INFN TIER1
(IT-INFN-CNAF)
“Concerns from sites” Session

LHC OPN/ONE

“Networking for WLCG” Workshop

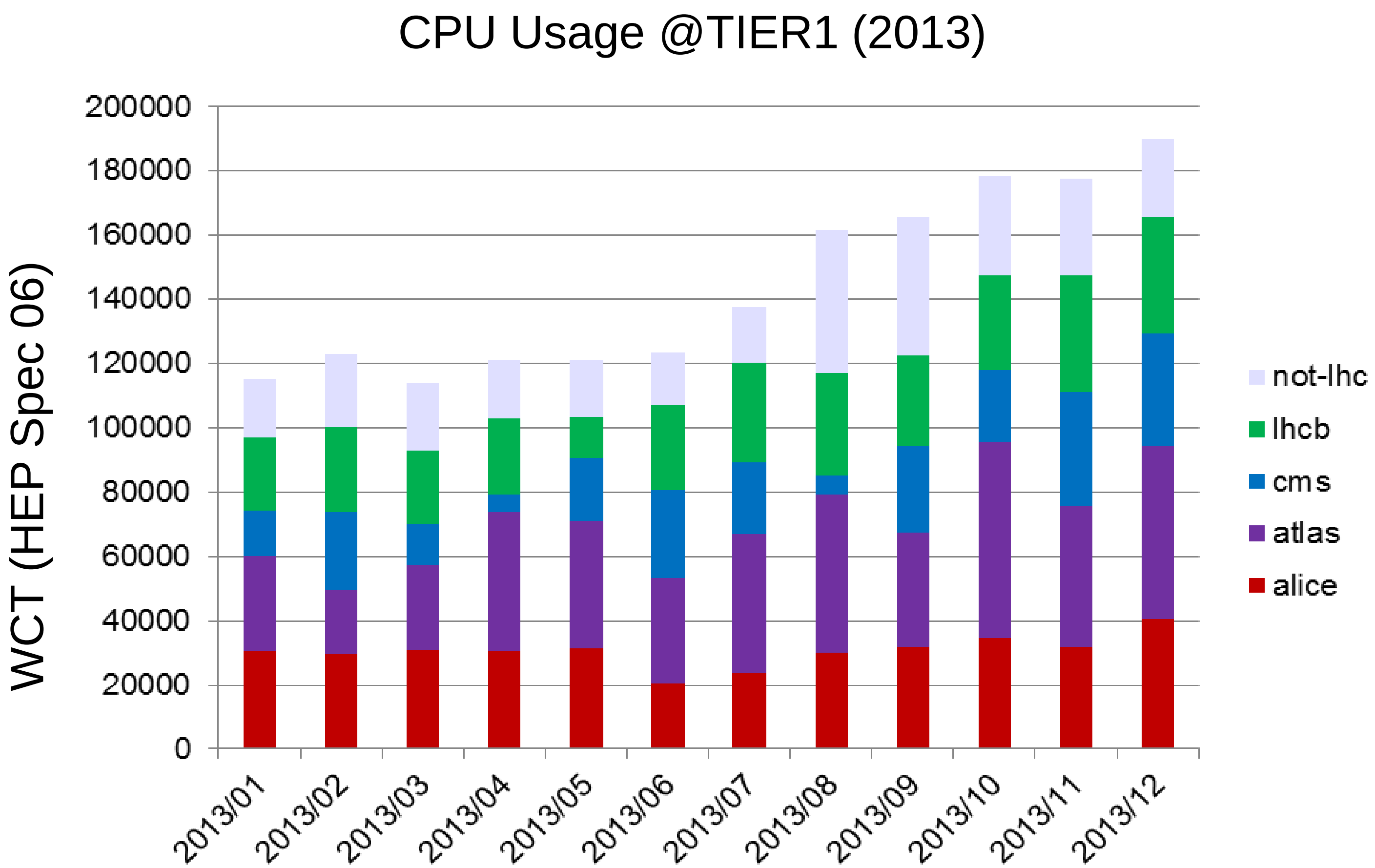
CERN, 10-2-2014

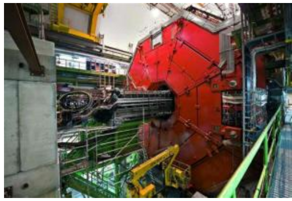
Stefano Zani

stefano.zani@cnaf.infn.it

INFN tier1 Experiment activity

- CNAF is a Tier1 center for all the LHC experiments and provides resources to many (20) other experiments like: AMS, ARGO, AUGER, Borexino, CDF, Pamela, Virgo, Xenon…





INFN TIER1
Resources Today and after LS1(pledge 2015)

(TwinSquare)
4 mainbords in 2 U (24Cores).

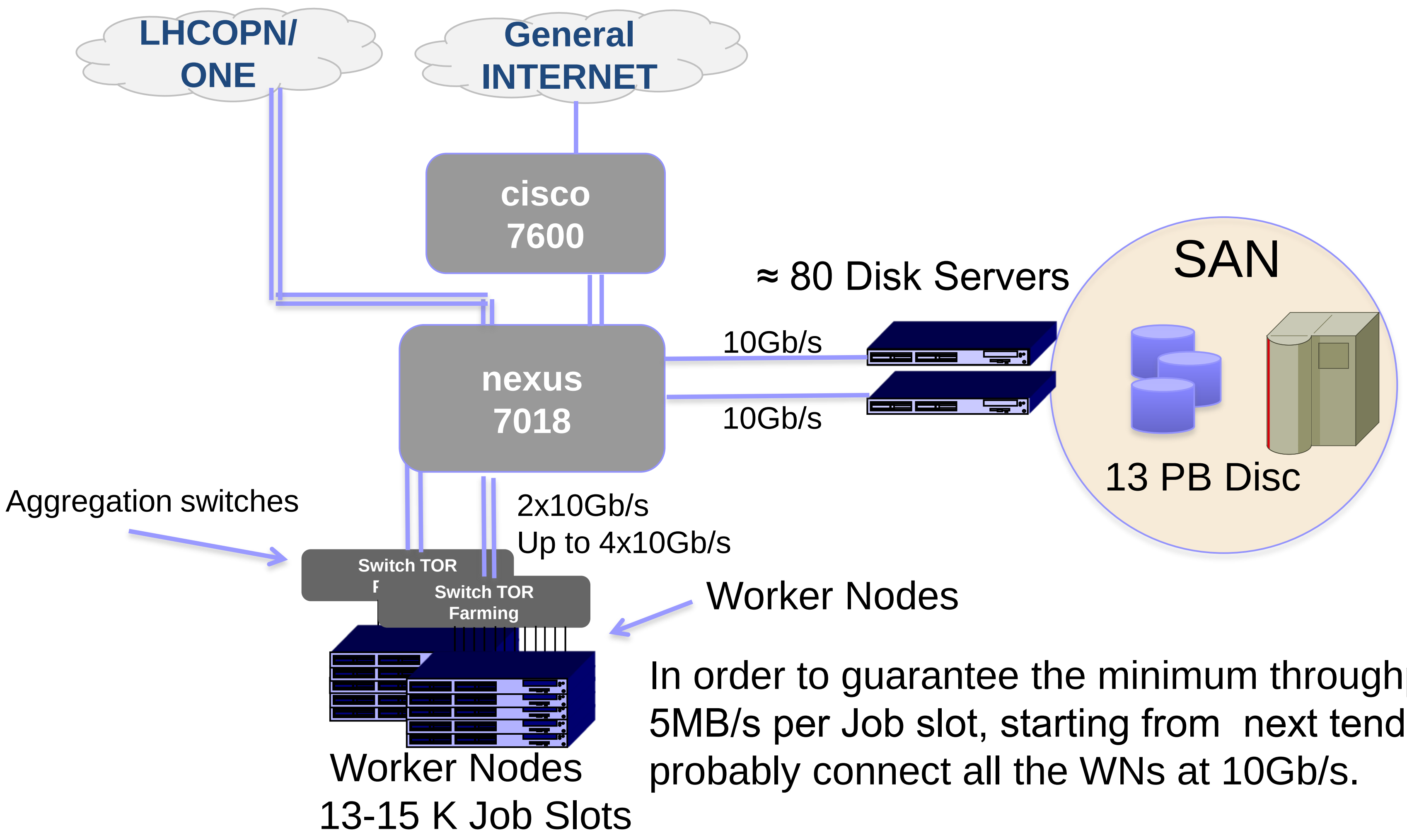
- Total computing resources
  - 195K HepSpec-06 (17K job slots)
- Computing resources for LHC
  - Current: 100KHepSpec-06,10000 Job Slot (pledged 2014)
  - After LS1: 130KHepSpec-06, ~13000 job slots (Pledged 2015)

