

Cloudscaping the Data Center

The Experience of the INFN National
Computing Center

SLAC, Feb 21, 2012

Davide Salomoni

([Davide.Salomoni@cnafr.infn.it](mailto: Davide.Salomoni@cnafr.infn.it))

Summary slide

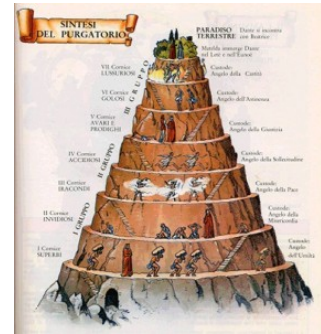


VS.



Agenda

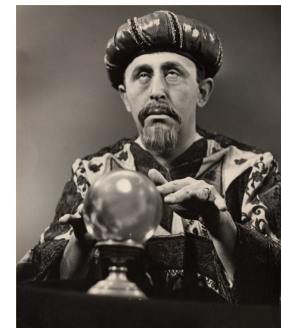
- Introduction: The Context



- The INFN Tier-1: The Status



- (More) Clouds at the horizon: The Challenges



About me

- Physics degree from U Bologna in 1990, then for 8 years with INFN
 - Working mostly with networks and network protocols
- At SLAC, from Jan 1999 to Feb 2001 😊
 - SCS Networking
- From 2001 to 2005 in the Netherlands
 - Working for the private and public sector; R&D with networking and distributed computing
- From 2006 with INFN (again)
 - Computing Manager at the INFN Tier-1 from 2006 to 2011
 - Computing Research Director at CNAF from fall 2011



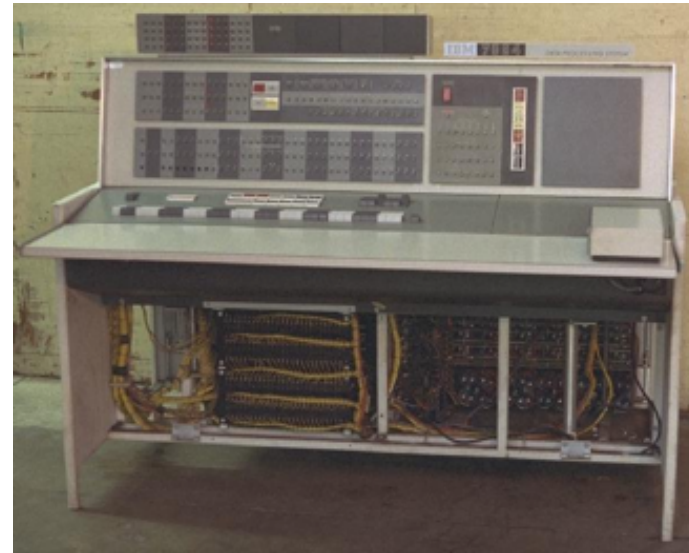
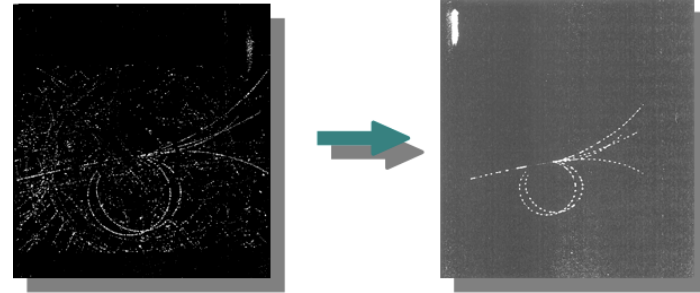
INFN

- **INFN : Italian National Institute of Nuclear Physics** – a research institution with the mission to study the fundamental constituents of matter.
- For more than 60 years (founded in 1951), INFN has been carrying on theoretical and experimental research in the fields of **subnuclear, nuclear, astro-particle physics and research and development in related technological areas.**
- Today: **19 sites** spread all over Italy, **4 national labs** (Frascati, Legnaro, Catania, Gran Sasso), and **1 national computing center** (CNAF, in Bologna).
- Currently about 2,000 employees, plus approx. 2,450 university researchers and professors, and approx. 1,300 students and research associates.



CNAF then...

- Created in 1962 in Bologna, for the purpose of **high-precision digitalization of pictures coming out of bubble chambers**
 - Hence the acronym, “**C**entro **N**azionale **A**nalisi **F**otogrammi”.
- One of the first adopters in Italy of the IBM 7090 – for what it was at the time called “large scale scientific applications”.



... and now

- **80's-90's:** the first definition and implementation of the Italian Internet
 - CNAF becomes the main driver of the new INFN national network, then migrated into GARR, the Italian Academic and Research Network. CNAF hosted the first GARR NOC (moved to Rome in 2001).
- **2000-today:** INFN creates and opens its **National Computing Center** at CNAF to serve scientific experiments and, in particular, those at the LHC.
- In the same years, the **Grid architecture** is defined and implemented. CNAF attracts a large number of physicists and computing experts working to define computing models and software frameworks.
- **Today: approx. 60 people (23 staff)**

The INFN Tier-1

- Called “Tier-1” to emphasize its role in the Worldwide LHC Computing Grid, but it also acts as Tier-0, Tier-2 or Tier-3 for other experiments.
 - The Tier-1 currently supports about 20 scientific international collaborations
 - It is rather young: officially opened in 2005, re-engineered in 2009.
- A 1,000m² computing room, with space for 120 racks and several tape libraries. 5 MVA of electrical power, 6 chillers, 2 power lines, 2 rotary UPS systems + diesel engine.
- More than 1,300 CPU servers, $O(10^4)$ cores, about 110 KHEP-SPEC06.
- Approx 9 PB of disk space, 10 PB of tape space.
- Connected to CERN and other computing centers with an aggregated networking capacity of about 40 Gbit/s.
- Redundant technological infrastructure for 24x7 operations.

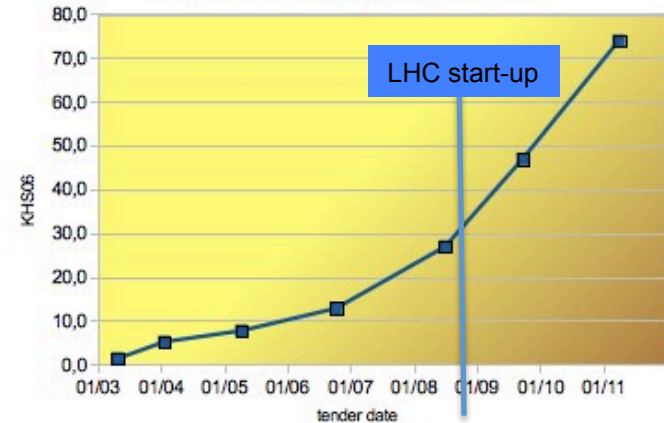
Resources at the INFN Tier-1

- **Exponential growth trend** in resource acquisition (2011 and 2012 tenders not considered in the plots)
 - Typically one tender per year
- Emphasize **resource sharing**

CNAF PLAN APRIL 2011

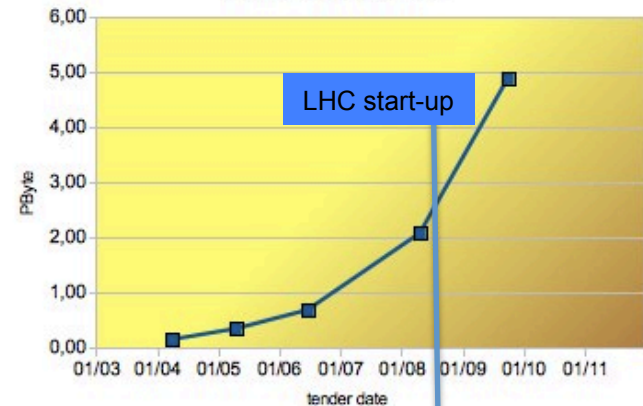
Experiment	2011			2012		
	CPU HS06	DISK TB-N	TAPE TB	CPU HS06	DISK TB-N	TAPE TB
ALICE	22200	1501	2400	25890	1749	3952
ATLAS	22600	2480	3000	25900	2700	3600
CMS	18300	2400	6500	18850	2860	6630
LHCB	9750	525	520	16950	1425	930
Total LHC TIER1	72850	6906	12420	87590	8734	15112
BaBar	2360	350	0	1600	350	0
SuperB (dal 2011)	2500	50	0	2500	50	0
CDF	7000	300	15	7000	300	15
LHCB TIER2	5400	0	0	7200	0	0
TOTALE GRUPPO I	17260	700	15	18300	700	15
AMS2	2457	143	50	2745	211	55
ARGO	800	160	752	800	184	1086
AUGER	1200	110	0	1200	110	0
FERMI/GLAST	1400	60	40	1400	60	40
MAGIC	450	30	50	500	30	60
PAMELA	600	60	80	600	48	64
Virgo	7500	469	348	7500	660	660
TOTALE GRUPPO II	14407	1032	1320	14745	1303	1965
All experiments	104517	8638	13755	120635	10737	17092
All w/ overlap factor	87098	7853	13755	100529	9761	17092
CNAF TOTAL (PLAN)	87098	7853	13755	100529	9761	17092
overlap mitigation				102098	9761	
Effective overlap				1.18	1.10	
CNAF to be procured with overlap mitigation	21171	1148	5294	13432	1558	3337
				15000	1558	

Growth of compute capacity at CNAF



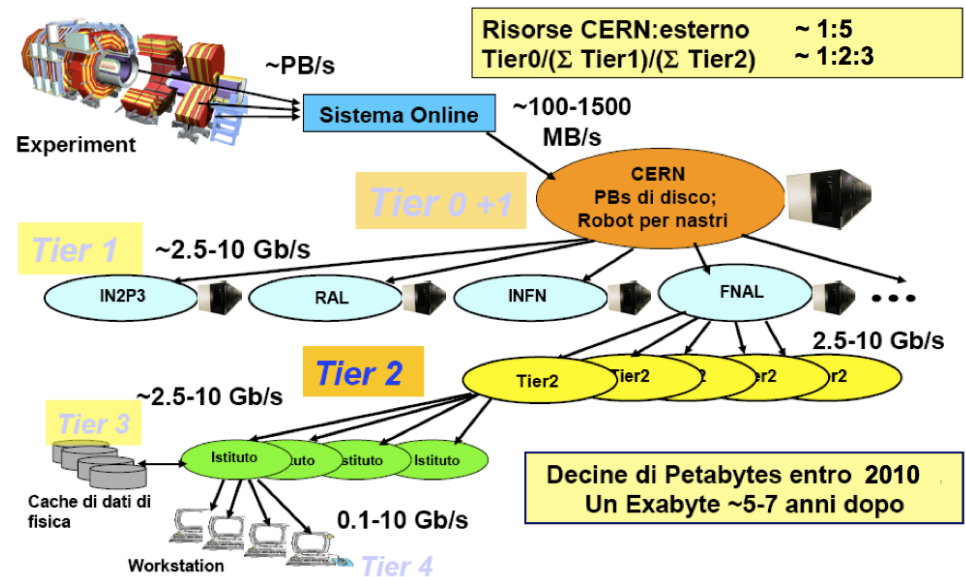
Growth of disk capacity at CNAF

> 6M€ in the past 8 years



Grid-based Processing

- The LHC case: originally, a rather rigid **hierarchical architecture** – See the Monarc model, circa 2000
- **Uniformity** of environments translates to simplification – but it is also a hindrance

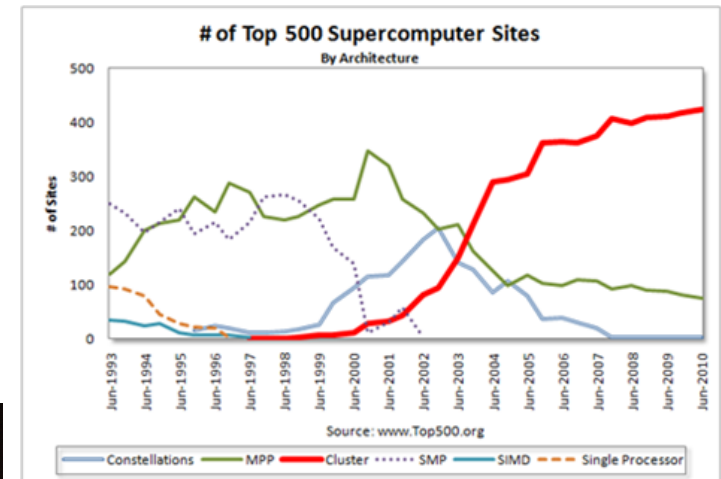


Distributed computing, main historical drivers

- **Cost:** the cost of a cluster made of many cheap computers can be lower than the cost of a single supercomputer
 - See also **energy costs**



- **Reliability:** avoid SPoF
- **Scalability / expansion:** modular architecture

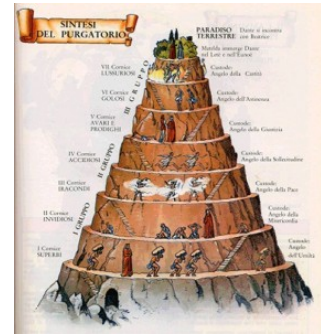


(intermezzo)



Agenda

- Introduction:
The Context



- The INFN Tier-1:
The Status



- (More) Clouds at the horizon:
The Challenges

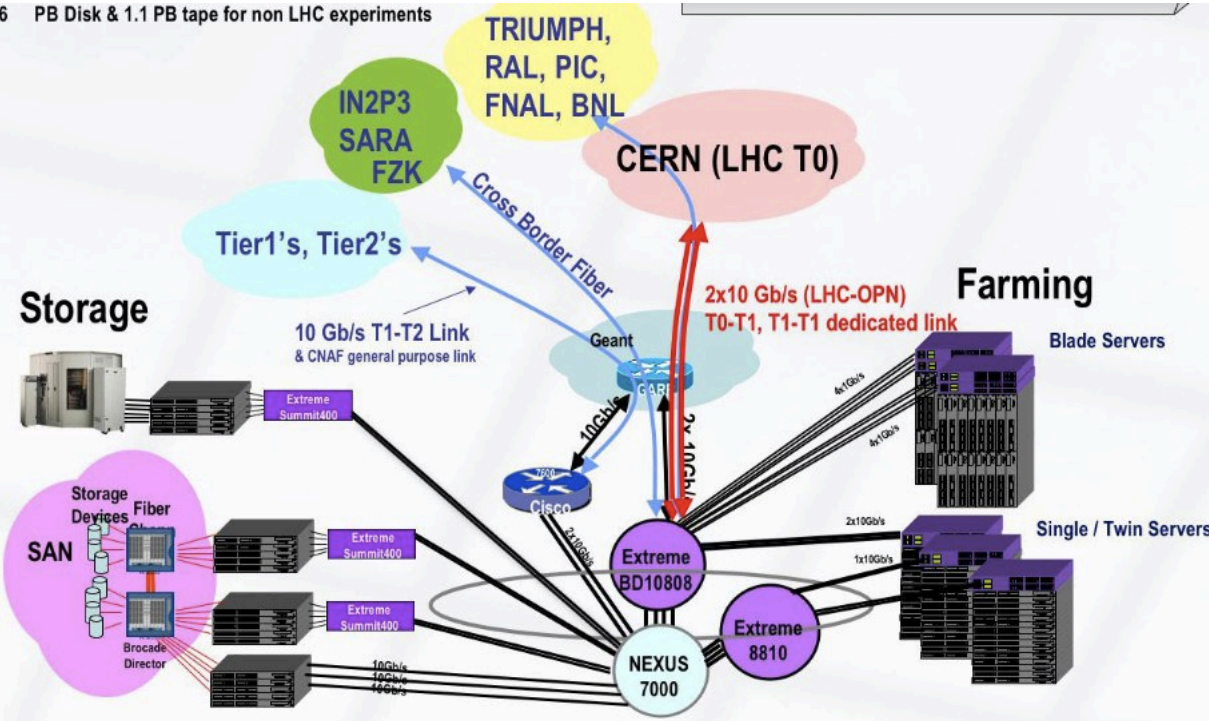


Main roles (relevant here) of CNAF

- **CNAF charter:** develop, implement, manage equipment / services, conduct technological R&D work serving the mission of INFN
 - Tier-1 director
 - Storage, Network, Farming, Infrastructure services
 - R&D director
 - New services, national/int'l research projects (e.g., IT/EU projects, WLCG R&D, Cloud & virtualization, SuperB, Intel MIC, etc.)
 - User support director
 - User support, outreach
- **The R&D and the Tier-1 parts are actually tightly coupled.**
 - Explicitly, we don't have nor want (anymore) any "R&D vs. operations" rigid distinction.

