SUBATECH Site Report
&
AAF Experience

Outline

- Site report :
  - Introduction
  - Site architecture
  - Events/Plans/Issues
- SAF : Subatech Analysis Facility
- Conclusion

Presentation

Nantes : West of France, south of Britanny. 80Kms to the beach, mild, oceanic climate...

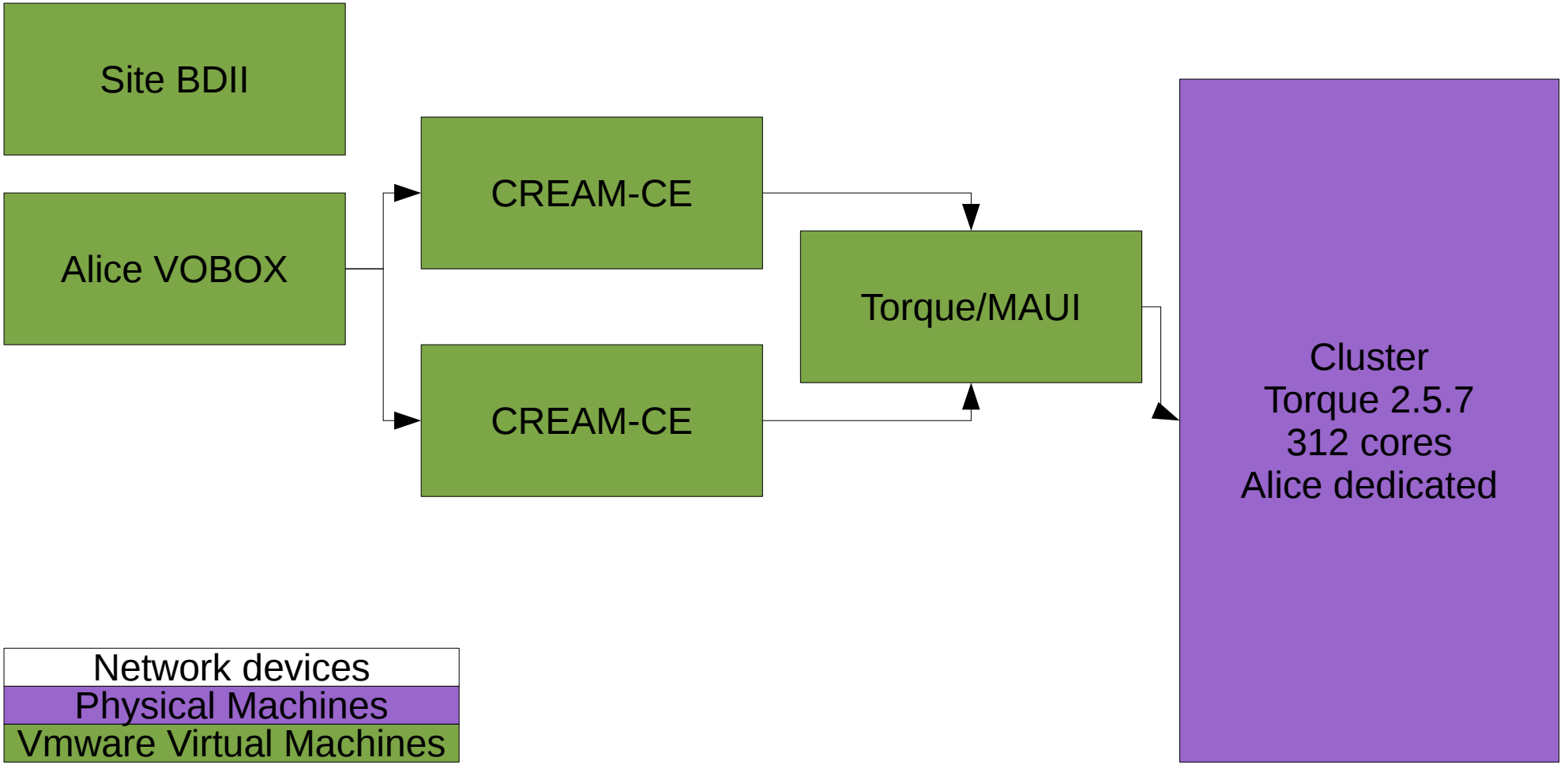
Ecole des Mines de Nantes :
- 5 teaching/research departments
- over 850 students-
- ~125 teachers/reserchers