DDN SFA 12K
connected to a SAN

- Storage Resources for LHC
  - Curent 11 PB Disk and 16 PB Tape (Pledged 2014)
  - After LS1: 13 PB Disk and 24 PB Tape (Pledged 2015)

Pledged numbers doesn't suggest any big increment of computing resources in 2015 (20-30%).

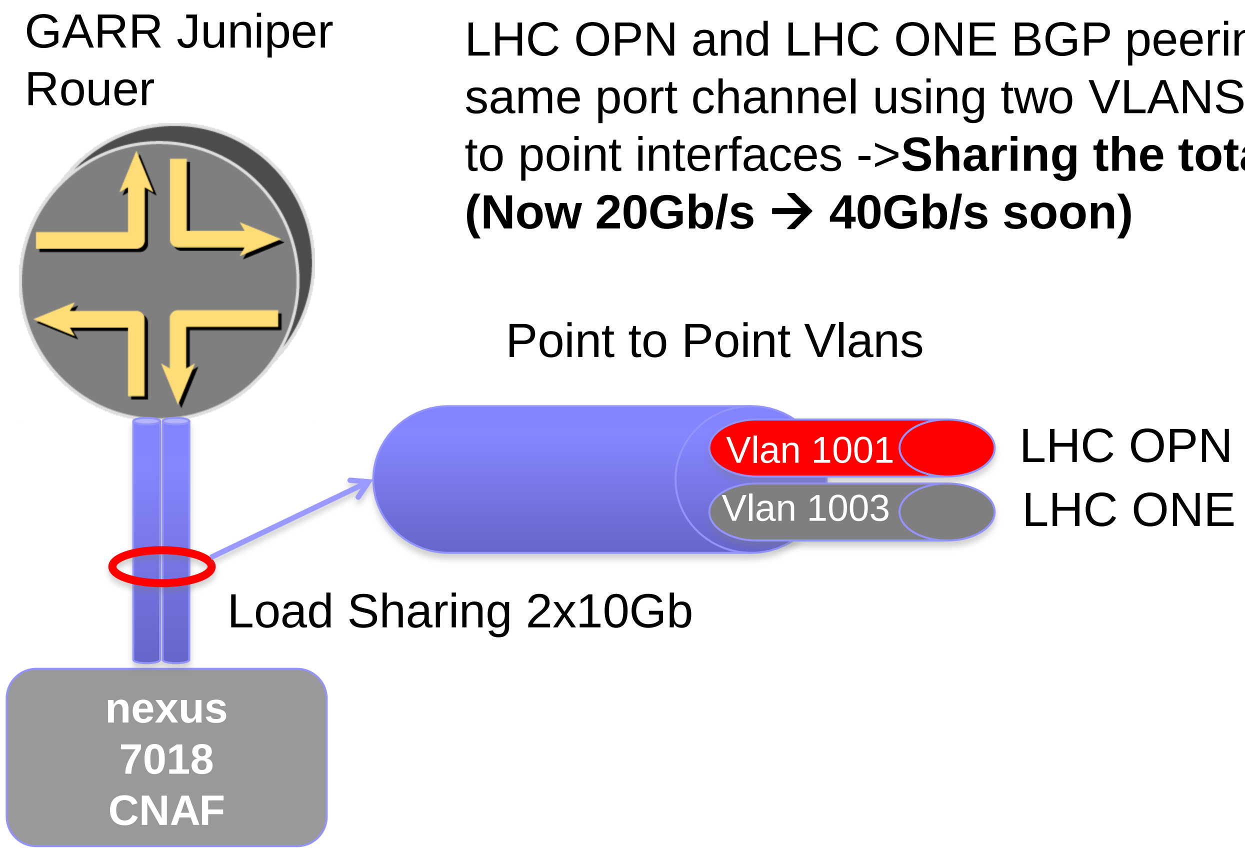
Farming and Storage local interconnection