The Tier-1 at a glance (Oct 2011)

1.6 PB Disk & 1.1 PB tape for non LHC experiments



Storage

NOW	2012
9 PB disk capacity (SAN)	12 PB disk capacity (SAN)
110 Disk Servers (50% 10Gbit)	110 Disk Servers (50% 10Gbit)
10 Tape TSM-HSM clients	12 Tape TSM-HSM Clients
1 FC Directors (core switches)	1 FC Directors
20 FC edge switches (peripheral)	20 FC edge switches
10 PB On-Line Tape Capacity	18 PB On-Line Tape Capacity

GrifFTP StoRM GPFS TSM

2 TSM server (one in stand-by)
110 GPFS server
6 StoRM instance
25 GridFTP Server
640 TB Disk & 620 TB tape for LHCb
2.2 PB Disk & 3.6 PB tape for CMS
2.4 PB Disk & 1 PB tape for ATLAS
1.35 PB Disk & 300 TB tape for ALICE
1.6 PB Disk & 1.1 PB tape for non LHC experiments

Network

Core Route/Switches	WAN Connections
4 Core Switches	2 x 10Gb/s T0 (CERN)-T1, T1-T1
200 x 10Gb/s Ports	1 x 10Gb/s T1-T2, General purpose
468x 1Gb/s Ports	1 x 10Gb/s T1-T1 (Karlsruhe, IN2P3, SARA)
Aggregation Switches	
78 Switches (1 Rack Unit)	
21 Blade Switches	
4300x 1Gb/s Ports	
100 x 100Mb/s Ports	
36 x 10Gb/s	

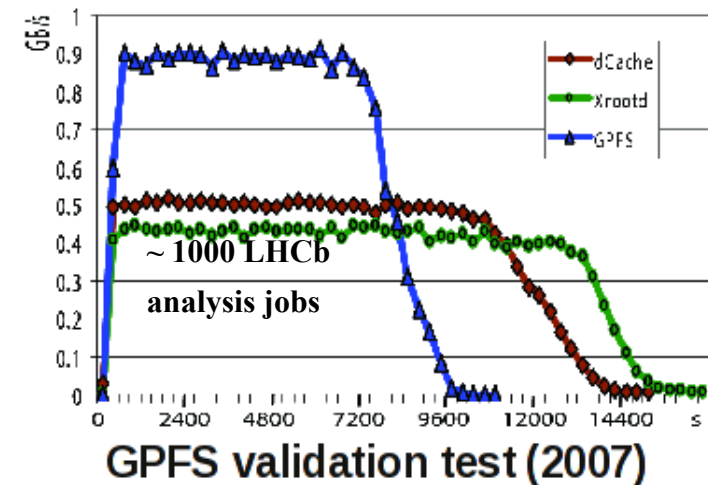
- ### Farming
- The Tier-1 common cluster computing power is 100,000 HEP-SPEC06 (to be brought to about 125,000 HEP-SPEC06 by 2012) with 10,000 CPU cores.
 - About 20 scientific international collaborations are using Tier-1 resources.
 - More than 50,000 computing jobs are executed every day on the computing farm.
 - Virtual Machines are transparently and dynamically provisioned as Virtual computing nodes using the INFN Worker Nodes on Demand Service (WNoDeS)

Infrastructure

Power supply	15000 V
Power transformers	3 (~2500 kVA)
Racks	> 120 units
Chillers	7 (~2740kVA)
UPSs	<ul style="list-style-type: none"> 2 rotary UPSs (high-mass spinning flywheel + diesel engine) providing redundant emergency power for computational, storage and network units up to 3400 kVA 1 diesel engine providing emergency power for chilling units up to 1200 kVA

Storage: MSS, 2003-2007

- **CASTOR** was the “traditional” solution for MSS at CNAF for all VO's since 2003
 - CMS has historically been the main CASTOR user (end Q3 2009: ~ 1 PB on tape)
- **Large number of issues**
 - At the set-up/admin and at the VO level (complexity, scalability, stability, support)
 - Still, successfully used in production, despite with sometimes large operational overhead
- In parallel to production, in 2006 CNAF started to **search for a potentially more scalable, performant and robust solution**
 - Q1 2007: GPFS (from IBM) adopted for disk-based storage after extensive comparison tests
 - outstanding I/O perf, stability and easiness of mgt
 - Q2 2007: StoRM (developed at INFN) implemented the SRM 2.2 specs
 - Q3-Q4 2007: StoRM/GPFS in production for D1T0 for LHCb and ATLAS
 - Clear benefits for both experiments (highly reduced load on CASTOR)
 - No major impact on CMS workflows (no large use of D1T0)
- **However, we were still looking for a complete MSS solution based on StoRM/GPFS**



Storage: MSS, 2007-now

End 2007: a project was started to define a **comprehensive grid-enabled HSM solution based on StoRM/GPFS/TSM**

- StoRM was extended to include the SRM methods required to manage data on tape
- GPFS specific features (available since version 3.2) were combined with TSM (also from IBM) and StoRM
- An interface between GPFS and TSM was implemented (not all needed functionalities were provided out of the box)

Q2 2008: First implementation (D1T1, i.e. w/o user driven recalls) in production for LHCb (CCRC'08)

Q2 2009: “GEMSS” (StoRM/GPFS/TSM) supporting a full HSM solution ready for production at CNAF

- Pre-production test-bed built to accommodate the scaling needs of CMS

Q3 2009: **CMS@CNAF moved from CASTOR to GEMSS**

GEMSS: GPFS/TSM/StoRM integration

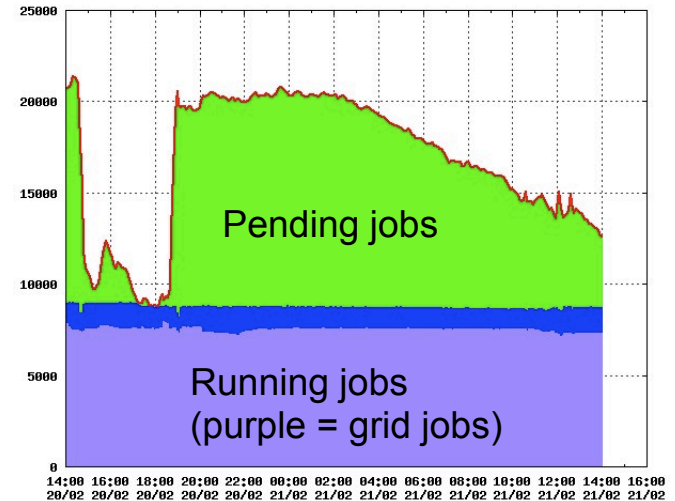
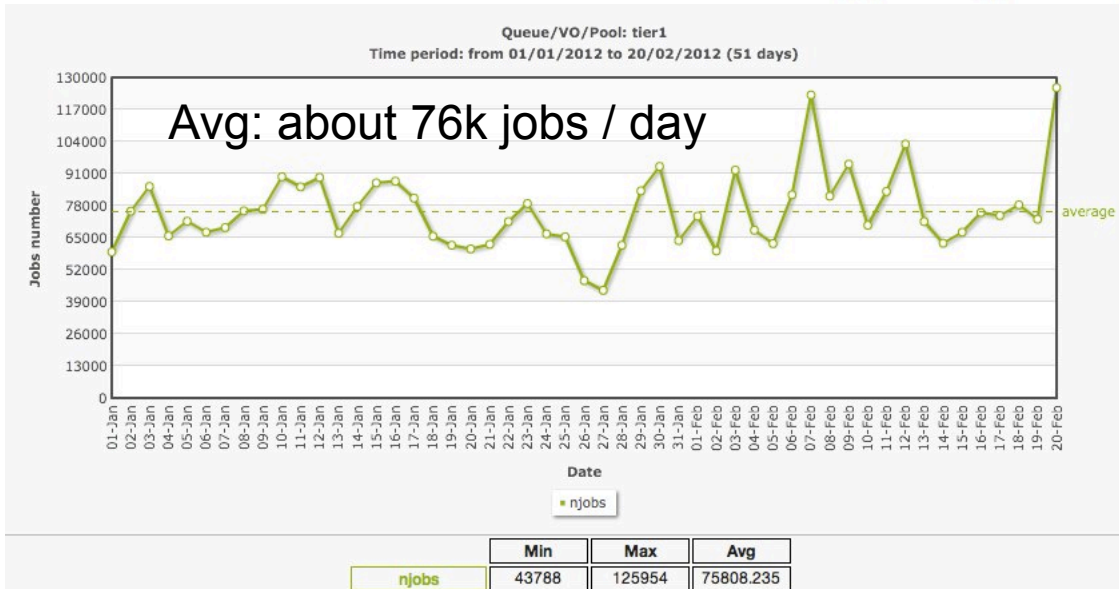
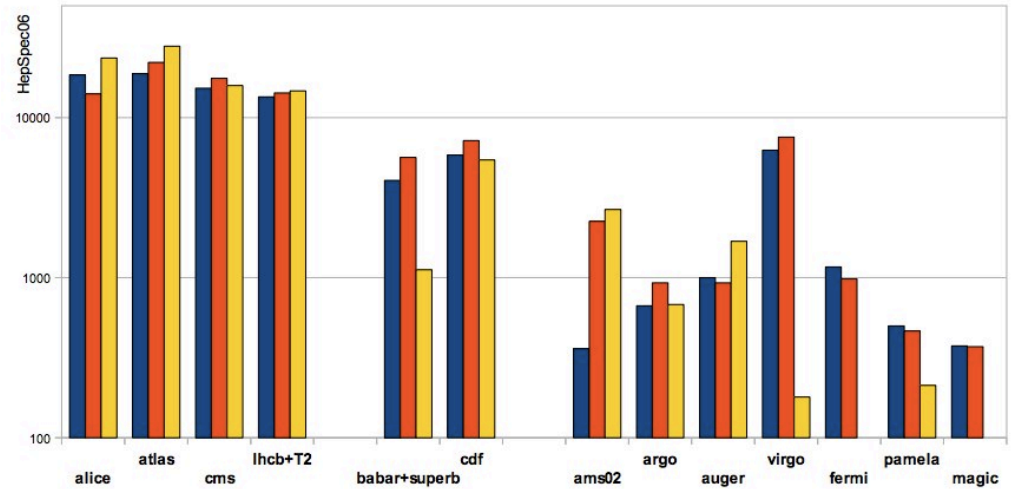
- StoRM (developed by INFN) implements SRM 2.2
 - In use at the INFN Tier-1 since 2007 and at other centers for T0D1 service challenges
 - Designed to leverage the advantages of parallel / POSIX file systems in a Grid environment
- We combined the features introduced in GPFS v3.2 (*now running 3.4*) and TSM with StoRM, to provide a transparent grid-enabled HSM solution.
 - The GPFS Information Lifecycle Management (ILM) engine is used to identify candidates files for migration to tape and to trigger the data movement between the disk and tape pools
- An interface between GPFS and TSM (named YAMSS) was implemented to enable tape-ordered recalls
 - For the ALICE experiment, an xrootd plug-in was developed
- GEMSS is now used by all the experiments supported at CNAF
- Future: while TSM licensing does not worry us too much, GPFS does.

Farming

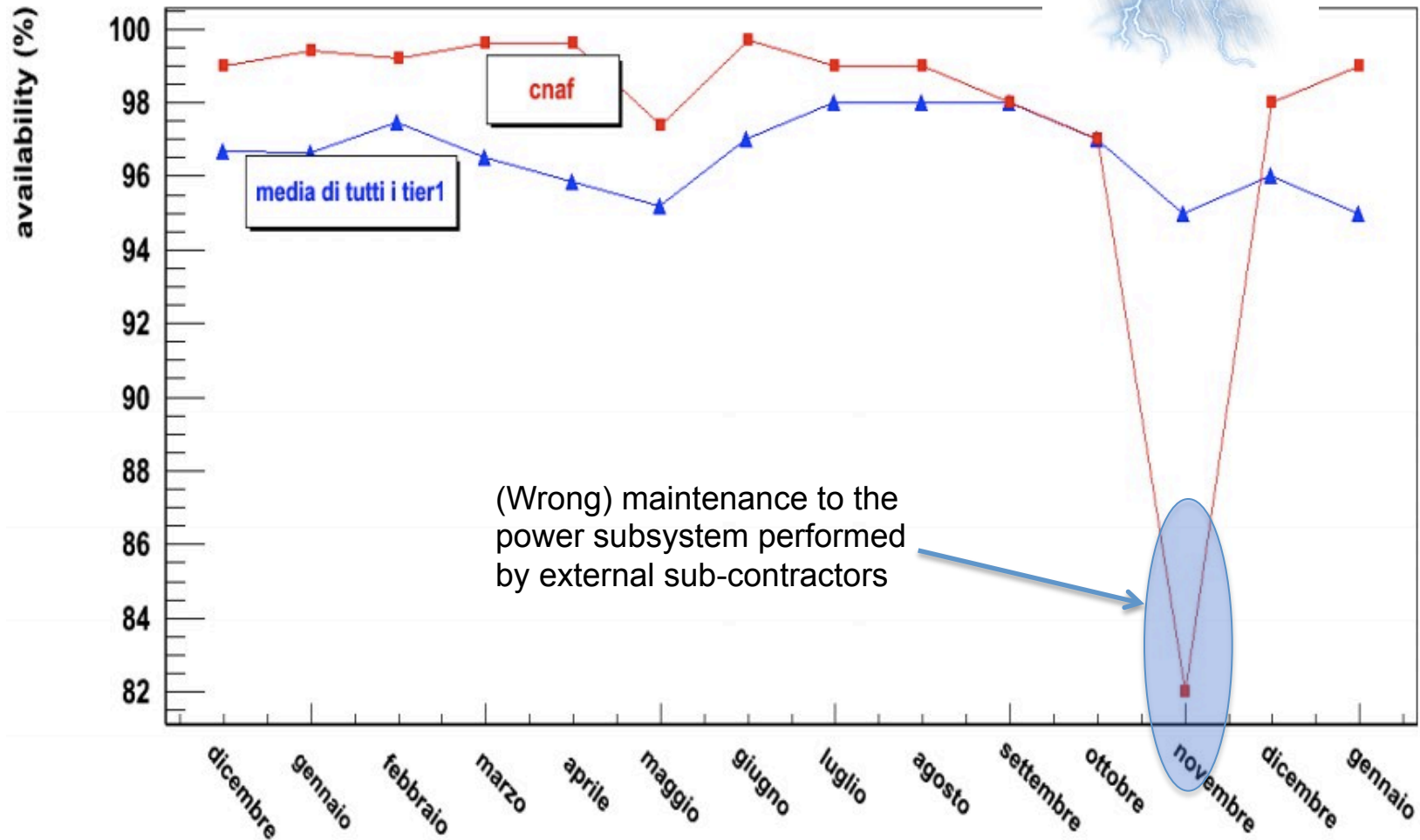
- Currently about 110 KHS06, **all servers installed in a single big cluster**
 - Heterogeneous hardware (from 8-core to 24-core systems, multiple vendors)
 - Very limited installation (and use) of GPU's
 - Very limited MPI requirements so far
 - **Compute nodes are now all redundant wrt power supply**
- **The LRMS is LSF (7.0.6)**
 - Hierarchical fairsharing
 - INFN-wide licenses
 - LSF licenses not playing very well with how we use dynamic virtualization (see later) – a similar licensing problem may arise with GPFS
 - Double support channel (local integrators and Platform/IBM)
 - Currently evaluating support for many-core requests (not MPI)
 - There is some growing interest in evaluating alternatives to LSF
- **O(60) machines used to support services** like squid servers, Grid computing elements, info systems, monitoring, accounting servers
 - Several on (static) VM's

