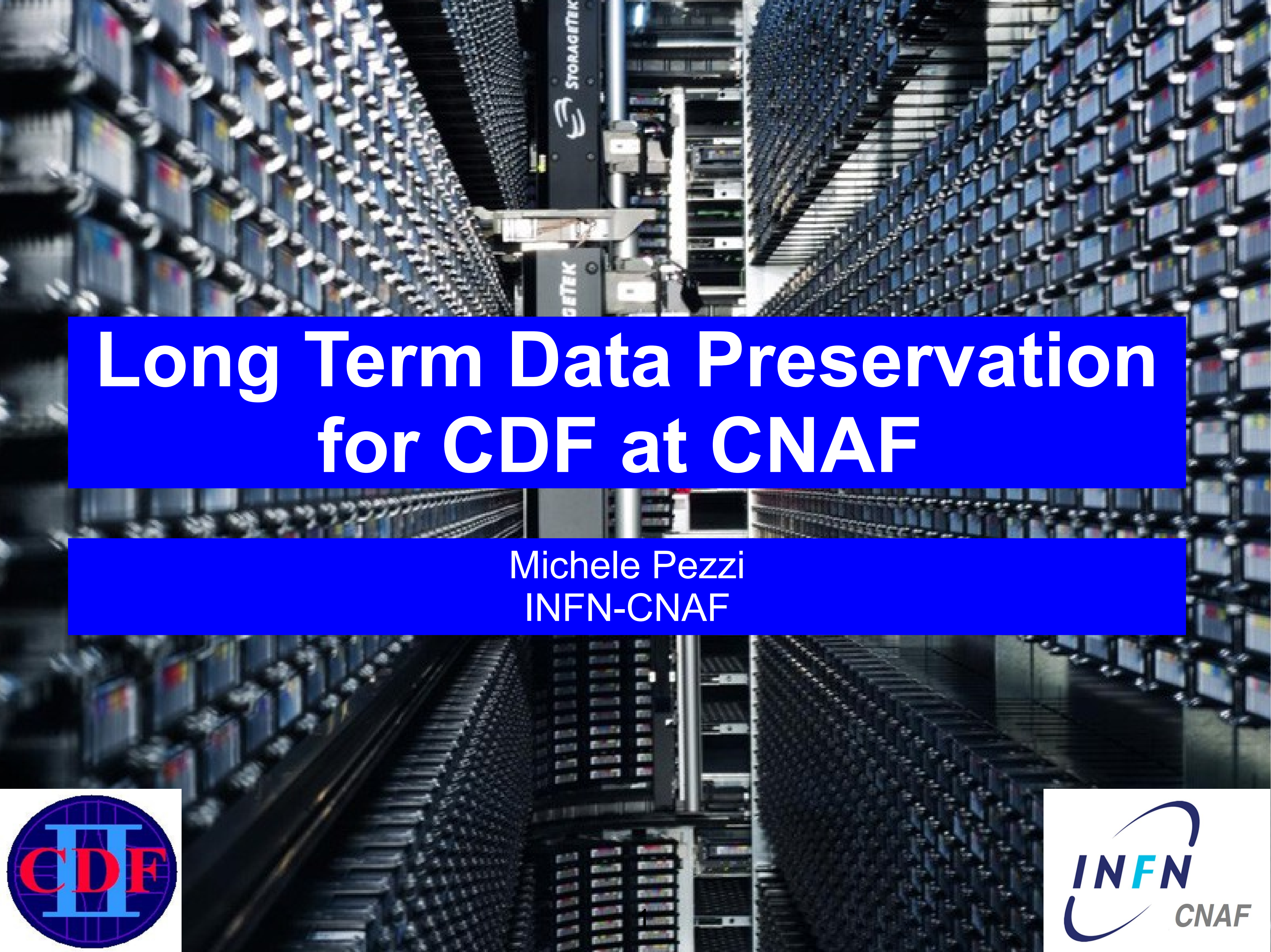


Long Term Data Preservation for CDF at CNAF

Michele Pezzi
INFN-CNAF

Outline

- Introduction to Data Preservation
- CNAF-CDF project
- Data transfer
  - network layout
  - transfer system
  - data archiving
  - monitoring tools
- Data analysis : present & future
- Conclusions

Outline

- Introduction to Data Preservation
- CNAF-CDF project
- Data transfer
  - network layout
  - transfer system
  - data archiving
  - monitoring tools
- Data analysis : present & future
- Conclusions

Introduction to Data Preservation

Data preservation refers to the series of managed **activities** necessary to **ensure** continued **access** to digital materials for **as long as necessary**.

Long-term data preservation can be defined as the ability to provide **continued access** to digital materials.

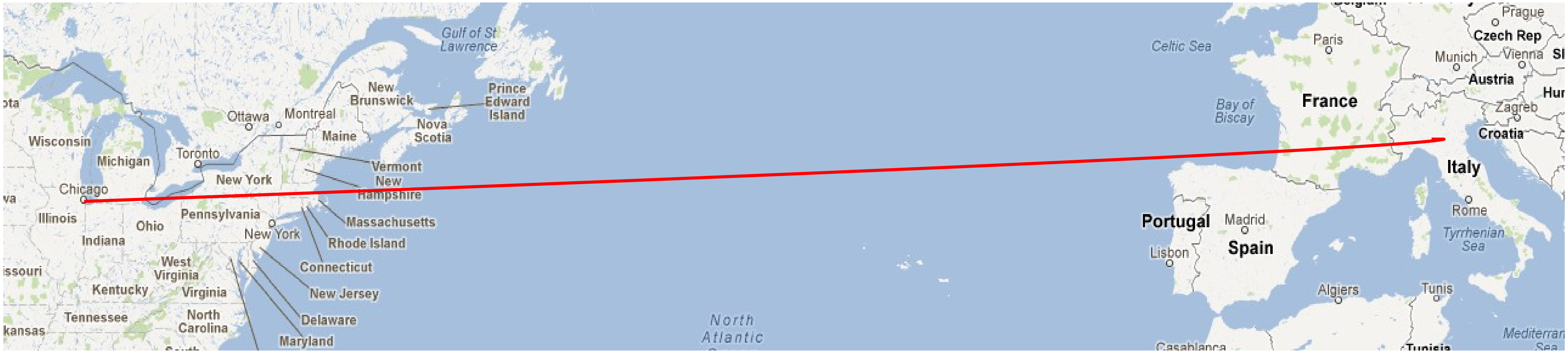
Data preservation is one of the areas targetted in **Horizon2020.**