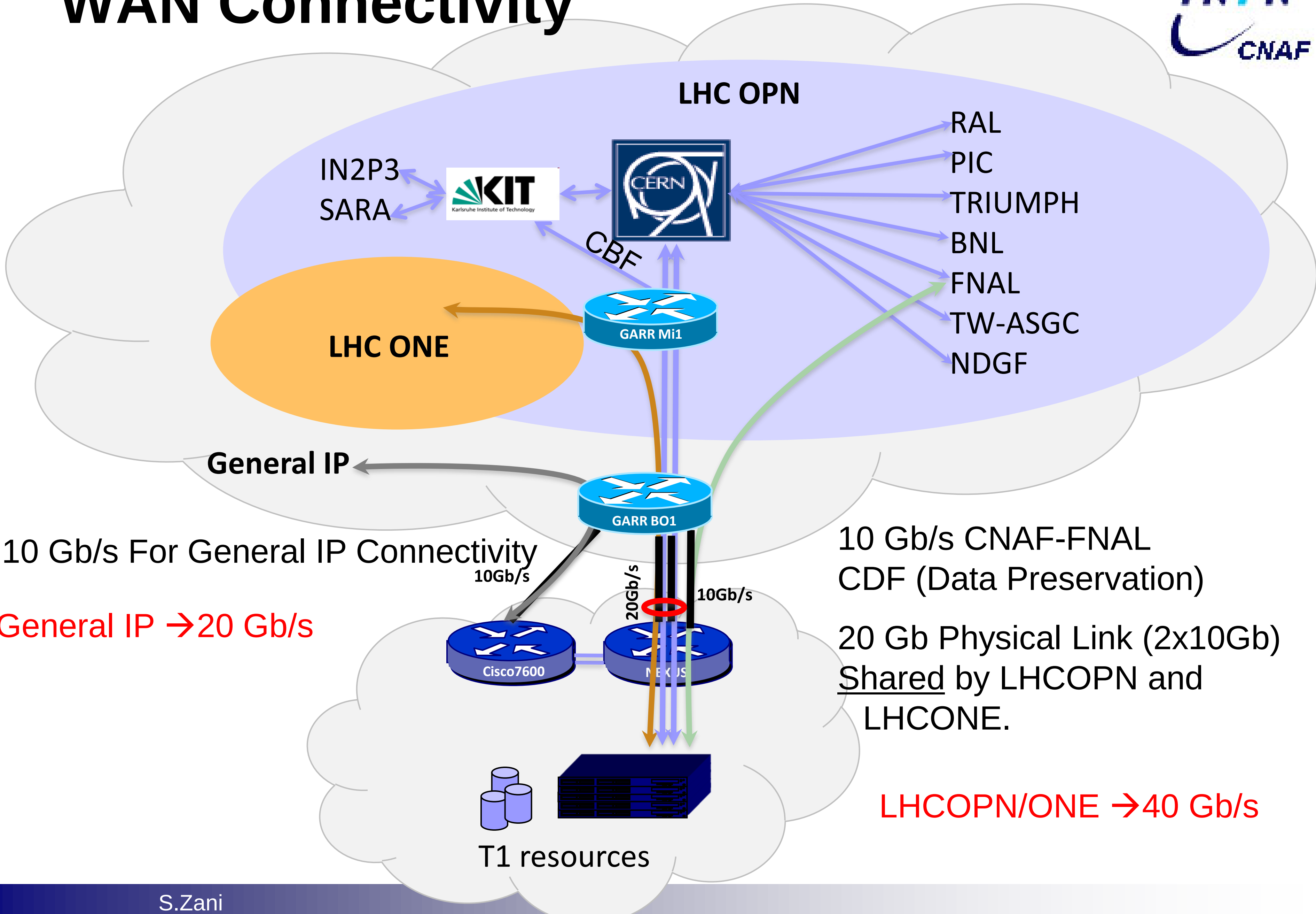
In order to guarantee the minimum throughput of 5MB/s per Job slot, starting from next tender, we'll probably connect all the WNs at 10Gb/s.

LHC OPN and ONE access link

LHC OPN and LHC ONE BGP peerings are made on the same port channel using two VLANS reserved for the point to point interfaces ->**Sharing the total access bandwidth (Now 20Gb/s → 40Gb/s soon)**



WAN Connectivity



10 Gb/s For General IP Connectivity

General IP →20 Gb/s

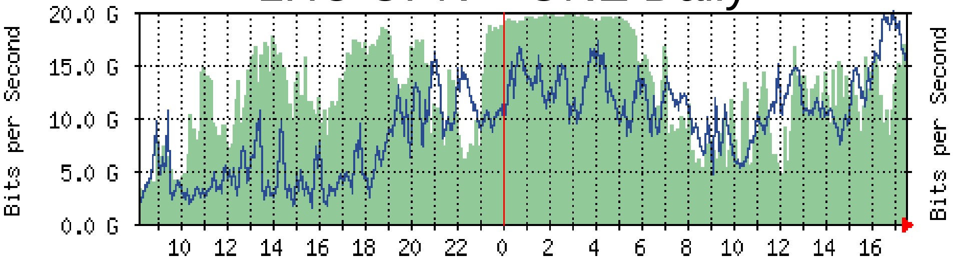
10 Gb/s CNAF-FNAL
CDF (Data Preservation)

20 Gb Physical Link (2x10Gb)
Shared by LHCOPN and LHCONE.

LHCOPN/ONE →40 Gb/s

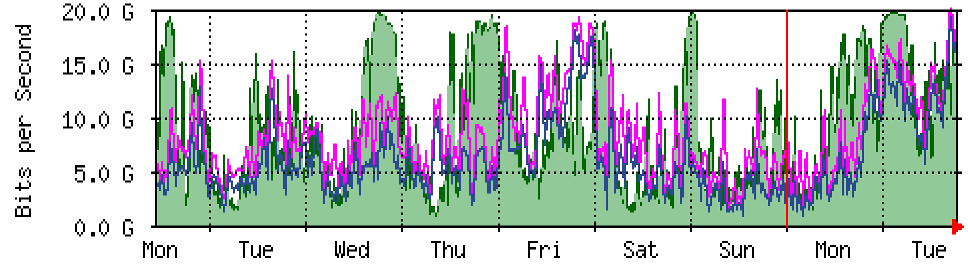
WAN LHC total utilization (LHC OPN+LHCONE) "How the network is used"

LHC OPN + ONE Daily

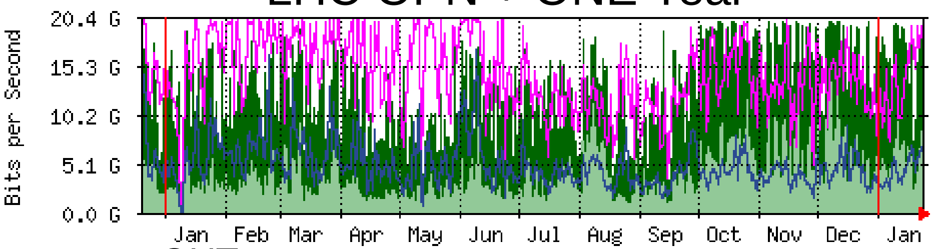


OUT IN

LHC OPN + ONE Weekly



LHC OPN + ONE Year



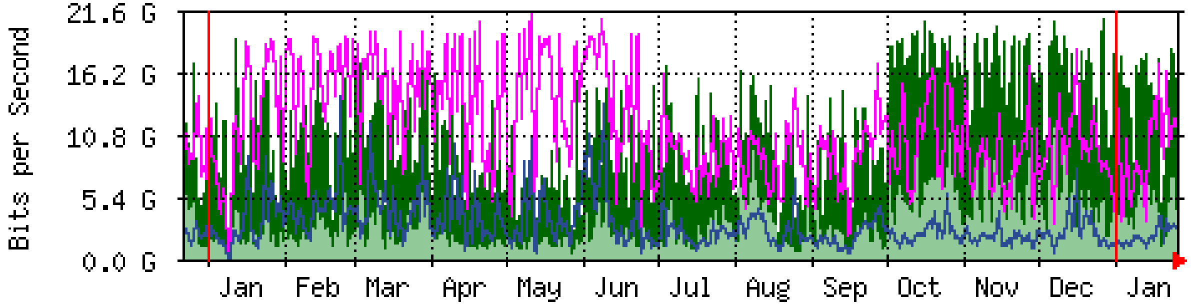
OUT IN OUT Peak IN Peak

AVERAGE IN: 4.1Gb/s AVERAGE OUT: 4.8 Gb/s
MAX IN:20Gb/s MAX OUT:20Gb/s

We are observing many peaks at the nominal maximum speed of the link.

