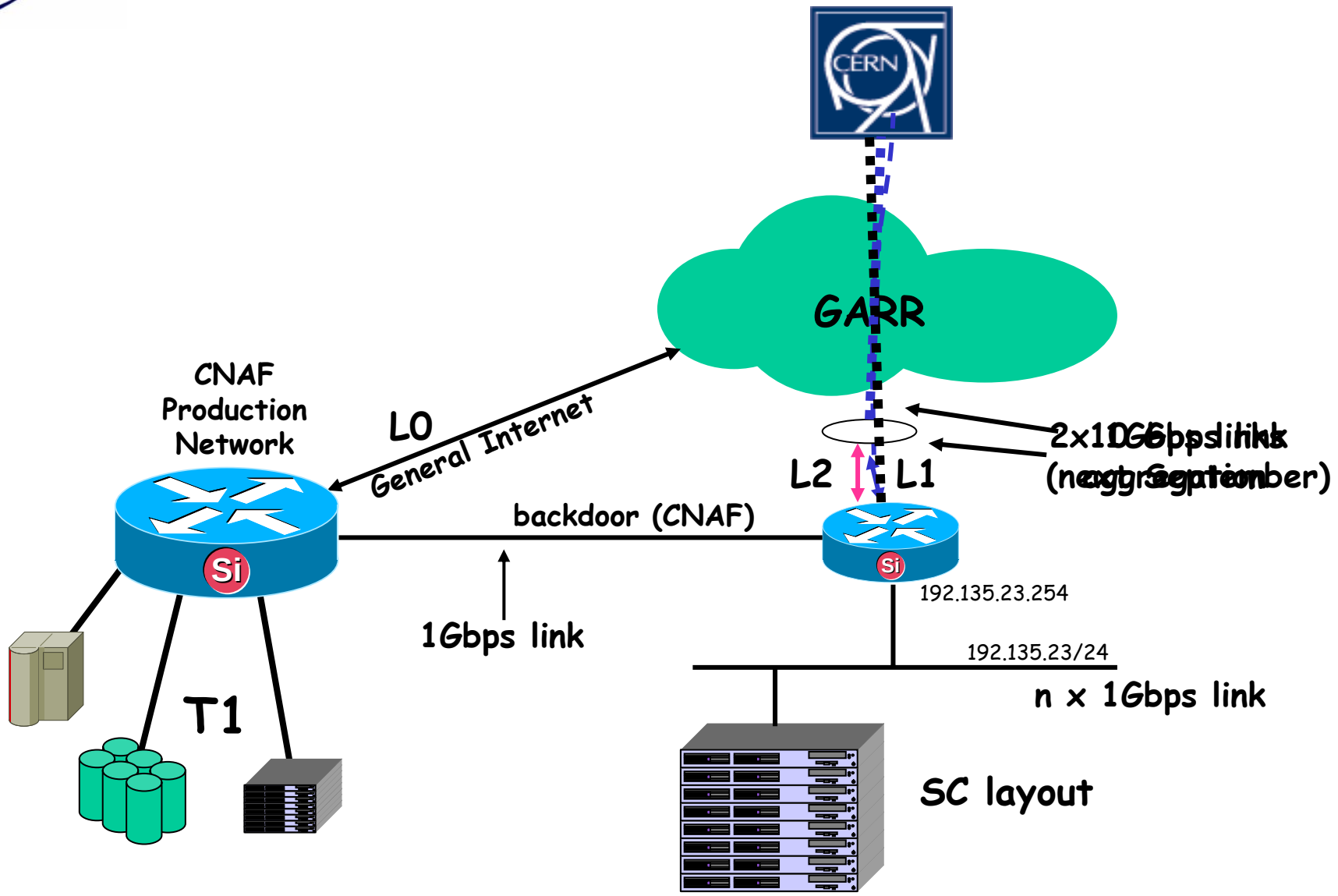
SC3@INFN

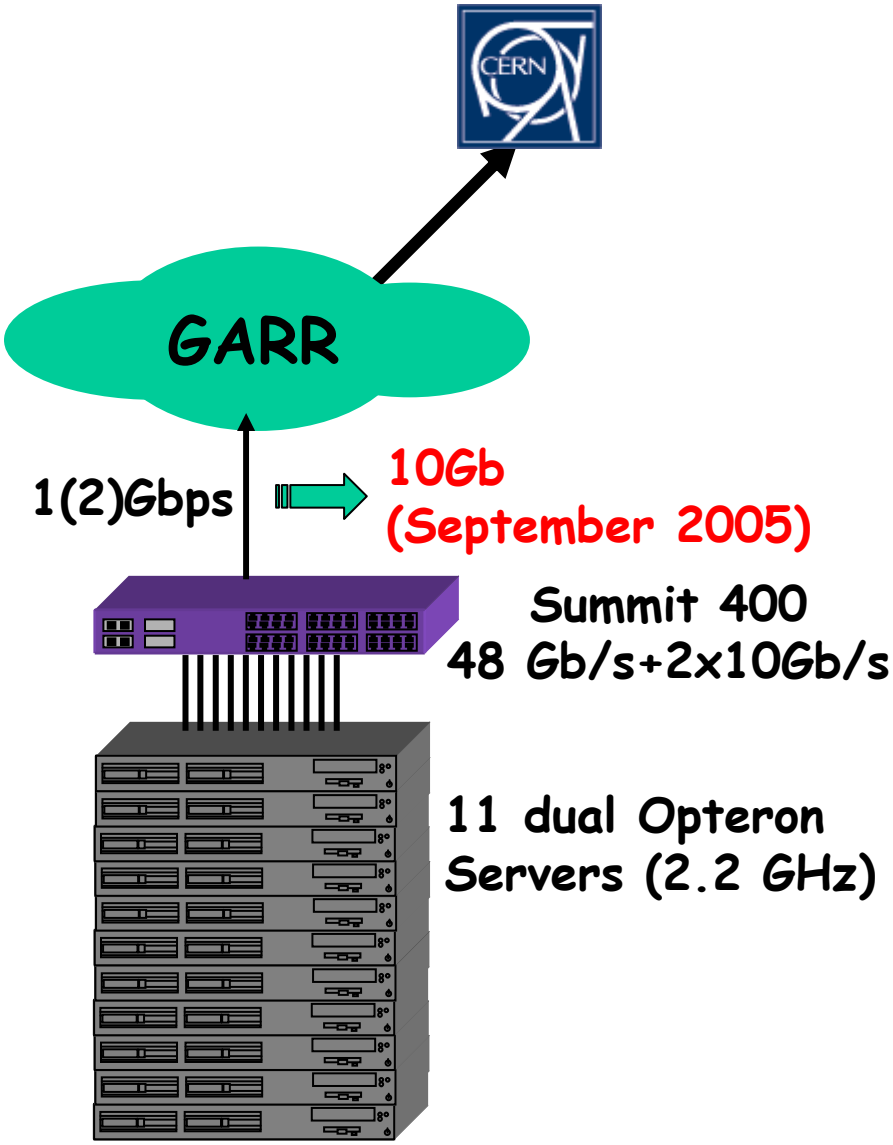
Luca dell'Agnello
