Service challenge for LHC

L'impatto del calcolo degli esperimenti LHC sulla infrastruttura di rete
locale, nazionale ed internazionale

Tiziana.Ferrari@cnaf.infn.it

on behalf of the SC INFN Team

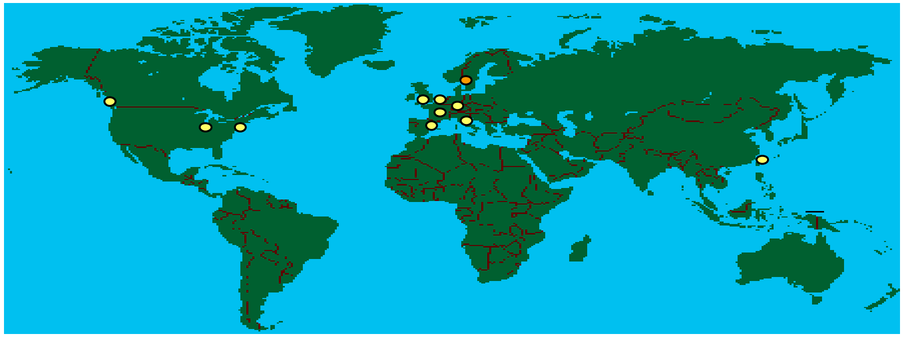
GARR Workshop, Roma, 18 Nov 2005

Outline

- **Introduction to the LHC Service Challenge**
- The LHC Optical Private Network
- Results:
  - CERN – CNAF: TCP and GridFTP performance
  - CNAF – INFN Tier-2: performance testing
  - Differentiated Services and LHC
  - 10 GigaEthernet testing
- Conclusions

The LHC computing model

- Data source: the Large Hadron Collider - CERN
- Data analysis:
  - based on the Grid computing paradigm
  - hierarchical organization of Grid computing sites distrbuted all over the world:
    - TIER 0 → CERN
    - TIER 1 → Academia Sinica (Taipei) , Triumf (CA), IN2P3 (FR), Forschungszentrum Karlsruhe (DE), CNAF (IT), NIKHEF (NL), PIC (SP), CLRC (UK), Brookhaven and FermiLab (US)
    - TIER 2 → Bari, Catania, Legnaro, Milano, Pisa, Torino, ... (around 100 sites in 40 counties)
    - TIER 3 → ...



Bandwidth requirements (CNAF)

- Nominal rate sustained: 200 MBy/s CERN disk → CNAF tape
  - *Raw figures* produced by multiplying e.g. event size x trigger rate
  - *Headroom*: a factor of 1.5 that is applied to cater for peak rates → 300 MBy/s
  - *Efficiency*: a factor of 2 to ensure networks run at less than 50% load → 600 MBy/s
  - *Recovery*: a factor of 2 to ensure that backlogs can be cleared within 24 – 48 hours and to allow the load from a failed Tier1 to be switched over to others → 1200 MBy/s

  - Total requirement: 10 Gb/s to/from every Tier-1 centre for reliable bulk data exchange
    - Tier-0 → Tier-1s for **raw** and **1st pass** reconstructed data
    - Tier-1 → Tier-0 and other Tier-1s for **reprocessed** data and **replication**

Service challenge: purpose

- Understand what it takes to operate a real Grid service – run for days/weeks at a time (outside of experiment Data Challenges)
- Trigger/encourage the Tier1 & large Tier-2 planning – move towards real resource planning – based on realistic usage patterns
- Get the essential Grid services ramped up to target levels of reliability, availability, scalability, end-to-end performance
- Data management, batch production and analysis by April 2007

CNAF

- From Nov 2005 to Oct 2006:
  - Data disk: 50 TBy (Castor front-end)
    - → 350 TBy (dCache, Castor2, StoRM on GPFS)
  - Tape: 200 TBy
    - → 450 TBy
  - Computing (farm is shared): min 1200 kSI2K - max1550 kSI2K
    - →2000 KSI2k, max 2300 KSI2k
  - Network connectivity: 2 x 1 GigaEthernet (dedicated)
    - →10 Gb/s guaranteed bandwidth CERN – CNAF (Nov 2005)
    - → additional 10 Gb/s: CNAF – Karlsruhe (backup), Tier-1 to Tier-2 connectivity

- Introduction to the LHC Service Challenge
- **The LHC Optical Private Network**
- Results:
  - CERN – CNAF: TCP and GridFTP performance
  - CNAF – INFN Tier-2: performance testing
  - Differentiated Services and LHC
  - 10 GigaEthernet testing
- Conclusions

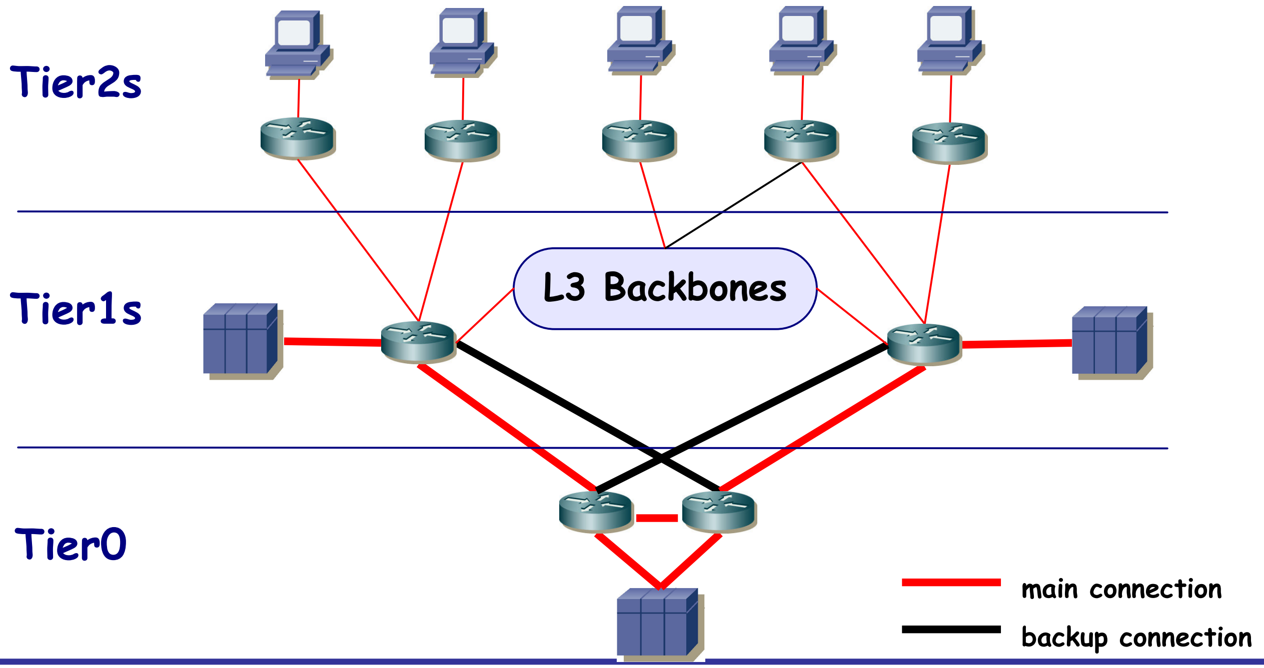
LHC Optical Private Network (OPN)