LHC-OPN vs LHCONE

LHC OPN Yearly

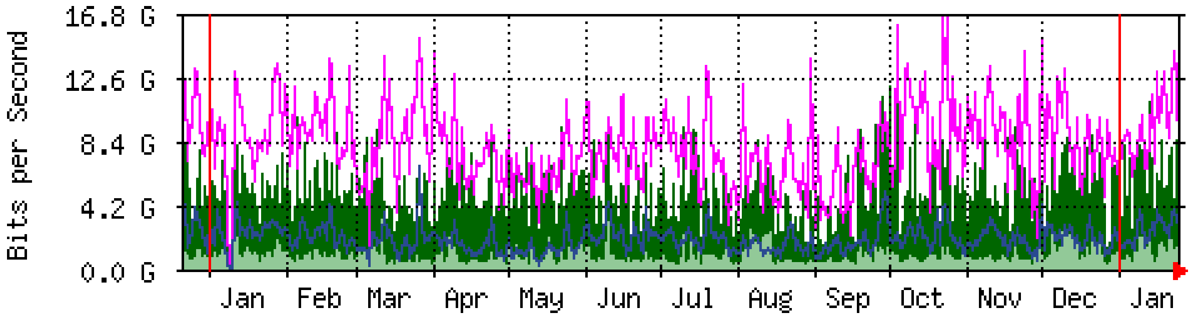


AVERAGE IN: 3.2Gb/s AVERAGE OUT: 2.9 Gb/s
MAX IN:20Gb/s MAX OUT:20Gb/s

Many Peaks up to 20Gb/s
TOP Apps:
53% Xrootd
47% GridFtp

TOP Peers: CERN, KIT, IN2P3…

LHC ONE Yearly



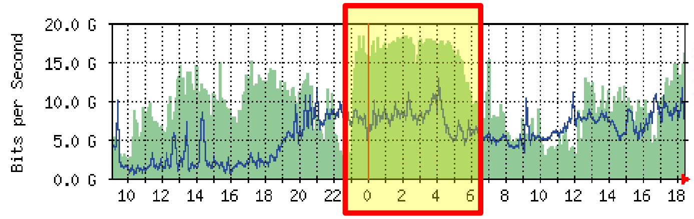
AVERAGE IN: 0.9 Gb/s AVERAGE OUT: 1.9 Gb/s
MAX IN:12 Gb/s MAX OUT:16 Gb/s

Some Peaks up to 16 Gb/s
TOP Apps:
70% GridFtp,
30% Xrootd

TOP Peers:SLAC, INFN MI, INFN LNL ..

<u>LHOPN traffic it is significantly higher than traffic on LHC ONE.</u>
The most relevant peaks of traffic are mainly on LHC-OPN routes.

“Analysis of a peak” (LHC-OPN Link)



IN OUT Custom Selection From: 2014-02-03

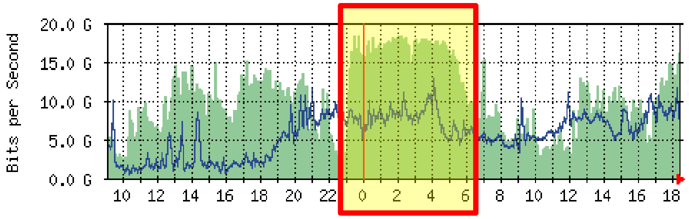
Application | Application Groups | Protocol Distribution | Showing 1 to 10 View per page 10

Application	% of total traffic
xrootd-1095	84%
GridFTP	14%
xrootd-1094	1%
Unknown_App [Show Ports]	<1%
http	<1%
lorica-in	<1%
irdmi	<1%
pcsync-https	<1%
https	<1%
pcsync-http	<1%

Top Traffic - ApplicationIN

In this case the application using most of the bandwidth was Xrootd

“Analysis of a peak” (LHC-OPN Link)



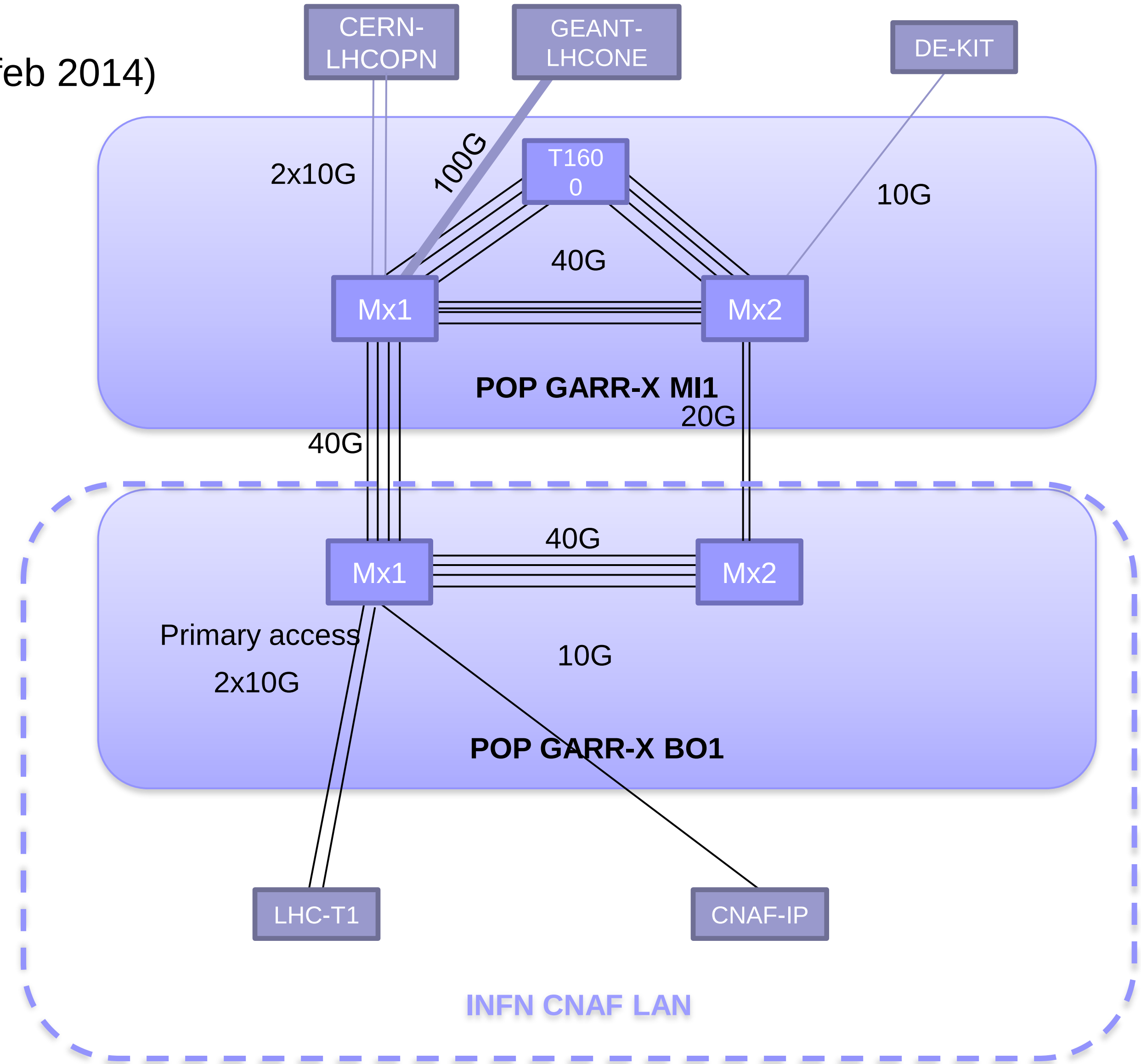
Show IP | Group by None Showing 1 to 50 View per page 50

