Status and plans of the T3 in Geneva

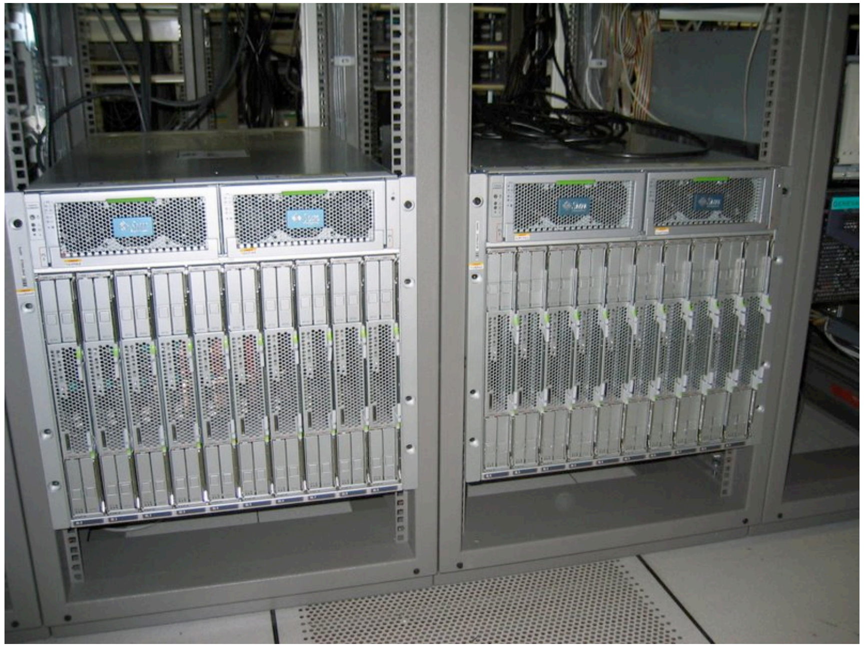
Szymon Gadomski
DPNC, University of Geneva

Swiss ATLAS Grid Working Group
Jan 7th, 2008

The cluster in Geneva (1)

The cluster in Geneva (2)

12 worker nodes in 2005



and 20 in 2007!

21 in 2006

The cluster in Geneva (3)

power and network cabling of worker nodes

three nodes for services (grid, batch, storage abstraction)

direct line from CERN

The hardware in numbers

- 61 computers to manage
  - 53 workers, 5 file servers, 3 service nodes
- 184 CPU cores in the workers
- 75 TB of disk storage
- can burn up to 30 kW (power supply specs)

A part of the hardware is already in production. More about the status later.

The functionality we need

- our local cluster computing
  - log in and have an environment to work with ATLAS software, both offline and trigger
    - develop code, compile,
    - interact with ATLAS software repository at CERN
  - work with nightly releases of ATLAS software, normally not distributed off-site but visible on /afs
  - We also use CERN afs accounts.
    - We get all advantages of lxplus.cern.ch (examples of ATLAS software…)
    - without disadvantages (much more disk space, fewer users).
  - disk space visible as normal linux file systems (cd, cp, …)
  - use of final analysis tools, in particular ROOT
  - a convenient way to run batch jobs
- grid computing
  - tools to transfer data from CERN as well as from and to other Grid sites worldwide
  - ways to submit our jobs to other grid sites
  - a way for ATLAS colleagues, Swiss and other, to submit jobs to us

