StoRM & GPFS

CMS and Offline Week 23/04/2009

I. Cabrillo Bartolomé
Instituto de Física de Cantabria (IFCA)
Spain

I.González Caballero
Oviedo University
Spain

F. Matorras Weinig
Instituto de Física de Cantabria (IFCA)
Spain

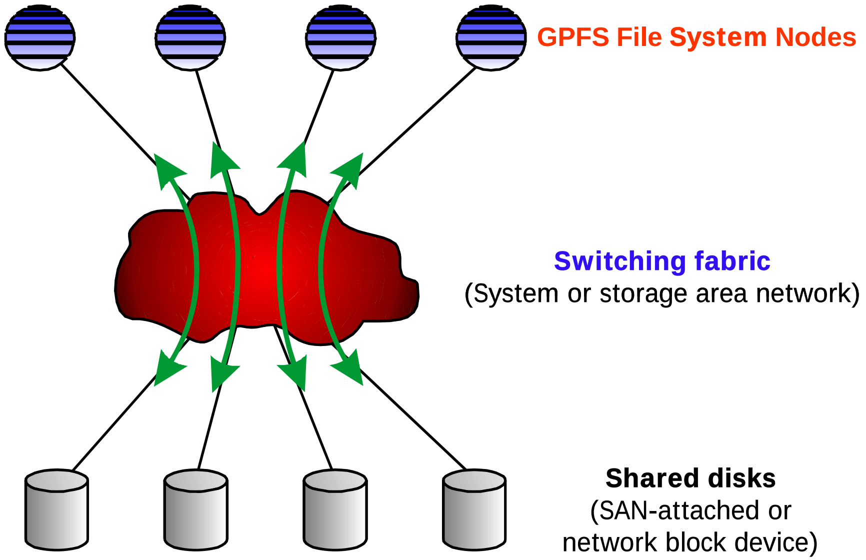
IFCA

- Pluridisciplinar: HEP, Astrophysics, Cosmology, Statistical Physics, …
  - It is involved in diferent Computing proyects:
    - Supporting CMS in LHC and other non-HEP communities (Plank in astrophysics, statistical physics, Biomedicine, …).
    - Bunch of GRID computing projects like NGI-ES, DORII, EGEE, EGI, EUFORIA, GRID-CSIC and INTEUGRID.
  - It was involved in other Grid projects like CROSSGRID and DATGRID

IFCA
Instituto de Física de Cantabria

GPFS Description

- Is an IBM high-performance scalable file management solution that provides fast, reliable access to a common set of file data from a single computer to hundreds of systems.
- Mixed server and storage components.
- Online storage management,
- Scalable (2000 nodes and haundred of PB)
- Direct I/O
- Replication
- Snapshots
- Quotas



Storage System at IFCA I (Hardware)