Src IP	Dst IP	Application	Port	Protocol	DSCP	Traffic(967.57 GB)	Percent
f01-070-138-e.gridka.de	ds-202-08-15.cr.cnaf.infn.it	xrootd-1095	1095	TCP	Default	40.74 GB	4%
f01-070-128-e.gridka.de	ds-202-08-13.cr.cnaf.infn.it	xrootd-1095	1095	TCP	Default	36.32 GB	4%
f01-070-128-e.gridka.de	ds-202-08-15.cr.cnaf.infn.it	xrootd-1095	1095	TCP	Default	30.72 GB	3%
f01-070-122-e.gridka.de	ds-202-08-15.cr.cnaf.infn.it	xrootd-1095	1095	TCP	Default	28.06 GB	3%
f01-070-130-e.gridka.de	ds-202-08-15.cr.cnaf.infn.it	xrootd-1095	1095	TCP	Default	24.46 GB	3%
f01-070-122-e.gridka.de	ds-202-08-13.cr.cnaf.infn.it	xrootd-1095	1095	TCP	Default	23.3 GB	2%
f01-070-138-e.gridka.de	ds-202-08-13.cr.cnaf.infn.it	xrootd-1095	1095	TCP	Default	19.77 GB	2%
f01-070-130-e.gridka.de	ds-202-08-13.cr.cnaf.infn.it	xrootd-1095	1095	TCP	Default	12.05 GB	1%
f01-081-126-e.gridka.de	wn-206-07-21-04-a.cr.cnaf.infn.it	xrootd-1095	1095	TCP	Default	2.95 GB	<1%
lxfsrf09c01.cern.ch	wn-206-01-05-02-b.cr.cnaf.infn.it	xrootd-1095	1095	TCP	Default	2.74 GB	<1%
lxfsrd37a02.cern.ch	wn-206-01-27-02-b.cr.cnaf.infn.it	xrootd-1095	1095	TCP	Default	2.39 GB	<1%
lxfsrd14c03.cern.ch	wn-205-10-06-01-b.cr.cnaf.infn.it	xrootd-1095	1095	TCP	Default	2.37 GB	<1%
lxfsrd12c03.cern.ch	wn-206-01-02-02-b.cr.cnaf.infn.it	xrootd-1095	1095	TCP	Default	2.36 GB	<1%
lxfsrk54c03.cern.ch	wn-206-04-11-01-a.cr.cnaf.infn.it	xrootd-1095	1095	TCP	Default	2.32 GB	<1%
lxfsrd38c03.cern.ch	wn-205-10-34-01-b.cr.cnaf.infn.it	xrootd-1095	1095	TCP	Default	2.31 GB	<1%
lxfsrd39a01.cern.ch	wn-204-11-05-02-b.cr.cnaf.infn.it	xrootd-1095	1095	TCP	Default	2.24 GB	<1%
lxfsre08b06.cern.ch	wn-205-03-22-02-b.cr.cnaf.infn.it	xrootd-1095	1095	TCP	Default	2.23 GB	<1%
lxfsrf15c07.cern.ch	wn-205-01-04-02-b.cr.cnaf.infn.it	xrootd-1095	1095	TCP	Default	2.22 GB	<1%
lxfsrd06c01.cern.ch	wn-205-03-27-01-b.cr.cnaf.infn.it	xrootd-1095	1095	TCP	Default	2.21 GB	<1%
lxfsrd56c03.cern.ch	wn-204-13-27-01-b.cr.cnaf.infn.it	xrootd-1095	1095	TCP	Default	2.2 GB	<1%
lxfsrf09c01.cern.ch	wn-200-04-09-02-a.cr.cnaf.infn.it	xrootd-1095	1095	TCP	Default	2.2 GB	<1%
lxfsre48c01.cern.ch	wn-200-01-05-01-a.cr.cnaf.infn.it	xrootd-1095	1095	TCP	Default	2.19 GB	<1%
lxfsrd38c03.cern.ch	wn-204-03-33-02-b.cr.cnaf.infn.it	xrootd-1095	1095	TCP	Default	2.18 GB	<1%
lxfsrk36c03.cern.ch	wn-204-13-07-02-b.cr.cnaf.infn.it	xrootd-1095	1095	TCP	Default	2.18 GB	<1%
lxfsra28a07.cern.ch	wn-204-08-08-02-b.cr.cnaf.infn.it	xrootd-1095	1095	TCP	Default	2.18 GB	<1%

Looking the Source IPs and Destination IPs
Xrootd: From KIT → CNAF (ALICE Xrootd Servers)
Xrootd: From CERN → CNAF (Worker Nodes)

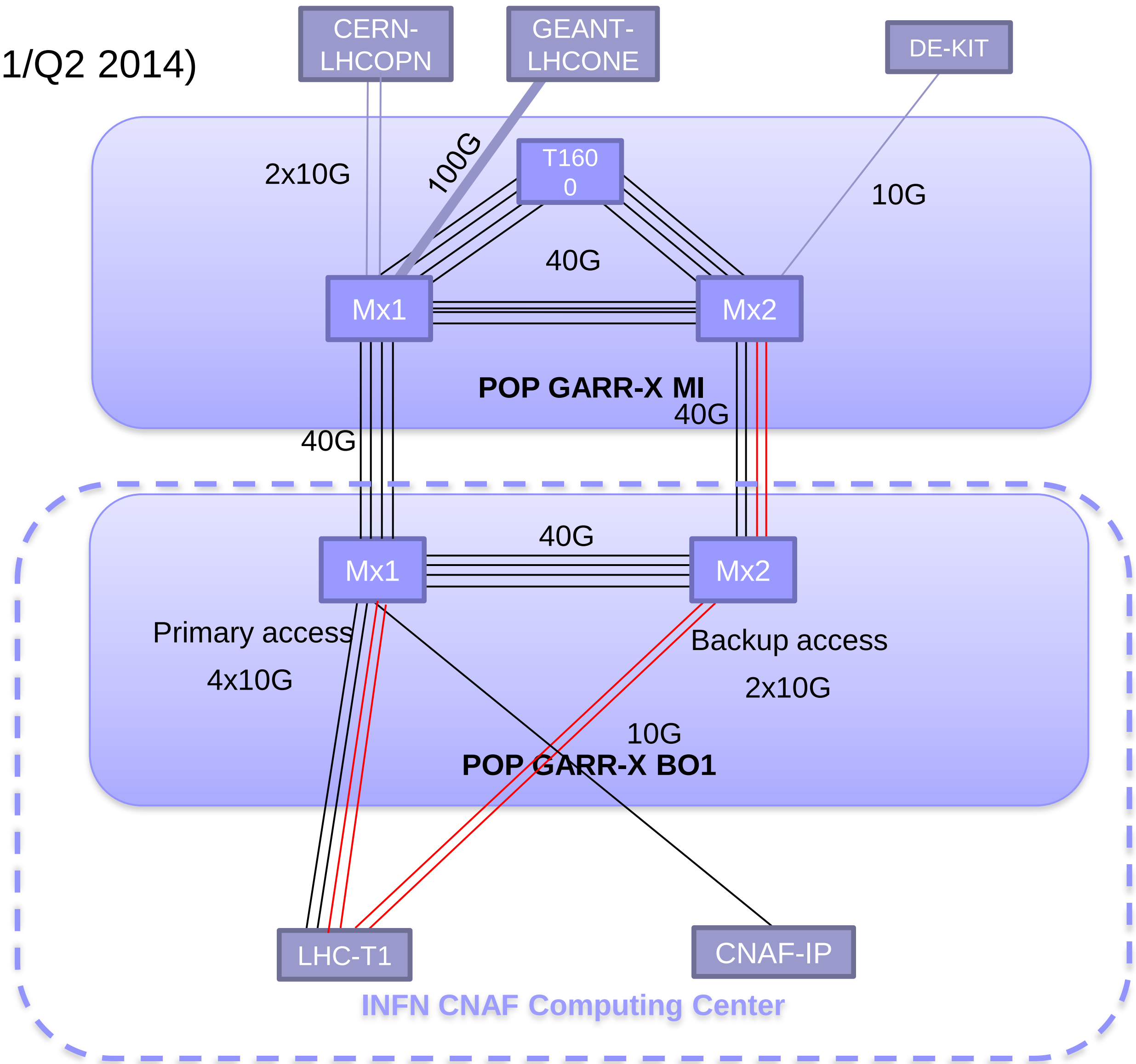
T1 WAN connection GARR side (NOW)