Some scientific areas, e.g. astrophysics, are well ahead in data preservation. HEP is narrowing the gap:

- **DPHEP:** Data Preservation in High Energy Physics
- Past experiments have already successful DP projects in place (e.g. Babar)
- All **LHC experiments** are devoting more efforts to data preservation

A data preservation project can be divided into two main areas:
1 - Bit preservation : how preserve data
2 - Analysis framework preservation : code preservation, virtualization …

CNAF-CDF project



Goal: preserve a complete copy of CDF data and MC samples at CNAF + services (access, data analysis capabilities)

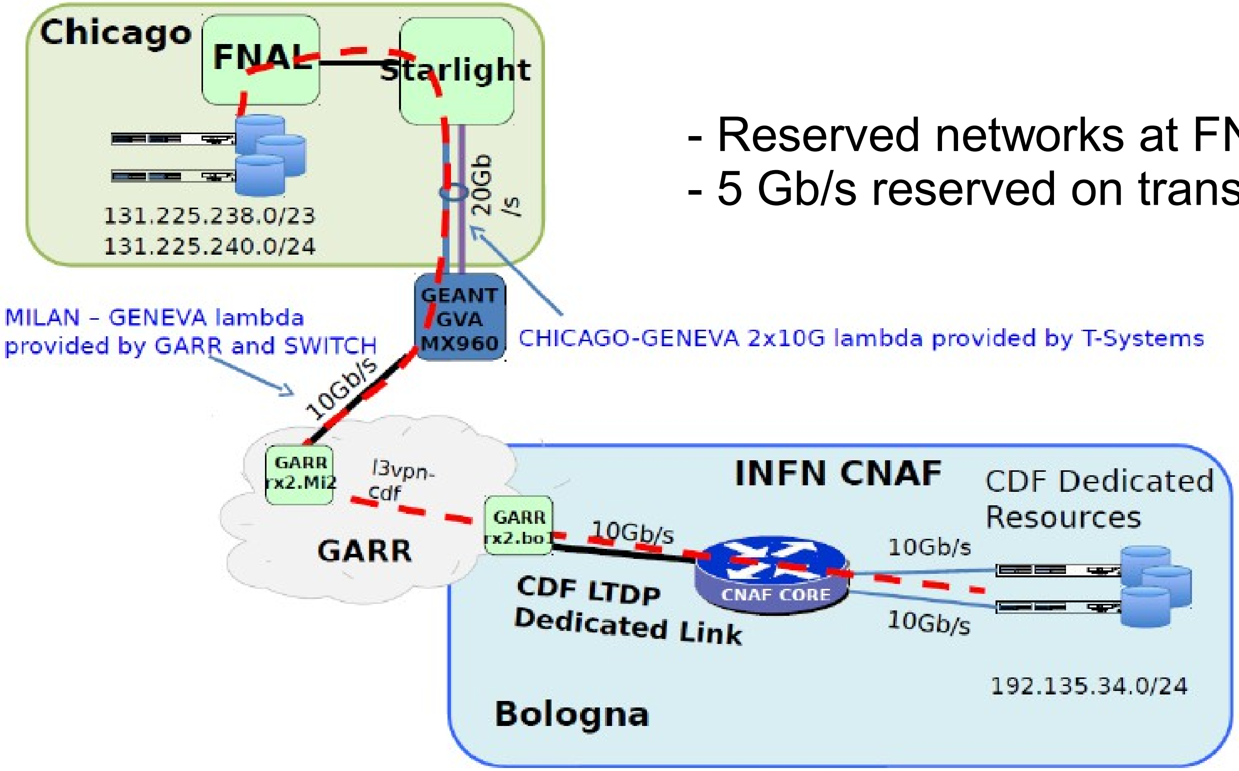
The copy will be splitted in two periods

- end 2013 - early 2014 → All data and MC user level n-tuples (2.1 PB)
- end 2014 → All raw data (1.9 PB) + Databases

During the 2014, development of the long term future analysis framework.

- Preserve data access
- Preserve CDF reconstruction and analysis software
- Give users resources to run CDF analysis (authentication, disk space, CPU)
- Documentation

Data transfer: network layout



- Reserved networks at FNAL and CNAF
- 5 Gb/s reserved on transtlantic link

Transfer layout optimization shows that :