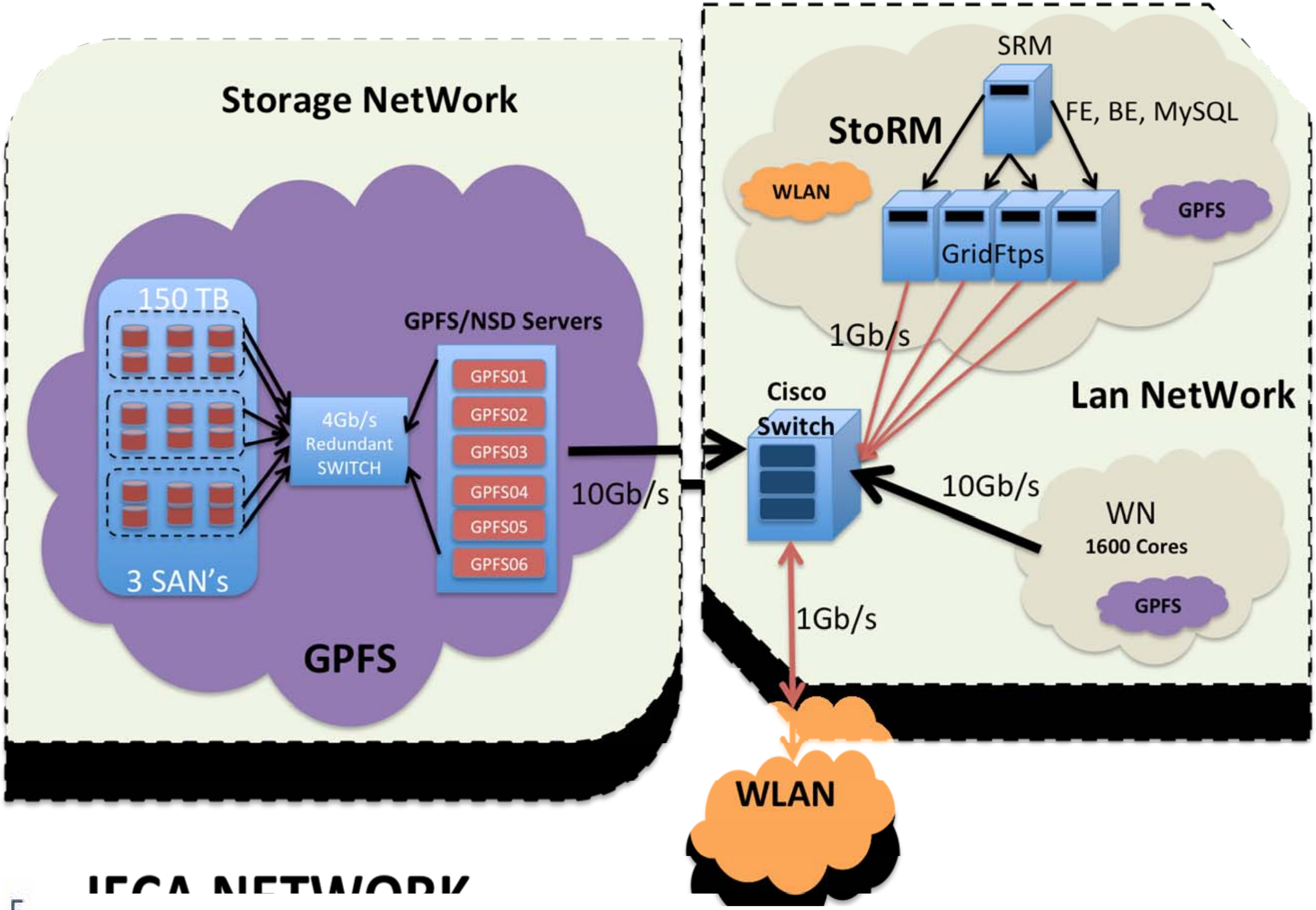
- 5 SAN's IBM (2 in production, 3 testing)
  - DS4700 Controllers and EXP810 expansion enclousures
    - Redundant FC 4 Gb/s connenction
    - FC and SATA HDD support (SATA for IFCA case)
    - Support For 112 HDD slots
    - RAID 0, 1, 5,6 (RAID5 in IFCA case)
- 6 GPFS Servers
  - X3650 IBM servers
    - RAID 1
    - Redundant 4 Gb/b FC connection
    - 10 Gb/s Network
    - MutiPath Driver (RDAC)
- StoRM
  - X3655 IBM server
    - RAID1
    - 1 Gb/s Network
  - 4 IBM X336 GridFtp's

Storage System at IFCA II

- GPFS
  - 6 GPFS/NSD Servers
  - Cluster with more than 200 nodes
  - 300 TB in 2 file systems
  - GPFS modules depend on the linux Kernel
  - Only Supported for RHEL and SE (a little modifications to use with SLC)
  - Losts of commands and Variables to be/set configured

- GPFS Storage Network
  - Deployed on top of a private LAN (to avoid security problems and to work with the WN).
  - Is able to export file systems through NFS or to create a “gpfs-cnfs” cluster (nfs fail over cluster through gpfs)
  - StoRM and GridFTP servers must have access to both networks.
    - One to Phedex or other srm Transfers
    - Other for access to Storage Network
  - GPFS has been installed on all the farm nodes, then all the WN can access through the usual POSIX commands (cp, rm, mv...) to the File System
  - Can be used as any other local file system.