Present (1 feb 2014)



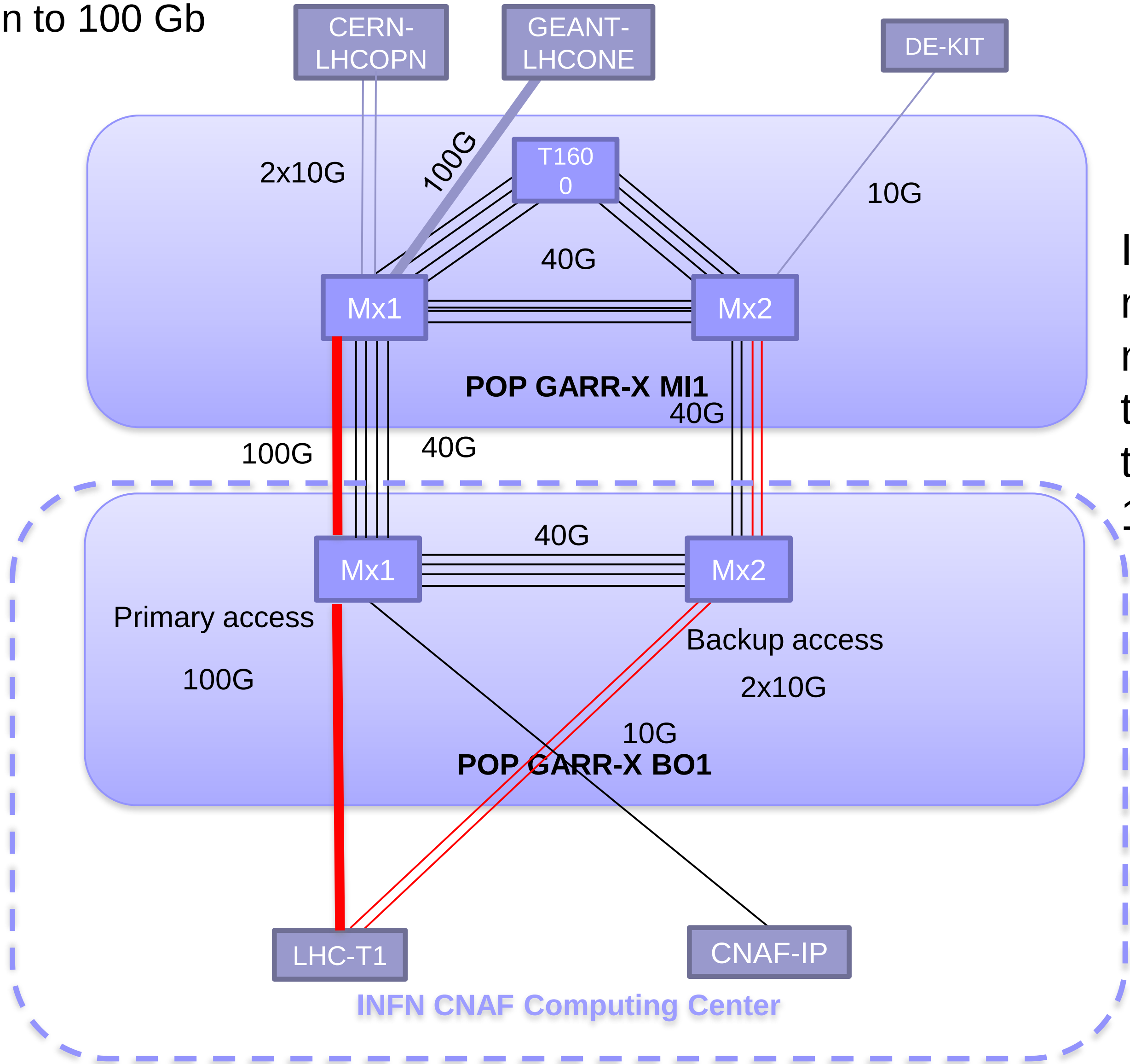
T1 WAN connection GARR side Next Step

Evolution (Q1/Q2 2014)



Next Steps in CNAF WAN Connectivity

Evolution to 100 Gb



If more bandwidth is necessary, GARR-X network can connect the TIER1 and part of the Italian TIER2s at 100Gb/s.

CNAF views and general concerns

- We are not experiencing real WAN Network problems → Experiments are using the center and the network I/O seems to be good even during short periods of bandwidth saturation… **But we are not in data taking..**

- Concern: **Direct access to data over WAN**, (for example analysis traffic) can potentially "Saturate" any WAN link → NEED TO UNDERSTAND BETTER THE DATA MOVEMENT OR ACCESS MODEL in order to provide bandwidth where it is necessary and "protect" the essential connectivity resources .

Open Questions

- Do we keep LHCOPN? Do we change it ?
- Do we keep the LHCONE L3VPN? Do we change it?

The answer to these questions is dependent on the role of the TIERs in next computing models.

- If **T0-T1 guaranteed bandwidth** during data taking **is still mandatory → We should keep LHCOPN (or part of it)** in order to have a “Better control” on the most relevant traffic paths and to have a faster troubleshooting procedures in case of network problems**.**

- If the data flows will be more and more distributed as a full mesh between Tiers→ a L3 approach on over provisioned resources dedicated to LHC (like LHCONE VRF) could be the best matching solution.

Services and Tools needed ?

- Flow analysis tools (Netflow/Sflow analizers) have to be improved by network admins at site level .