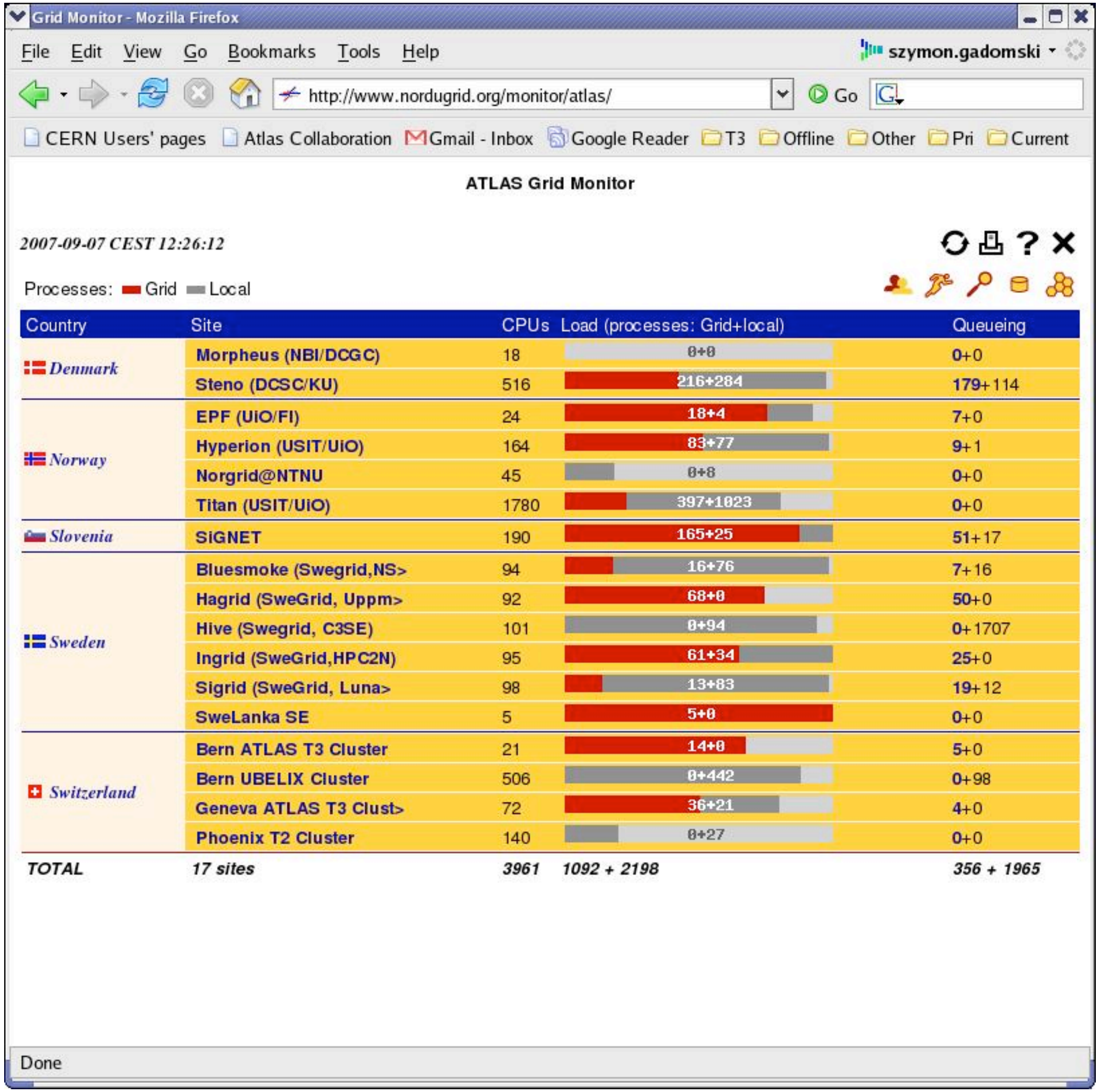
Our system in production

It changes. The status right now is:

- 41 TB of storage, available with NFS
  - 2 SunFire X4500 (Solaris), 1 old server (SLC3)
- 8 login machines
- 18 batch worker nodes
  - 18 SunFire X2200 (2×2 core)
- 1 grid front-end machine
  - SunFire X4500 with SLC4 (not used as a file server anymore)
- The direct line from CERN goes to three login machines and 2 file servers.

Our system in the NorduGrid

- Geneva is in NorduGrid since 2005
- In company of Berne and the T2 in Manno


Grid Monitor - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
szymon.gadomski
http://www.nordugrid.org/monitor/atlas/
Go
CERN Users' pages
Atlas Collaboration
Gmail - Inbox
Google Reader
T3
Offline
Other
Pri
Current
Done