Storage System at IFCA III

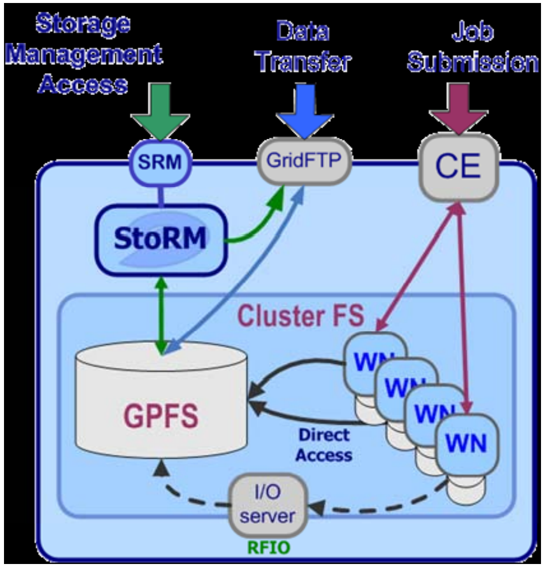


From DPM to StoRM

- DPM
  - 1 Head Node and 8 disk servers 30 TB (2006-2007)
    - Easy to Install and maintain
    - Unstable Hardware/Stable Software
    - Scaling Problems
    - FS non POSIX
    - Problems with dpm rfio libraries
- StoRM
  - 1 StoRM basic service (FE, BE, Mysql) and 4 Gridftp servers
    - Easy to Install and maintain
    - Vey Stable Hardware/Software
    - Good Scaling (only adding gridftp servers it maybe be virtuals)
    - Indepently of the FS. Other projects can work in the FS knowing anything about StoRM
    - No access problems(POSIX)
    - We had GPFS

StoRM I (Description)

- What is StoRM (Storage to Resource Manager)
  - StoRM is a grid Storage Resource Manager for disk based storage systems, it implements SRM interface version 2.x.
  - Designed to work over native parallel filesystems (Specially GPFS).
  - ACL support provided by the underlying file systems to implement the security models.



- Services
  - FrontEnd (FE): Get the transfer requests and register them into DB
  - Data Base (MySQL)
  - BackEnd (BE): Manage the SRM interface (access to FS)

StoRM II (Installation)

- Add the follwoing repositories into yum: ig, glite-generic and ig_gridftp for your arch.
- Install Java jdk
- installation:
  - ig_SE_storm_frontend (yum install ig_SE_storm_frontend)
  - ig_SE_storm_backend (yum install ig_SE_storm_backend)
- Install certificates
- Setup your site-info.def with the correct StoRM parameters
- Configure your nodes:
  - ig_yaim -c -s <your-site-info.def> -n ig_SE_storm_frontend
  - ig_yaim -c -s <your-site-info.def> -n ig_SE_storm_backend
    - This will also install the Gridftp Server
  - ig_yaim -c -s <your-site-info.def> -n ig_GRIDFTP
    - This is only needed if Gridftp service is not on the same machine that ig_storm_backend you must configure a node as gridftp (you can use other non ig_GridFtp)
  - Detailed instructions for different configurations may be found in :
    - Backend installation in a cluster
    - Frontend installation in a cluster

StoRM III (Improvements)

- All services (FE, BE, Mysql and Griftp) on same machine
  - System overloaded with more than 20 connections at the same time
    - High usage of RAM, cached and buffered memory
    - Swaping (sometimes reboot)
- 1 node with StoRM Basic services (FE, BE, MySQL) and 4 GridFTP servers
  - To avoid overload problems
  - To prepare for future external network upgrades
  - Balance/shared through DNS round robin (maybe upgrades to LVS balance)