Energy and Environment
Information Technology

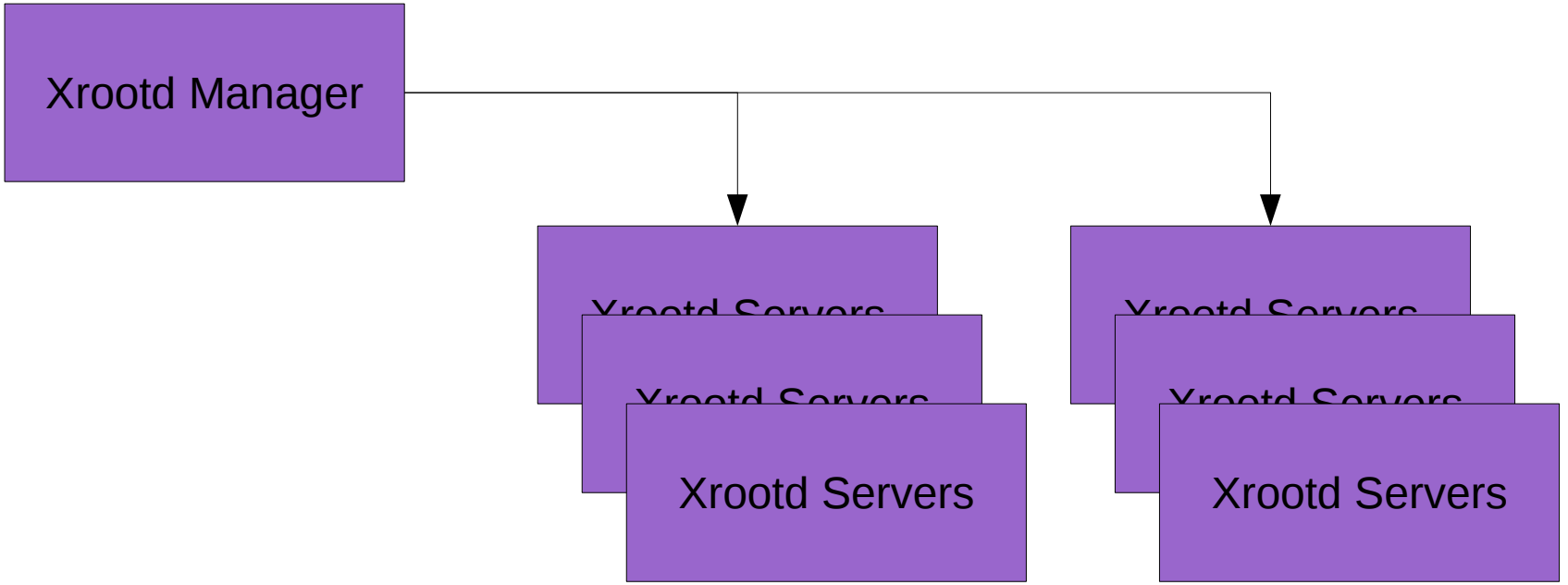
SUBATECH in short

- SUBATECH is a Joint Research Unit (UMR)
  - CNRS/IN2P3, Ecole des Mines, Nantes University
  - Over 200 staff including PHD students
  - IT service : 5 FTE (1.3 for the grid)
- Tier2 LCG-France , Alice VO only
  - http://lcg.in2p3.fr/wiki/index.php/Tier_2:Subatech