- Activities:
  - control the implementation plans of WAN connectivity, as requested from the various LHC Computing Models
  - ensure that individual agreements among T1s will provide a coherent infrastructure to satisfy the LHC experiment requirements
  - address the problem of management of the end-to-end network services
- First priority: to plan the networking for the Tier-0 and Tier-1 centers

  → The Service Challenges will test the overall system (from network to applications) up to full capacity production environment

LHC Optical Private Network: architecture

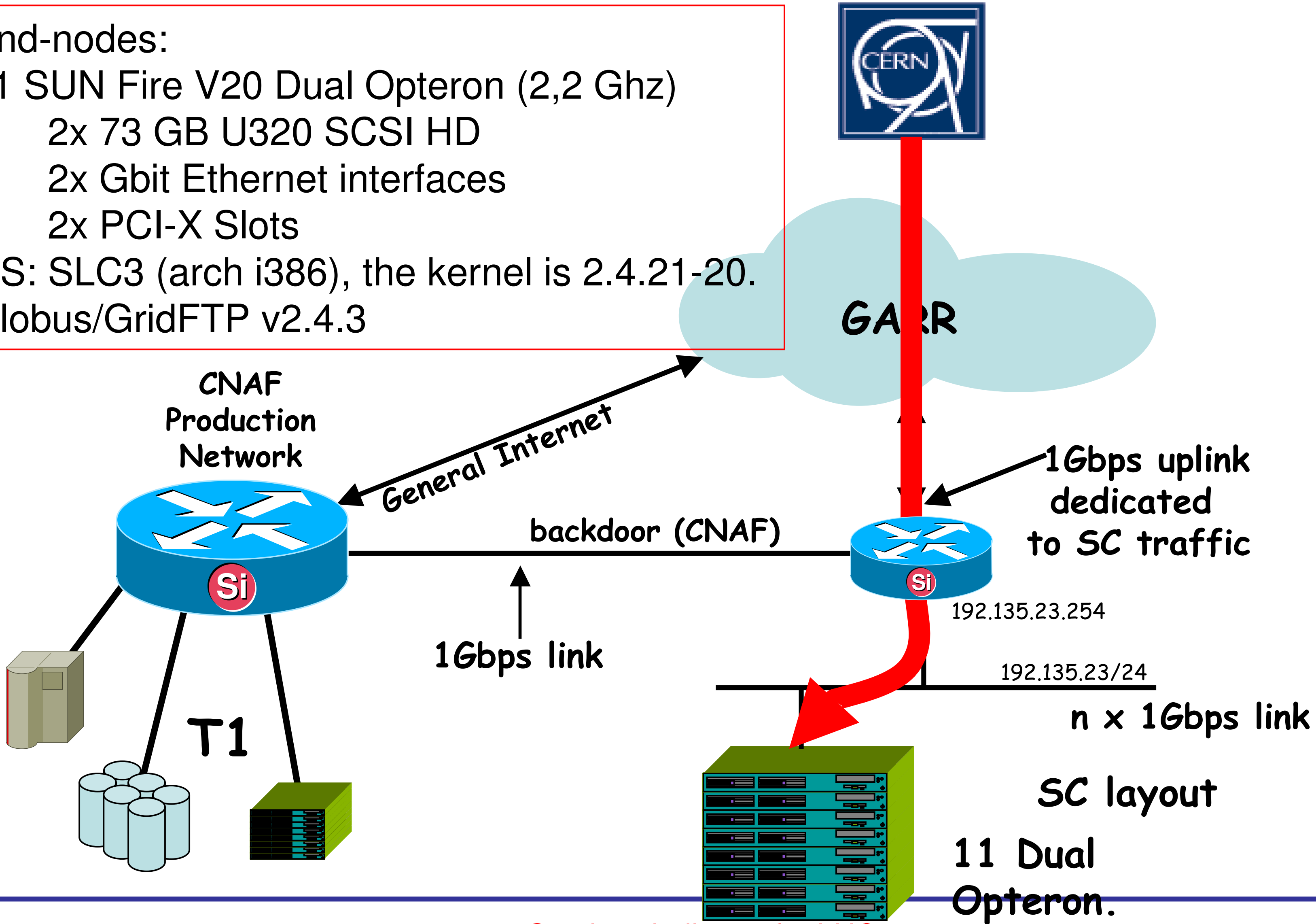
- At least one dedicated 10 Gbit/s light path between T0 and each T1
  - every **T0 -T1 link** should handle **only production LHC data**
  - T1 to T1 traffic via the T0 allowed BUT T1s encouraged to provision **direct T1-T1 connectivity**
  - **T1 - T2** and **T0 - T2** traffic handled by the normal **L3 connectivity provided by NRENs**
  - T2s usually upload and download data via a particular T1
- **Backup through L3 paths across NRENs discouraged** (potential heavy interference with general purpose Internet connectivity of T0 or the T1s)