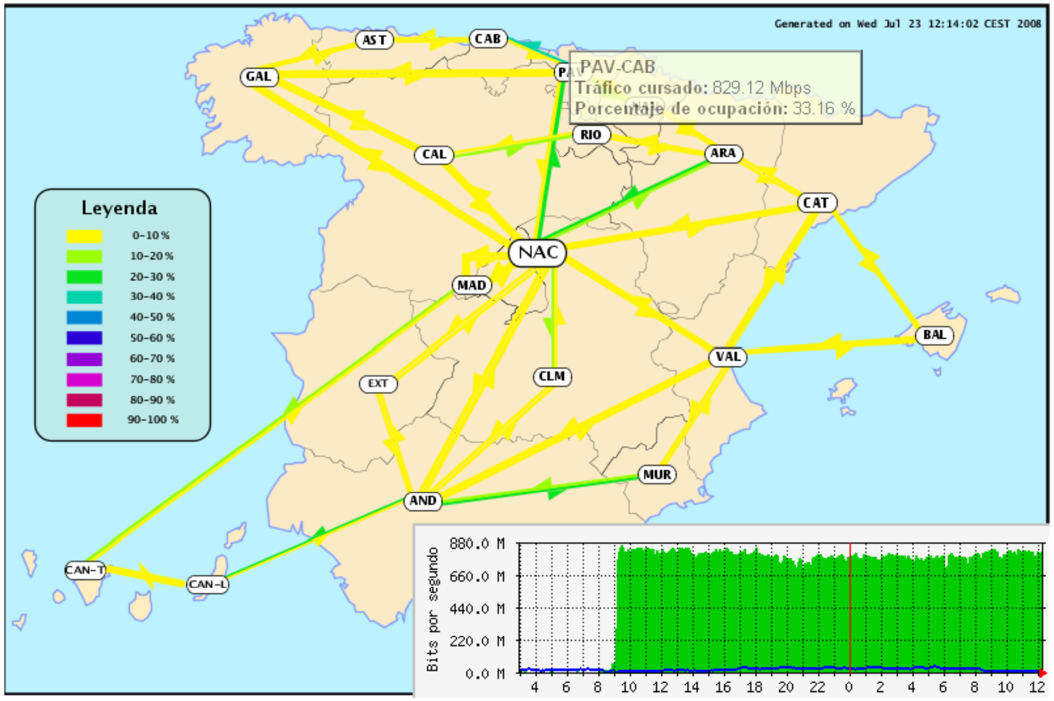
  - StoRM Machine works fine
  - Gridftp’s overloaded sometimes with more than 20 gridftp processes running
    - High usage of RAM, cached and buffered memory
    - Swaping (sometimes reboot)
    - Kernel parameter modification needed (most of them at Dcache Network Tunning)

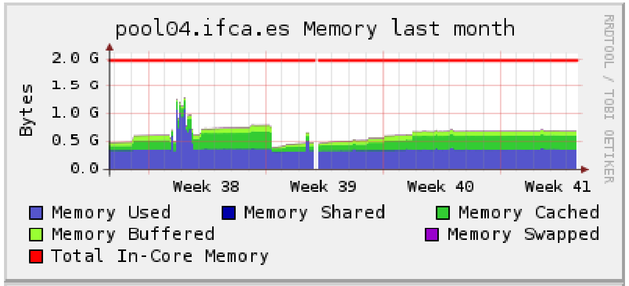
```
net.core.rmem max = 1048576
net.core.rmem default = 87380
net.core.wmem max = 131072
net.core.wmem default = 32768
net.ipv4.tcp rmem = 4096 87380 1048576
net.ipv4.tcp_wmem = 4096 32768 131072
net.ipv4.tcp_mem = 65536 87380 98304
```

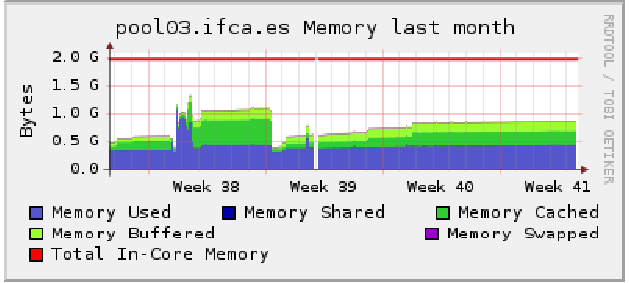
```
vm.min free kbytes = 65536
vm.overcommit_memory = 65536
vm.overcommit_ratio = 2
vm.dirty_ratio = 10
vm.dirty_background_ratio = 3
vm.dirty expire centisecs = 500
```

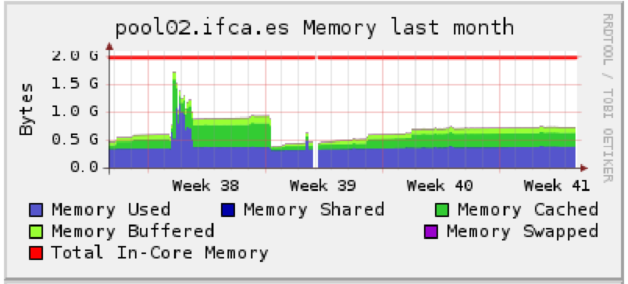
StoRM IV (improvements)