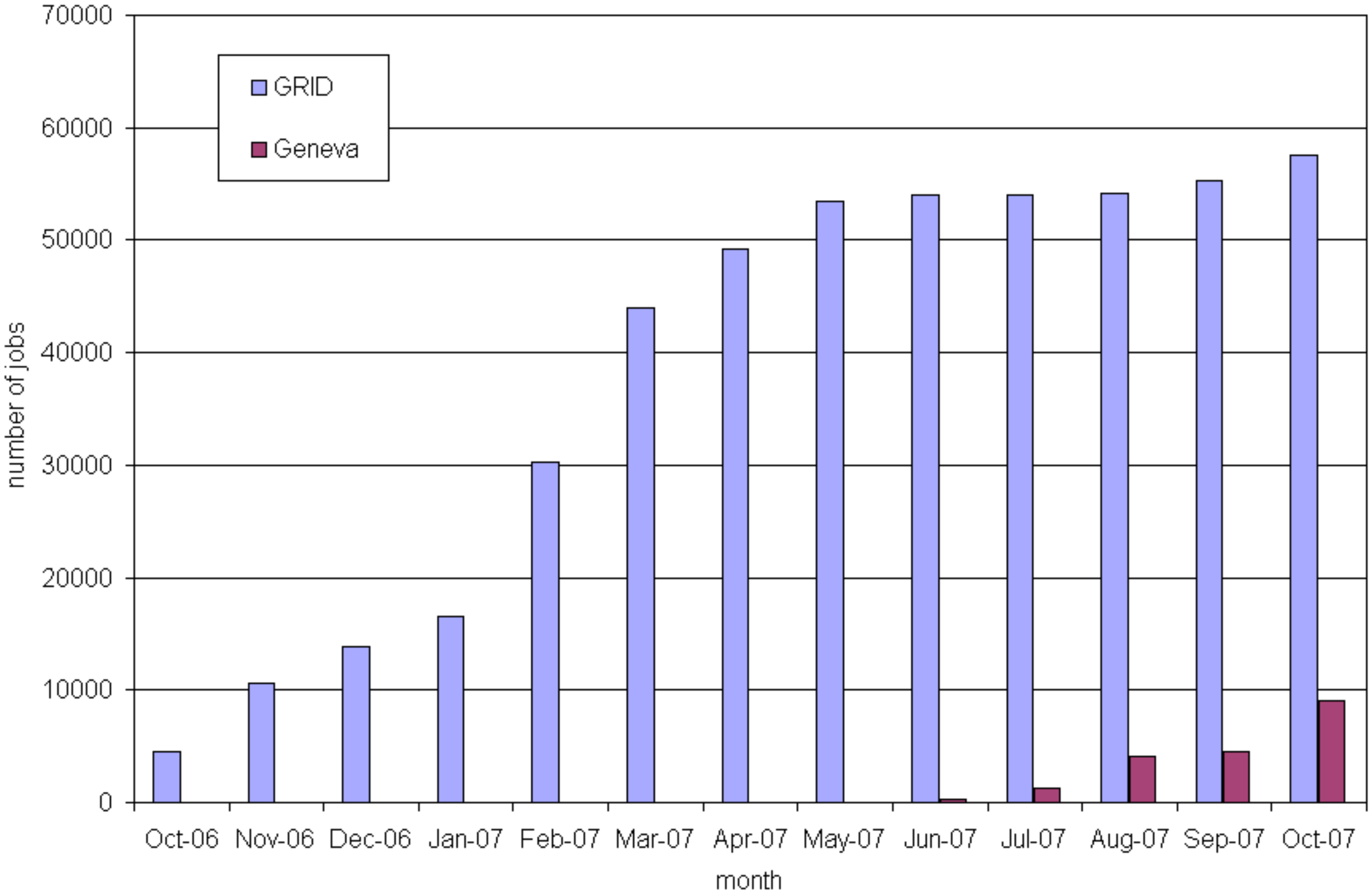
ATLAS Grid Monitor

2007-09-07 CEST 12:26:12

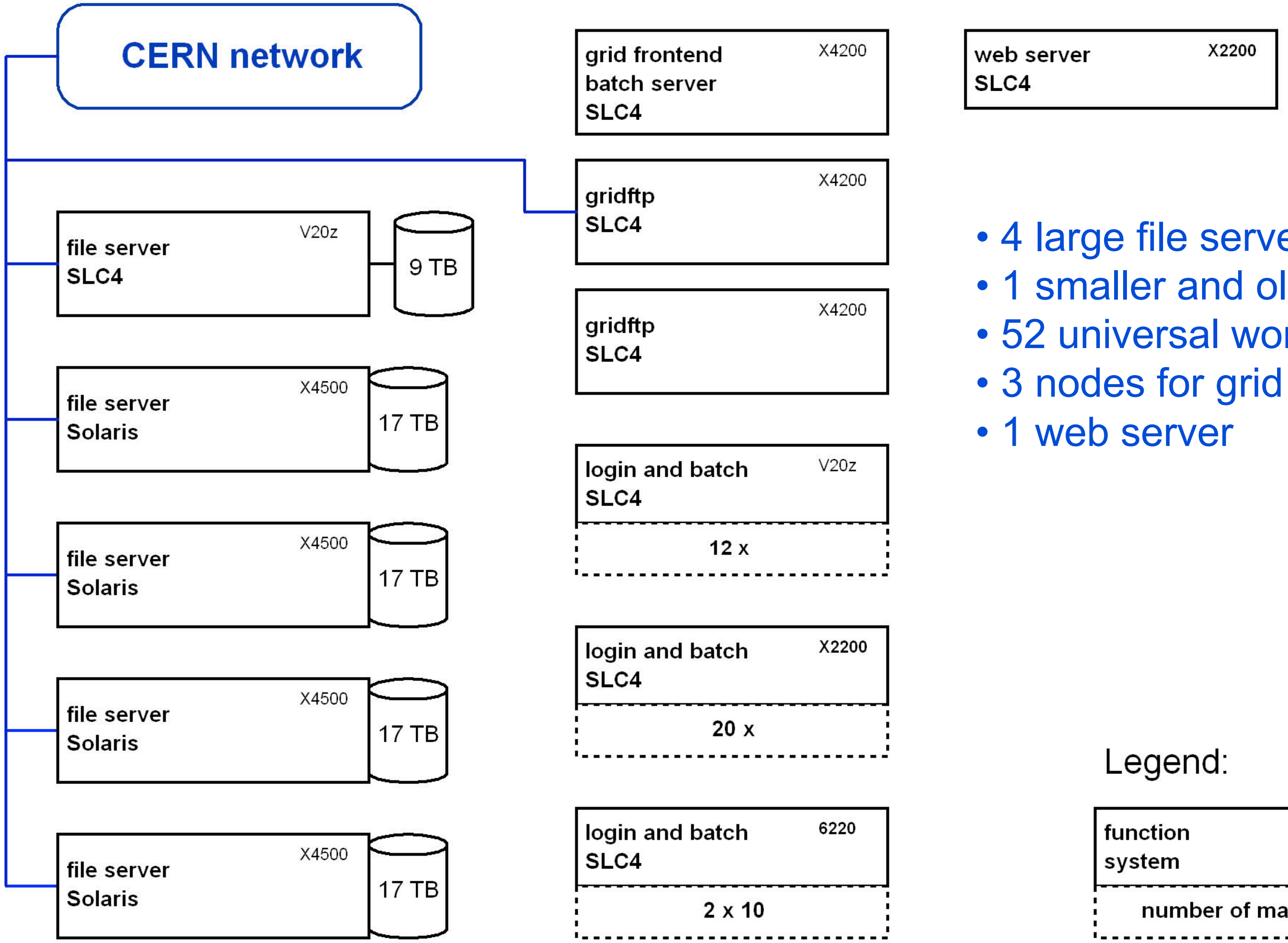
Processes: 🟥 Grid ▬ Local

Country	Site	CPUs	Load (processes: Grid+local)	Queueing
Denmark	Morpheus (NBI/DCGC)	18	0+0	0+0
	Steno (DCSC/KU)	516	216+284	179+114
Norway	EPF (UiO/FI)	24	18+4	7+0
	Hyperion (USIT/UiO)	164	83+77	9+1
	Norgrid@NTNU	45	0+8	0+0
	Titan (USIT/UiO)	1780	397+1023	0+0
Slovenia	SiGNET	190	165+25	51+17
Sweden	Bluesmoke (Swegrid,NS>	94	16+76	7+16
	Hagrid (SweGrid, Uppm>	92	68+0	50+0
	Hive (Swegrid, C3SE)	101	0+94	0+1707
	Ingrid (SweGrid,HPC2N)	95	61+34	25+0
	Sigrid (SweGrid, Luna>	98	13+83	19+12
	SweLanka SE	5	5+0	0+0
Switzerland	Bern ATLAS T3 Cluster	21	14+0	5+0
	Bern UBELIX Cluster	506	0+442	0+98
	Geneva ATLAS T3 Clust>	72	36+21	4+0
	Phoenix T2 Cluster	140	0+27	0+0
TOTAL	*17 sites*	*3961*	*1092 + 2198*	*356 + 1965*

Batch jobs statistics

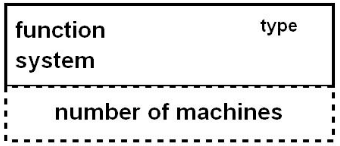


The system as planned for 1st data


CERN network
grid frontend batch server SLC4 X4200
web server SLC4 X2200
gridftp SLC4 X4200
gridftp SLC4 X4200
file server SLC4 V20z
9 TB
file server Solaris X4500
17 TB
file server Solaris X4500
17 TB
file server Solaris X4500
17 TB
file server Solaris X4500
17 TB
login and batch SLC4 V20z
12 x
login and batch SLC4 X2200
20 x
login and batch SLC4 6220
2 x 10

- 4 large file servers (68 TB)
- 1 smaller and older (7 TB)
- 52 universal workers (184 cores)
- 3 nodes for grid + batch server
- 1 web server

Legend:



(Re)installing everything. Why?