INFN-CNAF

LAN & WAN T1 connectivity

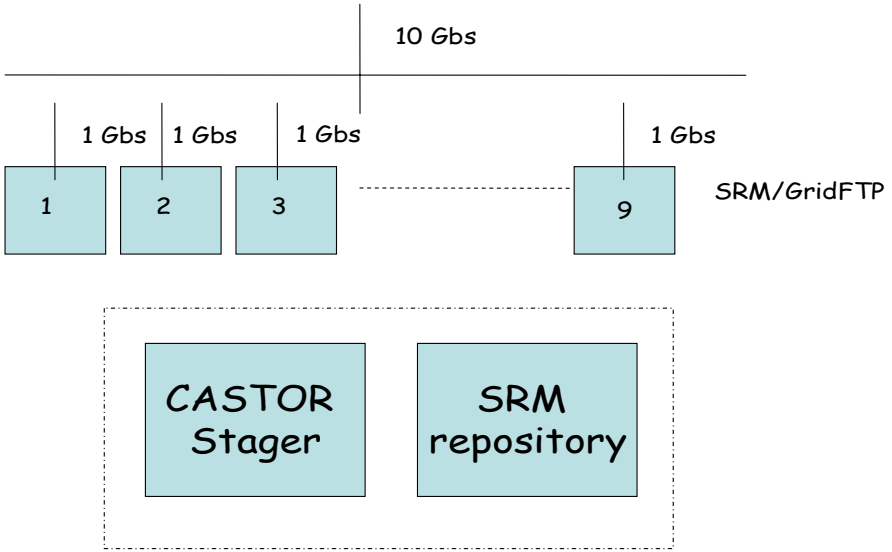


CNAF Transfer cluster (1)