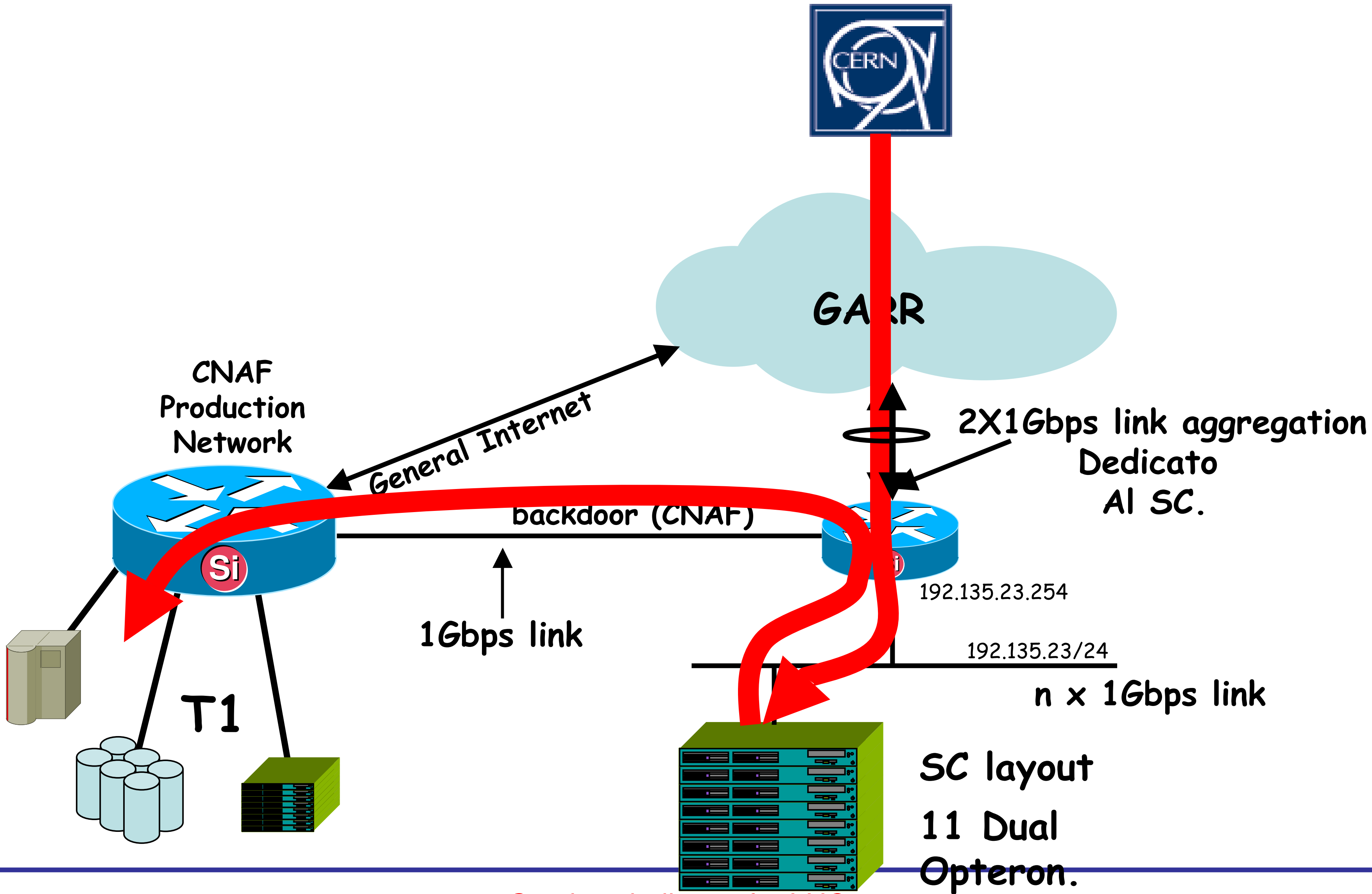
SC2: CNAF network layout

End-nodes:
11 SUN Fire V20 Dual Opteron (2,2 Ghz)
- 2x 73 GB U320 SCSI HD
- 2x Gbit Ethernet interfaces
- 2x PCI-X Slots

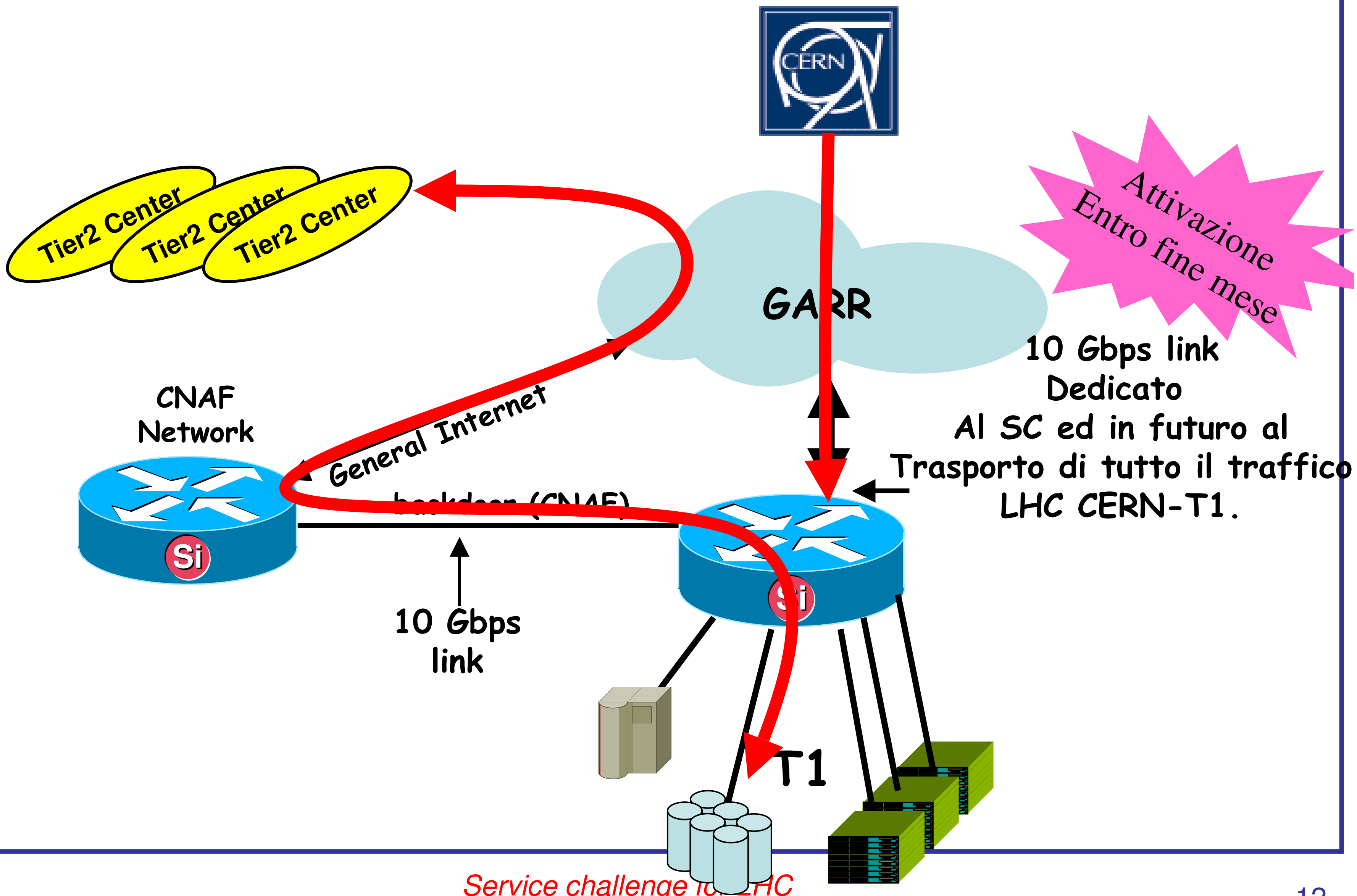
OS: SLC3 (arch i386), the kernel is 2.4.21-20.
Globus/GridFTP v2.4.3



SC3: CNAF network Layout



SC3 and 4: CNAF Network Layout



- Introduction to the LHC Service Challenge
- The LHC Optical Private Network
- Results:
  - **CERN – CNAF: TCP and GridFTP performance**
  - CNAF – INFN Tier-2: performance testing
  - Differentiated Services and LHC
  - 10 GigaEthernet testing
- Conclusions

TCP stack configuration 1/2

- Tuning: function of the available Round Trip Time (18.2 msec)
  - Network Interface **Transmission queue length**: 10000 packets (default = 1000)
  - Application **send/receive socket buffer**: ~ 3 Mby (doubled by kernel)
  - **Other sysctl TCP parameters** tuning
  - PCI slot of NIC: 64 bit

```
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: 3.0 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%
```