Site Architecture : Computing



Site Architecture : Storage

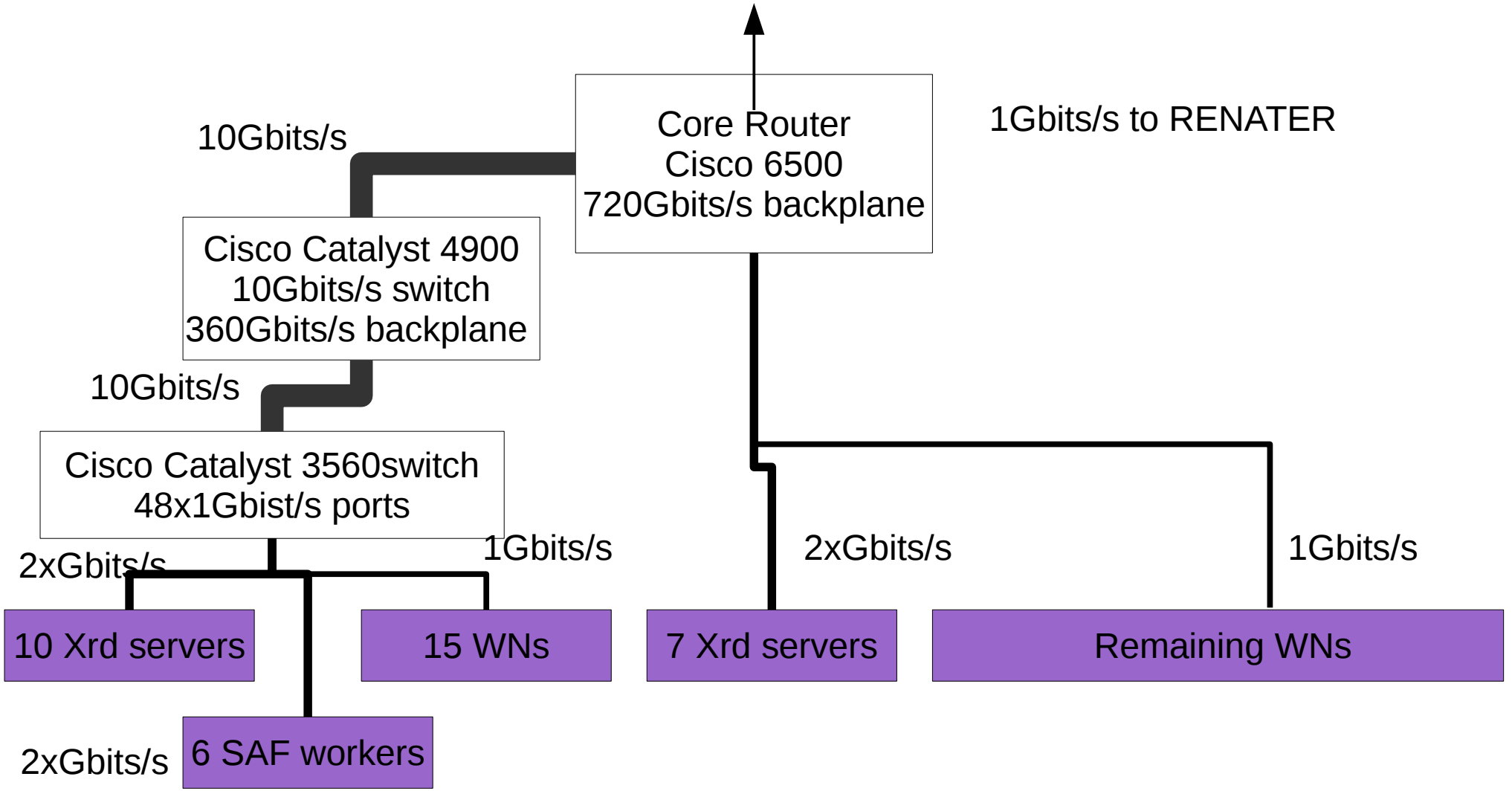


7 Combo Dell :
1 server, 2 MD1000 RAID6
24To * 7 = 168To

10 Alineos Boxes:
1 server, 1RAID6
12To * 10 = 120To

Total : ~ 270To

Site Architecture : Network



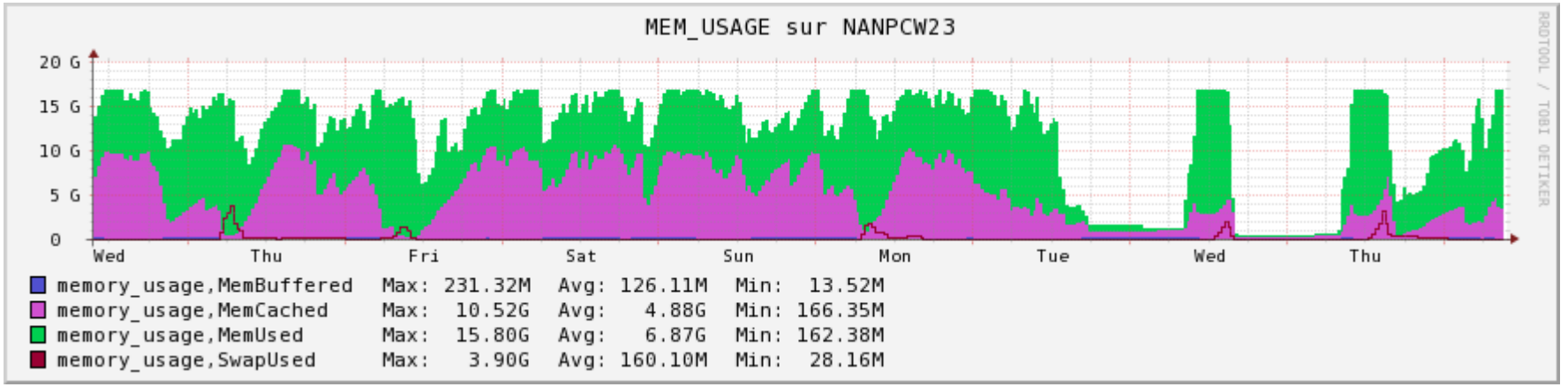
Recent events

- LCG-CE decommissionning
- Torque 2.5.7 update
  - Installed a new batch server, drain and reconfigure one CE, move worker nodes from the previous batch server to the new. Reconfigure 2nd CE when empty
- 10Gbits connection to RENATER