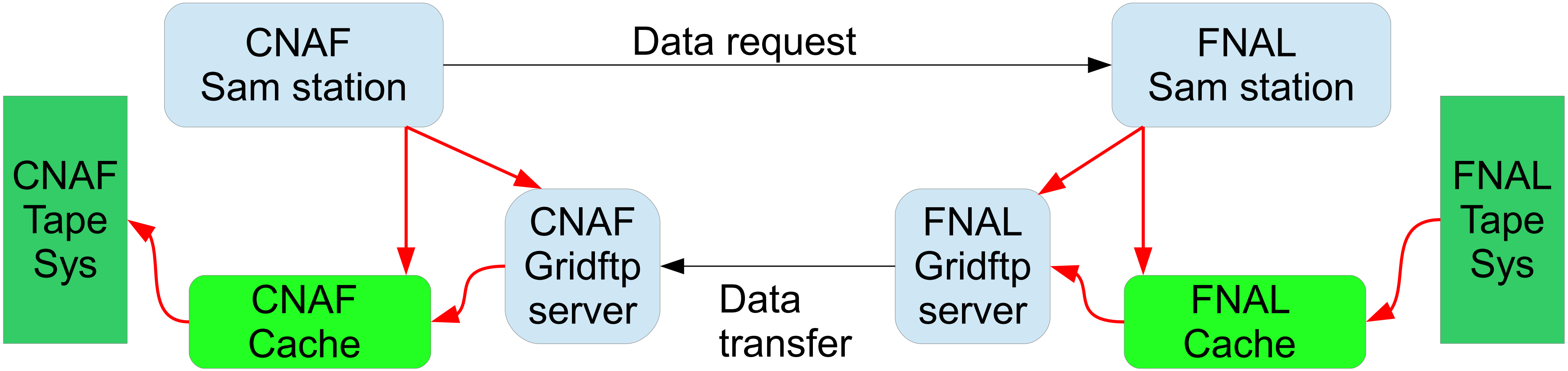
- Reached the peak of 5Gb/s
- Saturate the avaliable bandwidth

(80 simultaneous process each of which divided into 20 parallel streams)

Temporary limited due to tape shortage

Data transfer: transfer system

We use Sequential Metadata Access (SAM), developed at Fermilab, installed on a dedicated machine at CNAF, to pilot the data transfer.



1) CNAF requests data from FNAL that are staged at FNAL cache
2) Data are copied via gridftp protocol, in a third party transfer, and SAM control the checksum
3) Once the data are in the CNAF cache, they are automatically migrated to tape

Using custom integration of SAM and GridFTP command (bash script) to perform data transfer in a semi-automated way

The pre-staging of the data at FNAL is made in parallel with the data transfer at FNAL

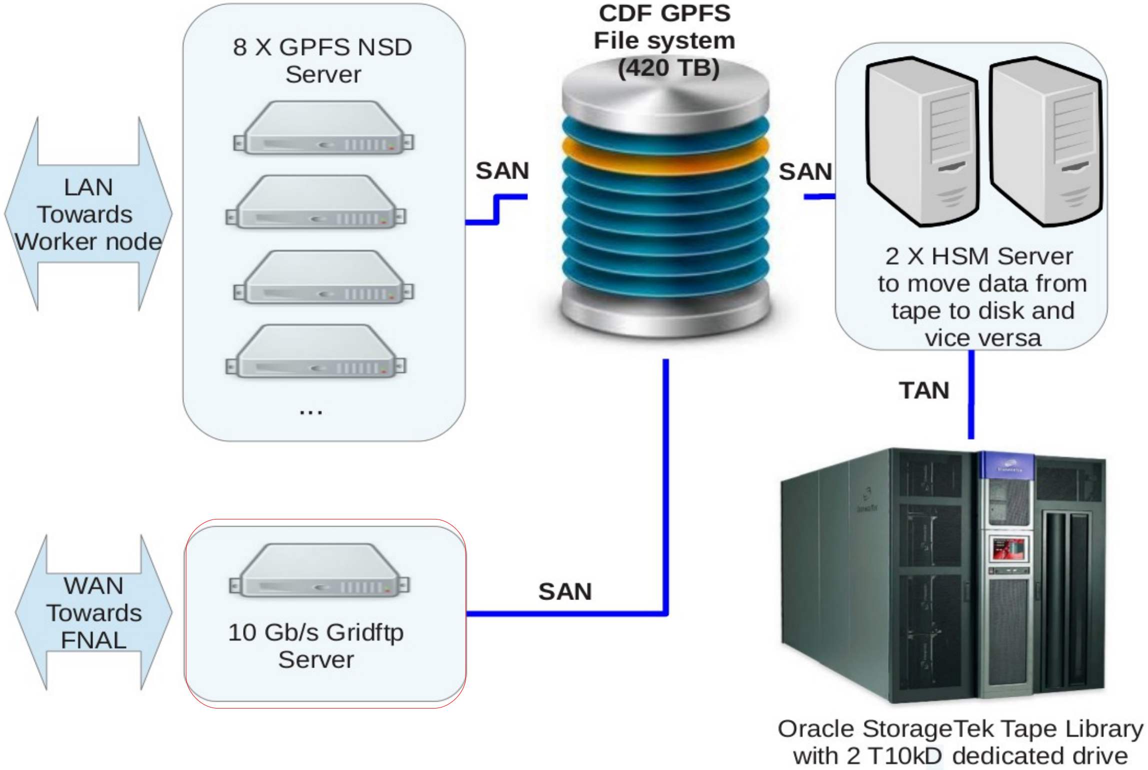
Data transfer: data archiving

1 - Data are copied from FNAL via gridftp.

2 - GEMSS moves data to tape (see next slide).

3 - Retrieval from tape using standard CDF commands (SAM station).

About 115 x 8,5 TB tapes (T10000D) written => ~ 1PByte



2 T10kD drive (130MB/s reading from the CDF GPFS Filesystem)

Actual DISK => TAPE (2 drives) bandwidth can reach 260MB/s

DISK => TAPE bandwidth will be improved with additional drives (2nd half 2014)

1 Single Point of Failure (single gridftp)

- Cold spare machine avaliable (can be configured in a couple of hours)
- since for the data transfer is not required a 24x7 availability is not necessary have redundancy

It's the same storage configuration that is used for other experiment

Data transfer : data archiving

GEMMS

Grid Enabled Mass Storage System (GEMSS) is a software layer for GPFS-TSM interaction for optimization and administration, providing a complete solution for storage access.

GEMSS is based on a custom integration between 3 software layers:

- Parallel filesystem (the IBM General Parallel File System, GPFS)

- Tivoli Storage Manager (TSM), which provides Hierarchical Storage Management (HSM) capabilities

- Grid Storage Resource Manager (StoRM), developed at CNAF, providing access to grid users through a standard SRM interface. It is not essential as in our case.