TCP performance issues on LAN

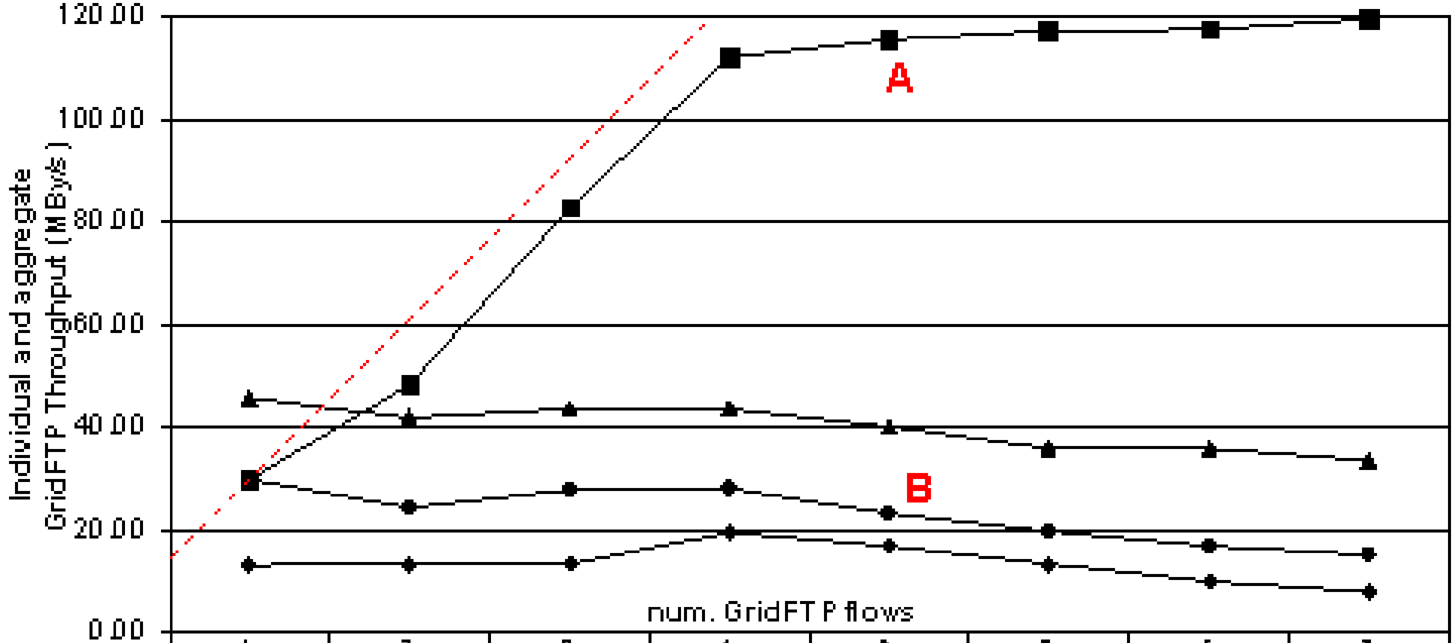
- CERN cluster:
  - Sporadic loss on LAN
  - non-zero errors/dropped/overruns counters on transmitting interface
  - Error counters increasing during throughput-intensive test sessions
  - In case of high-speed memory-to-memory data transfer sessions, packet loss is related to the concurrent running of monitoring processes, which collect network statistics by accessing system files such as /proc/net/tcp
  - Problem only affecting the CERN hosts

TCP performance on WAN 1/2

- WAN connections affected by sporadic packet loss in both directions
  - memory-to-memory throughput above 900 Mb/s only [83 – 90]% of the time
  - No use of dedicated network paths apart from the CERN/CNAF uplinks
  - Network performance CNAF → CERN often non-deterministic
  - Problem solving extremely complex
  - 24-hour memory-to-memory throughput:
    - avg: 900 Mb/s, max: 950 Mb/s

GridFTP on WAN (CERN → CNAF)

- Individual GridFTP performance disk – disk (moderate disk utilization):
  - extremely variable: [15, 40] MBy/s
- Minimum num of GridFTP sessions for saturation: 6 (single session for every couple of tx/rx nodes)



	1	2	3	4	5	6	7	8
min	12.99	13.16	13.16	19.29	16.87	13.16	9.20	7.89
avg	29.87	24.04	27.58	28.00	23.08	19.51	16.79	14.94
max	45.45	41.87	43.43	43.43	40.00	35.71	35.71	33.33
aggregate	29.87	48.08	82.89	112.01	115.40	117.08	117.52	119.50

Application tuning: the load-balancing problem

- Load balancing:
  - *Relies on a homogeneous distribution of sessions to destination servers based on the DNS*;
  - Requires destination servers to support *equal write performance* to disk, otherwise:
    - the number of open sessions **tends to increase on low-performance servers**
    - **the larger** the number of open sessions, **the lower** the overall performance of the server

→ **Results**:

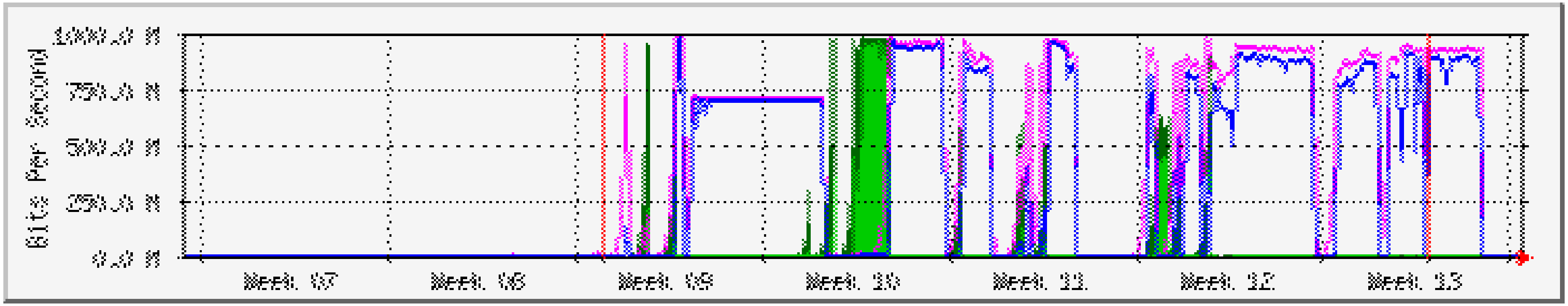
- black hole phenomenon
- the number of concurrent gridFTP sessions needed to saturate the available bandwidth grows
- uneven distribution of gridFTP sessions to serve

→ **Solution**:

- DNS + removal from cname of the busiest server
- load is a function of the number of pending gridFTP sessions

Overall SC2 throughput results

- daily average throughput of **500 MBy/s** achieved for approximately **ten days**
- INFN performance results:
  - Average throughput: 81.54 MBy/s
  - Overall amount of data moved from CERN: 67.19 TBy
- SC2 input/output traffic to/from CNAF of Service Challenge Phase 2 (CER → CNAF traffic: blue)



- Introduction to the LHC Service Challenge
- The LHC Optical Private Network
- Results:
  - CERN – CNAF: TCP and GridFTP performance
  - **CNAF – INFN Tier-2: performance testing**
  - Differentiated Services and LHC
  - 10 GigaEthernet testing
- Conclusions