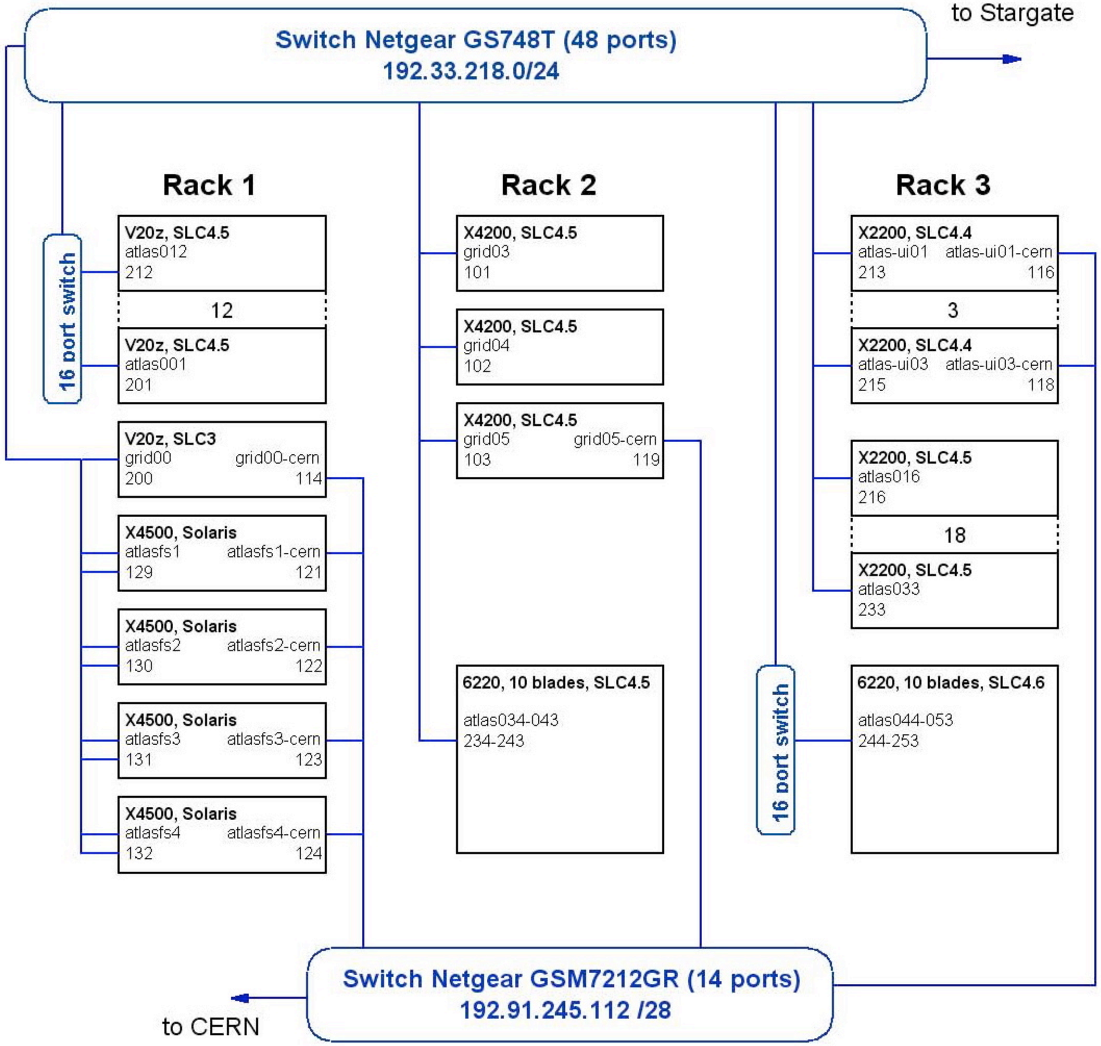
- three generations of hardware
  - 2005, SLC3, system getting obsolete
  - 2006, setup in record time, can’t be propagated anyway
  - 2007, recently delivered, need to be installed
  - we want the same version of SLC on all the workers
- not only a pure batch job facility anymore
- grid and batch services on dedicated nodes, no longer on file servers
- file servers under Solaris
  - SunFire X4500 is designed for this, zfs is attractive
  - I can get help at the Uni to have this done. I still have the remaining 57 machines to manage.

