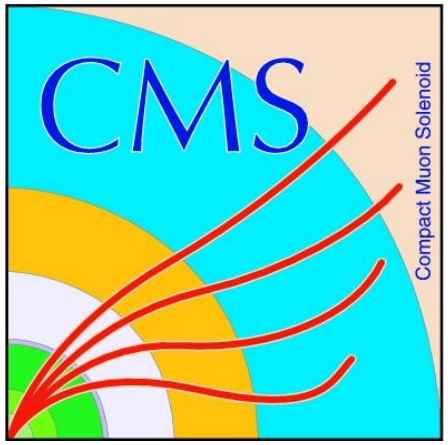
StoRM & GPFS

CMS and Offline Week 23/04/2009



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

I.González Caballero
Oviedo University
Spain

IFCA

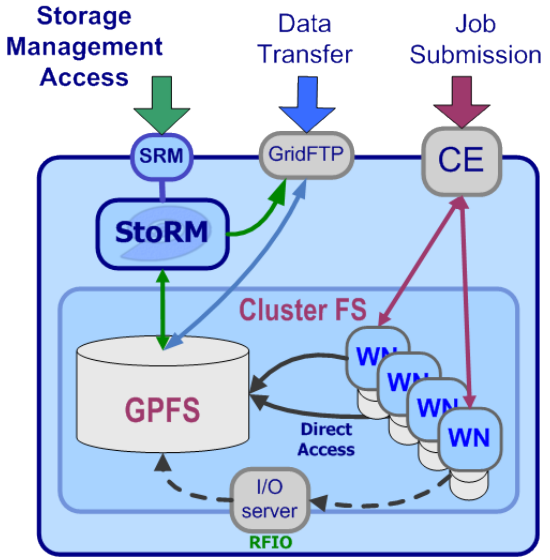
- IFCA
  - IFCA is a Pluridisciplinar Physisc Centre "Rellenar"
  - It's evolved in diferent Computing proyects:
    - Supporting CMS in LHC and other non-HEP communities (Plank in astrophysics, statistical physics, Biomedicine, ...).
    - Bunch of GRID computing projects like NGI-ES, DORII, EGEE, EGI, EUFORIA, GRID-CSIC and INTEUGRID.

- StoRM at IFCA to solve the large request for data transfer and availability of some projects, mainly CMS
- maching up with a mayor upgrade in the facilities.
- We have GPFS
- IFCA has deployed into production the StoRM SE system since March 2008.

Introduction I

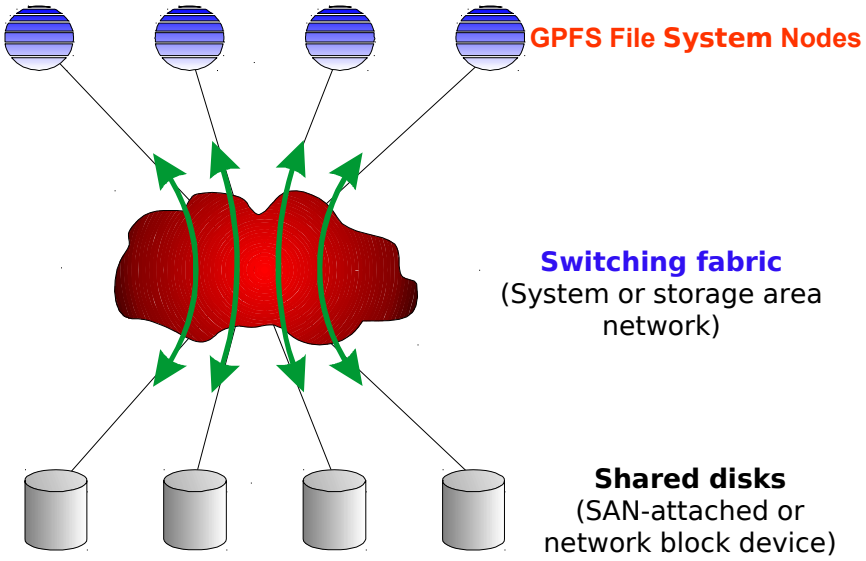
- What's StoRM (Storage to Resource Manager)
  - StoRM is a grid Storage Resource Manager for disk based storage systems, it implements SRM interface version 2.x.
  - Designed to work over native parallel filesystems (Specially GPFS).
  - ACL support provided by the underlying file systems to implement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

- Services
  - FrontEnd(FE): Get the transfer requests and register it into DB
  - Data Base (Mysql)
  - BAckEnd: Manage the SRM interface



Introduction II

- What's GPFS (General Parallel File System)
  - Is a 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
  - Memory mapped files
  - Direct I/O
  - Replication
  - Snapshots
  - Quotas
  - Extended attributes



Storage System at IFCA

- Configuration at IFCA
  - 1 node with StoRM services (FrontEnd, BackEnd and MySQL database)
  - 4 nodes hosting GridFTP servers to increase the transfer performance.

