG4	31
13 SUB-DETECTORS	35

-19 groups
-111 unique users

- Continuative history of CAF since May '06 shown on the MonALISA-based web repository
  (see poster #461 from C.Grigoras)
- Peak of 23 concurrent users

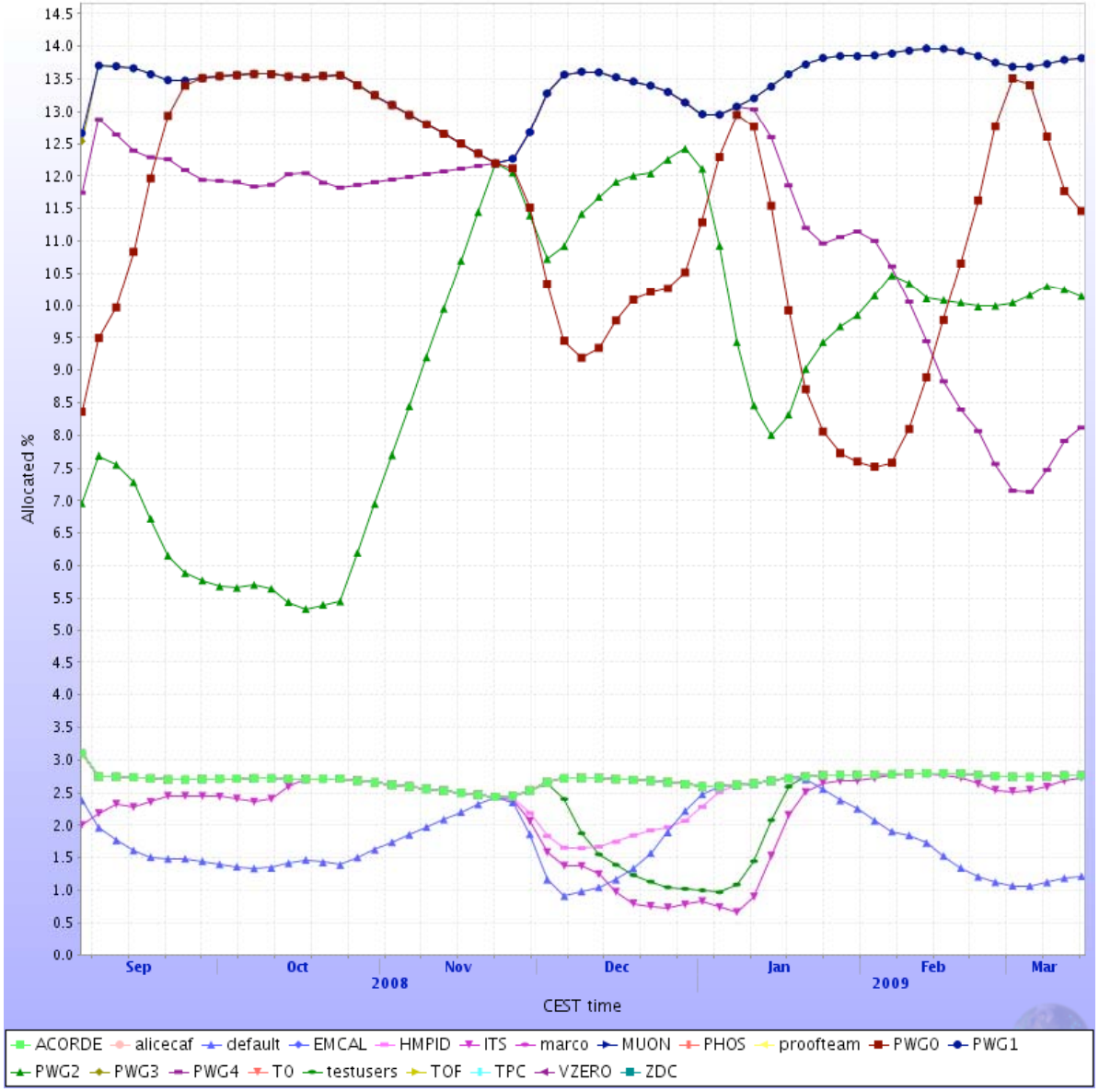
Data Processed

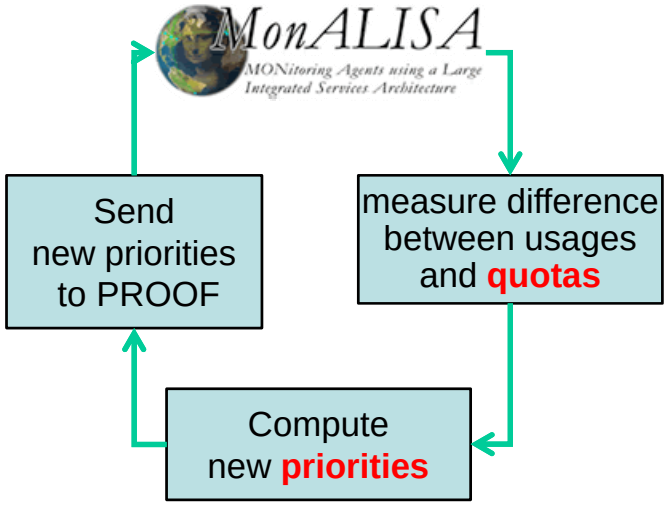


	CAF1	CAF2	+/-
Bytes read	266TB (14m)	83.3TB (5.5m)	-21%
Events	4.17G (14m)	1.95G (5.5m)	+17%
Queries	9000 (2.5m)	10600 (5.5m)	-47%