GEMSS is the standard solution used at the INFN CNAF Tier1 for archiving data for all the LHC and non-LHC experiments.

Data transfer : monitoring

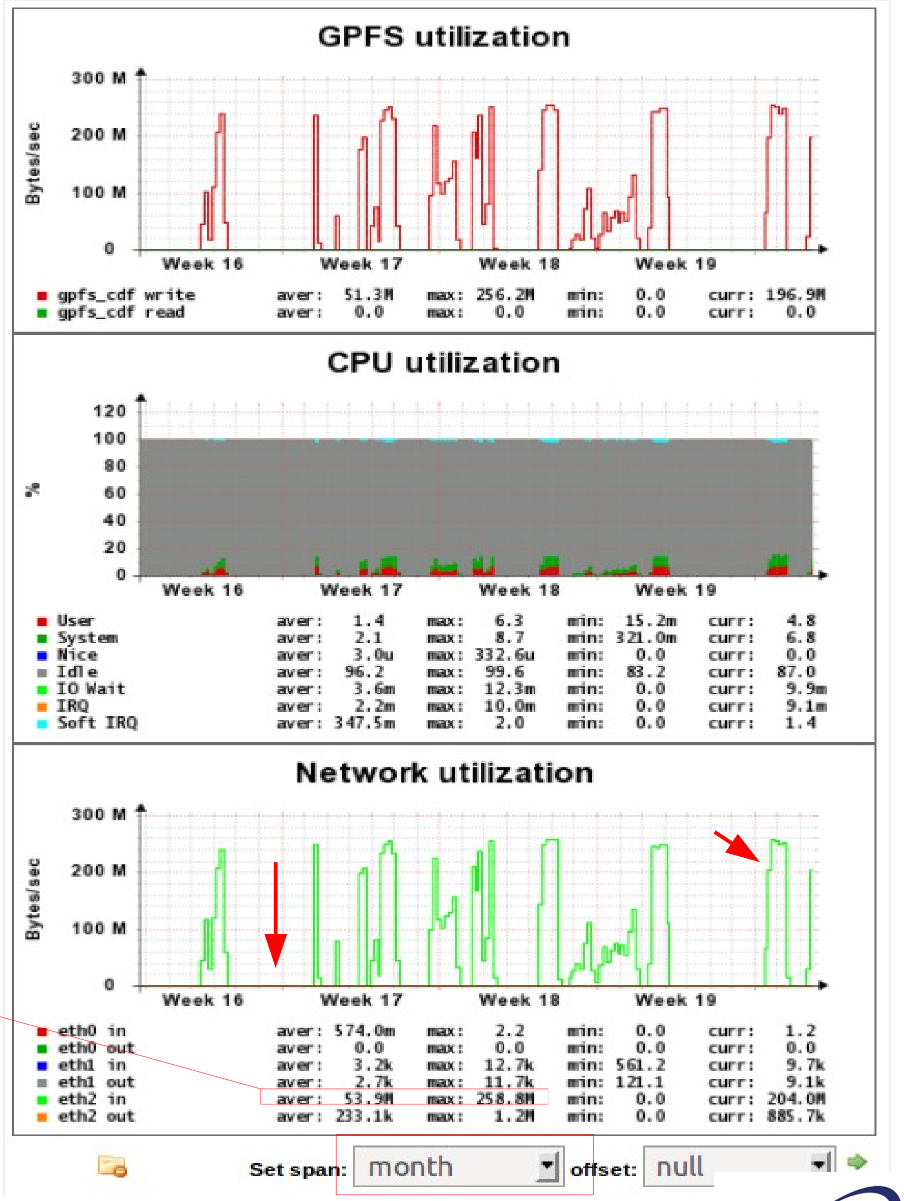
LEMON

Display statistical Graph with:
- Information about machine
- GPFS utilization
- Cpu utilization
- Network utilization
- other...

NAGIOS

This software can handle different type of alarm.

If one service for instance gridftp daemon crash, Nagios send e-mail to the admin or can perform active action (e.g. try to restart the services)



```
■ gpfs_cdf write    aver:  51.3M   max: 256.2M   min:   0.0    curr: 196.9M
■ gpfs_cdf read     aver:   0.0    max:   0.0    min:   0.0    curr:   0.0
```

```
■ User              aver:   1.4    max:   6.3    min:  15.2m   curr:   4.8
■ System            aver:   2.1    max:   8.7    min: 321.0m   curr:   6.8
■ Nice              aver:   3.0u   max: 332.6u   min:   0.0    curr:   0.0
■ Idle              aver:  96.2    max:  99.6    min:  83.2    curr:  87.0
■ IO Wait           aver:   3.6m   max:  12.3m   min:   0.0    curr:   9.9m
■ IRQ               aver:   2.2m   max:  10.0m   min:   0.0    curr:   9.1m
■ Soft IRQ          aver: 347.5m   max:   2.0    min:   0.0    curr:   1.4
```

```
■ eth0 in           aver: 574.0m   max:   2.2    min:   0.0    curr:   1.2
■ eth0 out          aver:   0.0    max:   0.0    min:   0.0    curr:   0.0
■ eth1 in           aver:   3.2k   max:  12.7k   min: 561.2    curr:   9.7k
■ eth1 out          aver:   2.7k   max:  11.7k   min: 121.1    curr:   9.1k
■ eth2 in           aver:  53.9M   max: 258.8M   min:   0.0    curr: 204.0M
■ eth2 out          aver: 233.1k   max:   1.2M   min:   0.0    curr: 885.7k
```

Set span: month ▾ offset: null ▾

Max: 258MB/s
Aver: 54 MB/s

Since the system is not fully automatic we have to launch the script to copy the data, which is why the graph is discontinuous