Network performance testing Tier-1 ←→Tier-2

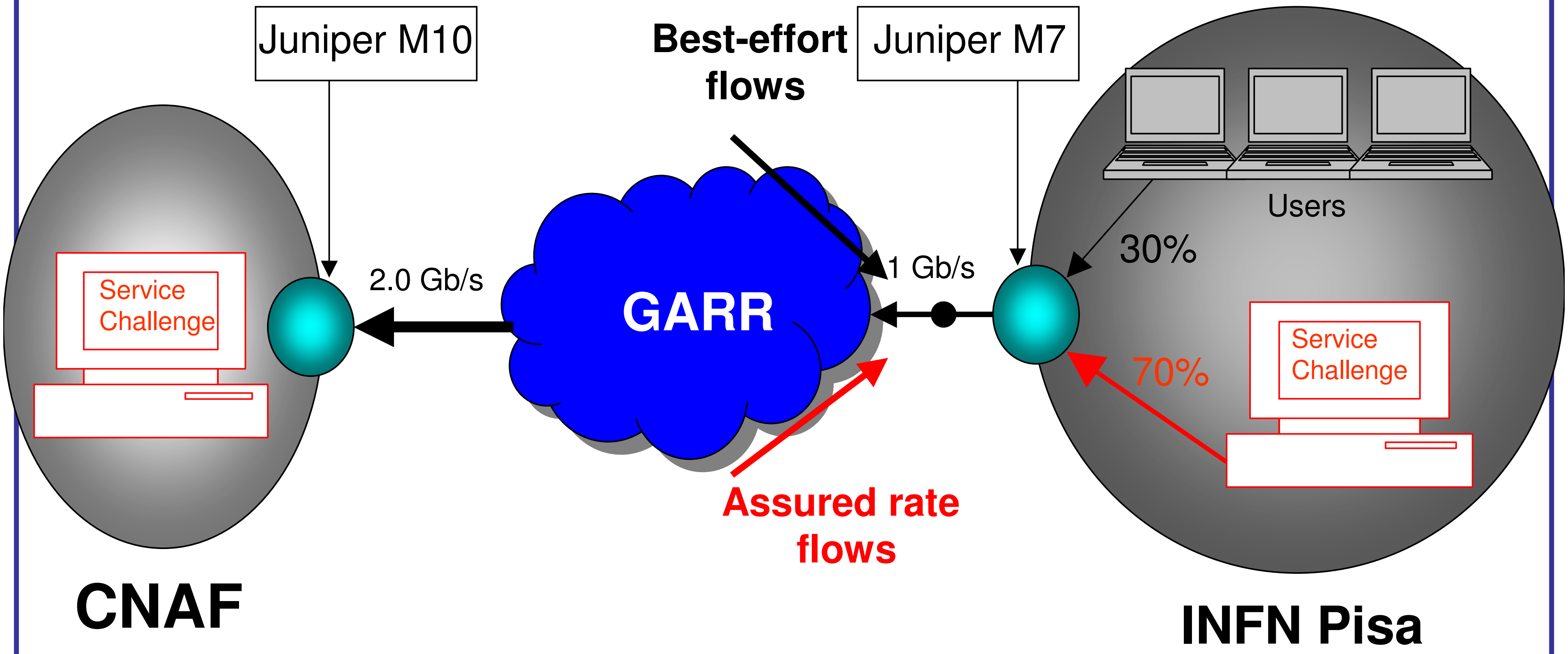
- Purpose of SC3:
  - Tuning of number of parallel streams per GridFTP session and of concurrent GridFTP transfers, from CERN to every Tier-1
  - Integration of Grid Data Management services with the application sw
  - Networking (INFN only):
    - Asymmetric performance: CNAF – INFN Pisa (1 GigaEthernet)
      → Network configuration fixed
    - High CPU utilization on CE router (buggy IOS version): CNAF – Torino (1 GigaEthernet)
    - IOS upgrade
    - Hardware and software issues on CE equipment:
      - CNAF – Bari (FastEthernet)
      - CNAF – Catania (1 GigaEthernet)

- Introduction to the LHC Service Challenge
- The LHC Optical Private Network
- Results:
  - CERN – CNAF: TCP and GridFTP performance
  - CNAF – INFN Tier-2: performance testing
  - **Differentiated Services and LHC**
  - 10 GigaEthernet testing
- Conclusions

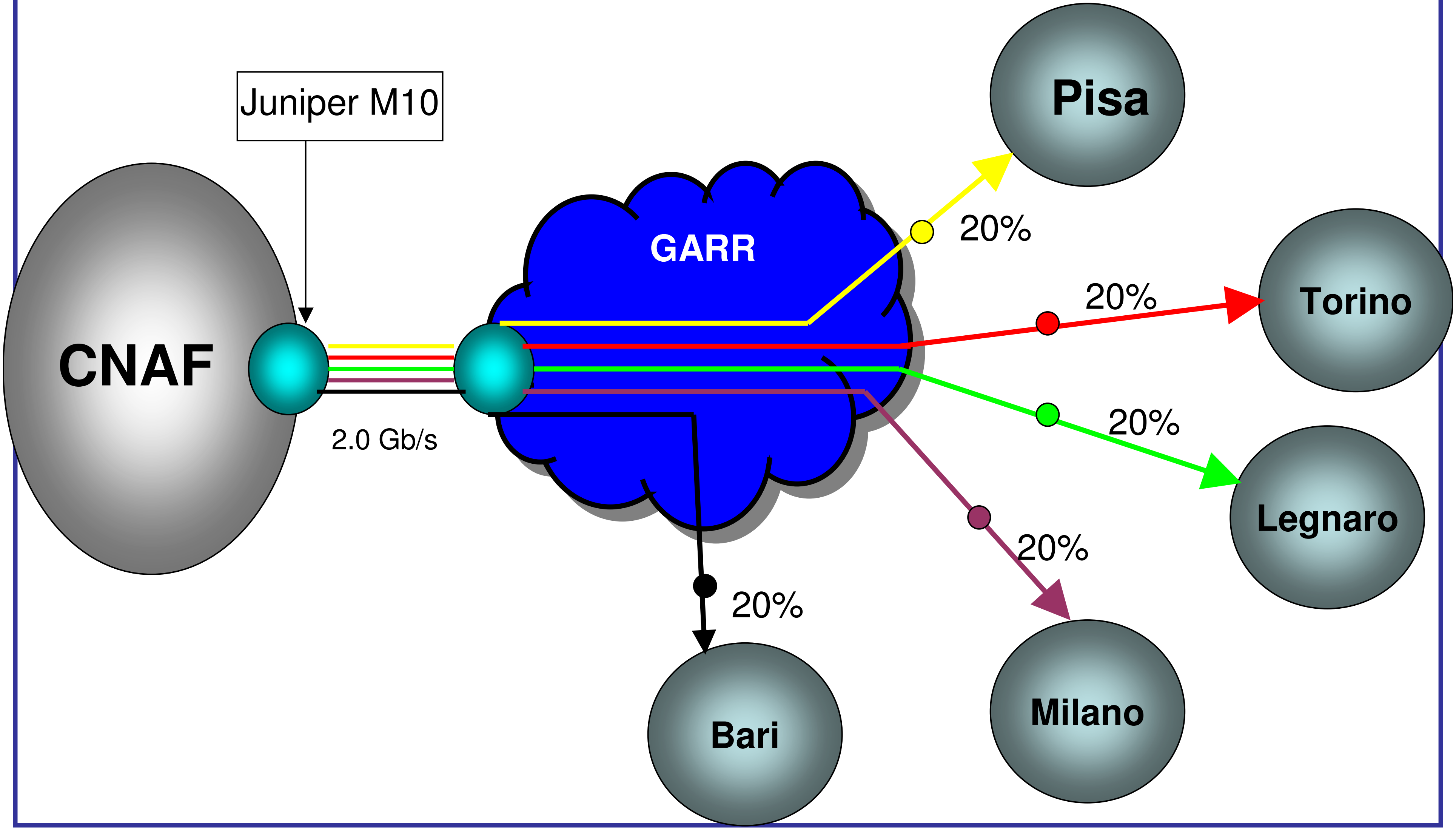
Guaranteed capacity on WAN Tier-1 ← → Tier-2

- Problems:
  - Interaction of incoming/outgoing bursty SC data traffic with legacy traffic
  - Fair distribution of incoming/outgoing bandwidth to/from CNAF from/to the INFN Tier-2 sites
- Differentiated Services → testing CNAF – Pisa
  - 2 traffic classes: guaranteed bandwidth (LHC) and best-effort (legacy)
  - Weighted Round Robin scheduling
  - LHC → 70% of link capacity (minimum)
  - Legacy traffic → 30% of link capacity