Towards the final system

- **What is done already**
  - Experience with the system in production
    - File servers used also as batch servers and grid front-ends
      - Both crashed last June! File servers will not do anything else anymore.
    - Power was lost three times last Summer.
  - Stable operation since September
    - >30 users rely on the cluster for daily work, interactive + batch
    - NorduGrid jobs are running ~all the time
  - All the new hardware is in the racks
  - The networks are done
  - Work on power supply is done (during the Christmas break).
  - The new “final” batch system is in place
    - Batch server on a SunFire X4200 machine
      - TORQUE and pbs_sched
    - 10 worker nodes on new blades (Sun Blade 6220)
    - Independent hardware, see the same file servers as the system in production

Networking for the final system

In place since last November.



Switch Netgear GS748T (48 ports)
192.33.218.0/24

to Stargate

16 port switch

Rack 1

V20z, SLC4.5
atlas012
212

12

V20z, SLC4.5
atlas001
201

V20z, SLC3
grid00 grid00-cern
200 114

X4500, Solaris
atlasfs1 atlasfs1-cern
129 121

X4500, Solaris
atlasfs2 atlasfs2-cern
130 122

X4500, Solaris
atlasfs3 atlasfs3-cern
131 123

X4500, Solaris
atlasfs4 atlasfs4-cern
132 124

Rack 2

X4200, SLC4.5
grid03
101

X4200, SLC4.5
grid04
102

X4200, SLC4.5
grid05 grid05-cern
103 119

6220, 10 blades, SLC4.5
atlas034-043
234-243

Rack 3

X2200, SLC4.4
atlas-ui01 atlas-ui01-cern
213 116

3

X2200, SLC4.4
atlas-ui03 atlas-ui03-cern
215 118

X2200, SLC4.5
atlas016
216

18

X2200, SLC4.5
atlas033
233

16 port switch

6220, 10 blades, SLC4.6
atlas044-053
244-253

Switch Netgear GSM7212GR (14 ports)
192.91.245.112 /28

to CERN

Towards the final system (2)

- **Next to do: finish "sysadmin" work**
  - Change scheduler to Maui on the new batch server
    - learn how to configure it…
  - Get machine certificates
    - Can we use GlobalSign, rather than SwissSign? CHIPP gives us this option now! Certificates are valid for three years…
  - Setup NorduGrid on a new front-end machine
  - Operate the new system in parallel with the one currently in production for some time, check that all is stable.
  - Setup a web server (my 1st…)
  - Setup monitoring of network and CPU use.
    - learn about Ganglia
  - Install all remaining hardware
    - 10 remaining blades, 5 currently unused machines from 2005
    - two more X4200 to be used for gridftp
  - Migrate worker nodes and login machines to the new system.
  - Retire grid02 as grid+batch

Towards the final system (3)

- Longer term things to do
  - Sysadmin follow-up
    - reduce dependence on grid00 (SLC3)
    - other optimizations and fixes
  - Data transfer exercises
    - Official data path: CERN > FZK > CSCS > (BE+GE)
      - Data can be pushed up to CSCS.
      - GE + BE can pull it from there
    - Backup path: CERN > GE > BE
      - Pull only. Small quantities of data.
    - Consider GE as a DDM service site
      - Push data like this: CERN > FZK > (CSCS + GE)
      - Pull data GE > BE
      - We need a file catalogue listing data in Geneva. Can it be at CSCS?
  - Tools to submit many batch jobs at once
    - GridPilot?
    - Ganga
  - Interactive use of many machines
    - PROOF

Summary

- The ATLAS cluster in Geneva is a large Tier 3
  - will have 184 worker's CPU cores and 75 TB soon
- A part of the system is in production
  - a Grid site since 2005, runs ATLAS simulation like a Tier 2, plan to continue that.
  - since Spring in constant interactive use by the Geneva group and friends, plan to continue and to develop further
- Expect to be busy installing machines and services for another few weeks
  - a long of history, three generations of hardware
  - need a uniform system with a rational architecture
  - we have learned from experience
- After that, I can start participating in data transfer exercises.
  - I will try to get some help for that in the Geneva group.