- GPFS Storage Network is deployed on top of a private LAN.
  - StoRM and GridFTP servers must have access to both networks.
  - All the farm is able to access GPFS through the usual POSIX commands (cp, rm, mv...)
  - Can be used as any other local file system.

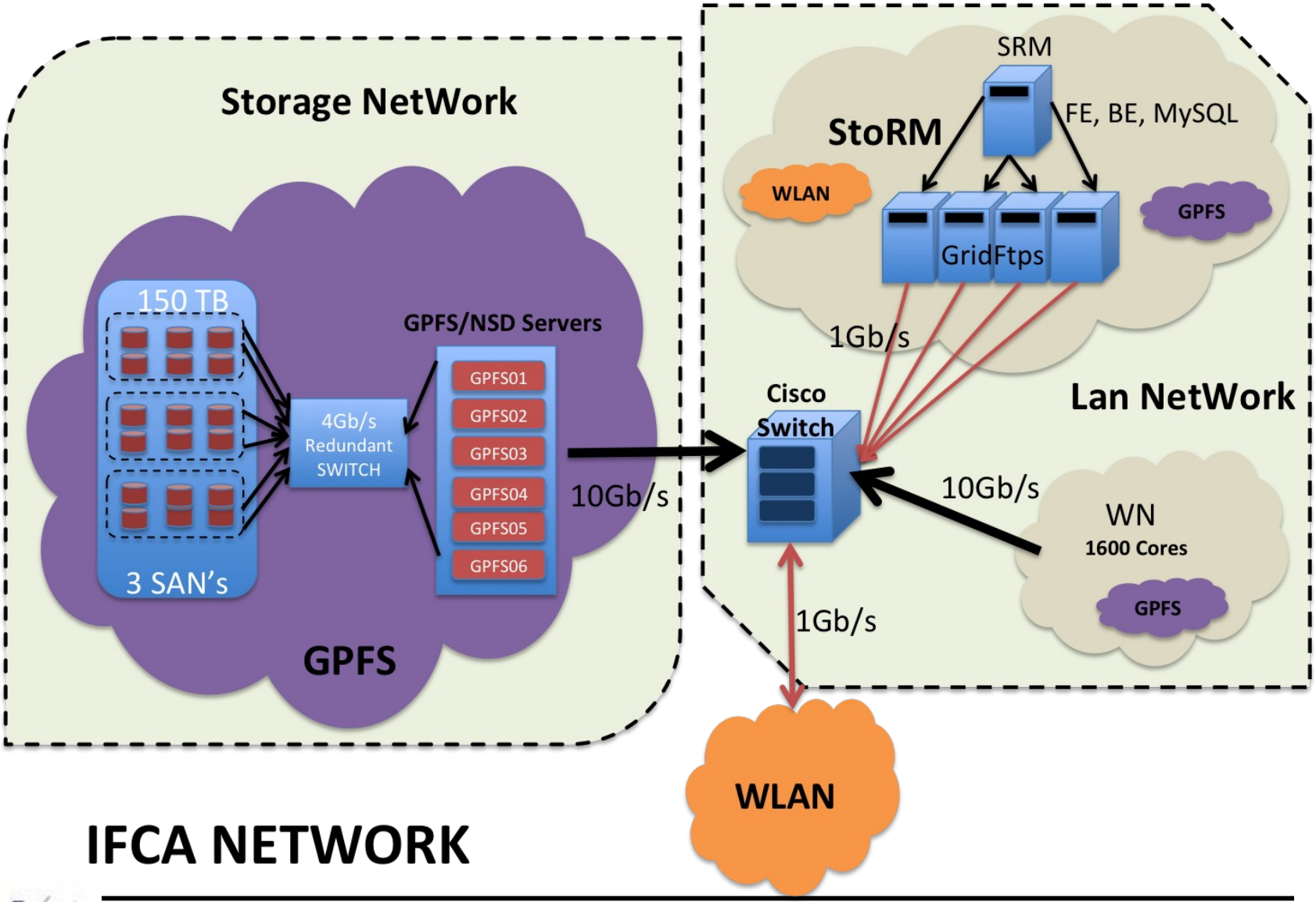
Storage System at IFCA I

- 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 Networ
    - 4 IBM X336 GridFtp's