- Extreme Summit 400 (48xGE+2x10GE) dedicated to Service Challenge
- 11 SUN Fire V20 Dual Opteron (2,2 Ghz)
  - 2x 73 GB U320 SCSI HD
  - 2x Gbit Ethernet interfaces
  - 2x PCI-X Slots
- OS: SLC3.03 (arch i386), the kernel is 2.4.21-20.
- LCG 2.5 (Globus/GridFTP v2.4.3, CASTOR SRM v1.4.3-1), Stager CASTOR v1.7.1.5.
  - Need 4or profiles to install LCG 2.5



CNAF Transfer cluster (2)



- 9 GridFTP/SRM servers
  - Load balancing though servers (DNS round-robin algorithm)
    - in order to avoid the "black-hole" effect the most loaded server is taken out from the CNAME (sc.cr.cnaf.infn.it) every 10 minutes
- 1 server for CASTOR Stager/SRM-repository/NAGIOS control system
- 1 FTS server (see next slide for more details)

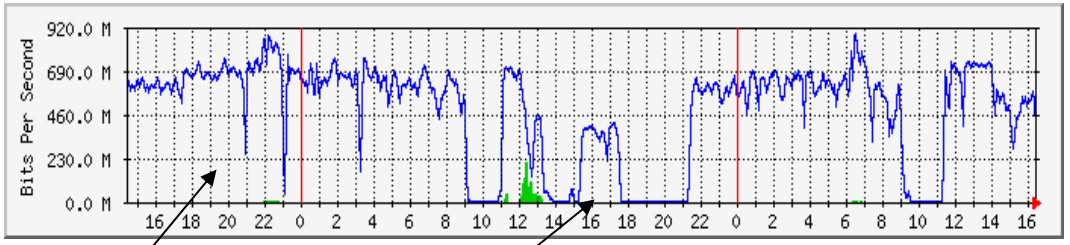
FTS server

- FTS server (sc2.cr.cnaf.infn.it) installed for access from T2s
  - 4 channels configured
    - INFN-LNL (dpm, CMS), INFN-TO (dcache, ALICE), INFN-MI (dpm, ATLAS), INFN-BA (dcache, CMS, ALICE)
  - Basic tests
    - Channel creation, parameters configuration (e.g. # of concurrent transfers, # of TCP streams/transfer)
    - Basic functionalities seem good, experienced many timeouts in the requests that FTS issues to the SRM endpoints, still to understand
  - T2s use the same SC link used for T0→T1 transfers
    - T2s → T1 tests only during T0 → T1 transfers stops
    - Upgrade to 2 Gbps link capacity this week

Storage

- ~ 13 TB Castor stager at T1
  - 2 Infortrend FC raid sets (~ 6 TB each)
  - ~ 600 GB on gridftp servers 2nd disk (used in SC2)
- ~ 35 TB on tape (*in the current set-up for SC3 throughput phase*) at T1
- Distinct Castor paths to ease debug
  - /castor/cnaf.infn.it/grid/lcg/dteam/sc3_notape
    - No migration to tapes (only during 150 MB/s transfers from T0)
  - /castor/cnaf.infn.it/grid/lcg/dteam/sc3
    - Data are migrated to tapes
- #3 9940B (out of 4) + #2 LTO2 will be used in disk → tape phase at T1
- From ~ 2 TB up to ~ 8 TB in each T2 dedicated for SC3