DiffServ deployment scenarios (1)



DiffServ deployment scenarios (2)



- Introduction to the LHC Service Challenge
- The LHC Optical Private Network
- Results:
  - CERN – CNAF: TCP and GridFTP performance
  - CNAF – INFN Tier-2: performance testing
  - Differentiated Services and LHC
  - **10 GigaEthernet testing**
- Conclusions

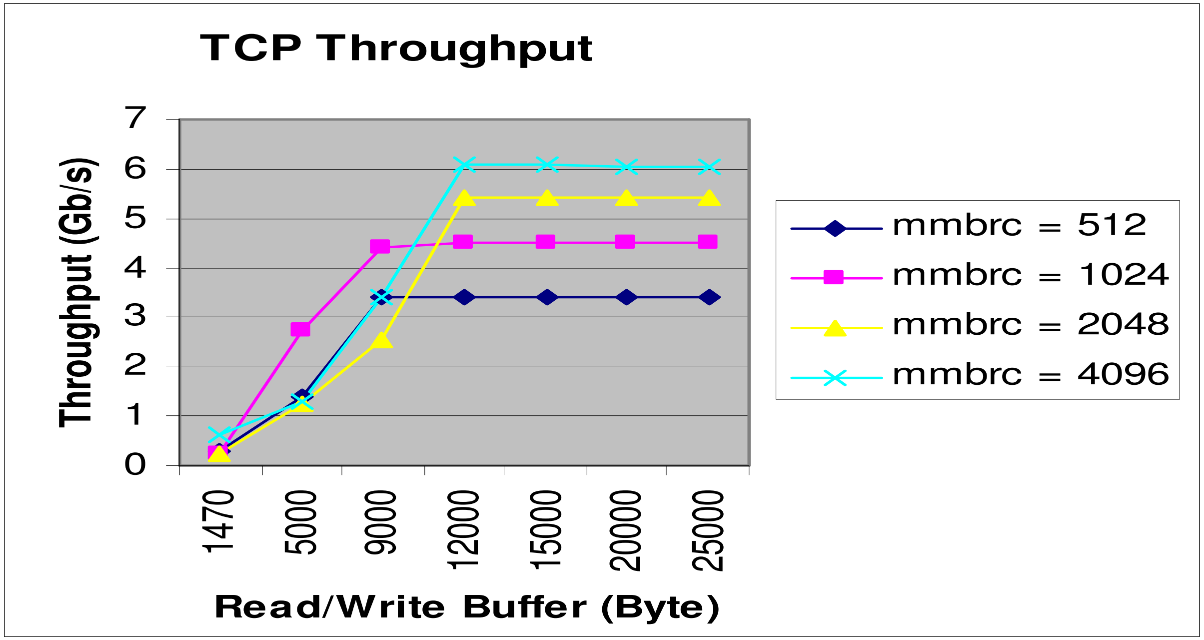
3. 10 GigaEthernet performance

- SUN Fire V20z Server
  - **Processors**: 2 single-core AMD Opteron 252 (2,6 GHz)
  - L2 **Cache** per Processor: 1 MB
  - **Memory**: 4 GB (4 * 1-GB DIMMS)
  - Two **64-bit PCI-X slots** : One full-length at 133 MHz; One half-length at 66 MHz
  - **Operating System**: Scientific Linux Kernel 2.4.21

- Intel Pro 10GE Server Adapter
  - Controller MAC PCI-X 10GE
  - Intel® 82597EX a **133 MHz/64-bit**
  - **16 KByte** maximum packet size (**Jumbo Frame**)
  - Conformity to PCI-X 1.0a and PCI 2.3

Application/kernel/hardware tuning

- Read/write buffer size (number of software interrupts) → 12000 by
- Send/receive socket size → window: 32 Mby
- NIC transmission queue/receive backlog: 100000 packets
- PCI mmbrc (max memory byte read count): part of the PCI-X Command Register, sets the maximim byte count the PCI-X device may use when initating a Sequence with one of the **burst read commands** (value range **512-4096 Byte**) → 4096
- MTU → 9216 by



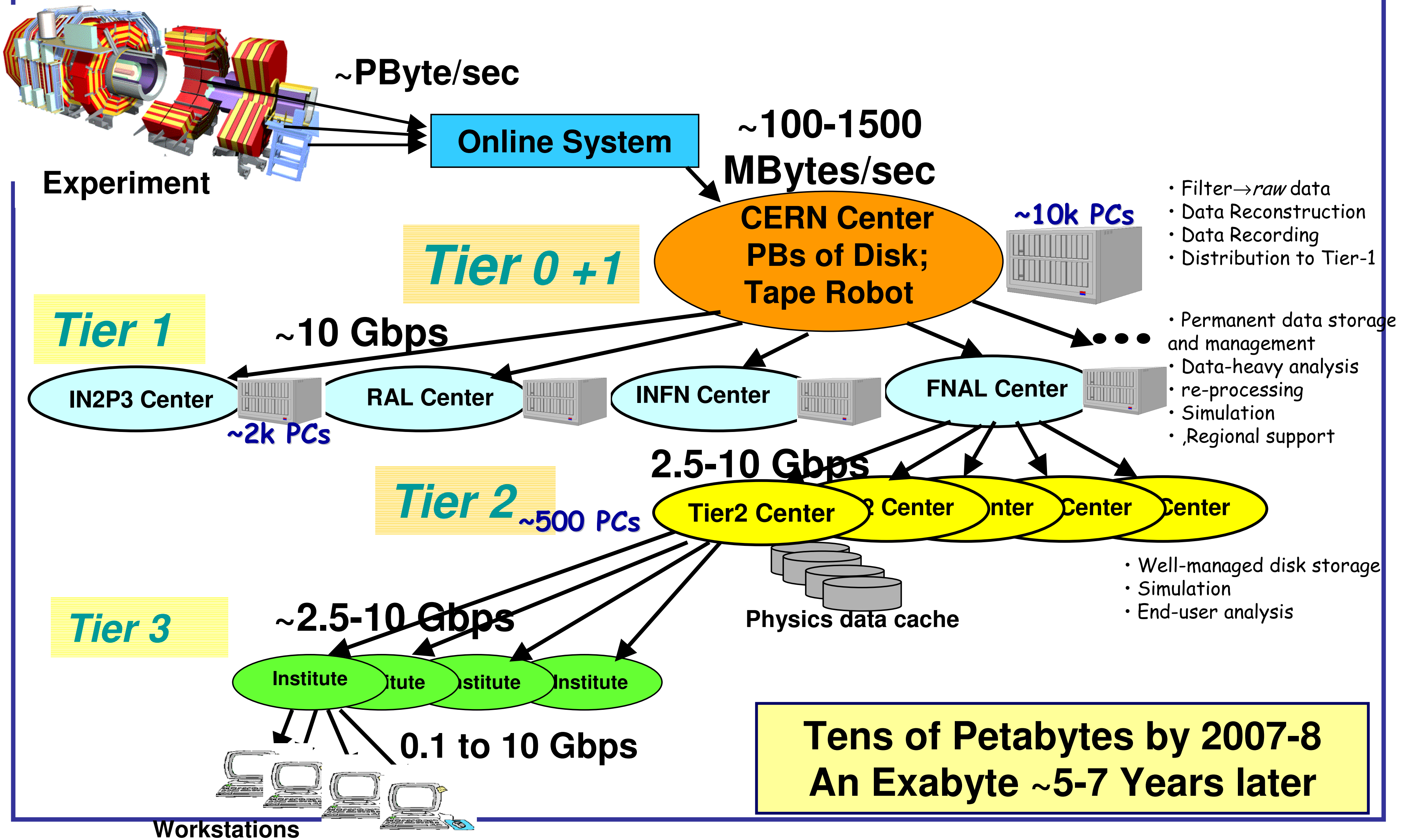
- Introduction to the LHC Service Challenge
- The LHC Optical Private Network
- Results:
  - CERN – CNAF: TCP and GridFTP performance
  - CNAF – INFN Tier-2: performance testing
  - Differentiated Services and LHC
  - 10 GigaEthernet testing
- Conclusions