Data analysis : present and future

Now we have already replicated at CNAF ≈ 1 PB of CDF data on tape

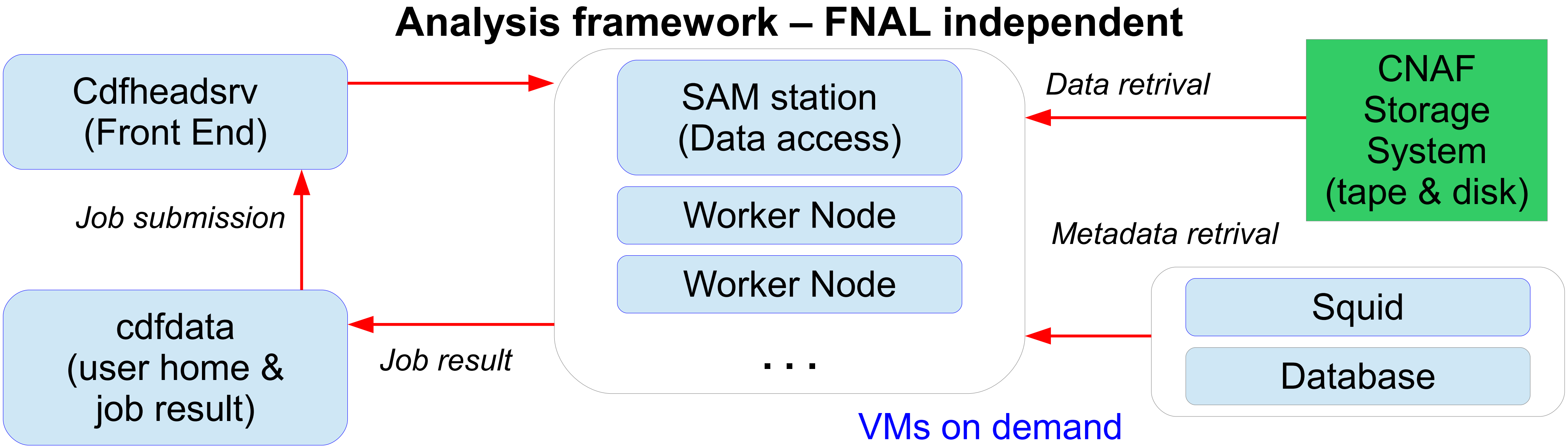
CDF data analysis in the long term future

To analyze CDF data stored at CNAF we need to implement:

- SAM station to retrieve data from tape: new version of SAM code in preparation at FNAL.
- CDF code volume, accessible via CVMFS
- Access to Oracle Databases (with local replicas at the INFN CNAF Tier1)
- Code preservation: CDF legacy software release (SL6) under test

In the long term future, CDF services and analysis computing resources can be instantiated on demand on pre-packaged VMs in a controlled environment.

Data analysis : future



This framework assumes limited use of CDF data in the long term future. Problems under discussion:

- Replication of Database

- Data access : QUESTION: How many years IBM will support GPFS on SL6?
  Possible solution could be NFS which provides greater compatibility with earlier version.

- Authentication : In the long future, access to the GRID resources will not be necessary.
  Possible solution could be job submission restricted to the local CNAF nodes.

- Data integrity : Cyclic reading of the data (eg. repacking) and checksum calculation.

Conclusion

A data preservation project can be divided into two main areas:

1 - Bit preservation : how preserve data

We are currently transferring data and we already have on tape about 1PB.
By the end of 2014 will end the transfer.

2 - Analysis framework preservation : code preservation, virtualization …

We are also studying solutions to build a analysis framework on demand.

First Data Preservation project of INFN

Could be the prototype for the other experiments that have data at CNAF

Thanks for your attention

Questions?

michele.pezzi@cnaf.ifnfn.it

Update

Data transfer : how

The copy is done in two steps:

Dataset import :

1) *sam get dataset* “Dataset name”
2) Check that all files have been imported

Update DB :

1)Lock file (to prevent the cacellation in the case in which the cache is full)
2)Check CRC
3)Move file to tape
4)Update SAM DB with tape location
5)Unlock file
6)Remove file from cache

- UpdateDB starts as soons as the Dataset import has finished
  (no more that 1 dataset import at a time)

- Multiple updateDB in parallel possible.

Every steps has a log file so is possible to keep track of every step of the transfer process

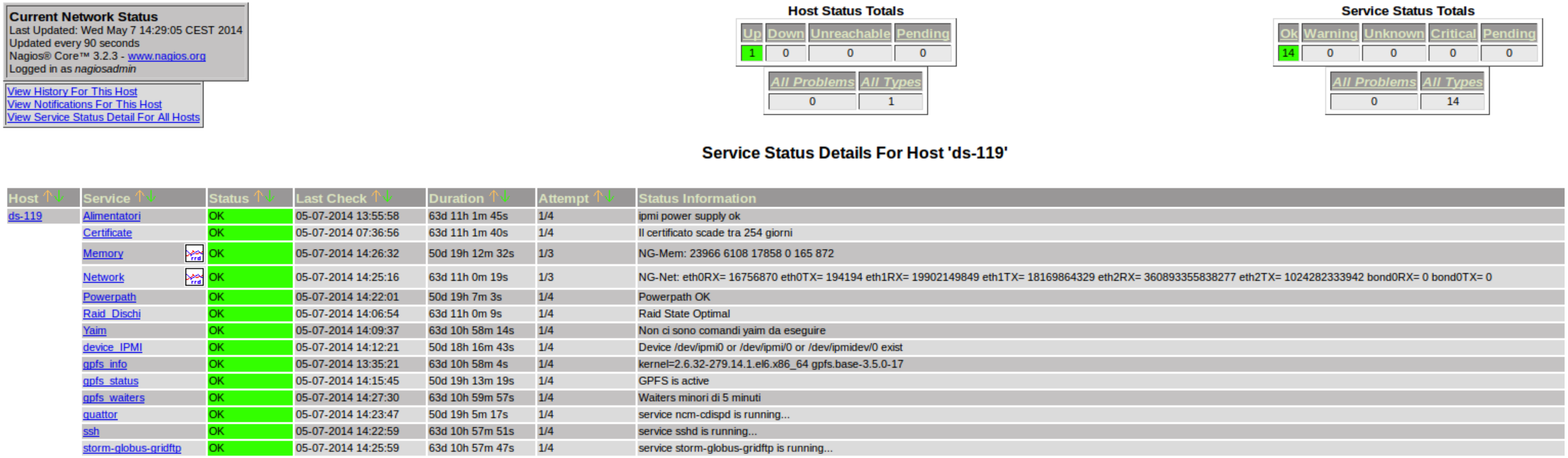
Data transfer : monitoring

NAGIOS

This software can handle different type of alarm.