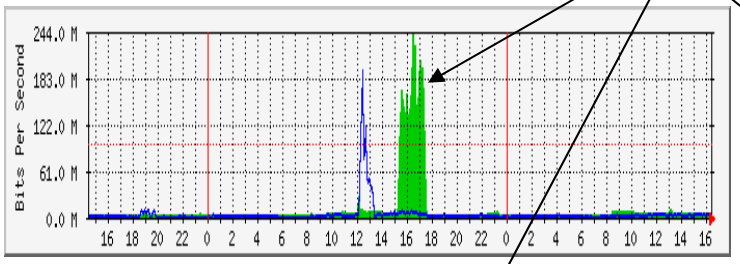
(FTS) Transfers to INFN T1



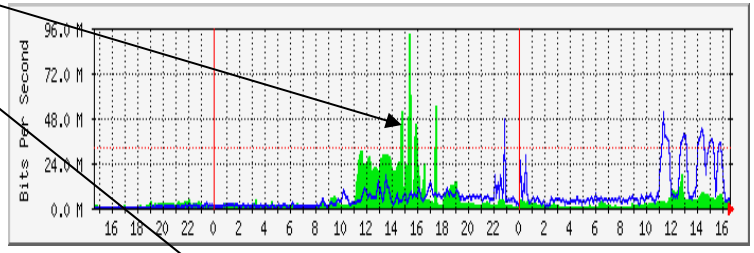
T0 → T1

T2s → T1

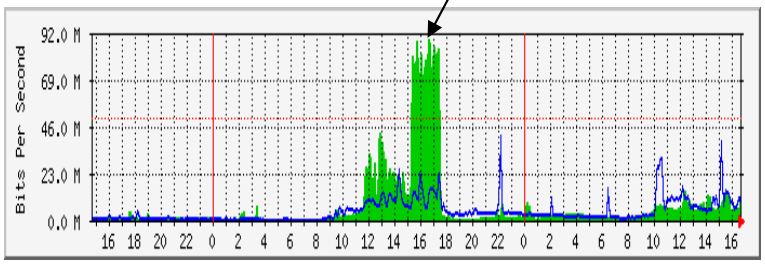
Traffic on INFN-T1 SC link



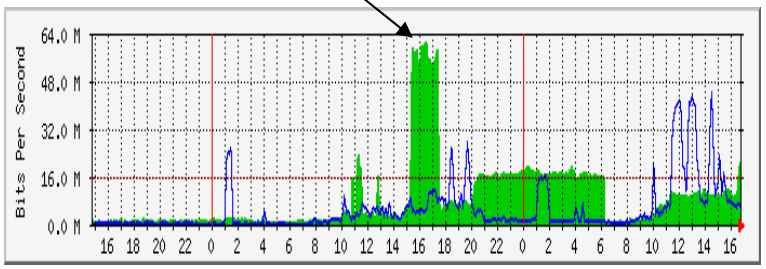
Traffic on INFN-LNL link



Traffic on INFN-MI link



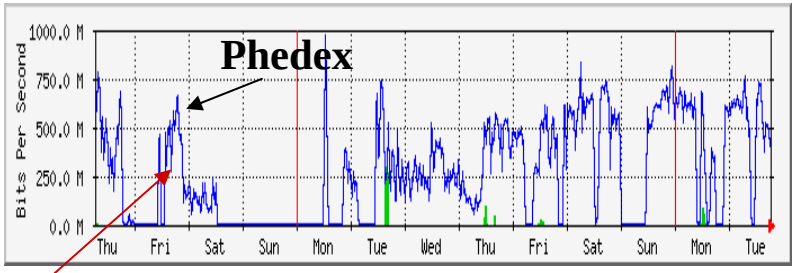
Traffic on INFN-TO link



Traffic on INFN-BA link

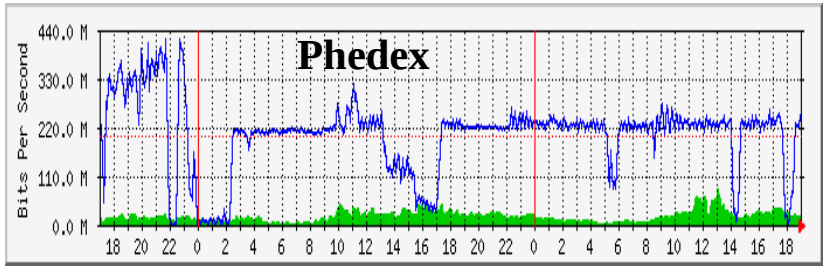
CMS

- Phedex is now testing on production (shared) infrastructure and not on SC3 (dedicated) infrastructure, since asked not to interfere with FTS transfers
  - 1 Gbps production link (but shared among all exp and all CMS activities)
  - 2 classic disk-only SE (no tapes)
  - 1 Castor SE (tapes)
- Up to ~80 MB/s with dedicated SC infrastructure
- ~25 MB/s ~ sustained with production infrastructure
  - To be investigated