- Services (like for example FTS) used to “Optimize and tune” the main data transfers (and flows) from and to a site could be very useful to the experiments and the sites too ..

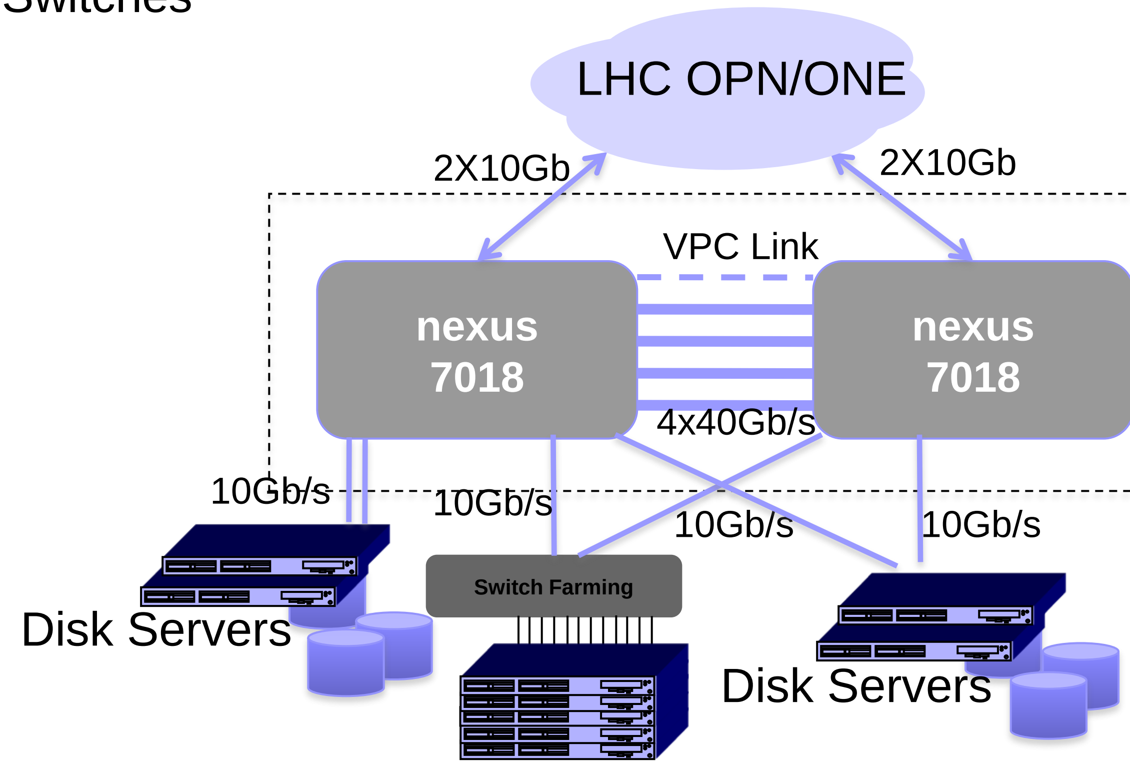
Thank You!

Backup Slides

Internal network possible evolution

Redundancy:
2 Core Switches

Scalability:
Up to 1536 10Gb/s Ports.



550Gb / Slot

CNAF Network Devices

4 Core Switch Routers (fully redundant)

Cisco Nexus 7018 (TIER1 Core Switch and WAN Access Router)

- 208 10Gb/s Ports
- 192 Gigabit ports

Extreme Networks BD8810 (Tier1 Concentrator)

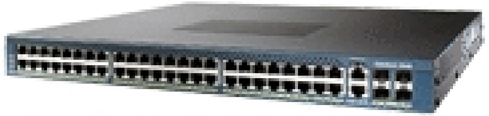
- 24 10Gb/s Ports
- 96 Gigabit ports

Cisco 7606 (General IP WAN access router)

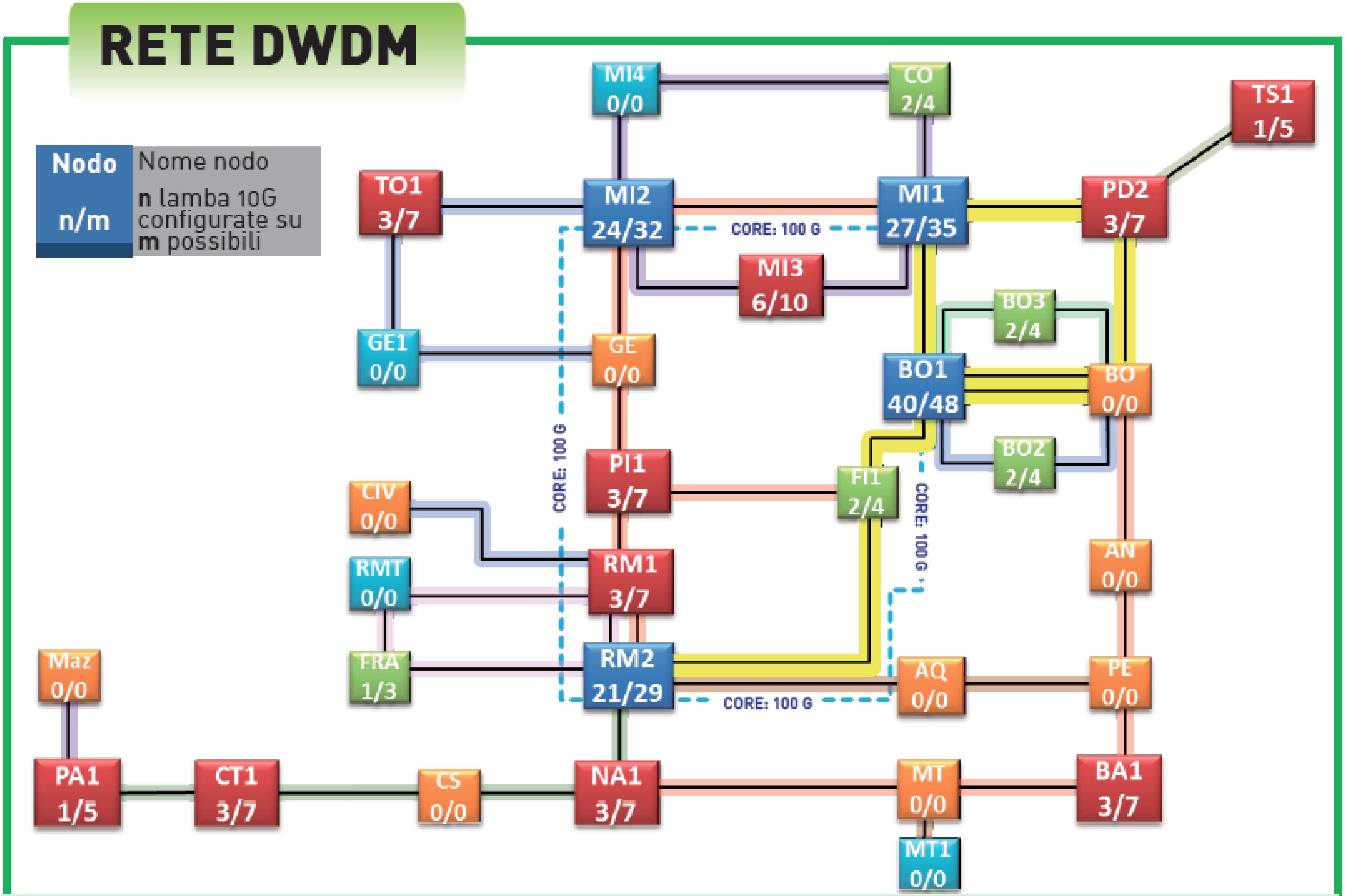
Cisco 6506 (Offices core switch)