If one service for instance gridftp daemon crash, Nagios send e-mail to the admin or/and can perform active action (e.g. try to restart the services)

Depending on the type of service the control is done at different time intervals (e.g. for the certificate every 24h whereas for sshd every 5 minutes)



Current Network Status
Last Updated: Wed May 7 14:29:05 CEST 2014
Updated every 90 seconds
Nagios® Core™ 3.2.3 - www.nagios.org
Logged in as *nagiosadmin*

View History For This Host
View Notifications For This Host
View Service Status Detail For All Hosts

Host Status Totals

Up	Down	Unreachable	Pending
1	0	0	0

All Problems	All Types
0	1

Service Status Totals

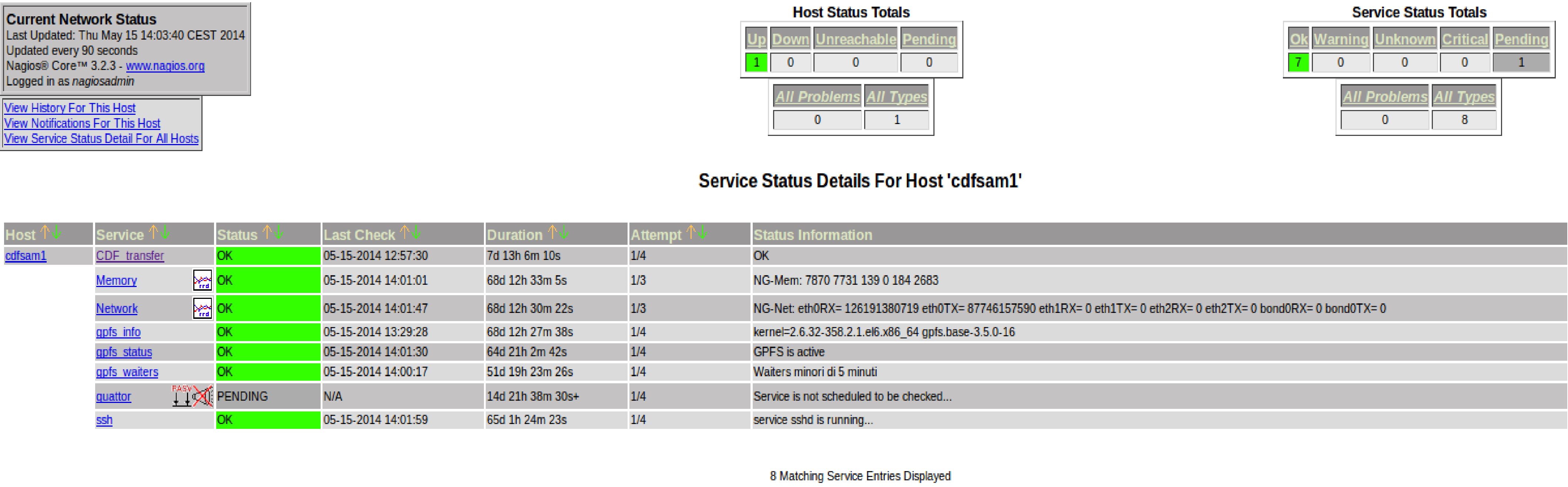
Ok	Warning	Unknown	Critical	Pending
14	0	0	0	0

All Problems	All Types
0	14

Service Status Details For Host 'ds-119'

Host ↑↓	Service ↑↓	Status ↑↓	Last Check ↑↓	Duration ↑↓	Attempt ↑↓	Status Information
ds-119	Alimentatori	OK	05-07-2014 13:55:58	63d 11h 1m 45s	1/4	ipmi power supply ok
	Certificate	OK	05-07-2014 07:36:56	63d 11h 1m 40s	1/4	Il certificato scade tra 254 giorni
	Memory	OK	05-07-2014 14:26:32	50d 19h 12m 32s	1/3	NG-Mem: 23966 6108 17858 0 165 872
	Network	OK	05-07-2014 14:25:16	63d 11h 0m 19s	1/3	NG-Net: eth0RX= 16756870 eth0TX= 194194 eth1RX= 19902149849 eth1TX= 18169864329 eth2RX= 360893355838277 eth2TX= 1024282333942 bond0RX= 0 bond0TX= 0
	Powerpath	OK	05-07-2014 14:22:01	50d 19h 7m 3s	1/4	Powerpath OK
	Raid_Dischi	OK	05-07-2014 14:06:54	63d 11h 0m 9s	1/4	Raid State Optimal
	Yaim	OK	05-07-2014 14:09:37	63d 10h 58m 14s	1/4	Non ci sono comandi yaim da eseguire
	device_IPMI	OK	05-07-2014 14:12:21	50d 18h 16m 43s	1/4	Device /dev/ipmi0 or /dev/ipmi/0 or /dev/ipmidev/0 exist
	gpfs_info	OK	05-07-2014 13:35:21	63d 10h 58m 4s	1/4	kernel=2.6.32-279.14.1.el6.x86_64 gpfs.base-3.5.0-17
	gpfs_status	OK	05-07-2014 14:15:45	50d 19h 13m 19s	1/4	GPFS is active
	gpfs_waiters	OK	05-07-2014 14:27:30	63d 10h 59m 57s	1/4	Waiters minori di 5 minuti
	quattor	OK	05-07-2014 14:23:47	50d 19h 5m 17s	1/4	service ncm-cdispd is running...
	ssh	OK	05-07-2014 14:22:59	63d 10h 57m 51s	1/4	service sshd is running...
	storm-globus-gridftp	OK	05-07-2014 14:25:59	63d 10h 57m 47s	1/4	service storm-globus-gridftp is running...