Farming usage

- Fairsharing, example from Jan 1, 2012 to Feb 16, 2012
- Labels: **HS06 pledged** vs. **HS06 available** (installed) vs. **HS06 actually used**

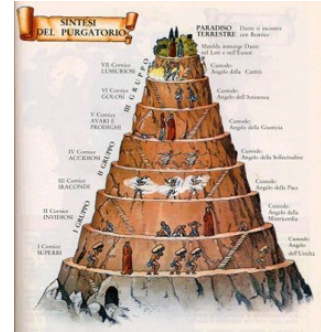


Tier-1 availability, Jan 2011-Jan 2012



Agenda

- Introduction: The Context



- The INFN Tier-1: The Status



- (More) Clouds at the horizon: The Challenges



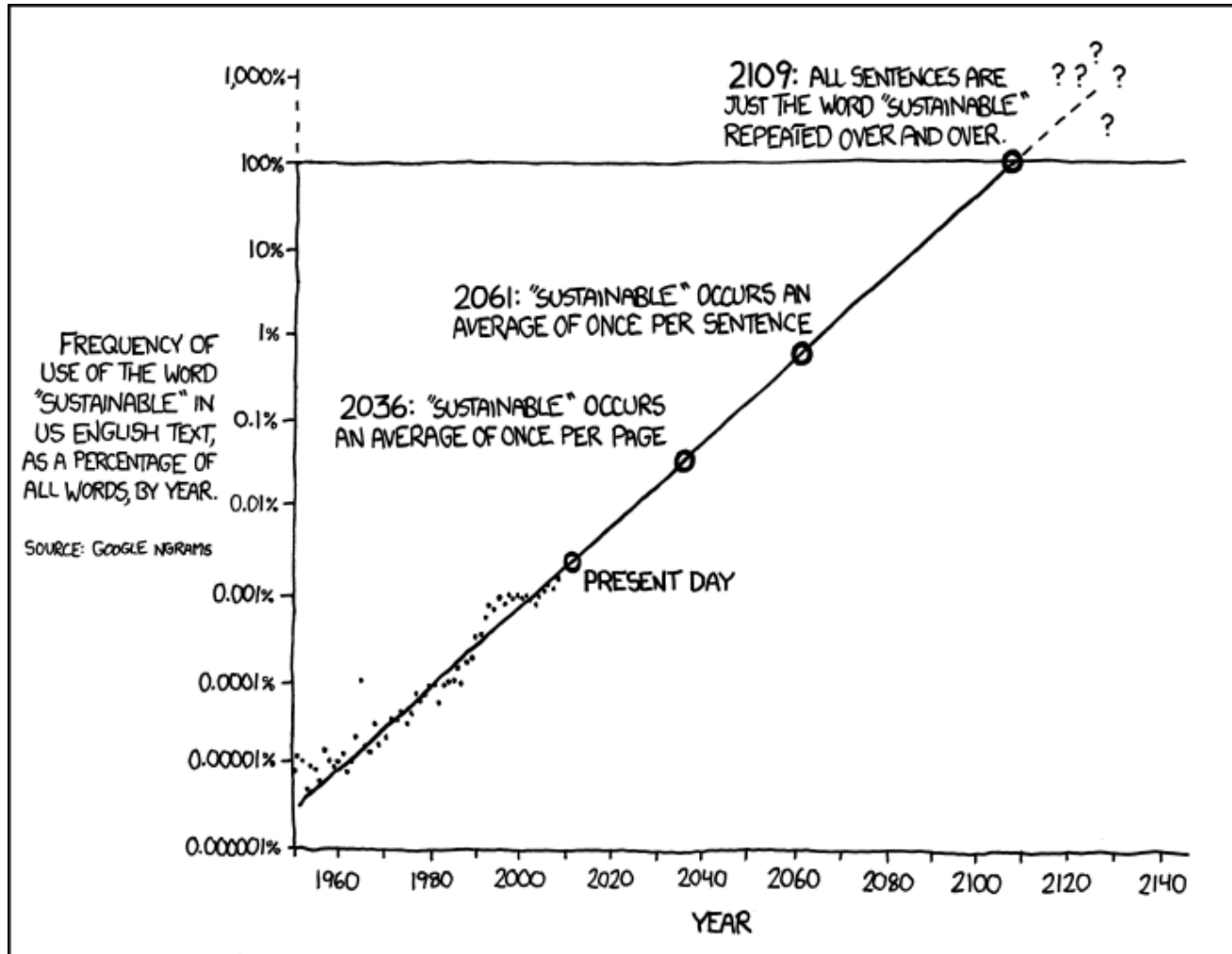
Challenges



One-size-does-not-fit-all

- A first issue: with several tens of different scientific collaborations, we need to **flexibly adapt to their needs**
 - The “one middleware, one O/S for everybody” idea just doesn’t cut it (anymore)
 - Hence, we have a somewhat detailed R&D program regarding (for example) efficient dynamic virtualization and service provisioning
- **More in general**, our current main customer (LHC) will stop taking data in a few years
 - This is not to say there won’t be the need to continue working on LHC data, but... **sustainability** and **protecting valuable INFN know-how** are important issues.

(intermezzo)

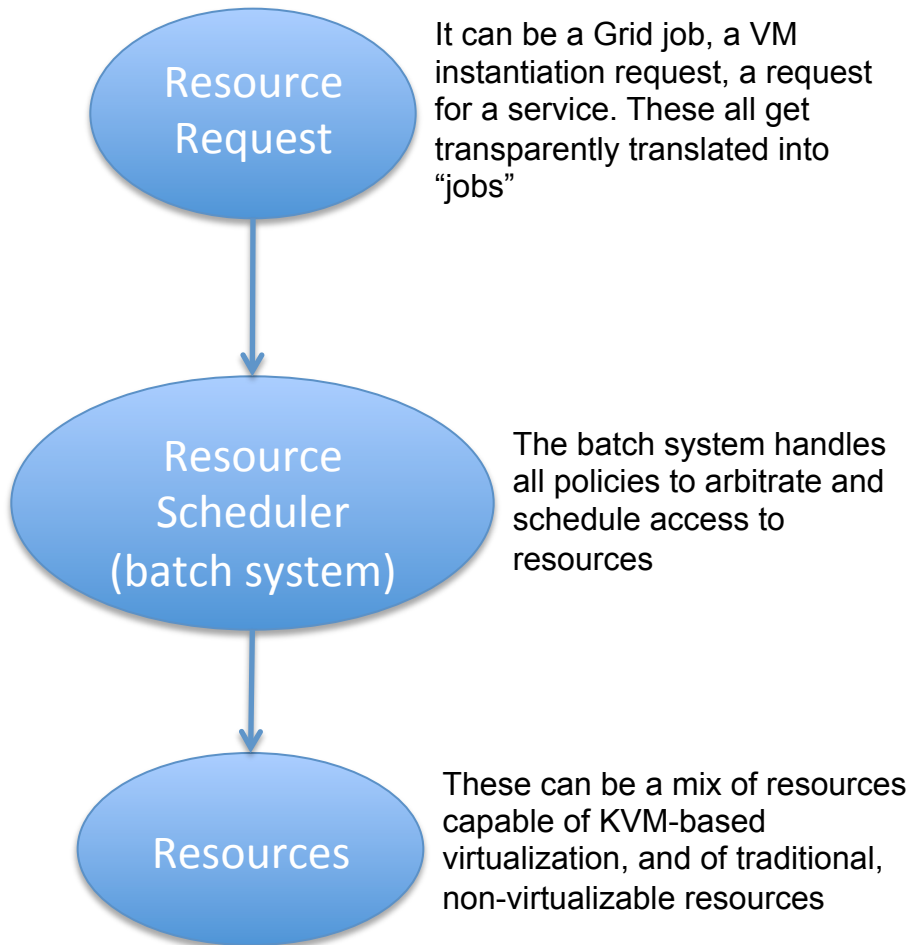


THE WORD "SUSTAINABLE" IS UNSUSTAINABLE.

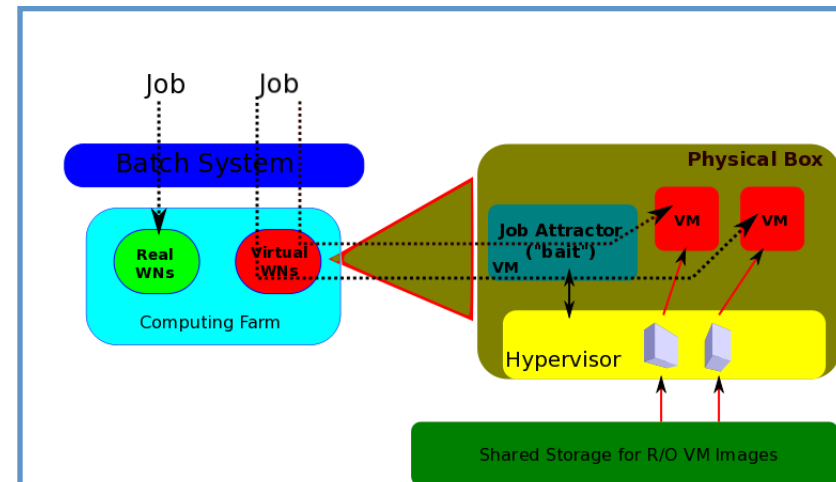
WNoDeS

- A software framework created by INFN to **integrate Grid and Cloud provisioning**
 - **Key feature**: all resources (presented via Grid, Cloud, or else) are taken from a *common pool* to avoid static partitioning
- **Scalable and reliable** – it is **in production** at several Italian centers, including the INFN Tier-1 since November 2009
 - Currently managing about 2,000 on-demand Virtual Machines (VMs) there
- Totally **transparent** for both users of Grid services and for users of traditional Computing Centers
- Supporting a **native Cloud** interface
 - OCCI (Open Cloud Computing Interface) compliant
 - A Cloud Web portal
- Integrating **authentication, policy and accounting**
- **Leveraging proven open source software** technologies like Linux KVM, Torque/Maui (Platform LSF also supported; SLURM support being considered), EMI gLite middleware
- Easily **expandable** in Python

WNoDeS, a synthetic architectural overview



General schema to handle a VM/service/job instantiation request



Every piece of hardware runs a specialized VM called "bait" whose purpose is to arbitrate access to dynamically created local VMs

The need for a “bait”

- A **bait** in WNoDeS is an LSF client sitting inside a VM (one bait VM per physical hardware) used to **attract jobs** from the LRMS.
- There is **no fundamental reason to have the bait process in a VM** and not in the hypervisor, *except* for the following reasons:
 - With the bait, no jobs can ever run on the hypervisor. This may be regarded as (mild) additional security. The hypervisor can have private IP addresses and be inaccessible from the outside.
 - Since LSF uses licenses based on the number of detected cores in a client, if we ran the bait on the hypervisor we’d need N licenses for the hypervisor itself, plus other N licenses for the VM’s → **an N-core system would eat $O(2N)$ LSF licenses.**
- **Beta** versions of WNoDeS offer the **option to run the bait either stand-alone or on the hypervisor** (see later why).

Caveats

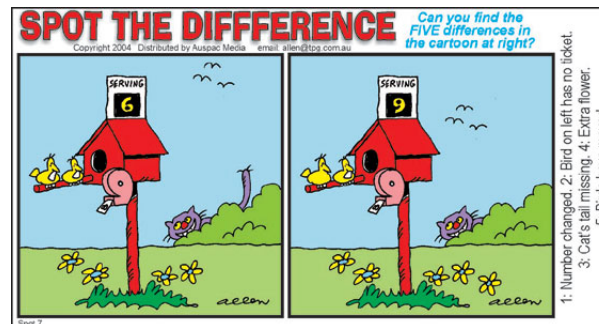
- Small/medium size VM deployments are not difficult. With $O(10^3)$ VM's *in production*, however, we observed a few possible issues.
 - Every VM (typically, one VM per physical core) becomes a GPFS client for data access. This leads to a very large GPFS cluster, needing special tuning.
 - GPFS is very sensitive to how a VM is shut down. No data is lost, but with very large numbers of VM's, taking down multiple VM's at a time abruptly may slow down the file system.
 - Similarly, every VM becomes an LRMS client. For example, beyond 4,000 LSF client, we had serious issues with job dispatching → solved with proper tuning.
- **Solutions** (beyond tuning GPFS and LRMS for large clusters):
 - With Cloud computing, there is no need (and actually no desire) to run an LRMS on a VM. WNoDeS only runs an LRMS client on a VM's when this VM has to handle traditional batch jobs. One can also reduce the size of an LRMS cluster e.g. with LSF MultiCluster.
 - WNoDeS VM's may be configured to avoid accessing a shared storage directly. For example, we have now GPFS on the hypervisors only (dramatically reducing the GPFS cluster size); the hypervisor then exports the GPFS file systems via NFS to its own VM's only and effectively is a GPFS/NFS gateway. VM's only need an NFS client. Performance figures are available.

Extending WNoDeS to Grid computing

- The use case is to let **Grid jobs request and use VM's** and in WNoDeS it is a simple extension of normal VM provisioning. Two possibilities:
 - All jobs belonging to certain Virtual Organizations (VOs) can be directed to pre-packaged VMs. **This is completely transparent for users.**
 - Grid users can explicitly specify **which VM they want their jobs to run on.**
 - Using standard EMI (European Middleware Initiative) job management tools.