Conclusions

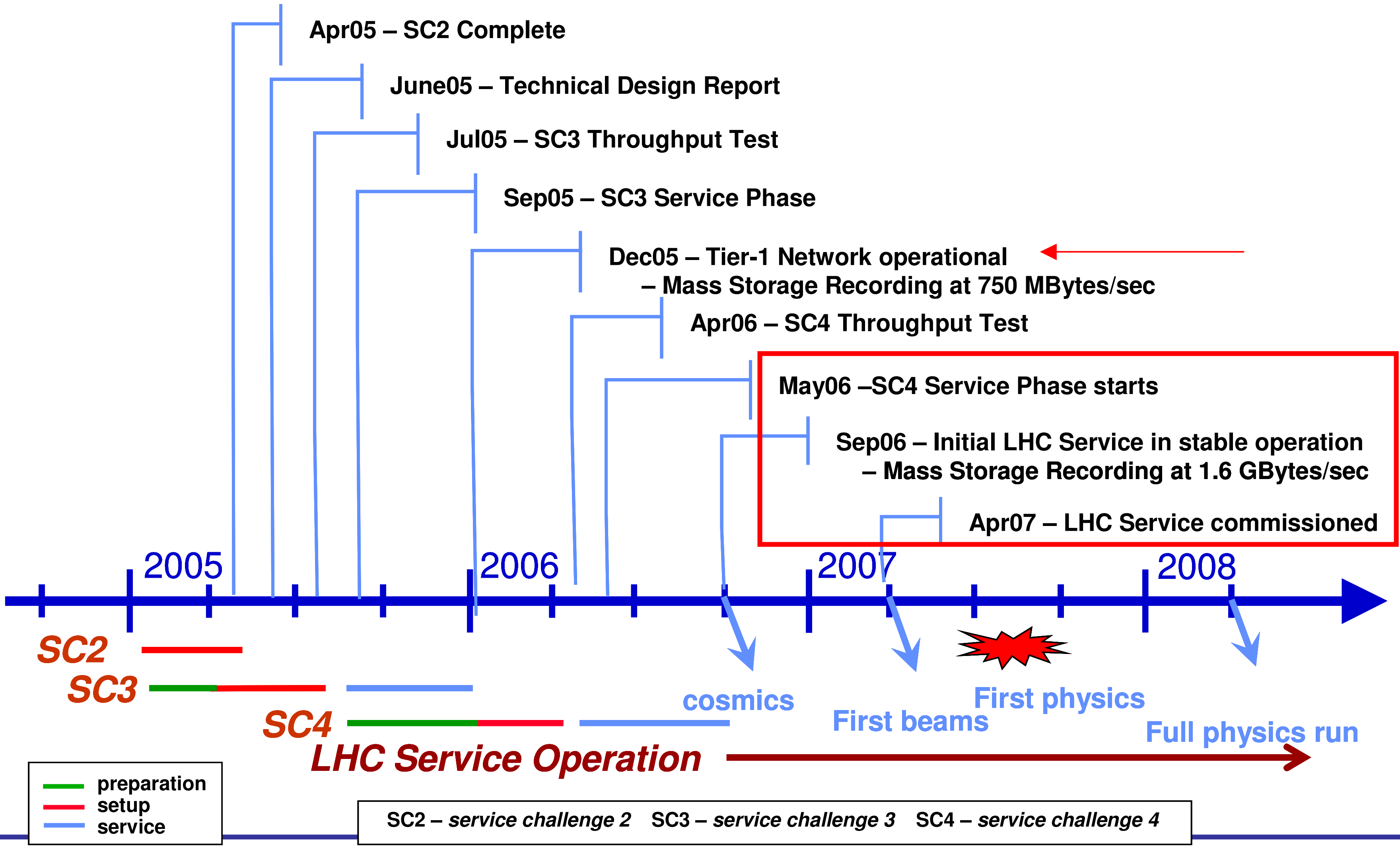
- Overlay private networks and high-speed connectivity in the LAN (10 GE) are becoming reality
- Differentiated Services still useful for bandwidth control at the customer – provider connection point, and for L3 WAN guaranteed bandwidth
- Importance of proper application/kernel/hardware tuning
- High-performance bulk data transfer is (also) strongly affected by:
  - hw/sw reliability
  - End-system hardware configuration
  - Efficiency of Grid data management software
  - Disk and tape read/write performance
- High-speed connectivity in the LAN (10 GE) becoming reality