Current issues

- RAM usage on worker nodes
  - Deployed nagios probes on worker nodes in order to detect abnormal swap usage and make comparisons with IRFU

Stacked values

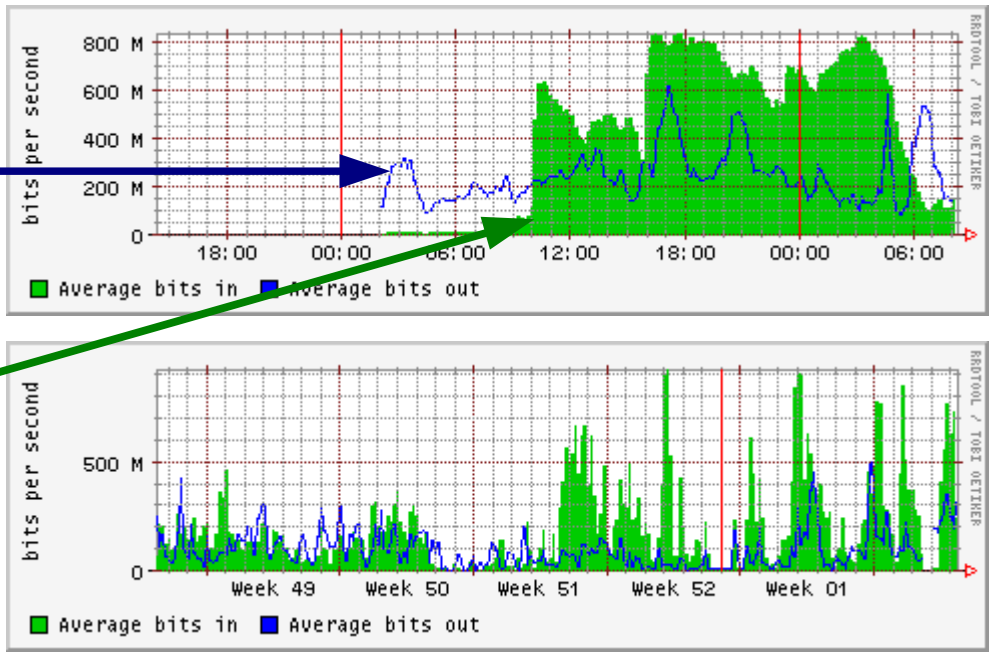


To be continued ?

Current issues (2)

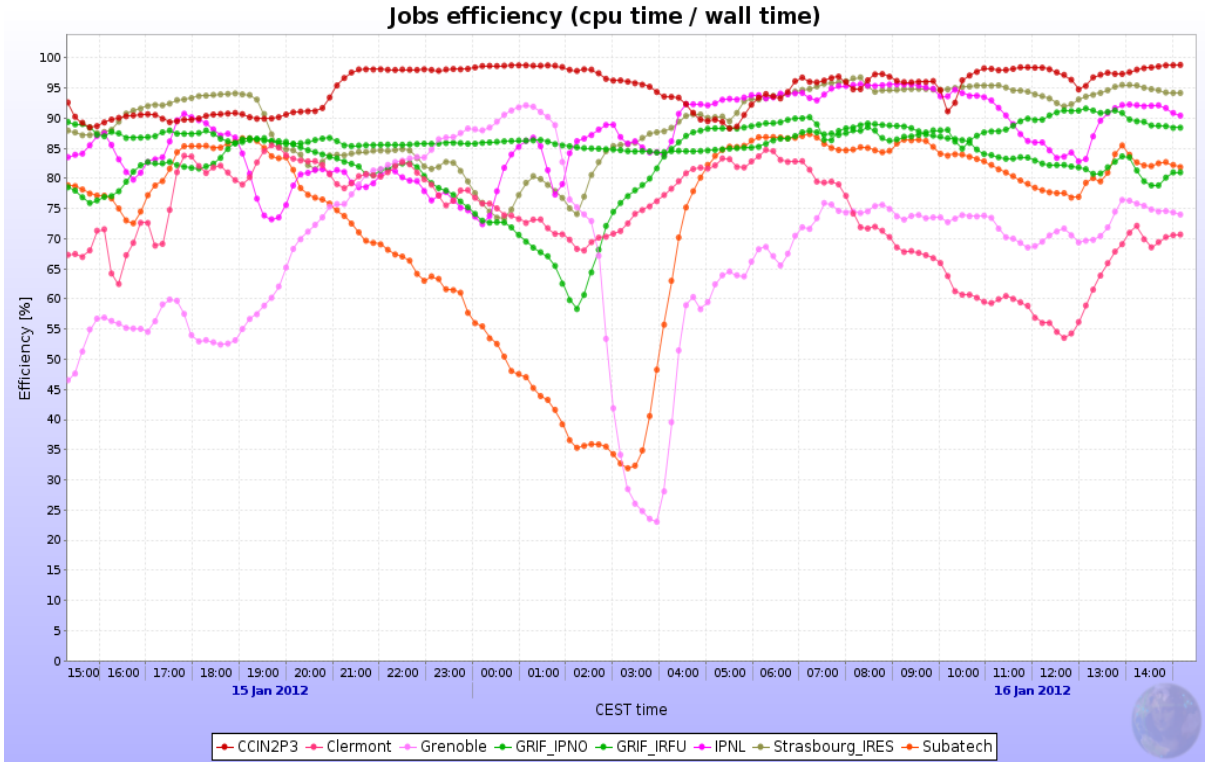
- Network : close to fill all the avail. 1Gbits/s BW