Data transfer : monitoring



Current Network Status
Last Updated: Thu May 15 14:03:40 CEST 2014
Updated every 90 seconds
Nagios® Core™ 3.2.3 - www.nagios.org
Logged in as *nagiosadmin*

View History For This Host
View Notifications For This Host
View Service Status Detail For All Hosts

Host Status Totals

Up	Down	Unreachable	Pending
1	0	0	0

All Problems	*All Types*
0	1

Service Status Totals

Ok	Warning	Unknown	Critical	Pending
7	0	0	0	1

All Problems	*All Types*
0	8

Service Status Details For Host 'cdfsam1'

Host ↑↓	Service ↑↓	Status ↑↓	Last Check ↑↓	Duration ↑↓	Attempt ↑↓	Status Information
cdfsam1	CDF_transfer	OK	05-15-2014 12:57:30	7d 13h 6m 10s	1/4	OK
	Memory	OK	05-15-2014 14:01:01	68d 12h 33m 5s	1/3	NG-Mem: 7870 7731 139 0 184 2683
	Network	OK	05-15-2014 14:01:47	68d 12h 30m 22s	1/3	NG-Net: eth0RX= 126191380719 eth0TX= 87746157590 eth1RX= 0 eth1TX= 0 eth2RX= 0 eth2TX= 0 bond0RX= 0 bond0TX= 0
	gpfs_info	OK	05-15-2014 13:29:28	68d 12h 27m 38s	1/4	kernel=2.6.32-358.2.1.el6.x86_64 gpfs.base-3.5.0-16
	gpfs_status	OK	05-15-2014 14:01:30	64d 21h 2m 42s	1/4	GPFS is active
	gpfs_waiters	OK	05-15-2014 14:00:17	51d 19h 23m 26s	1/4	Waiters minori di 5 minuti
	quattor	PENDING	N/A	14d 21h 38m 30s+	1/4	Service is not scheduled to be checked...
	ssh	OK	05-15-2014 14:01:59	65d 1h 24m 23s	1/4	service sshd is running...

8 Matching Service Entries Displayed

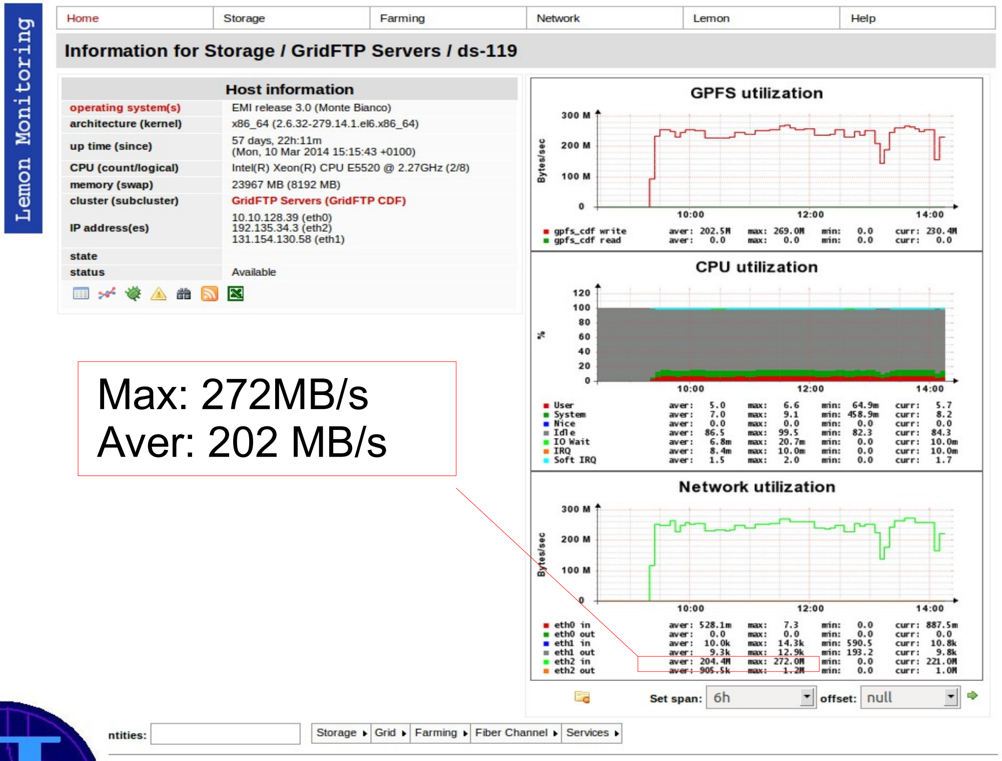
Data transfer : monitoring

LEMON

Display statistical Graph with:
- Information about machine
- GPFS utilization
- Cpu utilization
- Network utilization
- other...