Backup slides

Data distribution model

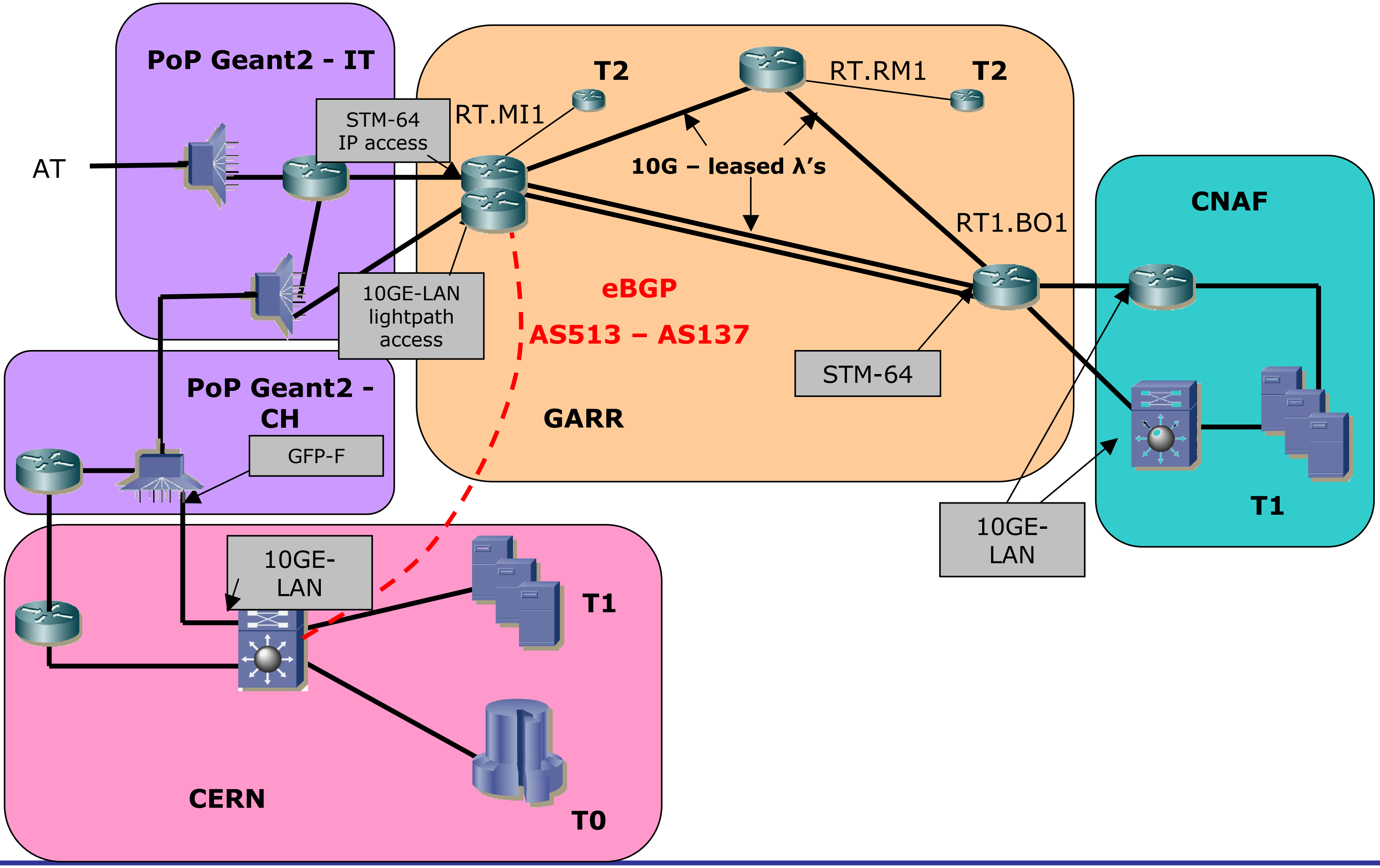


Service challenge: roadmap



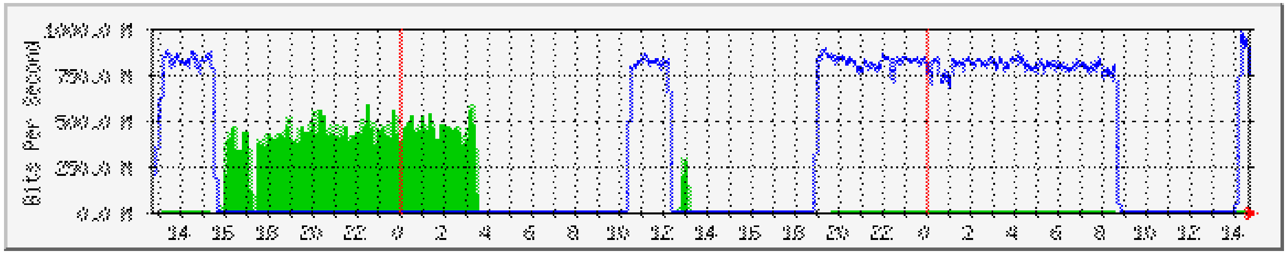
SC2 – *service challenge 2* SC3 – *service challenge 3* SC4 – *service challenge 4*

Italian LHC OPN Architecture

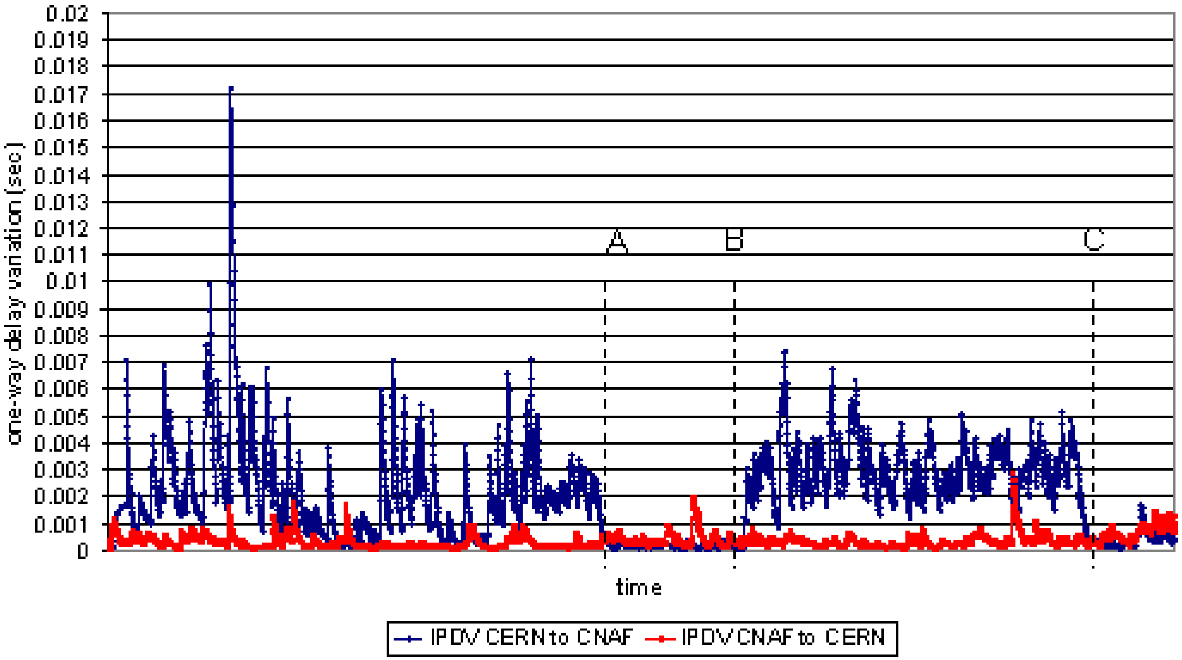


TCP performance on WAN 2/2

- Network load asymmetry:
  - Green: CNAF → CERN, blue: CERN → CNAF



- Variable queuing time under load (around 1 Gb/s)



Write-to-tape performance 1/4

- four *tape servers* and four *LTO2 IBM drives* accessing a *tape library StorageTek 5500*
- expected performance of one LTO2 IBM drive: around 15-20 MBy/s in case of “real” data (root files, with internal compression)
- *CASTOR stager* v. 1.7.1.5 (on sc1.cr.cnaf.infn.it)
- *Nagios* configured in order to handle alarms about CASTOR services (stager and *rfiod*) and about disk/tape space occupation

→ Target: 60 Mby/s to tape

Write-to-tape performance 2/4

- Two concurrent gridFTP transfer sessions CERN → CNAF sufficient for 60 Mby/s to tape sustained
- 24 h test
- Size of files written to tape: 1 Gby (favourable scenario)
- Observations:
  - two long down-times of the Radiant database storing the job data transfer queue at CERN
  - hardware failures of the tape system at CNAF:
    - One LTO2 IBM drive crashed
    - two tapes marked “disabled” during the tests

    → from 60 MBy/s to approximately 55 MBy/s (24-hour average)
  - CASTOR disk pool tends to run out of space (write-to-tape performance is the bottleneck)

Write-to-tape performance 3/4

- Observations (cont):
  - Tape servers configured to generate large streams → if the amount of available space on each tape is not sufficient to store an entire stream, even if the overall amount of free space on tape pool is sufficient for the stream, CASTOR cannot dynamically resize streams or distribute an individual stream to different tapes
  - Write-to-tape freeze → increase of capacity through addition of new tapes
  - overall amount of data written: 4 TB
  - Avg throughtput: 55 Mby/s
- Conclusions:
  - Quality of tape drives is fundamental
  - Tape stream size needs careful tuning