Jul/08/05

Traffic on T1 SC link



Traffic on T1 production link

Services status (summary)

- *SE and SRM 1.1:*
  - Castor at T1 (but other SRM under evalutation)
  - DPM at INFN-LNL, INFN-MI
  - Dcache at INFN-TO, INFN-BA
- *File Transfer Service:*
  - client+server INSTALLED (Oracle backend) at T1
  - client installed at T2s
- *PhEDEx*: INSTALLED at T1 and INFN-LNL. Tests: prod-quality with globus-url-copy, testing SRM+Castor end June / beginning July
- *CMS local file-catalogue*:
  - POOL local file catalogue: INSTALLED at T1 and INFN-LNL
- *LFC*:
  - 2.4.0 INSTALLED but not yet tested
  - Migration to LCG 2.5.0 version scheduled for tomorrow
- VO box setup for CMS, ALICE so far
  - In collaboration with experiments
- Myproxy: running

Issues

- Decrease in performance and stability in respect to SC2
  - 1 Gbps link never saturated
  - Problems under investigation at CERN
- Performing FTS tests with "our" T2s
  - Need not to interfere with T0 → T1 trasfers
    - Upgrade of SC3 dedicated link this week
- Phedex is working but not on the SC3 dedicated link


INFN

TCP stack configuration 1/2

- Tuning: function of the available Round Trip Time (18.2 msec)
  - Network Interface **Transmission queue length**: 10000 packets (deafult = 1000)
  - Application **send/receive socket buffer**: ~ 3 Mby (doubled by kernel)
  - **sysctl TCP parameters** tuning

```
net.ipv4.ip_forward = 0
net.ipv4.conf.default.rp_filter = 1
kernel.sysrq = 0
kernel.core_uses_pid = 1
net.ipv4.tcp_timestamps = 0
net.ipv4.tcp_sack = 0
net.ipv4.tcp_rmem = 1048576 16777216 33554432
net.ipv4.tcp_wmem = 1048576 16777216 33554432
net.ipv4.tcp_mem = 1048576 16777216 33554432
net.core.rmem_max = 16777215
net.core.wmem_max = 16777215
net.core.rmem_default = 4194303
net.core.wmem_default = 4194303
net.core.optmem_max = 4194303
net.core.netdev_max_backlog = 100000
```

TCP stack configuration 2/2

iperf TCP Throughput (-w: 2.75 MBy)

Number of Throughput instances extracted: 60

Min/Avg/Max Throughput (Mbit/sec): 90.7 / 878.11 / 951

Variance: 32590.37 Standard deviation: 180.53

Frequency distribution (bins in Mbit/sec):

Bins	N. instances	Percentage
0 , 100 :	1	1.67%
100 , 200 :	0	0.00%
200 , 300 :	0	0.00%
300 ,12 400 :	2	3.33%
400 , 500 :	1	1.67%
500 , 600 :	2	3.33%
600 , 700 :	1	1.67%
700 , 800 :	2	3.33%
800 , 900 :	1	1.67%
900 , 1000 :	50	83.33%

iperf TCP Throughput (-w: 2.75 MBy)

Number of Throughput instances extracted: 61

Min/Avg/Max Throughput (Mbit/sec): 22.3 / 923.51 / 952

Variance: 15572.91 Standard deviation: 124.79

Frequency distribution (bins in Mbit/sec):

Bins	N. instances	Percentage
0 , 100 :	1	1.64%
100 , 200 :	0	0.00%
200 , 300 :	0	0.00%
300 , 400 :	0	0.00%
400 , 500 :	0	0.00%
500 , 600 :	0	0.00%
600 , 700 :	1	1.64%
700 , 800 :	1	1.64%
800 , 900 :	2	3.28%
900 , 1000 :	56	91.80%