Introducing Clouds

- The essence of the [**Grid definition**] can be captured in a simple checklist, according to which a Grid is a system that:
 - coordinates resources that are **not subject to centralized control...**
 - ... using **standard, open, general-purpose protocols and interfaces...**
 - ... to deliver nontrivial **qualities of service.**
 (I. Foster, What is the Grid? A Three Point Checklist, 2002)
- **Cloud Computing** is a model for **enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources** (e.g., networks, servers, storage, applications, and services) that can be **rapidly provisioned and released** with minimal management effort or service provider interaction.
 (NIST Working Definition of Cloud Computing)



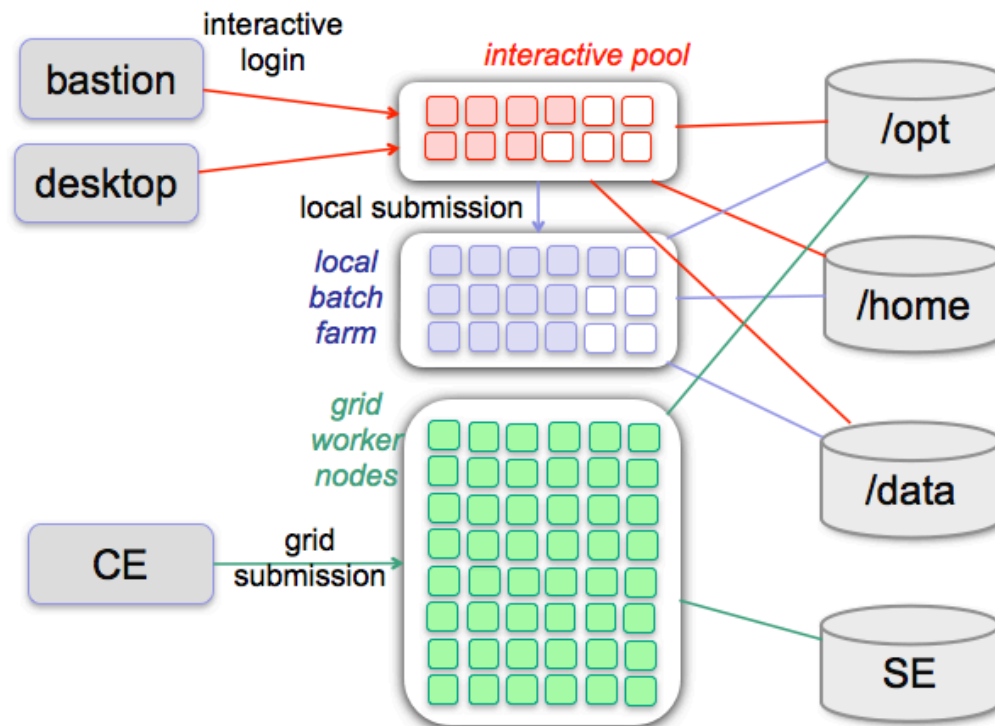
Virtual Interactive Pools (VIP)

- **Self-allocation of systems** by **users of a traditional computing center**
 - Systems are provisioned **from a common pool of resources** so that users can log on to them with their local account (no root access).
 - Users may specify characteristics such as VM image, number of CPUs, amount of RAM, local file systems to be mounted.
 - These systems can be employed by users for instance to create pools of machines for interactive analysis or to instantiate ad-hoc services.

This is a kind of **cloud computing applied to a traditional computing center** designed to efficiently offer new services, without incurring the overhead to dedicate resources for this purpose.

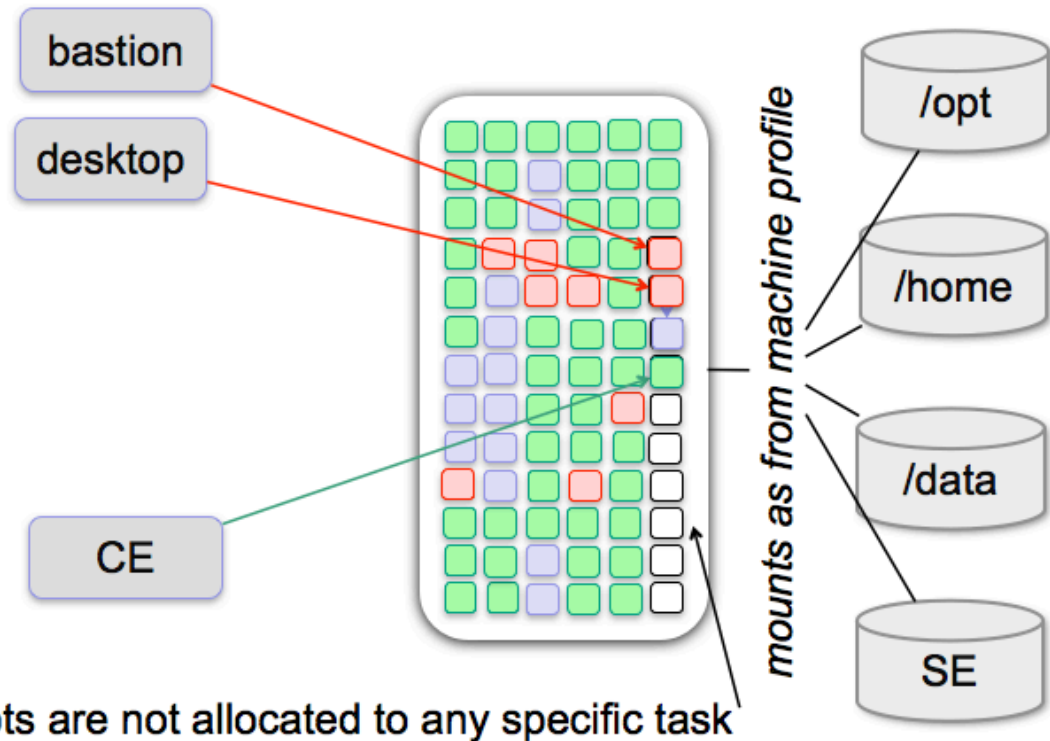
VIP in practice

Classical model



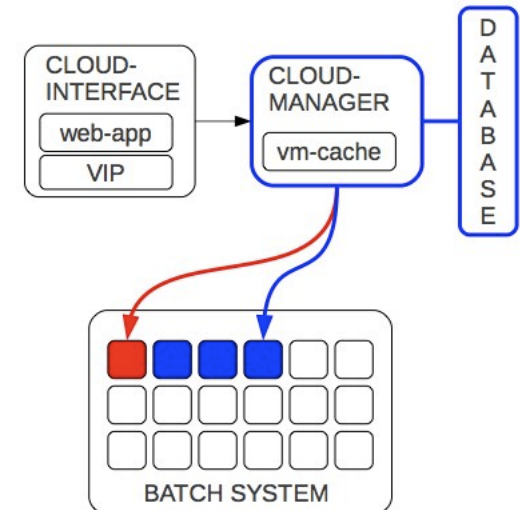
VIP in practice

Dynamic model



The WNoDeS Cache Manager

- VIP instantiations are constrained by the scheduling process of the LRMS. While this is normally not an issue for batch jobs, it may well be for interactive requests.
- The **WNoDeS Cache Manager** **pre-allocates** a configurable number of VM's so that they can be run straight away.



WNoDeS Cloud support

- **Cloud computing** can be supported by WNoDeS via:
 - VIP (a “kind of” cloud computing)
 - A Web portal
 - Both VIP and the Web portal can use the cache manager
 - The OCCI (Open Cloud Computing Interface) API
- If desired, VM’s can be put into different VLANs (more on this later)

The WNoDeS Cloud Web Portal




WNoDeS
 Grid Resources via Cloud Interface

MY RESOURCES **NEW RESOURCE** CONTACT US

/C=IT/O=INFN/OU=Personal Certificate/L=CNAF/CN=Davide Salomoni 

Create a New Virtual Machine

[Need Support?](#)

Current VO: cms (Change VO)

- 1** Hardware
- 2 Operating System
- 3 Keys
- 4 Create

Select the preferred configuration between the existing, [contact us](#) if you need more customization

- SMALL 1 core, 1.7 GB RAM, 50 GB HD, 100 Mb/s throughput
- MEDIUM 2 cores, 3.5 GB RAM, 100 GB HD, 200 Mb/s throughput
- LARGE 4 cores, 7 GB RAM, 200 GB HD, 400 Mb/s throughput
- EXTRA-LARGE** 8 cores, 14 GB RAM, 400 HD, 800 Mb/s throughput



Web App Integration

- The WNoDeS Web Portal will be integrated into a general scientific portal
- Work supported by the Italian Grid Initiative (IGI)
 - For which WNoDeS is the reference architecture for Grid/Cloud integration.



A portal for an easy access to the IGI grid infrastructure

Marco Benicivenni, Paolo Veronesi, Giuseppe Misurelli, Andrea Ceccanti, Francesco Giacomini, Vincenzo Glaschini, Marco Cecchi, Luciano Galdo, Riccardo Brunetti, Daniele Andreotti, Davide Salomoni, Diego Micheletto



FEATURES

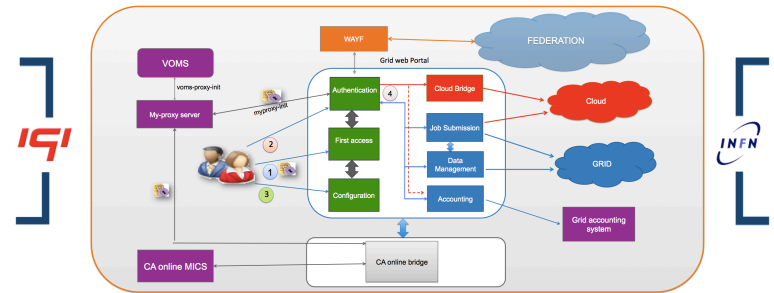
- Federated identity system for user authentication
- Interaction with ON-line CA to transparently request X.509 certificates on behalf of the user
- Personal certificates upload for skilled users
- Possibility to select a VO membership or request new VO membership on behalf of user
- Community related portal views and JDL customization for job submission
- Implementation of workflow submission

GOALS

- Grid job submission via web.
- Provisioning of a Cloud environment via web.
- Making easier the request and management of X.509 certificates and the request for a VO membership.
- Minimizing the job failure rate


IMPLEMENTATIONS

- Web portal based on Liferay framework
- Services implemented by ad hoc portlets (JRS 168 – 286)
- Secure communications with external services using shibboleth and encrypted protocols
- SAML delegation mechanism for X.509 certificate request
- Integration with existing monitoring and accounting system




1 - FIRST ACCESS

- The portal receive a delegation token
- CA bridge module requests to a CA-online a certificate on behalf of the user
- The user digit a passphrase for private key encryption
- The certificate is used to store a long-term proxy on a myproxy server (the private key encrypted will be conserved on my proxy server and the passphrase will be not conserved)



2 - AUTHENTICATION

- The portal redirects user to the his idp login page. Once the proper IDP has authenticated the user he will be automatically logged into the portal
- The portal will ask him the passphrase in order to retrieve the proxy from myproxy server.
- At the same time contact the VOMS server in order to sign the proxy with VO extension.

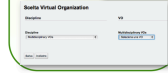


3 - CONFIGURATION

The user once registered can set his options

- Upload a new certificate (one is a default)
- Add new VO memberships (one is a default)
- Request for a new VO membership


For each VO specify the ICAN



4 - GRID / CLOUD ACCESS

At the moment for job submission and data management the portal uses WS-Grade (SZTAKI)

- Other solution under investigation is JSAGA (IN2P3)
- For cloud resources provisioning the portal is interfaced with WNoDeS (INFN-CNAF)
- The accounting portlet provides information for both environments



Testing Cloud access

- Through the [European Grid Infrastructure \(EGI\) Federated Cloud Working Group](#)
- This is a task force established by EGI (Sep 2011-Mar 2013). Main goal is to write a [blueprint document](#) for EGI resource providers that wish to securely federate and share their virtualized environments.
- As part of that goal, the WG deploys a test bed to evaluate the integration of virtualized resources across multiple EGI providers.
- WNoDeS participates to the TF as a Technology Provider, with a test-bed set up at CNAF running on top of PBS

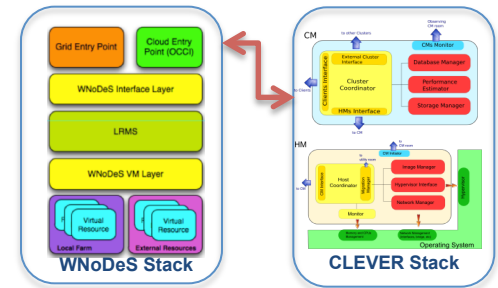


Needs for Cloud computing in WLCG experiments

- See some of the preliminary results of the WLCG Workload Management Technical Evolution Group (WM TEG) at <https://twiki.cern.ch/twiki/bin/view/LCG/WorkloadManagementTechnicalEvolution>
- **Transparent exploitation of public IaaS (Infrastructure as a Service), e.g. Amazon EC2 in addition to “private” resources**
 - E.g. ATLAS integrated its PanDA framework into EC2-provisioned appliances.
- Interest in **adding an EC2 or OCCl “cloud entry point” to traditional resource centers**
 - With “standard” node types across multiple providers
- **Authentication, authorization and billing** need to be clearly defined
 - E.g. fairsharing should take care of both local, Grid and Cloud instantiations.

Hybrid Clouds

- This is in general the **interconnection of Cloud Computing resource centers**
 - It seems sensible for us to capitalize on the multi-year experience in interconnecting resources centers via Grid infrastructures.
 - **Plan: WNoDeS integration with Virtual Infrastructure Management services and with Cloud Manager services provided by the CLEVER research project.**
- CLEVER (a project from University of Messina) defines an **inter-cloud communication protocol over XMPP** (IETF's Extensible Messaging and Presence Protocol) – peer-to-peer, in-band registration, open source.
- The **CLEVER-WNoDeS collaboration** foresees an **integrated architecture and VM scheduling** through a scalable resource brokering mechanism derived from EMI's WMS.



The WNoDeS *mixed mode*

- *In the real world*, resource providers would like to:
 - Support virtualization so that new use cases can be satisfied; **but also**
 - Run some payloads on physical nodes because virtualization penalties are sometimes not acceptable, or because certain environments are not amenable to be easily virtualized (e.g., GPU's); **but also**
 - Avoid static partitioning of resources.
- This is (to be) addressed by **WNoDeS *mixed mode***:
 - Let jobs run on an hypervisor, and allow also the creation of VM's on the same hypervisor for other jobs (or for cloud services). Here, the “bait” is actually on the hypervisor.
 - This allows resource providers to start introducing new services without the need to statically set resources aside.

```

[root@wn-205-06-26-01-b ~]#
[root@wn-205-06-26-01-b ~]#
[root@wn-205-06-26-01-b ~]# wnodes_manager -s "*"
Bait      : wn-205-06-26-01-b;
Bait status : ['CLOSED_FULL', "Resource ['CPU'] is less than the MIN value", 1329382706.092396, 0, 0, {'MEM': 2381, 'BANDWIDTH': 600, 'STORAGE': 171, 'CPU': 0}]

JobId  JobStatus  JobType  vmID  VM          Owner  vmImage  vmResources          vmStatus  TS              TimeSpentToReachLastStatus
3258959 RUN        BATCH_REAL NoId  wn-205-06-26-01-b aleita NoImg  [cpu:2480 mem:1 disk:30]  16/02-09:51  0(sec)
3259217 RUN        BATCH_REAL NoId  wn-205-06-26-01-b aleita NoImg  [cpu:2410 mem:1 disk:30]  16/02-09:58  0(sec)
3259214 RUN        BATCH_REAL NoId  wn-205-06-26-01-b aleita NoImg  [cpu:2480 mem:1 disk:30]  16/02-09:58  0(sec)
3259212 RUN        BATCH_REAL NoId  wn-205-06-26-01-b aleita NoImg  [cpu:2440 mem:1 disk:30]  16/02-09:58  0(sec)
3258960 RUN        BATCH      2      vwn-02223 aleita vwn_sl5_emi [cpu:2440 mem:1 disk:30]  NEW         16/02-09:59  520(sec)
3259218 RUN        BATCH      3      vwn-02225 aleita vwn_sl5_emi [cpu:2450 mem:1 disk:30]  NEW         16/02-10:03  319(sec)
3259215 RUN        BATCH      4      vwn-02226 aleita vwn_sl5_emi [cpu:2460 mem:1 disk:30]  NEW         16/02-10:03  323(sec)
3259213 RUN        BATCH      5      vwn-02229 aleita vwn_sl5_emi [cpu:2480 mem:1 disk:30]  NEW         16/02-10:04  356(sec)

Bait      : wn-205-06-26-02-b;
Bait status : ['CLOSED_FULL', "Resource ['CPU'] is less than the MIN value", 1329382706.1024261, 0, 0, {'MEM': 2381, 'BANDWIDTH': 600, 'STORAGE': 81, 'CPU': 0}]

JobId  JobStatus  JobType  vmID  VM          Owner  vmImage  vmResources          vmStatus  TS              TimeSpentToReachLastStatus
3258929 RUN        BATCH_REAL NoId  wn-205-06-26-02-b aleita NoImg  [cpu:2450 mem:1 disk:30]  16/02-09:45  0(sec)
3258930 RUN        BATCH      33     vwn-00099 aleita vwn_sl5_emi [cpu:2470 mem:1 disk:30]  NEW         16/02-09:54  517(sec)
3259206 RUN        BATCH_REAL NoId  wn-205-06-26-02-b aleita NoImg  [cpu:2480 mem:1 disk:30]  16/02-09:58  0(sec)
3259210 RUN        BATCH_REAL NoId  wn-205-06-26-02-b aleita NoImg  [cpu:2410 mem:1 disk:30]  16/02-09:58  0(sec)
3259208 RUN        BATCH_REAL NoId  wn-205-06-26-02-b aleita NoImg  [cpu:2440 mem:1 disk:30]  16/02-09:58  0(sec)
3259209 RUN        BATCH      35     vwn-02224 aleita vwn_sl5_emi [cpu:2470 mem:1 disk:30]  REGENERATE  16/02-10:03  312(sec)
3259207 RUN        BATCH      36     vwn-02227 aleita vwn_sl5_emi [cpu:2460 mem:1 disk:30]  NEW         16/02-10:03  332(sec)
3259211 RUN        BATCH      37     vwn-02228 aleita vwn_sl5_emi [cpu:2460 mem:1 disk:30]  NEW         16/02-10:04  350(sec)
  