13/01/2012 : Network becomes available at 2:30 following a long cut. Storage is immediately used from outside while incomming traffic starts only after the AliEn services are restarted and jobs start in the cluster.



Other issues

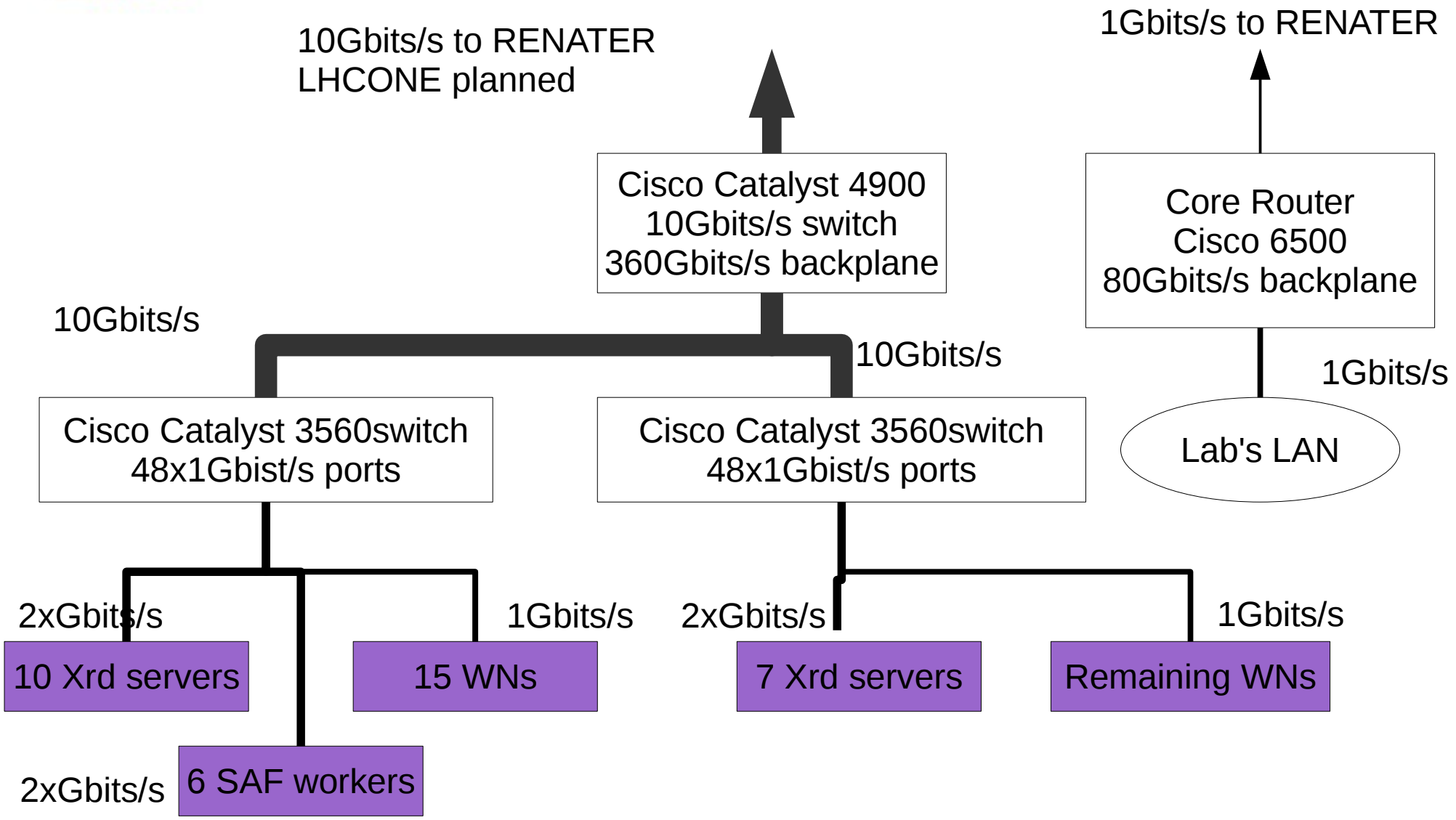
- AliEn-CE submission algorithm (minor)
- CPU efficiency can be sometimes really low !