CAF1 CAF2

CPU Fairshare



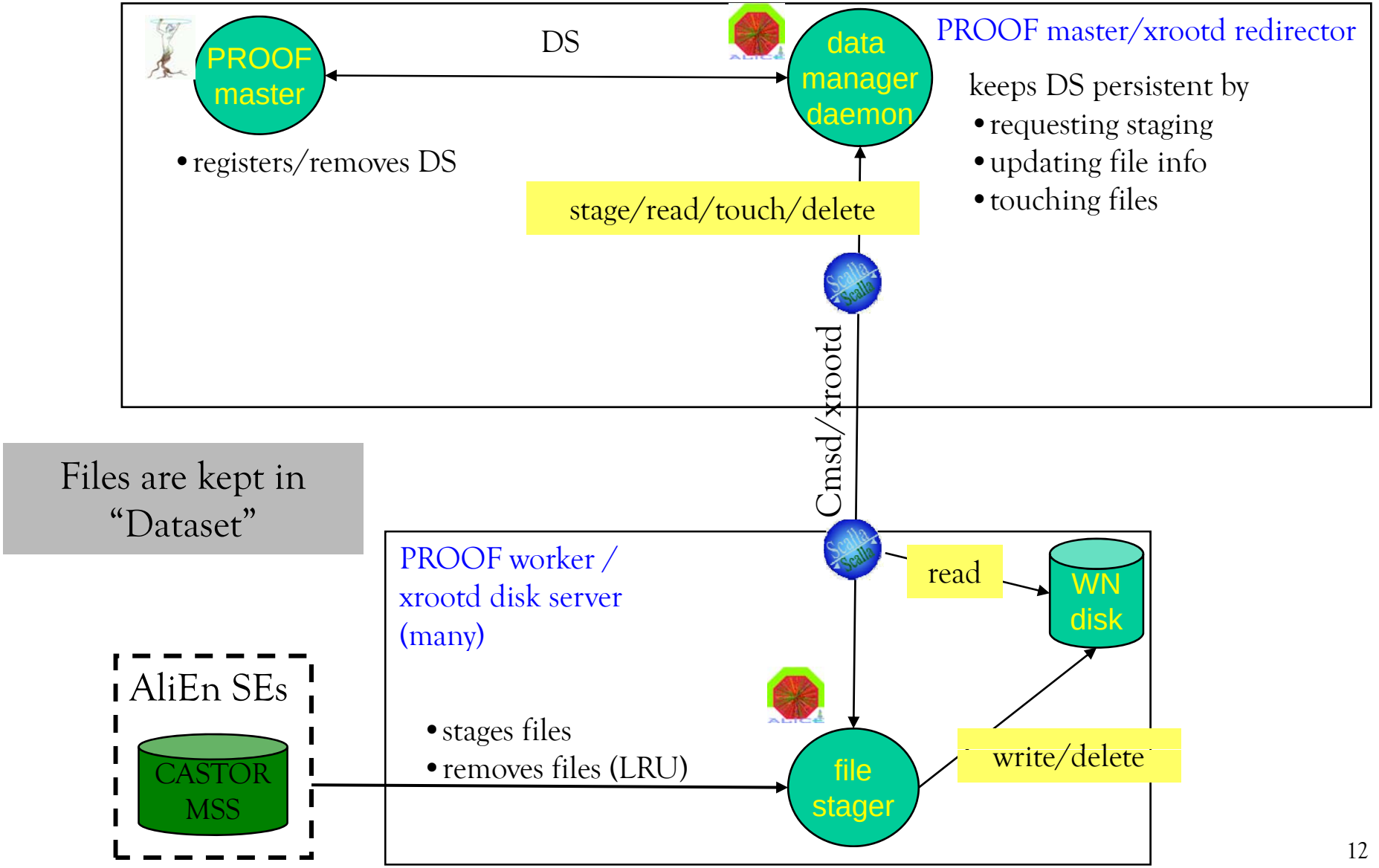


Default group quotas:
- Detectors: x
- PWGs: 5x

Dataset

- Datasets
  - are used to stage files from the Grid
  - are lists of files registered by users for processing with PROOF
  - may share same physical files
  - allow to keep file information consistent
  - files are uniformly distributed by the xrootd data manager
- The DS manager
  - takes care of the disk quotas at file level
  - sends monitoring info to MonALISA:
    - The overall number of files
    - Number of new, touched, disappeared, corrupted files
    - Staging requests
    - Disk utilization for each user and for each group
    - Number of files on each node and total size

Dataset Staging



Resource Monitoring

- A full CAF status table is available on the MonALISA repository

	Machine status					Storage	CPU		Memory		Networking		Hosted files	
Machine	Online	xrootd	cmsd	Load	Proof users	Staging (%)	usr	sys	Used	Free	IN	OUT	Files	Size
1. lxfsrd0506				0.15	3	4	0.848	0.407	2.453 GB	13.21 GB	6.322 Kbps	9.027 Kbps	-	-
2. lxfsrd0507				2.05	4	11	13.23	2.21	1.88 GB	13.78 GB	315.3 Kbps	7.259 Mbps	4285	241.9 GB
3. lxfsrd0508				1.05	4	10	21.13	1.097	2.205 GB	13.46 GB	2.823 Mbps	159.5 Kbps	4141	231.6 GB
4. lxfsrd0509				1.75	4	11	13.06	2.25	2.17 GB	13.49 GB	309 Kbps	6.858 Mbps	4353	242.6 GB
5. lxfsrd0510				2.92	6	10	24.47	1.526	2.239 GB	13.42 GB	1.417 Mbps	2.701 Mbps	4244	232.8 GB
6. lxfsrd0513				3.12	6	11	25.62	1.984	2.219 GB	13.44 GB	968.2 Kbps	2.299 Mbps	4208	233 GB