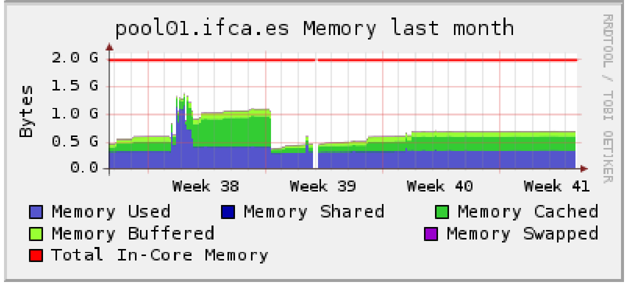
- Optimal results from these modifications
  - No gridFTP server over 1GB RAM
  - No more swaping
  - 830 Mb PhEDEx incoming Data during 26 h (1Gb max througput)





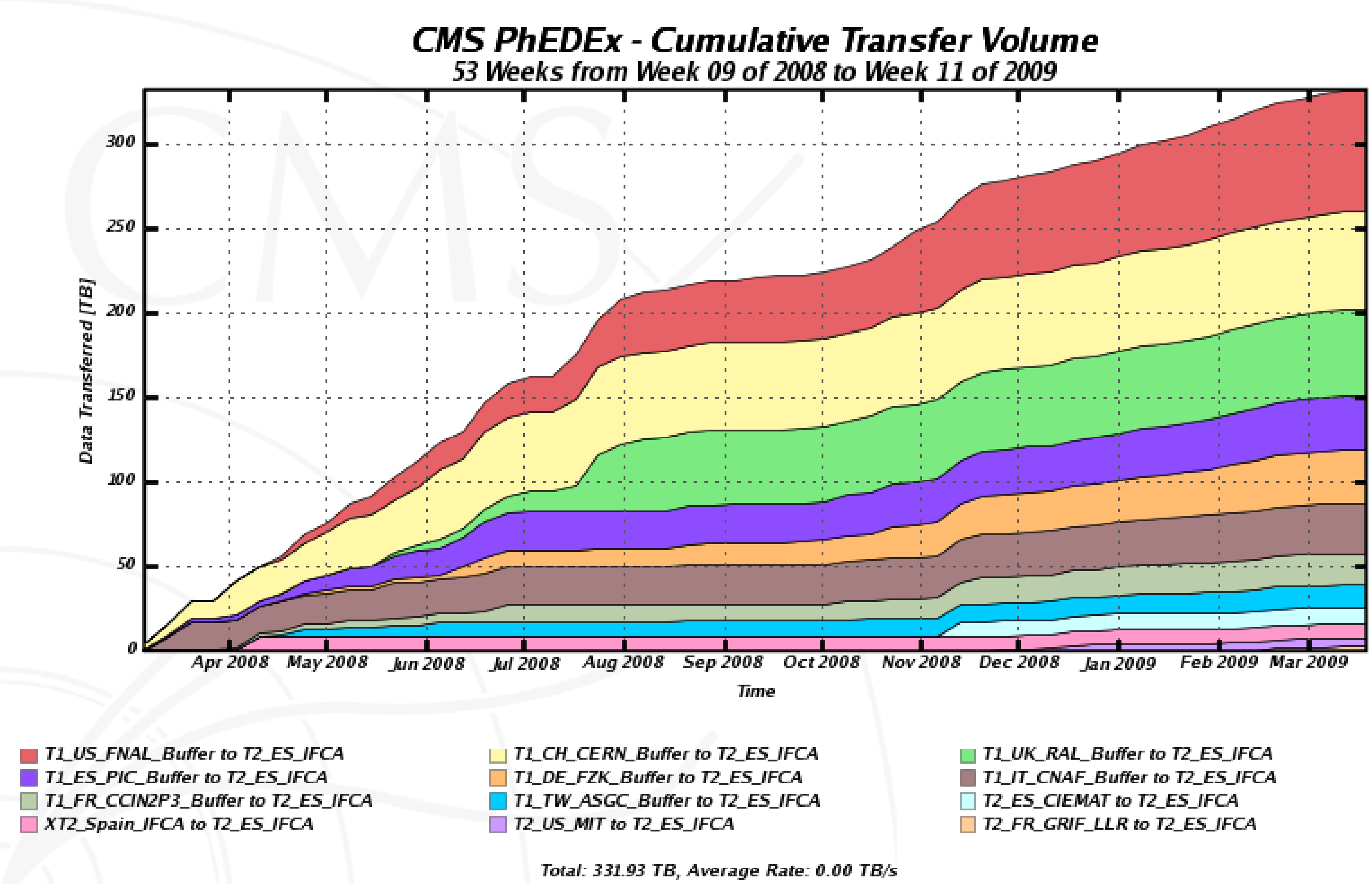






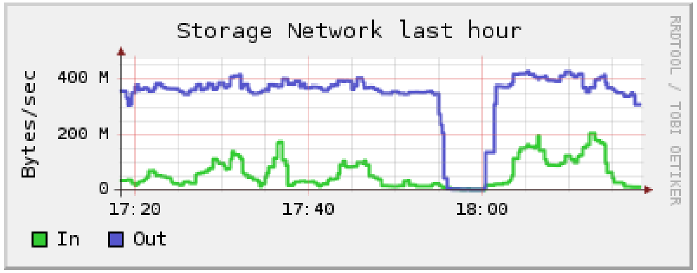
StoRM V

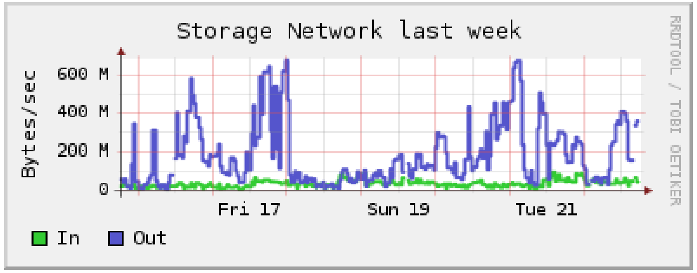
- More than 300TB transfered since we started (Prod+Debug)



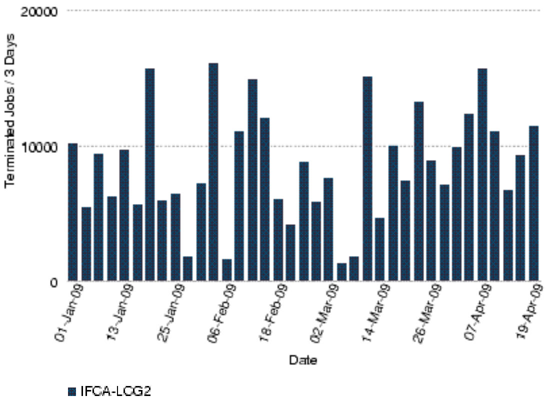
Dataflow

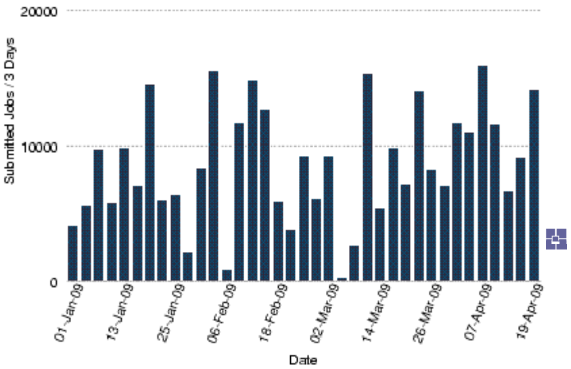
- GPFS
  - Is mounted as a local FS on all the WN
  - WN can read/write the FS directly
  - Now limited to 8Gbps (2 x 4Gbps FC SAN access)
  - Soon to be upgraded to 20 Gbps
- WN
  - 1800 Cores in ~ 230 nodes
    - All chasis blades have 10 Gbps external network
    - Each 4 chasis blades (56 nodes) have 10 Gbps access to Storage Network





- Storage Network usage during a common load of 500 jobs (production and Analisys) doesn't fully occupy the total BW (8 Gbps) but sometimes we are near this limit





- Terminated and Scheduled Jobs/3 days since 01/2009

Conclusions

- StoRM
  - Easy to install and easy to maintain
  - Stable (most problems caused by the FS)
  - Need some improvements in the user manage
  - **Thanks all people at StoRM Support (Luca, Riccardo,…)**
- GPFS
  - POSIX access. Do not need to implement other access methods
  - Difficult to optimize
  - Very Stable (needs optimization)
  - Good I/O needs optimization)
  - Dependency of GPFS modules on the Linux kernel. A recompilation of the modules is needed for each kernel upgrade

The End

¡Thank you very much!

CMS

Compact Muon Solenoid

IFCA

Instituto de Física de Cantabria