Plans

- Install pledged storage (40To more = 310 To)
  - Scarcity of funding forced a choice between more CPU and more storage.
- Re-organize network layout
  - Grid services on a separate LAN with 10Gbit/s bandwidth to RENATER
  - May impose a short period of unavailability

New Network Arch.



System Administration : tools & methods

- Redundancy (cooling, power, servers)
- Virtualization (cloning, snapshot, backtracking)
- Monitoring : custom Nagios probes with graphs (to see trends or evolution of a parameter that led to an incident)
- Logbook (elog) to record modifications and share knowledge among team members

SAF : Subatech Analysis Facility

- History
- Hardware setup
- AAF Installation
- AAF and the sys.admin.
- Results ?

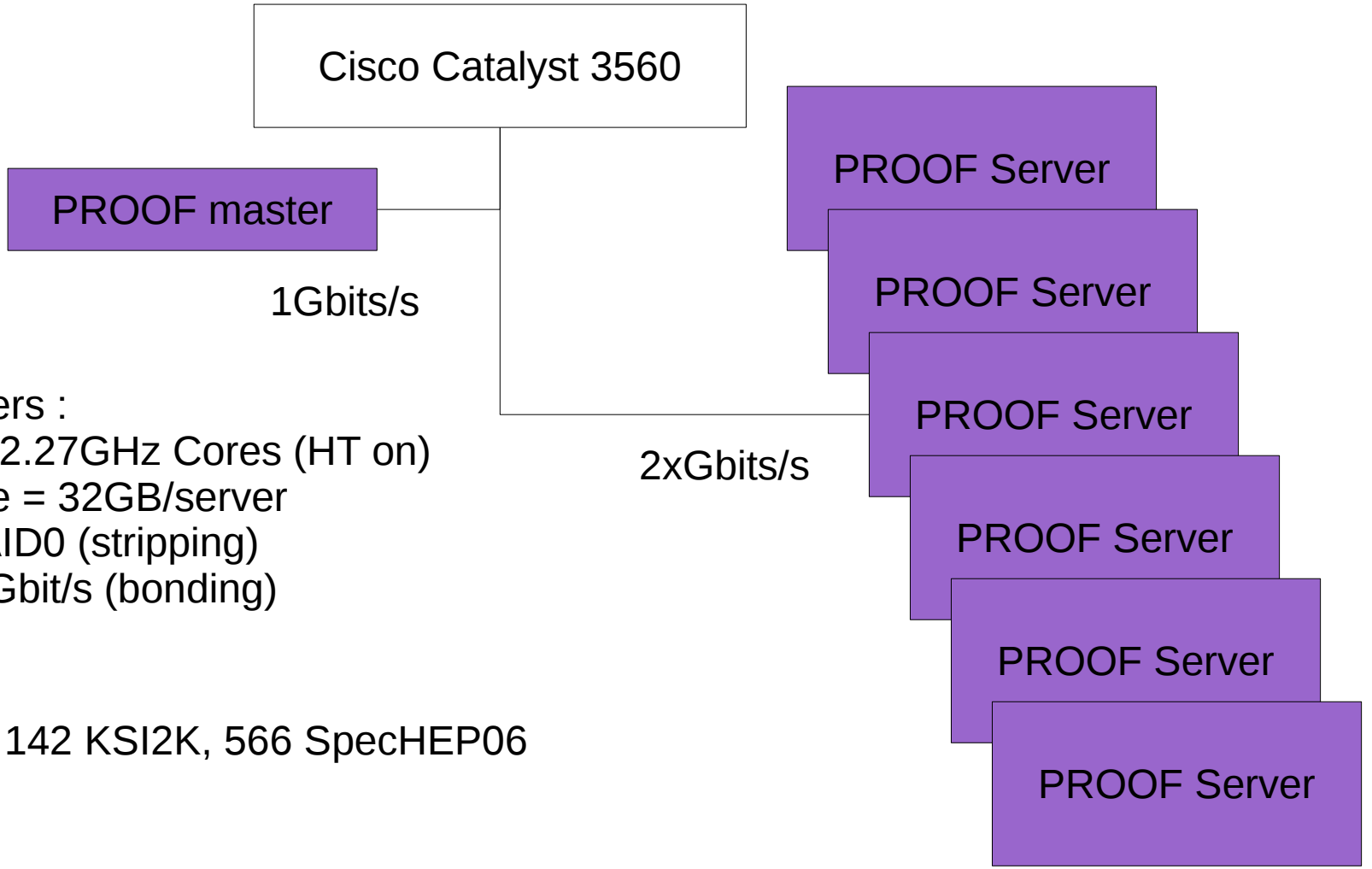
History

- June 2009 : Standard PROOF cluster
- April-summer 2010 : Work with Dario Berzano on the automatic staging daemon (afdsmgrd)
- July 2010 : Reinstall the cluster using AAF scripts from Martin Vala

Considerations on hardware setup

- Storage and CPU on the same servers
- Storage performance : security less important than access rate => RAID0
- Network BW between machines in the cluster
- RAM : 4GB/core = 32GB/server
- What is required for the PROOF master ? (we use a retired worker-node)

SAF current architecture



6 PROOF Servers :
CPU : 8 E5520/2.27GHz Cores (HT on)
RAM : 4GB/core = 32GB/server
2To Storage RAID0 (stripping)
Network : 2 x 1Gbit/s (bonding)

Total :
CPU : 48 cores, 142 KSI2K, 566 SpecHEP06
192GB RAM
Storage : 12To

AAF Installation (simplified)

- Generate configuration file (aaf.cf) (Web Interf.)
- Have a host certificate for the master
  - Same certificate on the workers
- Run install script :
  - Worker : aaf-installer –install-worker
  - Master : aaf-installer –install-master
- Create users accounts
  - Local accounts with same name as at CERN

AAF and the sys.admin 1/2

- Software packages management (Packman) :