```

Job packing

- One of the experiments we support (Auger, a 3000 km² cosmic ray observatory located in Argentina) wants a special configuration, where compute nodes need **read-only access to a mysql-based condition database to perform detector simulation from hundreds of compute nodes concurrently.**
- Auger run their jobs in WNoDeS-managed VM's, which include the mysql db.
- However, it is more efficient if the db is installed on the hypervisors, rather than on the VM's. A VM would then access the db on its hypervisor.
- But then **you would like to minimize the number of physical nodes with Auger VM's.**

WNoDeS job packing (work just started) **allows one to pack jobs (or cloud requests) onto a minimal set of physical resources.**

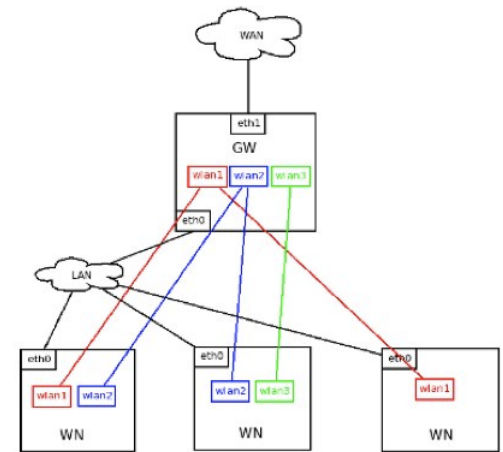
- Can also be used to implement selective power down of idle nodes **to save energy (and costs).**

Dynamic Virtual Networks

- When allocating VM's to users for cloud services, one would normally like to do this so that:
 - There is **traffic isolation** between customers (or groups of them)
 - **The system can scale** to several thousands of customers
- The WNoDeS DVN (Dynamic Virtual Networks) is a R&D project with the goal to define how to **dynamically add, delete and monitor logical networks to VM's** *with minimal or no reconfiguration of the underlying physical network layer.*
 - At CNAF, we have a L2 network topology with about 200 switches, from several vendors. It is totally impractical (and dangerous) to reconfigure the network by hand every time we need to add a Cloud customer.
 - We also do not like having automated reconfiguration procedures (even if it were theoretically possible) of the L2 network.

Setting up DVNs

- A couple of theoretical possibilities: IEEE 802.1ad (*802.1 QinQ*) and RFC 3069 (*private VLANs*)
 - Adoption / constraints in the real world?
- We are currently testing a **hub-and-spoke overlay topology for DVNs based on the GRE protocol**
- DVNs are defined through a **Policy Enforcement Service**, used to collect/distribute the traffic policies
 - A meta-language that eventually translates into e.g. iptables commands or router ACLs
- **The first tests**, done with a simple, single Linux-based central GW show good scaling properties for what regards CPU usage, network throughput
 - Consolidated results will be shown at CHEP (NY, May 2012)



WNoDeS Status

- WNoDeS is **licensed** under the European Union Public License (EURL), the first EU Free/Open Source license.
- WNoDeS 1 is **in production** at several Italian sites. See <http://web.infn.it/wnodes> or send email to wnodes@lists.infn.it for details.
- **WNoDeS 2** (introducing some of the features described here and support for PBS/Torque) will be **released as part of the European Middleware Initiative (EMI) EMI-2 release** at the end of April 2012.

Future work

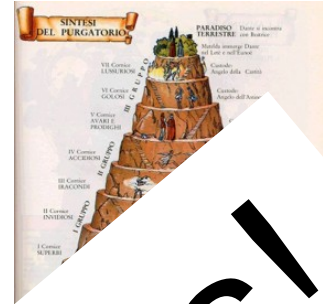
- Many thinks related to WNoDeS are still in the works. These include:
 - **Service provisioning** (rather than simple IaaS). At the INFN Tier-1, we have recently been working on cooperation between research and industry for the provisioning of dynamic compute services (with us as resource/technology providers, a relatively new path for INFN)
 - **Long-term data access/preservation**
 - **Certified, site-independent VM images**
 - **Support of other LRMS or virtualization technologies**
 - **Cloud storage**

Agenda

- Introduction:
The Context

- The INFN Tier-1:
The Status

- (More) Cloud. The horizon:
The Challenges



Thanks!

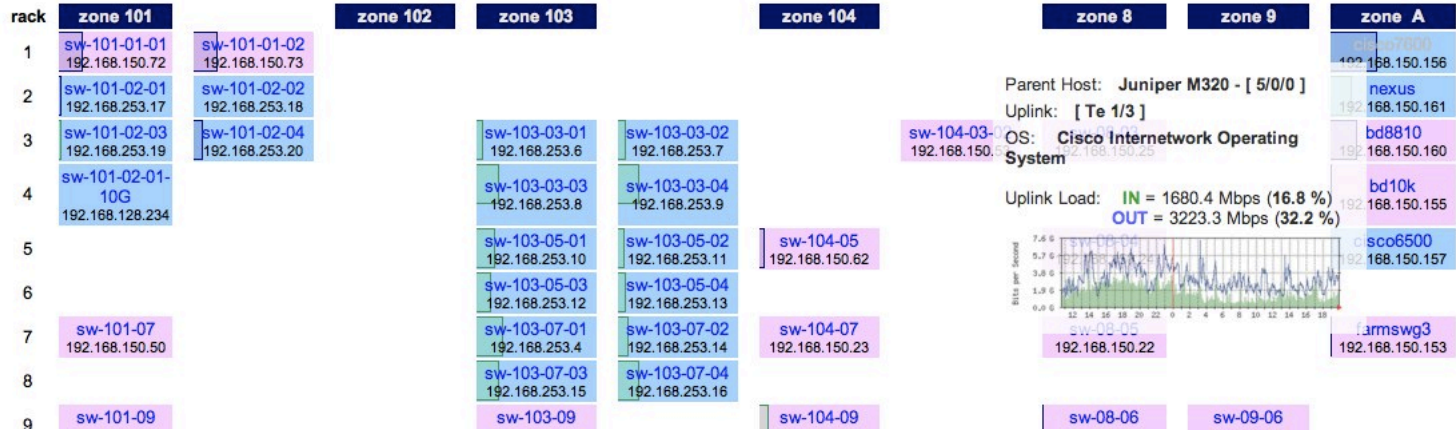


Back-up Slides

Network Monitoring

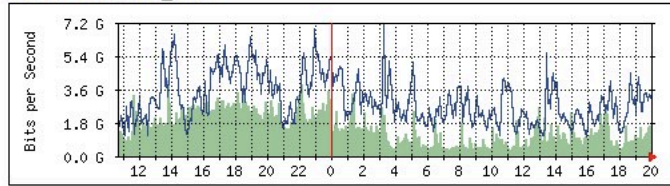
backbone | ROOM 1 | ROOM 2 | 10G connections

ROOM 2

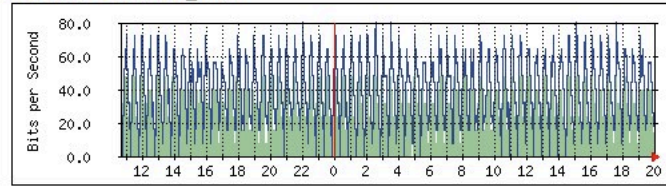


NEXUS 7018

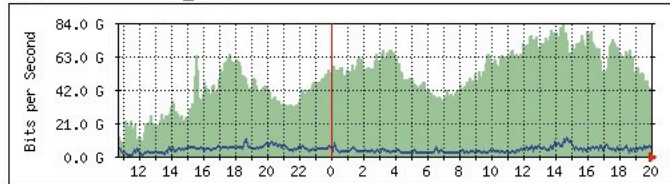
Vlan1 -- NEXUS_7018



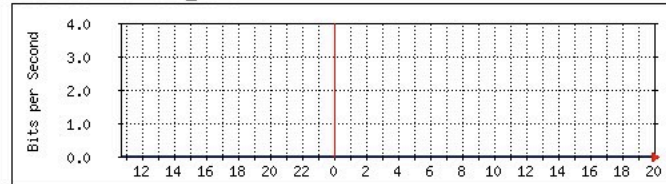
Vlan111 -- NEXUS_7018



Vlan128 -- NEXUS_7018



Vlan136 -- NEXUS_7018



Vlan140 -- NEXUS_7018

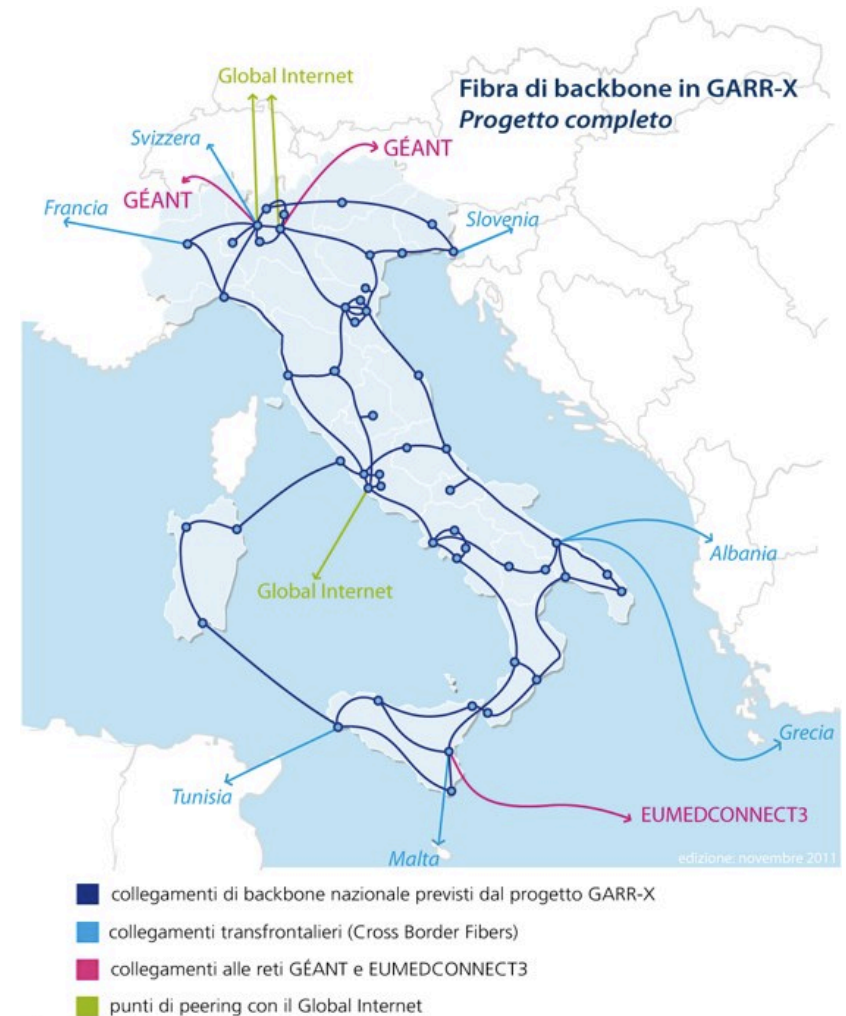


Vlan144 -- NEXUS_7018



The GARR Network

- GARR-X, the new DWDM-based network fully dedicated to Italian Universities and Research Institutions
 - Entirely managed by the GARR Consortium
 - Backbone being activated in 2012

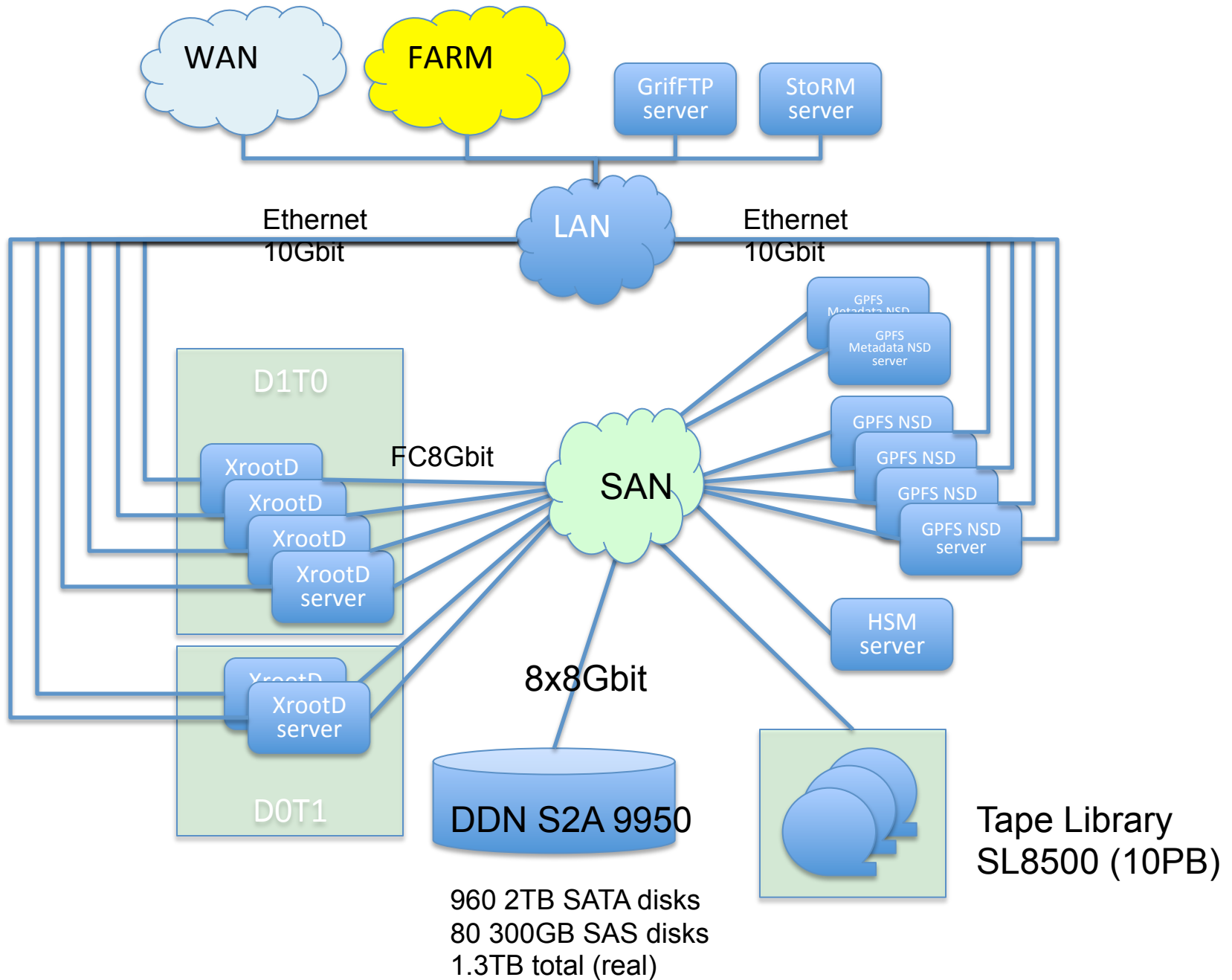


Some details about storage at CNAF

- Mostly using DDN S2A9950 and EMC CX-390, CX-480 (approaching end of life), with Fujitsu equipment to be delivered (2011 tender)
- SATA disks, connected to servers via FC
- GPFS metadata disks on separate SAS disks
- Tape library: Oracle/Sun StorageTek SL8500 with 20 x T10KB drives (1TB tapes), 10 x T10KC to be delivered (5TB tapes)
 - Replace current 1TB with 5TB tapes → from 10 to 50 PB
- For VO's requesting xrootd, n x xrootd servers (e.g. 4 for ALICE) connected at 10 Gbit/s accessing the GPFS file systems.