Storage System at IFCA II

- Software
  - 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
  - StoRM
    - Installed following the INFN instructions (very simple to install)
      - Two kind of config files:
        - one per StoRM machine (FE, BE, MyslQ, Gridftp)
        - one for the 4 Gridftp Servers

Storage System at IFCA III



StoRM Improvements I

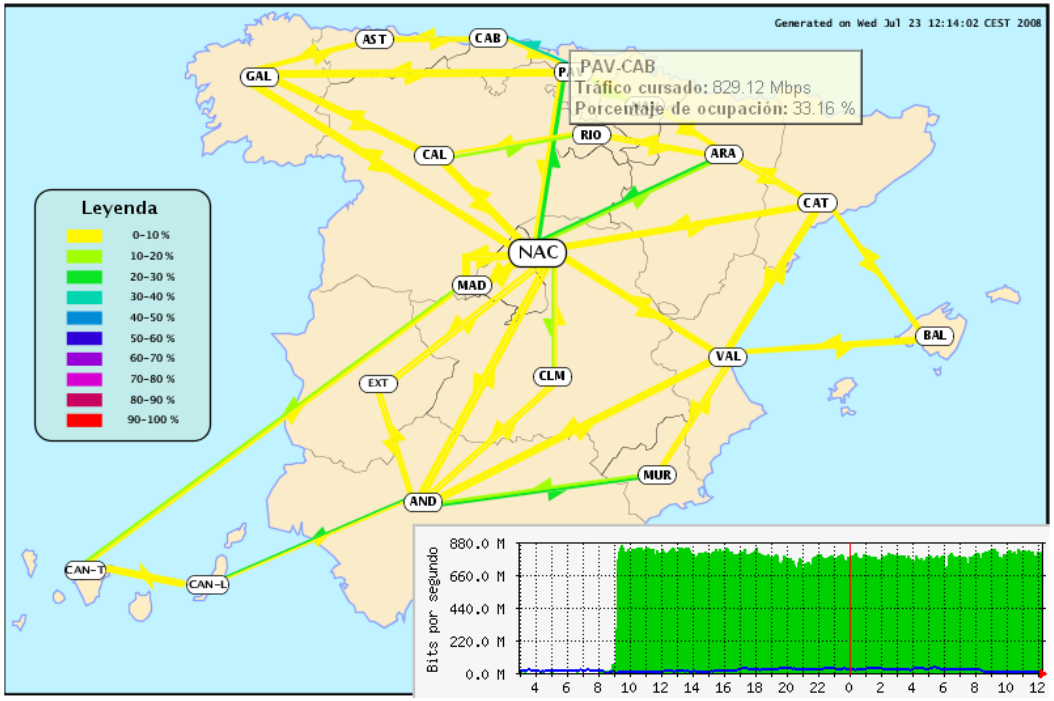
- All services (FE, BE, Mysql and Griftp) on same machine
  - System overloaded with more than 20 connections at the same time
    - High usage of RAM, cached and buffered memory
    - swaping
- 1 node with StoRM Basic services (FE, BE, Mysql) and 4 Gridftp servers
  - StoRM Machine works fine
  - Gridftp's overloaded with more than 20 gridftp processes running
    - High usage of RAM, cached and buffered memory
    - swaping
    - Kernel parameter modification needed