Subject: [Subatech #10648] Packages on SAF
Salut Jean-Michel,
mardi sur la SAF on avait le paquet AliRoot::v5-02-16-AN, que j'ai utilise' sans soucis.
Aujourd'hui le paquet est disparu.

Subject: [Subatech #10648] Packages on SAF
Salut Jean-Michel,
Tuesday on SAF we had package AliRoot::v5-02-16-AN, that I used without problem.
Today the package is gone away.

=> This is the job of a the cron : aaf-installer.sh –sync-alien-packages-cron
sometimes it unexpectively removes packages...

AAF and the sys.admin 2/2

- Change management
  - Changes are needed but the process have to be mastered
    - Have a testbed locally ? Not very realistic
    - Know what is being changed (components, config)
    - Be able to return to previous state and restore service
  - Tried to identify AAF components and config items. Not an easy task.

AAF components / config

AliEn

AliEn user certificate

Host certificate
(1 for the entire cluster ?)

Operating System
Libraries (libxml2, zlib,...)
Utilities : cmake

AAF Config file aaf.cf

PROOF binaries
from (which ?) ROOT package

PROOF config file(s)

xrootd binaries
compiled in situ, version defined in ?

Xrootd config file(s)

Xrootd scripts (including xrddm)

AFDSUtils macro

AliROOT/ROOT/Geant3 packages

Installer script : aaf-installer

Understand the engine

- Identify what can change, when, from which source :
  - Packages (Impact of Automatic package synchronization ?)
    - User packages (AliRoot,etc)
    - PROOF packages (ROOT, aaf-*)
  - Items from googlecode (scripts, etc.)
  - Configuration aaf.cf (not much supposed to change)

Revert to a known state

- See how each change can be reverted
  - Ex: new ROOT version brings a new PROOF :
    - New version of ROOT installed (package sync)
    - Root symlink to this new version part of aaf-aliroot post-installation
    - A restart of PROOF daemons starts the new version.
  - How do we revert to the previous PROOF in case of problem ?

Worth the effort ?