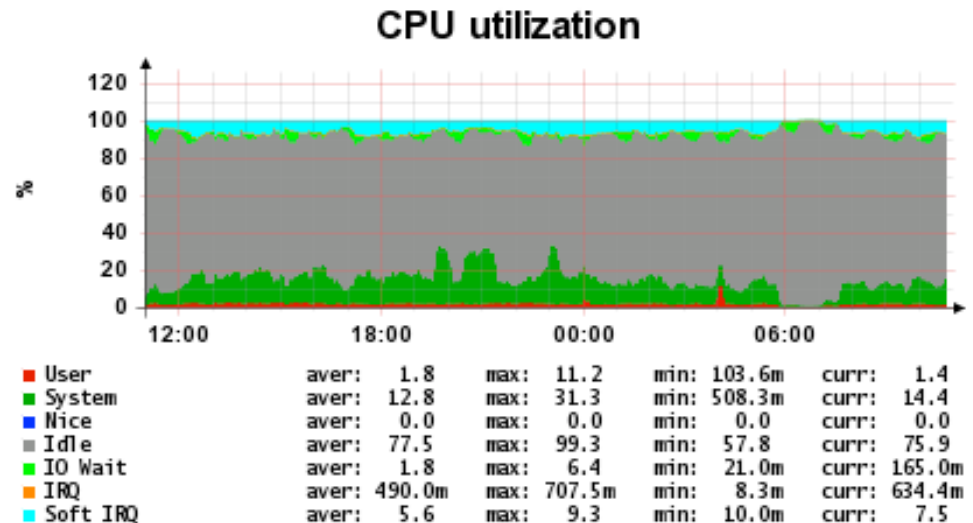
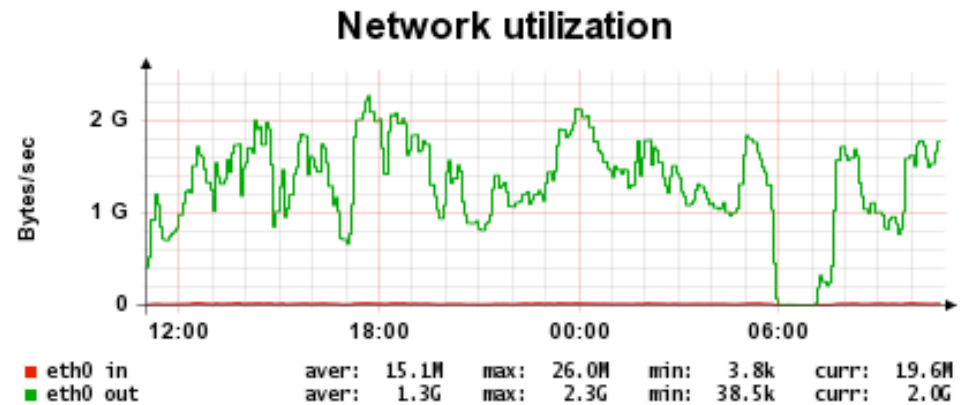
Alice cluster

- 4 XrootD servers for D1T0, 2 XrootD servers for D0T1
 - 8 core 2.2GHZ
 - 10Gbit ethernet
 - 2x8Gbit FC
 - 24GB RAM
 - All connected to the same (shared) file system (GPFS)
- 4 NSD servers (same as above)
 - Two of them to be converted to XrootD servers
- Storage
 - DDN S2A 9950,
 - 1.3TB net space
 - Two GPFS filesystems
 - 960TB disk-only (D1T0)
 - 385TB cache for tape (D0T1)
- Tape
 - Custom plug-in to interface XrootD with GEMSS (CNAF's MSS)
 (modified method XrdxFtsOfsFile::open in XrootD library)

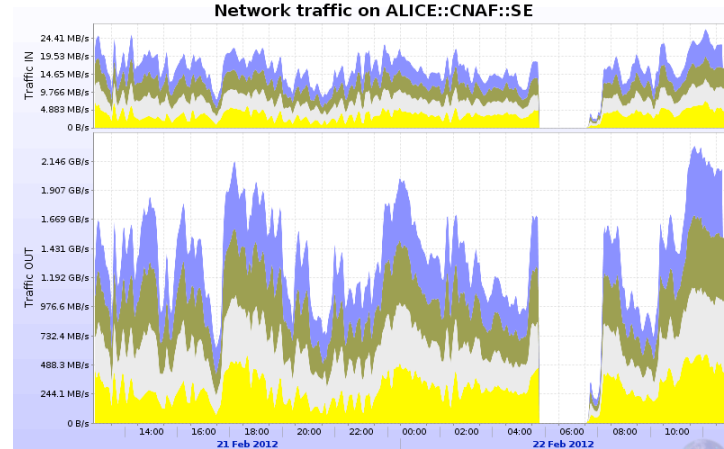


Performance and Some observations

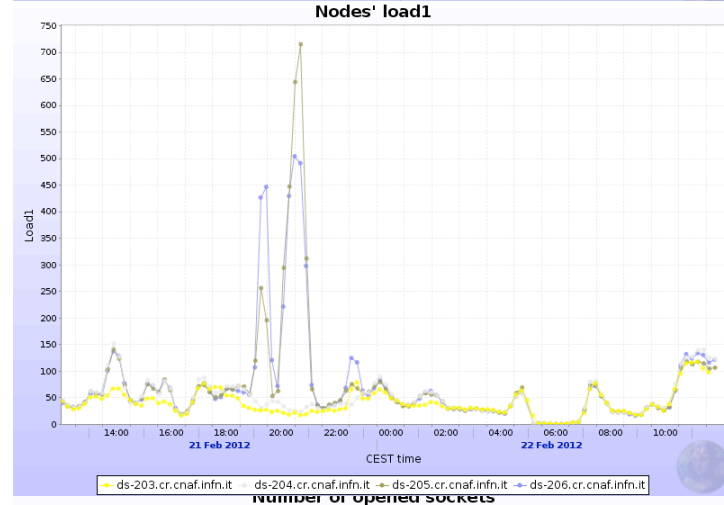
- With 4 XrootD servers we are limited by CPU power (or OS limitations ???)
- Huge number of open files/sockets (2-3K)
- Small blocks I/O while file system's BS=1MB
 - Overhead in network (no saturation on 10Gbit while with the same servers we are easily saturating 10Gbit on GPFS NSD)