7. lxfsrd0514				2.05	4	11	19.86	1.689	2.021 GB	13.64 GB	2.646 Mbps	202.9 Kbps	4267	241.5 GB
8. lxfsrd0701				1.05	6	11	9.374	0.911	2.013 GB	13.65 GB	1.746 Mbps	92.12 Kbps	4270	240.9 GB
9. lxfsrd0702				1.75	4	11	21.09	0.898	2.059 GB	13.6 GB	2.253 Mbps	137.9 Kbps	4160	233.9 GB
10. lxfsrd0705				3.66	4	10	24.61	2.621	2.165 GB	13.5 GB	1.092 Mbps	2.992 Mbps	4126	227.3 GB
11. lxfsrd0706				0	0	30	0.015	0.064	5.893 GB	9.77 GB	48.58 Bps	76.8 Bps	13714	685.4 GB
12. lxfsrd0906				1.03	4	11	8.721	0.569	2.257 GB	13.41 GB	1.677 Mbps	82.28 Kbps	4220	236.7 GB
13. lxfsrd1101				1.75	6	11	21	1.407	2.261 GB	13.4 GB	2.379 Mbps	141 Kbps	4329	245.7 GB
14. lxfsrd1111				1.76	4	11	19.69	0.734	2.252 GB	13.41 GB	2.301 Mbps	153.6 Kbps	4211	235.4 GB
15. lxfsrd1114				0.9	4	11	9.352	0.797	2.171 GB	13.49 GB	1.515 Mbps	89.4 Kbps	4221	233.9 GB
Total	**15**									**198.7 GB**	**21.41 Mbps**		**68749**	**3.675 TB**
Average		**1**	**1**	**3.503**	**4.2**	**11.6**	**15.47**	**1.278**	**2.417 GB**	**3.2**	**1.427 Mbps**	**1.543 Mbps**	**4910**	**268.8 GB**

- Many more parameters are available
  - Staging queue, usage of root and log partitions
  - CPU nice and idle status
  - Memory consumption details
  - Number of network sockets

In Conclusion...

- CAF is operational since three years
- More than 100 users are registered and ~10 per day use CAF
- Interactive analysis with PROOF is a good addition to local analysis and batch analysis on the Grid

...but if PROOF + Grid...