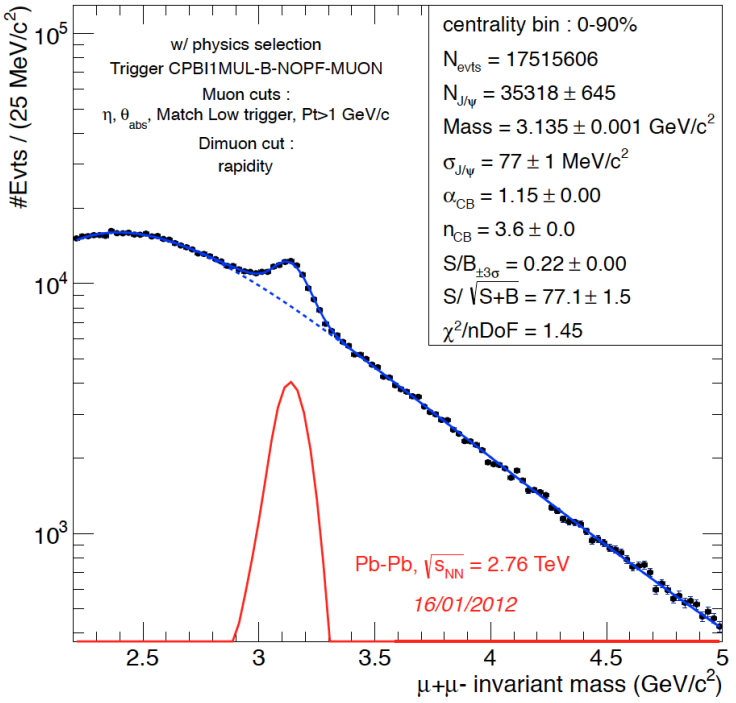
- You bet ! Lots of data analyzed fast on SAF.
- => 7.5 TB used ~ 60% used
- LHC10h [~ FULL STAT]→ x-check of 2010 PbPb J/Ψ analysis
- LHC11c,d,e [as much as was produced] → Υ (and J/Ψ / Ψ') analysis in p+p
- +sim ESDs of private production of single particles for eff x acc. corrections (LHC11d)
- + ESDs for a few runs (LHC11a, LHC11c) for MCH tracking resolution studies

It just works

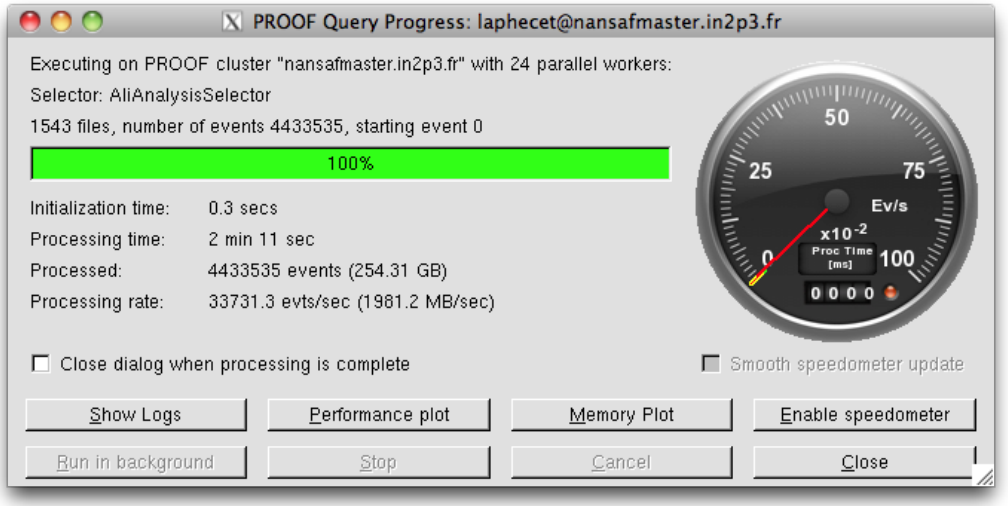
LHC11h [~ FULL STAT]

→ made possible a very quick analysis done right after the end of the 2011 PbPb run

→ as for LHC10h, will be used for J/Ψ analysis and, this year, for J/Ψ v2, single μ, etc...



Speed



- Rate ~ 30Kevents/s
- Full 2011 PbPb period (LHC11h) ~ 130Mevents => ~ 1 hour to process all the data (*) (**) !
- (to be compared to day(s) on the Grid)

(*) actually takes longer as there's currently a limit (proof bug ? Aliroot bug ?) on the number of datasets that can be processed at once, so there's an overhead of startingsession + merging phase

(**) note though that it's mainly due to the small size of the data type (muon AODs) we're looking at

Conclusion

- AAF has already proven being a very valuable tool
- Still a bit difficult to understand and maintain though