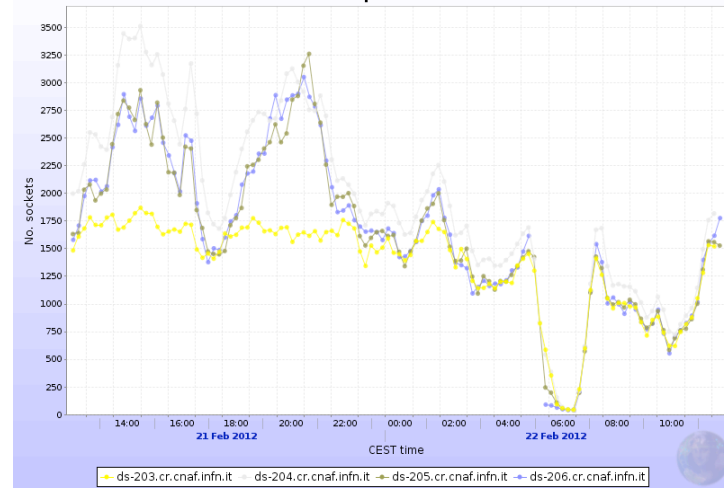
From ALICE monitoring:
Network traffic



CPU load



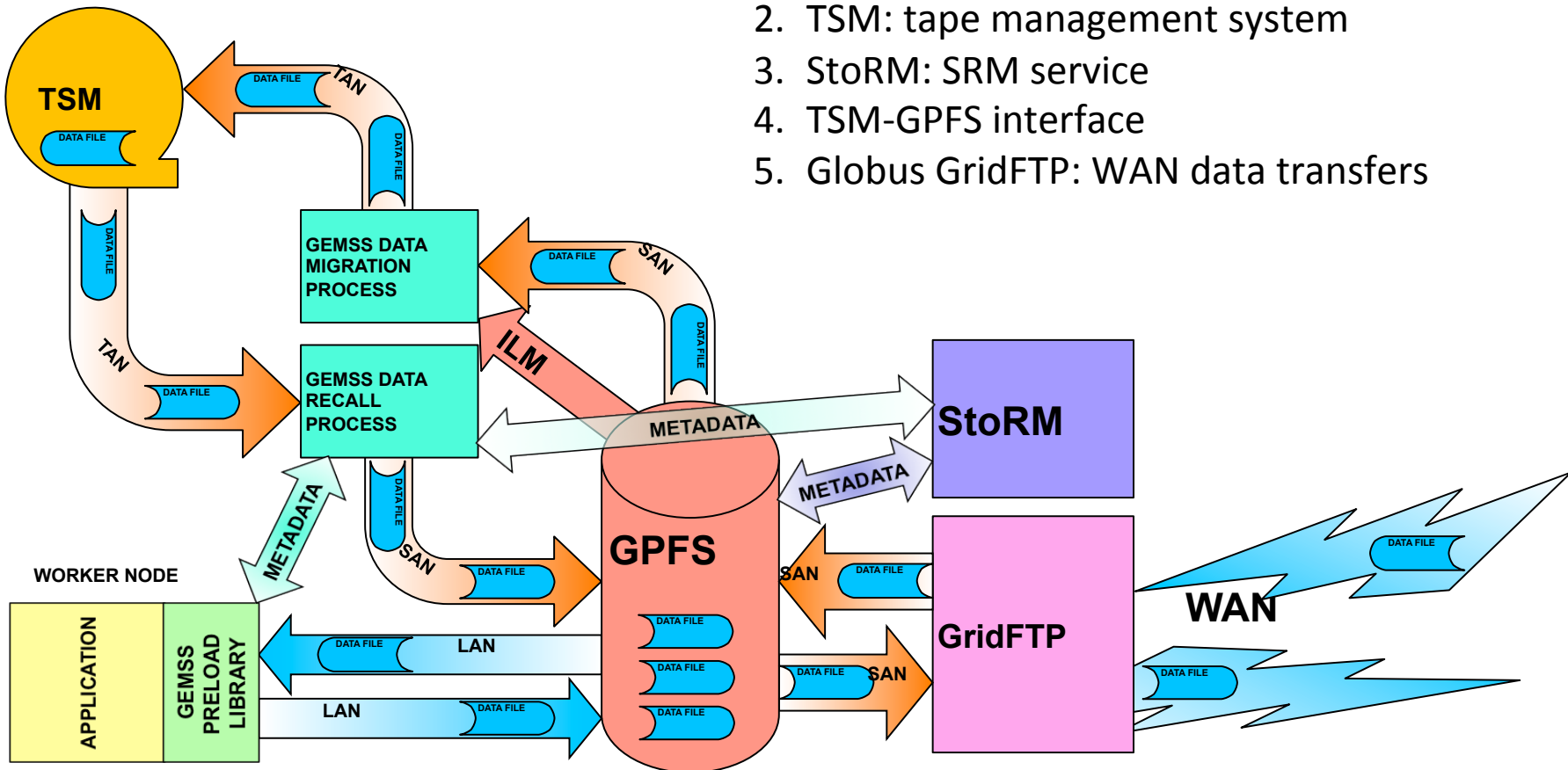
Open sockets



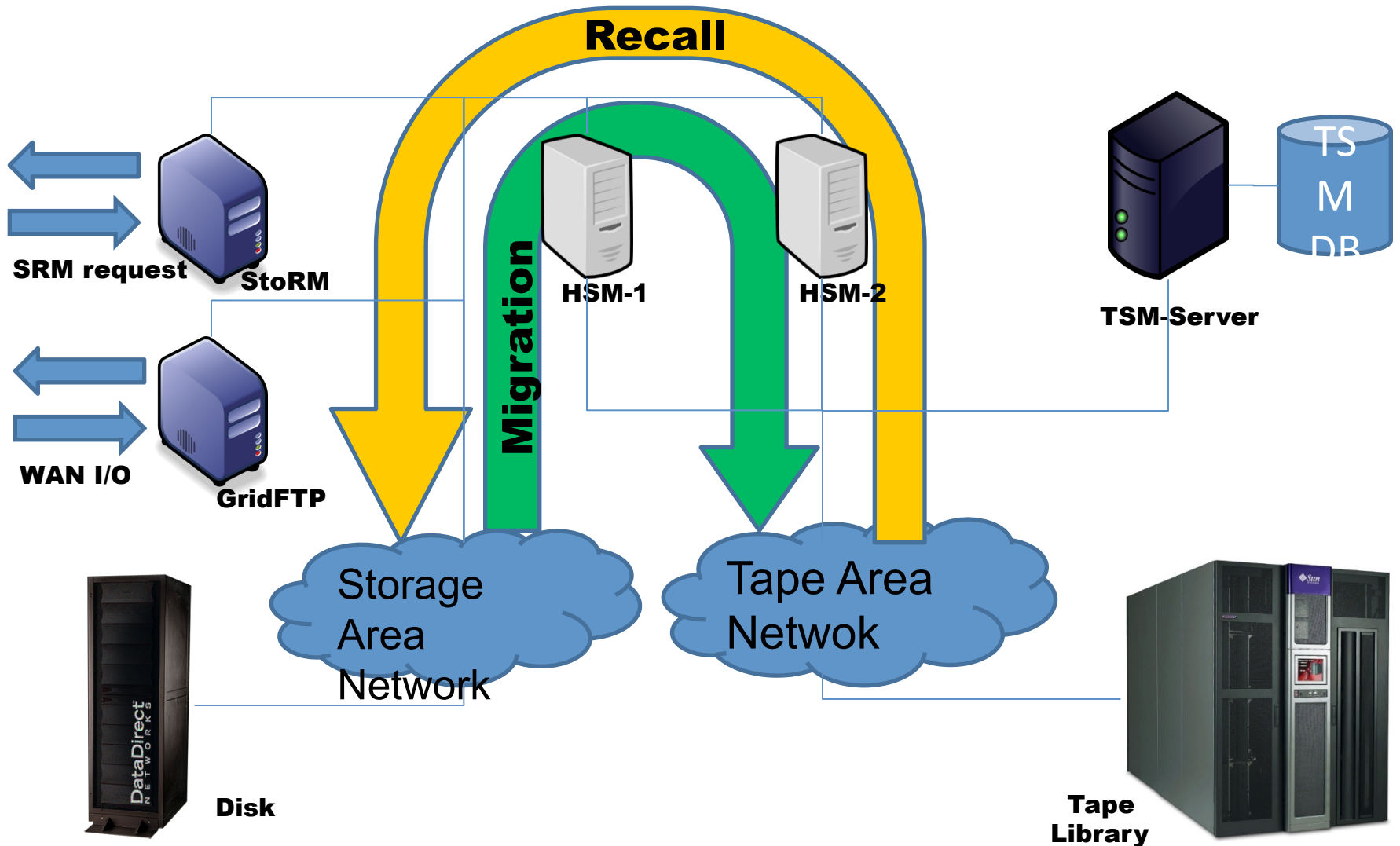
Building blocks of GEMSS system

Disk-centric system with five building blocks

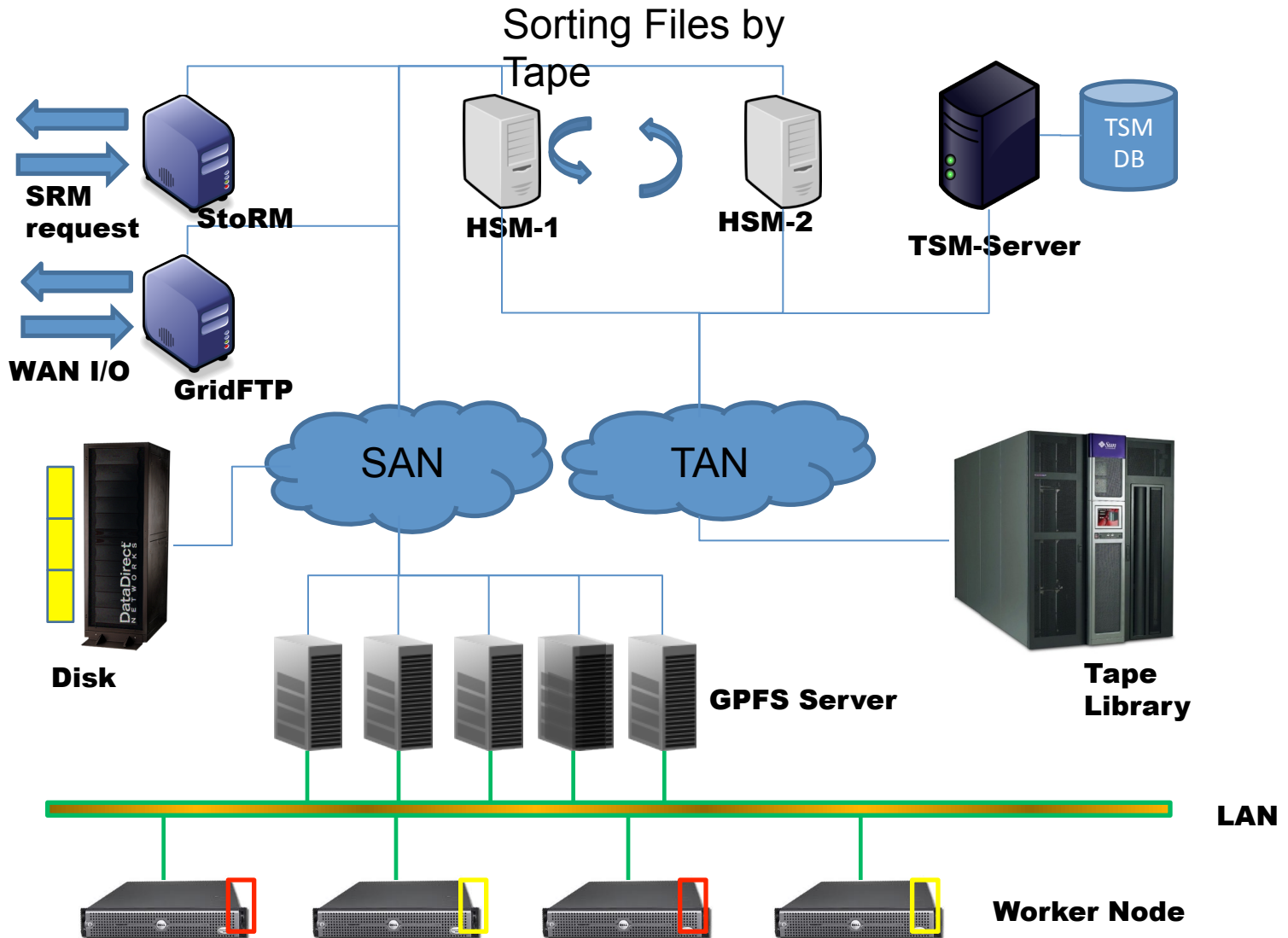
1. GPFS: disk-storage software infrastructure
2. TSM: tape management system
3. StoRM: SRM service
4. TSM-GPFS interface
5. Globus GridFTP: WAN data transfers



GEMSS layout @INFN-CNAF 1/2



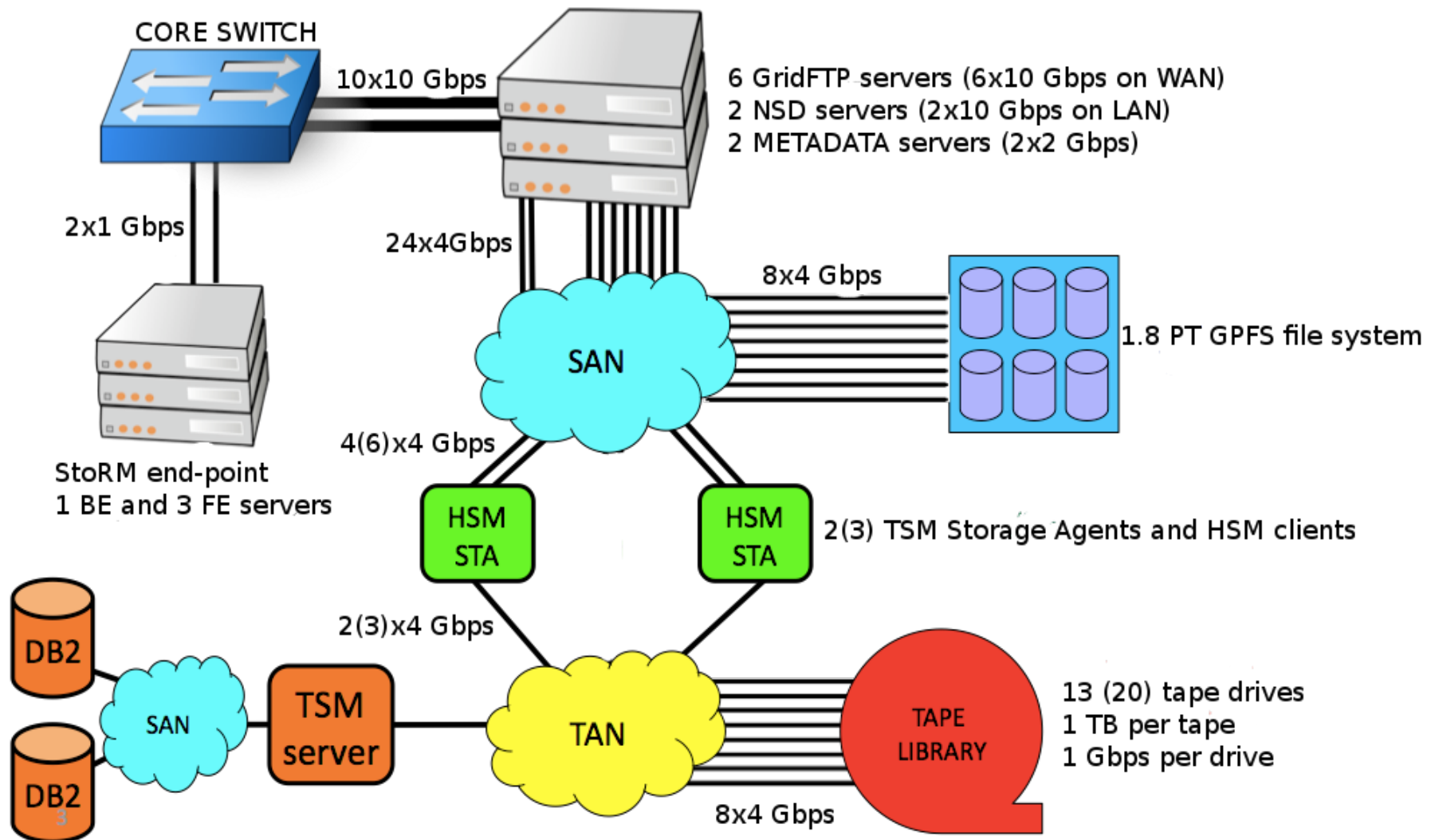
GEMSS layout @INFN-CNAF 2/2



GEMSS HA

- TSM DB is stored in a CX on the SAN
- TSM DB is backed up every 2 hours on a different CX disk and every 12 hours on tape with a persistency of 6 days
- TSM-SERVER have a secondary server in stand-by
 - It's possible to move the DB on the CX directly to the secondary server
 - With a floating IP all client are redirect to the new server
- We have 2 or 3 TSM-HSM clients for VO for failover
- GPFS servers, StoRM FEs and GridFTP servers are in cluster
 - StoRM BE is, by design, a single point of failure (cold spare ready)

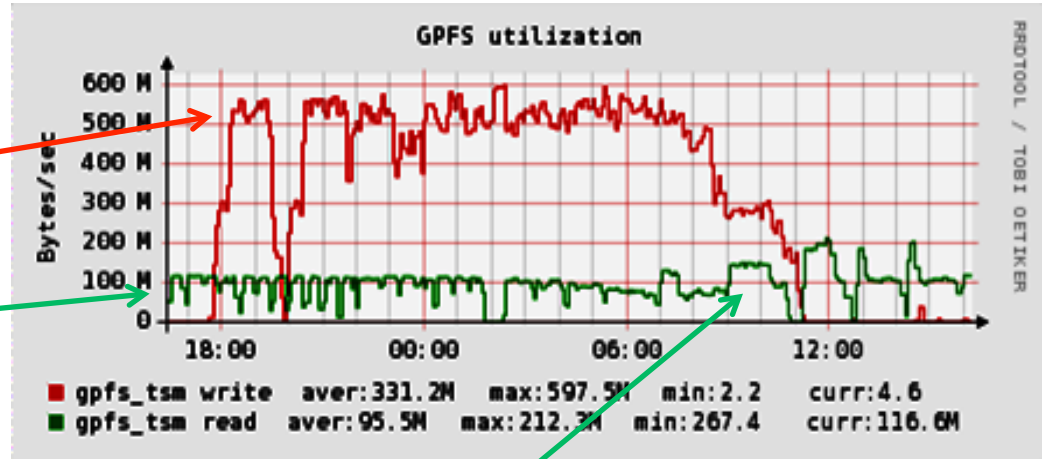
GEMSS layout for a typical LHC Experiments at INFN Tier-1



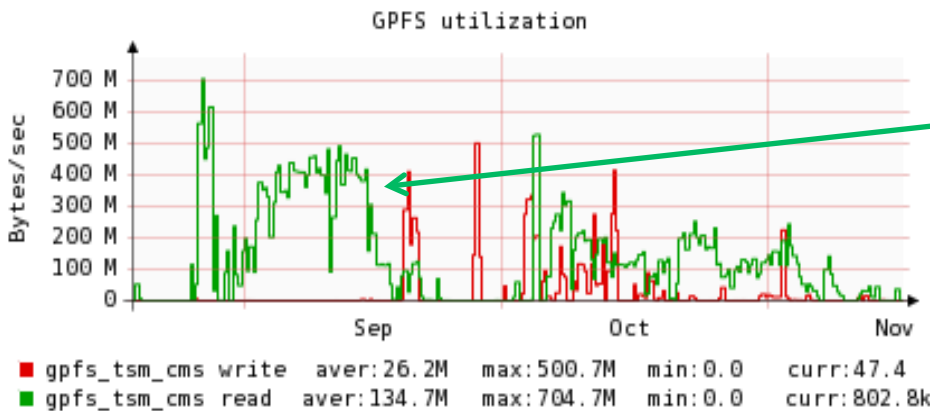
CMS tests for local access to TSM

Summer 2009 tests

- Manual recall from tape
 - 550 MB/s
- Migration to tape
 - 100 MB/s
- Local access to data on TSM
- from the batch farm nodes



higher throughput due to jobs output



Migration of data from Castor to TSM

- ◆ ~ 1 PB
- ◆ In parallel to production activities

— From tape — To tape

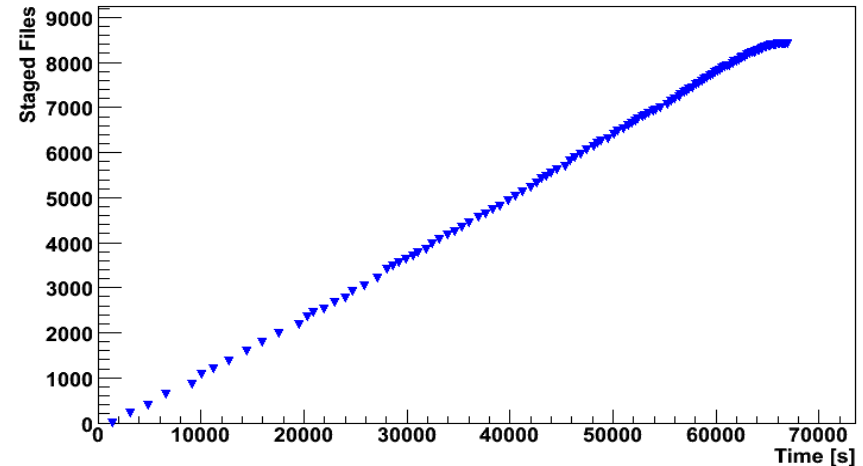
CMS tests of manual recall from tape

- 24 TB (8000 files) randomly spread over 100 tapes recalled in 19 hours
 - Peak measurements done with no overlap with other recalls
 - Quite some other activities running at the same time though (see plot below)

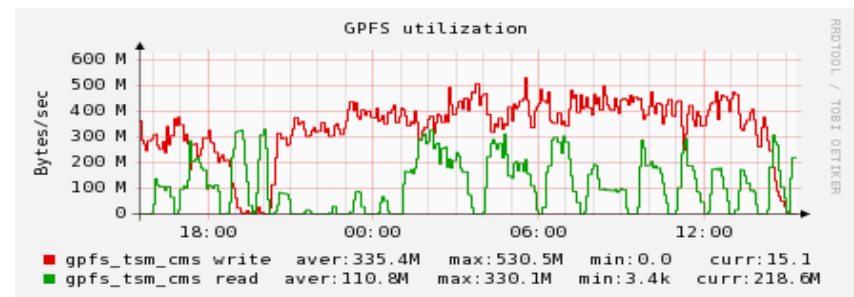
- 400 MB/s average throughput

- Peak at 530 MB/s
 - (using resources as from previous slides)
 - 85% of nominal tape drive throughput