PROOF on the Grid

- This is an ongoing development
- It combines PROOF and the ALICE Grid middleware (AliEn)

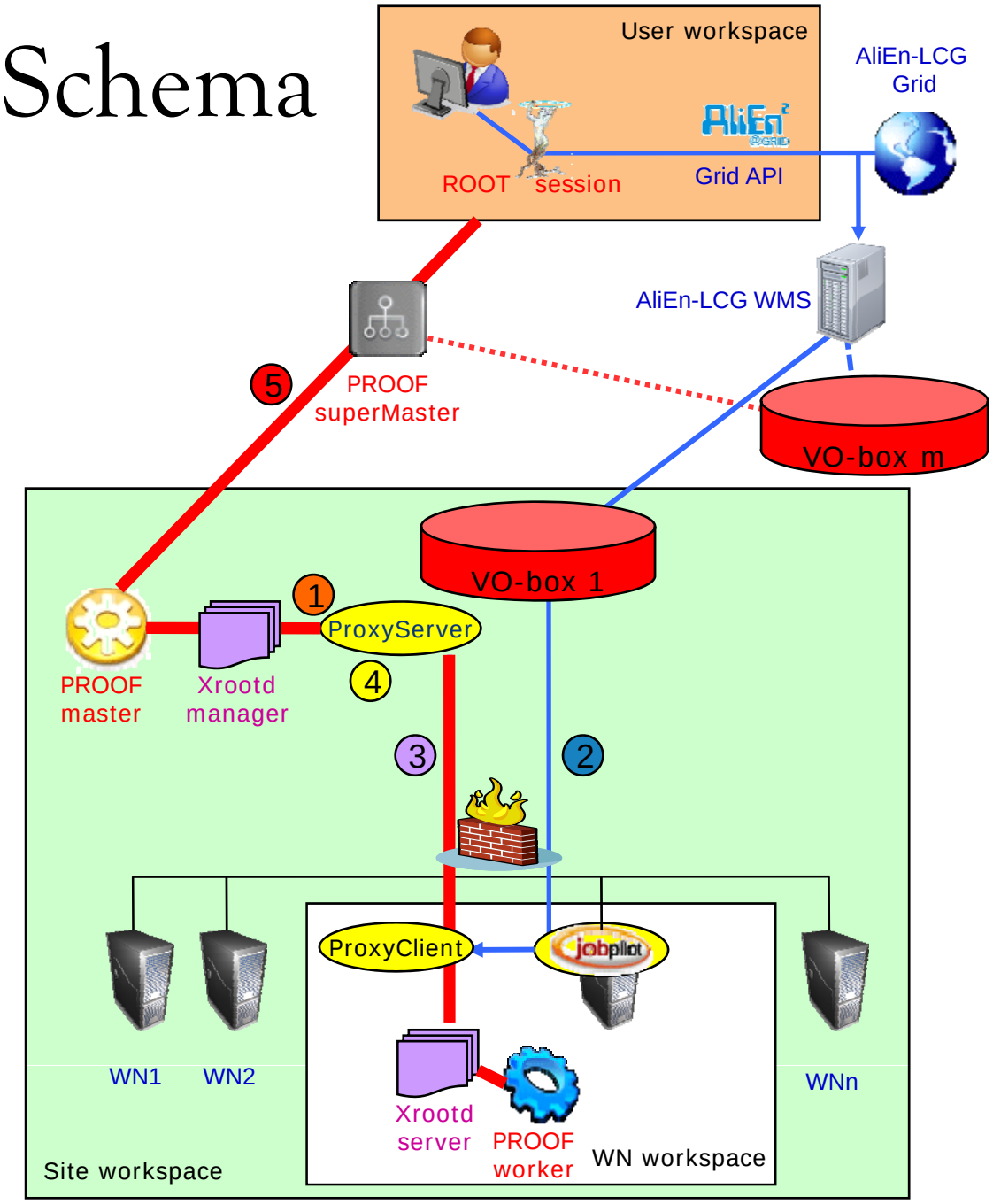
Reasons:

1. Cluster size:
   CAF can only hold a fraction of the yearly reco and sim data (~1PB)
2. Data store:
   not feasible financial and support wise in a single computing centre
3. Resources:
   The Grid provides lots of resources
4. Data location:
   bring the kB to the PB and not the PB to the kB

ToDo:

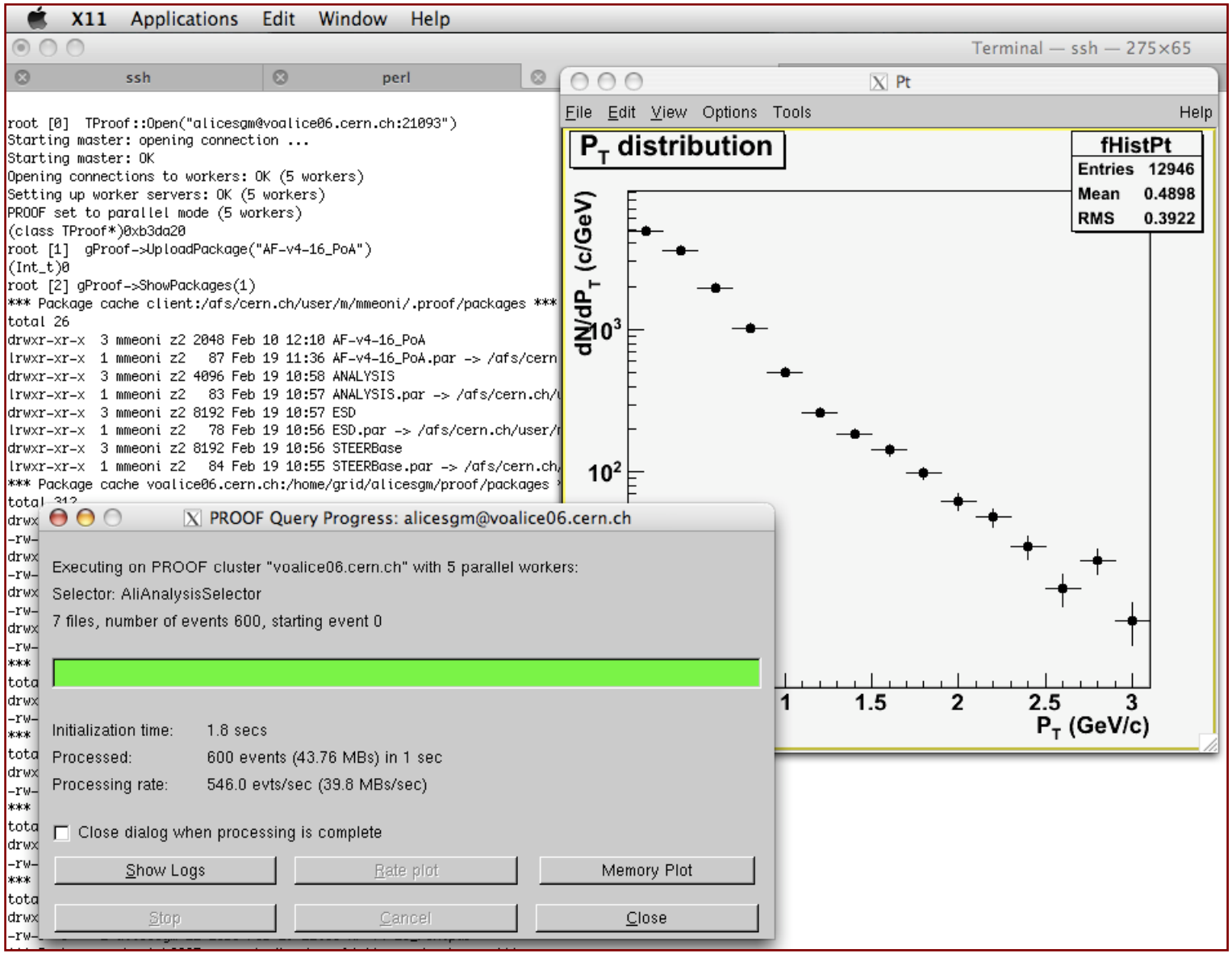
1. Cluster connectivity:
   Interconnection of Grid centres
2. Tasks and Data co-location:
   Execute tasks where data is
3. Protected access:
   WNs must connect to the Master
4. Dynamic Scheduler:
   Dynamic allocation of workers
5. Interactivity:
   Hiding of Grid latency

Schema



1. A ProxyServer service starts Xrootd and PROOF
2. Pilot Grid jobs are submitted to the Grid to start ProxyClients where user data is stored
3. A ProxyClient starts an Xrootd server and registers to the ProxyServer
4. A ProxyServer keeps the list of all the workers running at the WNs
5. A User PROOF session connects to the superMaster that, in turns, starts PROOF workers

As a PROOF of Concept



Summary

- ALICE uses PROOF on a local cluster (CAF) for quick interactive parallel processing
  - Prompt and pilot analysis
  - Calibration/Alignment
  - Fast Simulation and Reconstruction
- CAF in production since May 2006, HW and SW upgrade at the end of 2008
- Monthly tutorials at CERN (500+ users so far)
- Active collaboration with ROOT team
  - Contribution from ALICE to PROOF development
  - Implementation of dataset concept and CPU quotas
- Ongoing developments
  - Adaptation of PROOF to the Grid