Granularity = 5 minutes

Plot data for different time intervals: from 24h to 10 years



Information for Storage / GridFTP Servers / ds-119

Host information	
operating system(s)	EMI release 3.0 (Monte Bianco)
architecture (kernel)	x86_64 (2.6.32-279.14.1.el6.x86_64)
up time (since)	57 days, 22h:11m (Mon, 10 Mar 2014 15:15:43 +0100)
CPU (count/logical)	Intel(R) Xeon(R) CPU E5520 @ 2.27GHz (2/8)
memory (swap)	23967 MB (8192 MB)
cluster (subcluster)	GridFTP Servers (GridFTP CDF)
IP address(es)	10.10.128.39 (eth0) 192.135.34.3 (eth2) 131.154.130.58 (eth1)
state	
status	Available

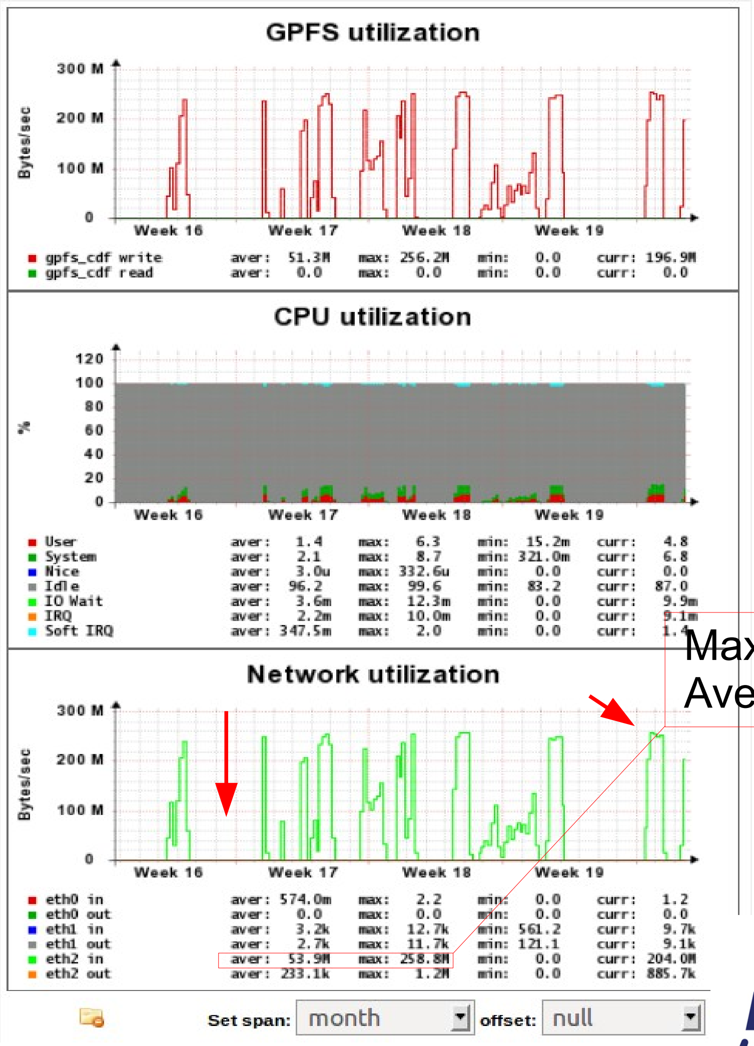
```
■ gpfs_cdf write      aver: 202.5M   max: 269.0M   min:   0.0    curr: 230.4M
■ gpfs_cdf read       aver:   0.0    max:   0.0    min:   0.0    curr:   0.0
```

```
■ User        aver:   5.0    max:   6.6    min:  64.9m   curr:   5.7
■ System      aver:   7.0    max:   9.1    min: 458.9m   curr:   8.2
■ Nice        aver:   0.0    max:   0.0    min:   0.0    curr:   0.0
■ Idle        aver:  86.5    max:  99.5    min:  82.3    curr:  84.3
■ IO Wait     aver:   6.8m   max:  20.7m   min:   0.0    curr:  10.0m
■ IRQ         aver:   8.4m   max:  10.0m   min:   0.0    curr:  10.0m
■ Soft IRQ    aver:   1.5    max:   2.0    min:   0.0    curr:   1.7
```

```
■ eth0 in     aver: 528.1m   max:   7.3    min:   0.0    curr: 887.5m
■ eth0 out    aver:   0.0    max:   0.0    min:   0.0    curr:   0.0
■ eth1 in     aver:  10.0k   max:  14.3k   min: 590.5    curr:  10.8k
■ eth1 out    aver:   9.3k   max:  12.9k   min: 193.2    curr:   9.8k
■ eth2 in     aver: 204.4M   max: 272.0M   min:   0.0    curr: 221.0M
■ eth2 out    aver: 905.5k   max:   1.2M   min:   0.0    curr:   1.0M
```

Set span: 6h ▾ offset: null ▾

Max: 272MB/s
Aver: 202 MB/s



```
■ gpfs_cdf write      aver:  51.3M   max: 256.2M   min:   0.0    curr: 196.9M
■ gpfs_cdf read       aver:   0.0    max:   0.0    min:   0.0    curr:   0.0
```

```
■ User        aver:   1.4    max:   6.3    min:  15.2m   curr:   4.8
■ System      aver:   2.1    max:   8.7    min: 321.0m   curr:   6.8
■ Nice        aver:   3.0u   max: 332.6u   min:   0.0    curr:   0.0
■ Idle        aver:  96.2    max:  99.6    min:  83.2    curr:  87.0
■ IO Wait     aver:   3.6m   max:  12.3m   min:   0.0    curr:   9.9m
■ IRQ         aver:   2.2m   max:  10.0m   min:   0.0    curr:   9.1m
■ Soft IRQ    aver: 347.5m   max:   2.0    min:   0.0    curr:   1.4
```

```
■ eth0 in     aver: 574.0m   max:   2.2    min:   0.0    curr:   1.2
■ eth0 out    aver:   0.0    max:   0.0    min:   0.0    curr:   0.0
■ eth1 in     aver:   3.2k   max:  12.7k   min: 561.2    curr:   9.7k
■ eth1 out    aver:   2.7k   max:  11.7k   min: 121.1    curr:   9.1k
■ eth2 in     aver:  53.9M   max: 258.8M   min:   0.0    curr: 204.0M
■ eth2 out    aver: 233.1k   max:   1.2M   min:   0.0    curr: 885.7k
```

Set span: month ▾ offset: null ▾

Max: 258MB/s
Aver: 54 MB/s