Number of staged files as function of time



Net GPFS disk throughput on the GEMSS data movers

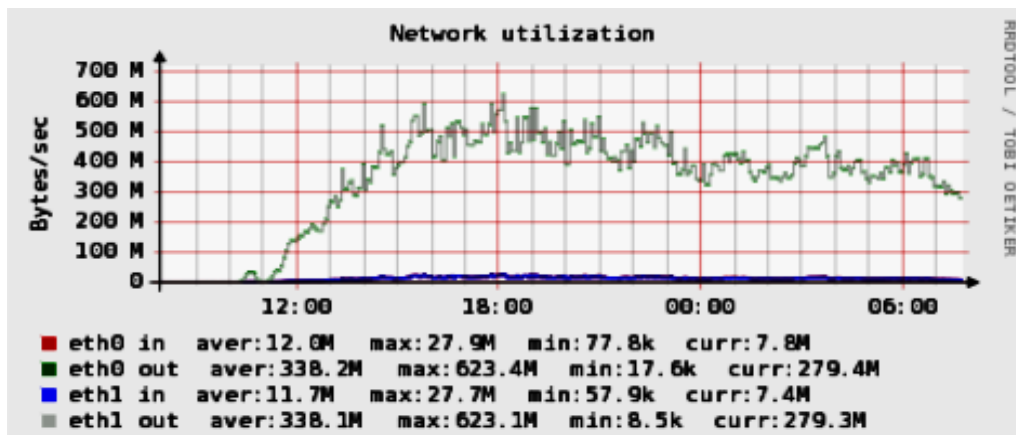
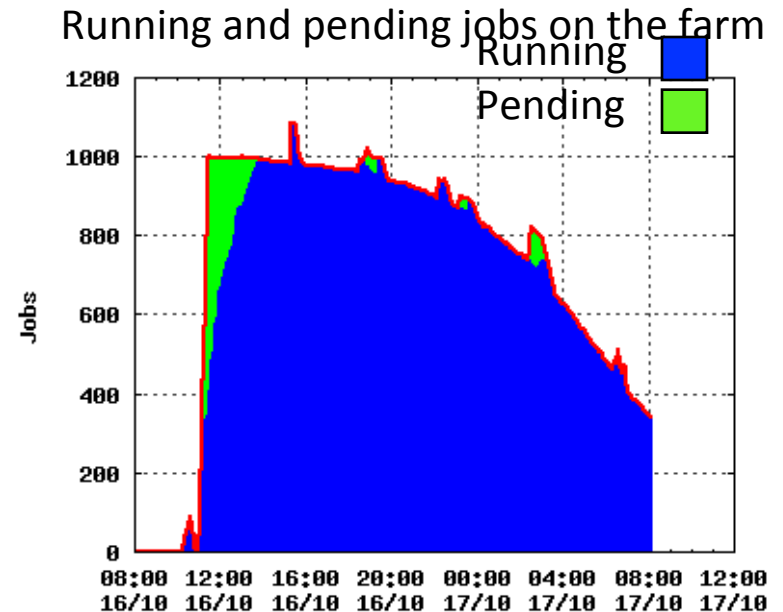


CMS tests on processing from the farm nodes

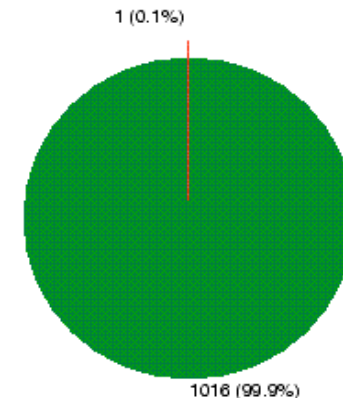
Using the file protocol

Up to 1000 concurrent jobs
 recalling from tape 1930 files

- 100% job success rate
- Up to 1.2 GB/s from the disk pools to the farm nodes



Job success rate on the farm

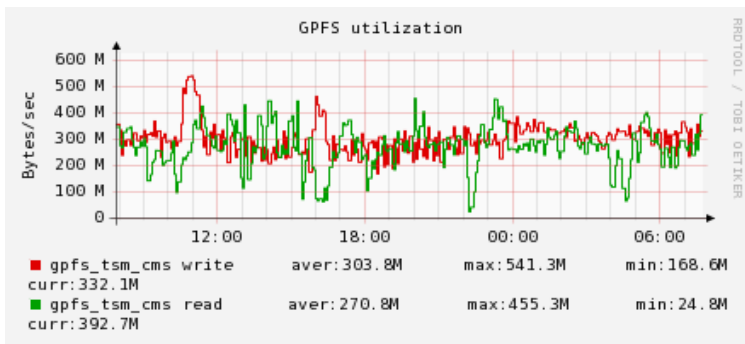
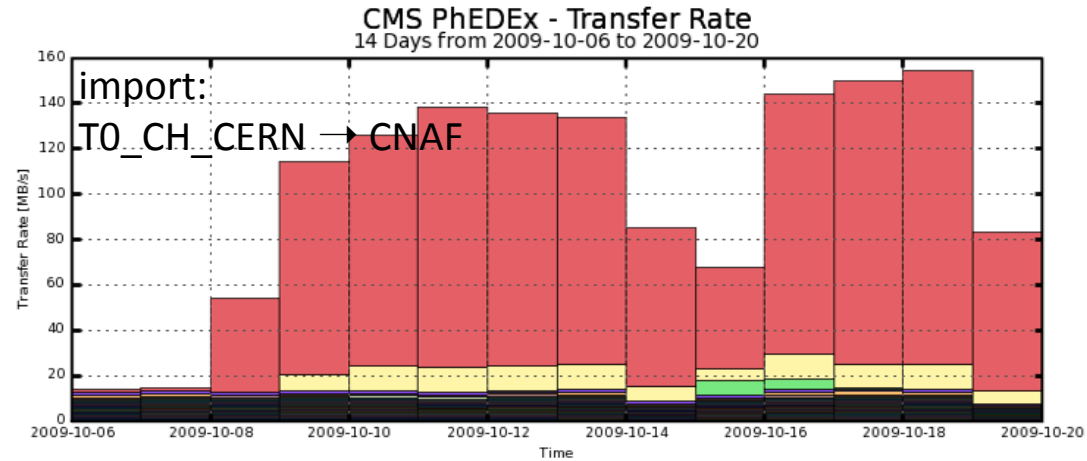


Traffic on one of the two network cards of the GPFS server

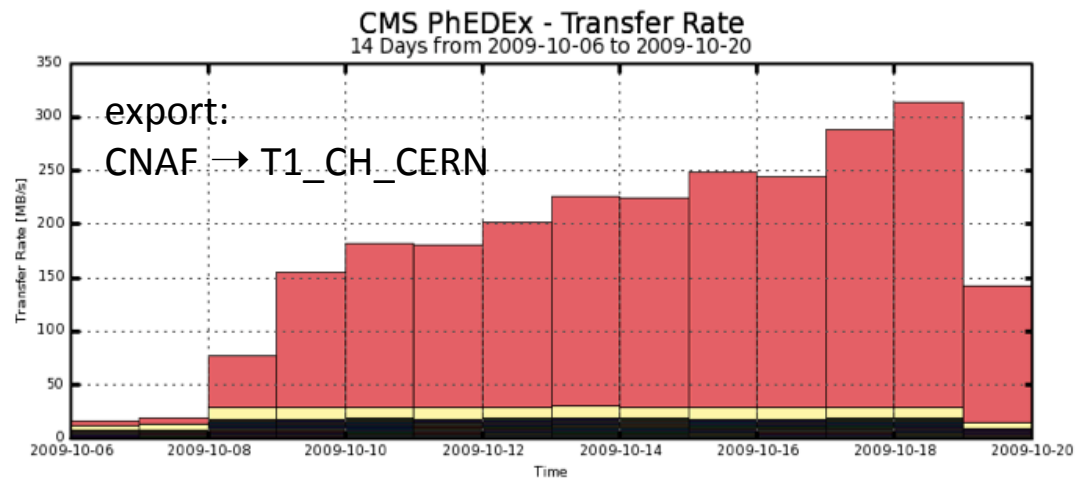
WAN transfer tests

Using the PhEDEx
 'LoadTest' infrastructure

- Up to ~160 MB/s import
- Up to ~300 MB/s export
- 80 MB/s background
- during other tests



Traffic on the gridFTP servers



GEMSS in production for CMS

GEMSS went in production for CMS in October 2009

◆ w/o major changes to the layout

- only StoRM upgrade, with checksum and authz support being deployed soon

Good-performance achieved in transfer throughput

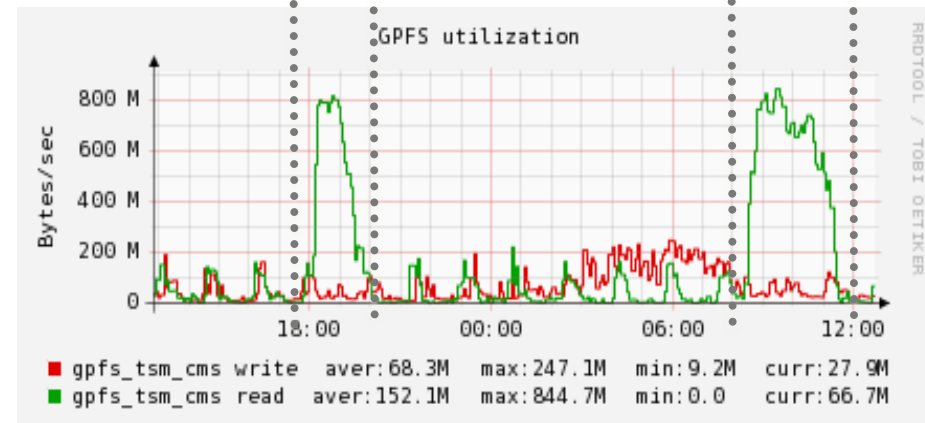
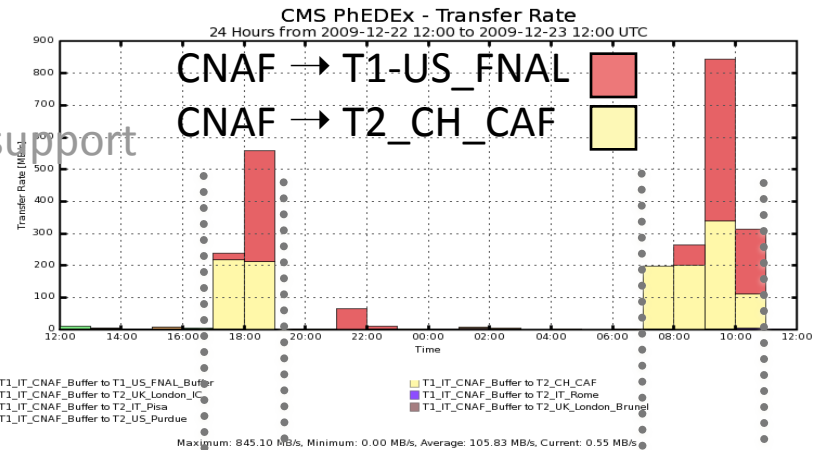
- High use of the available bandwidth
- (up to 8 Gbps)

Verification with Job Robot jobs in different periods shows that CMS workflows efficiency was not impacted by the change of storage system

- “Castor + SL4” vs “TSM + SL4” vs “TSM + SL5”

As from the current experience, CMS gives a very positive feedback on the new system

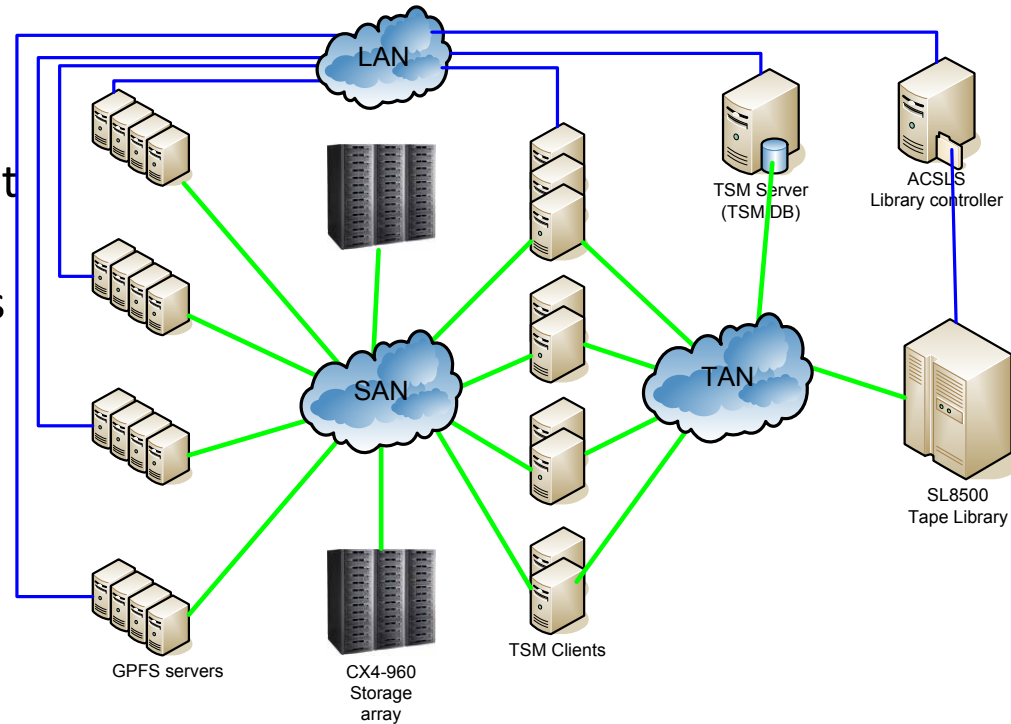
- Very good stability observed so far



TSM

TSM building blocks:

- The Server - provides backup, archive, and space management services to the clients. It uses a database to track information about server storage, clients, client data, policy, schedules
- The Client Storage Agent - enables LAN-free FC data movement for client operations
- Hierarchical Storage Management (HSM) - provides space mgt services for workstations.
- TSM for Space Management automatically migrates files that are less frequently used to server storage, freeing space on disk.



In production at CNAF since
 CCRC'08 for LHCb (D1T1)



GPFS + TSM

GPFS performs file system metadata scans according to ILM policies specified by the administrators

- The metadata scan is very fast (it is not a find) and is used by GPFS to identify the files which need to be migrated to tape
 - Possible to use Extended Attributes

The list of files obtained is passed to an external process which is run on the HSM nodes and it actually performs the migration to TSM

- This part in particular is the one implemented at CNAF

Recalls can be done passing a list of files to TSM

- This list will be tape-ordered by TSM itself

GPFS and the HSM nodes completely decoupled

- possible to shutdown the HSM nodes without interrupting the file system availability

All components of the system have intrinsic redundancy (GPFS failover mechanisms)

- No need to put in place any kind of HA features apart from the unique TSM server with the internal db
 - Backup and failover of TSM db tested

Recent press release with Intel

CASE STUDY

Intel® Xeon® processor 5600 series
 Enterprise Server
 Virtualization



Putting Italy on the cutting edge of scientific computing

Virtualization-based solution developed by INFN-CNAF brings the Grid and Cloud models closer

The Italian National Institute for Nuclear Physics (INFN) operates an organization in Bologna known as CNAF – the National Center for Research and Development in Information and Data-Transmission Technologies. CNAF is responsible for the management and development of the most important information and data transmission services to support INFN's high-energy physics research at a national level. Its research activities are divided into five scientific categories: accelerator physics, astroparticle physics, nuclear physics, theoretical physics and technological development.



CHALLENGES

- **Enhance infrastructure:** Provide the INFN community with a scalable and flexible solution for high-performance scientific computing
- **Guarantee continuity:** Deliver operating system support and scientific data availability for long-term data access at sustainable total cost of ownership (TCO)
- **Expand the customer base:** Offer new and enhanced services

SOLUTIONS

- **Integrated framework:** Implemented on-demand grid/cloud framework for scientific computing, based on open-standard technologies
- **Performance penalties minimized:** Physical and virtual environments have fine-tuned hardware and software solutions and efficient access to large-scale storage systems

IMPACT

- **World-first:** One of the first proven, OS-based implementations to achieve excellent scalability and flexibility in providing shared access to resources and integration between Grids and Clouds – without the need to partition resource pools
- **National use:** The INFN Worker Nodes on Demands Service* (wNoDeS*) framework is the production solution being offered for Grid and Cloud integration by the Italian Grid Initiative (IGI)



<http://www.intel.com/content/dam/doc/case-study/virtualization-xeon-5600-infncnaf-study.pdf>