More than **100 TOR switches** (About 4800 Gigabit Ports and 120 10Gb/s Ports)

- 40 Extreme summit X00/X450 (48x1Gb/s+4Uplink)
- 11 3Com 4800 (48x1Gb/s+4Uplink)
- 12 Juniper Ex4200 (48x1Gb/s+2x10Gbs Uplink)
- 14 Cisco 4948 (48x1Gb/s+4x10Gbs Uplink)
- 20 Cisco 3032 – DELL Blade
- 4 DELL Power Connect 7048 (48x1Gb/s+4x10Gb/s)



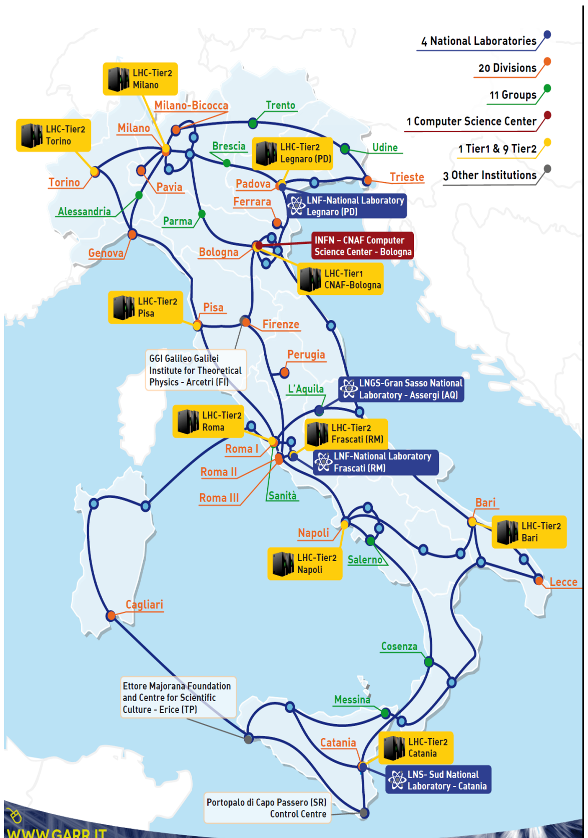
GARR-X: Topologia della rete trasmissiva

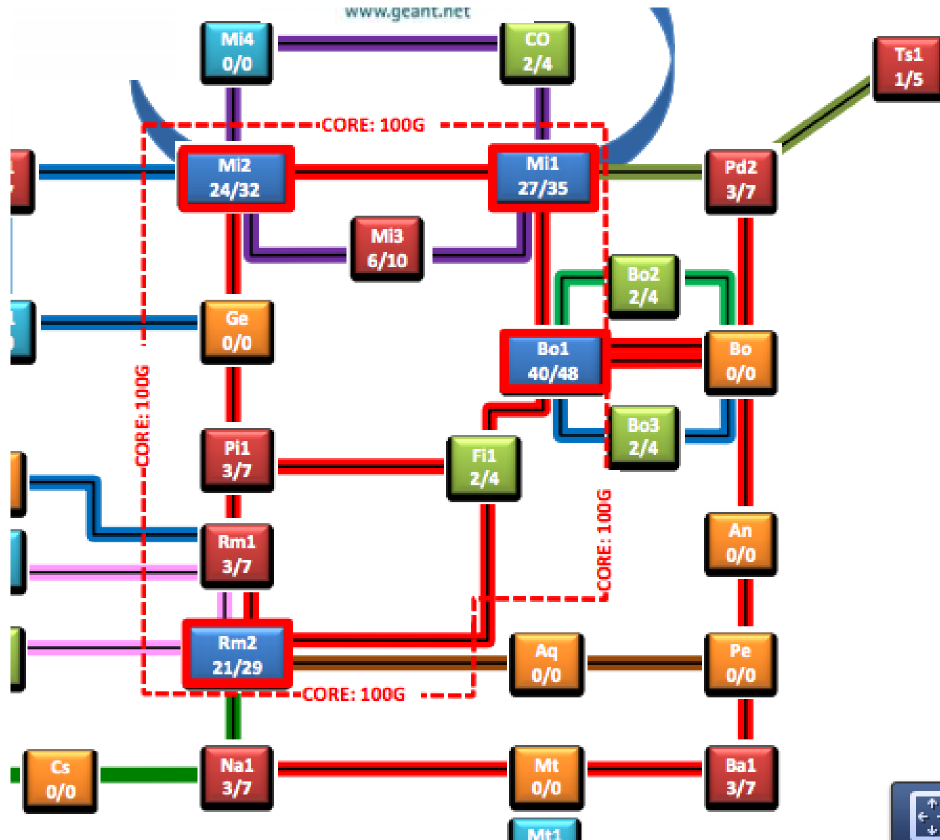


Le sedi INFN in GARR-X

GARR-POP

- One of the main GARR POP is hosted by CNAF and it is inside TIER1 computing center → **Tier1's wan connections are made in LAN (Local patches).**
- The GARR Bo1 POP today can activate up to 160 Lambdas and it is possible to activate the first 100Gb/s links.





T1-LHCOPN

Action(s) | More Reports | Dashboards

Traffic | **Application** | Source | Destination | QoS | Conversation | Multicast | Medianet

IN OUT Custom Selection From: 2014-02-03 23:23 To: 2014-02-04 05:36 Show

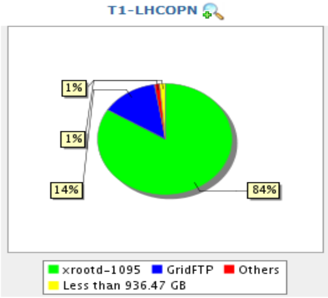
Application | Application Groups | Protocol Distribution | Showing 1 to 10 View per page 10

Application	Traffic(Total: 46.82 TB)	% of total traffic
xrootd-1095	39.25 TB	84%
GridFTP	6.44 TB	14%
xrootd-1094	300.13 GB	1%
Unknown_App [Show Ports]	117.15 GB	<1%
http	71.86 GB	<1%
lorica-in	52.29 GB	<1%
irdmi	10.17 GB	<1%
pcsync-https	6.21 GB	<1%
https	2.13 GB	<1%
pcsync-http	768.17 MB	<1%

Showing 1 to 10

Top Traffic - ApplicationIN

T1-LHCOPN



Next Steps in CNAF WAN Connectivity

GARR Backbone is already connected to GEANT with 2 x 100Gb/s links.

If more bandwidth will be necessary, GARR-X network can connect the TIER1 and part of the Italian TIER2s at 100Gb/s.