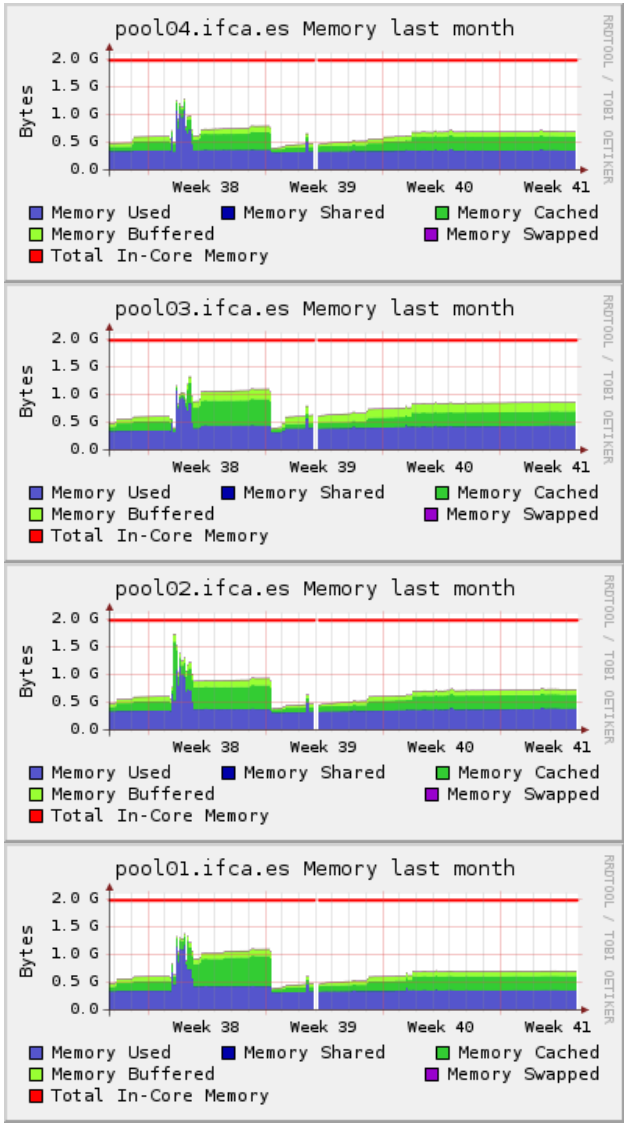
```
net.core.rmem_max = 1048576
net.core.rmem_default = 87380
net.core.wmem_max = 131072
net.core.wmem_default = 32768
net.ipv4.tcp_rmem = 4096 87380 1048576
net.ipv4.tcp_wmem = 4096 32768 131072
net.ipv4.tcp_mem = 65536 87380 98304
```

```
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 improvements II

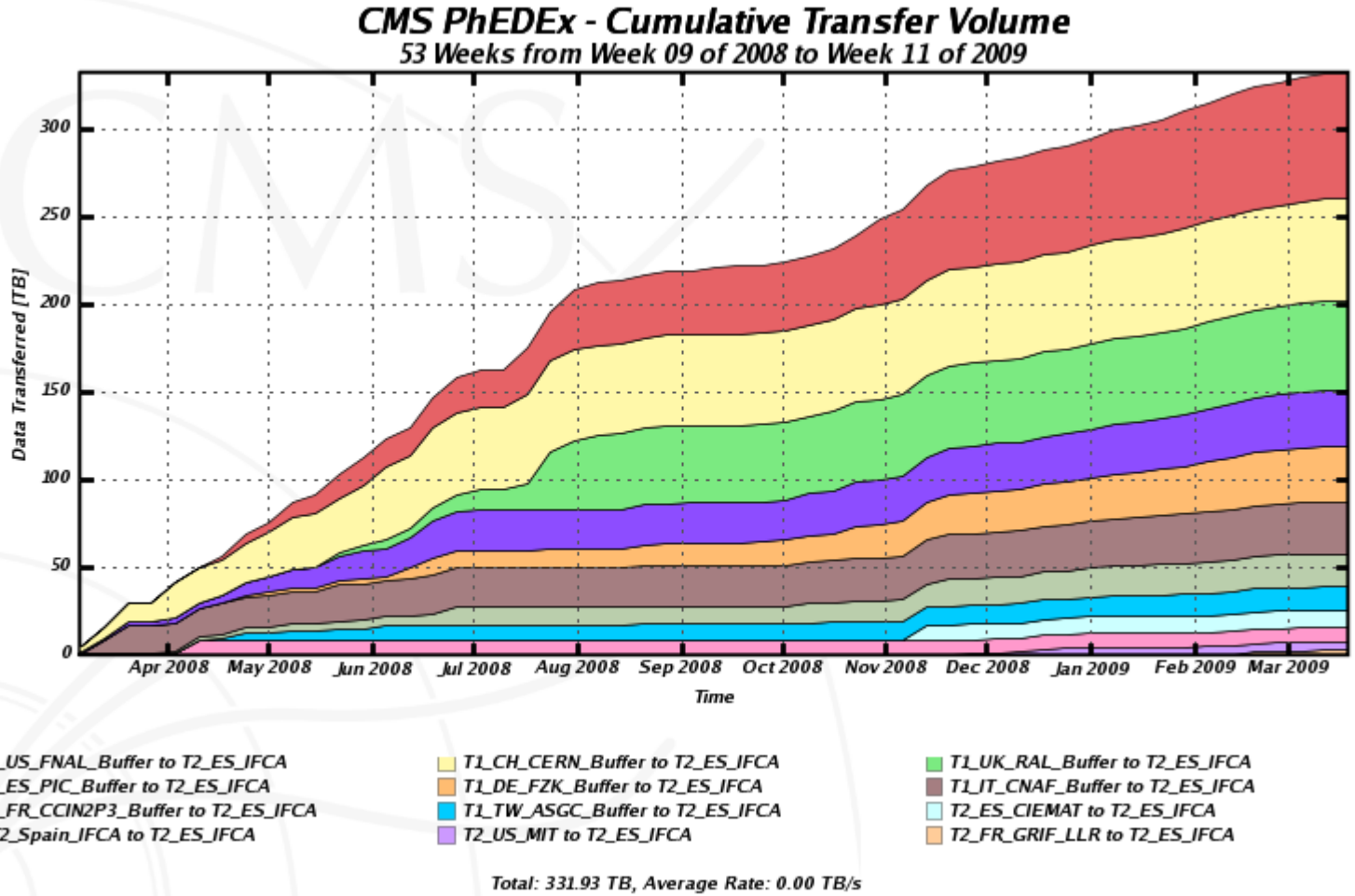
- Optimal results from these modifications
  - No gridftp server over 1GB RAM
  - No more swaping
  - 830 Mb PhEDEx incoming Data during 26 h (1Gb max througput)





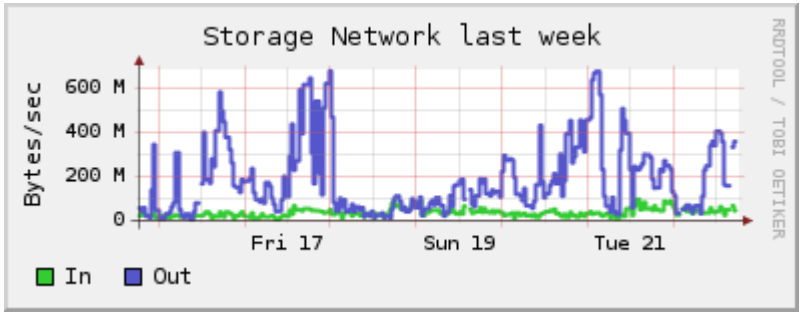
StoRM improvements III

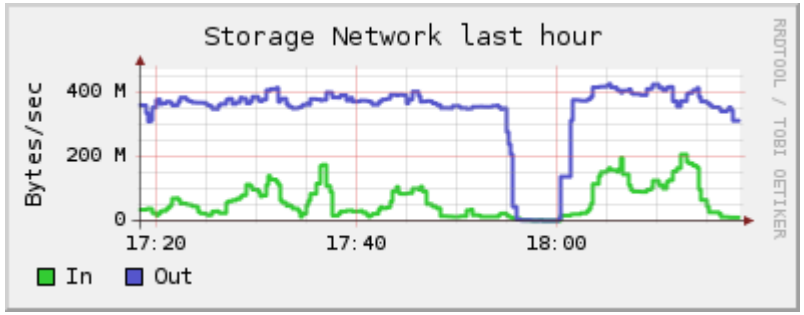
- More than 300TB transfered since we started



Dataflow

- GPFS
  - Is mounted as a local FS on all the WN
  - WN can read/write the FS directly
  - Now limited to 8Gb (2 x 4Gb FC SAN)
  - Soon to be upgraded to 20 Gb/s
- WN
  - 1800 Cores in ~ 230 nodes (10 Gb each 56 nodes) to access GPFS FS





Results

- StoRM
  - Easy to install and easy to maintain
  - Stable (most problems caused by the FS)
  - Need some improvemets in the user manage

- GPFS
  - POSIX access. Do not need to implement other access methods
  - Dificult to optimize
  - Very Stable (when optimized)
  - Good I/O Rate(when optimized)
  - GPFS modules depend on the linux Kernel, each kernel actualization we need to recompile the modules.

The End

¡Thank you very much!