ALICE Tier1 – Tier2 Workshop

ITALY STATUS AND PLANS

INFN

GENERAL OBSERVATIONS

- Domenico Elia replaced Massimo Masera as Italian Computing Coordinator
  - Welcome Domenico!
  - Thanks Massimo!
- Italian Tier-2 coordination not as brilliant as it used to be
  - My fault, mostly
  - Overhauling foreseen also because of upcoming activities (see last slides)
- For at least next two years, funding will be more "creative"
  - Constraints in resources distribution
  - Manpower issue
- Our referees keep pointing out that old data should be deleted
  - Nothing new here...

RESOURCES AVAILABLE FOR ALICE

- Tier-1 at CNAF, Bologna
  - Shared with other LHC experiments and a large number of others
- 4 “official” Tier-2 centres
  - “Official” means directly funded by INFN according to plans and official pledges
  - Torino, Catania, Bari and Padova/LNL
  - The last two shared with CMS
- Cagliari, Bologna, Trieste
  - Local resources, different creative funding
- CyberSar (CA) and TriGrid (CT)
  - Both projects ended, resources becoming obsolescent

THE TIER-1

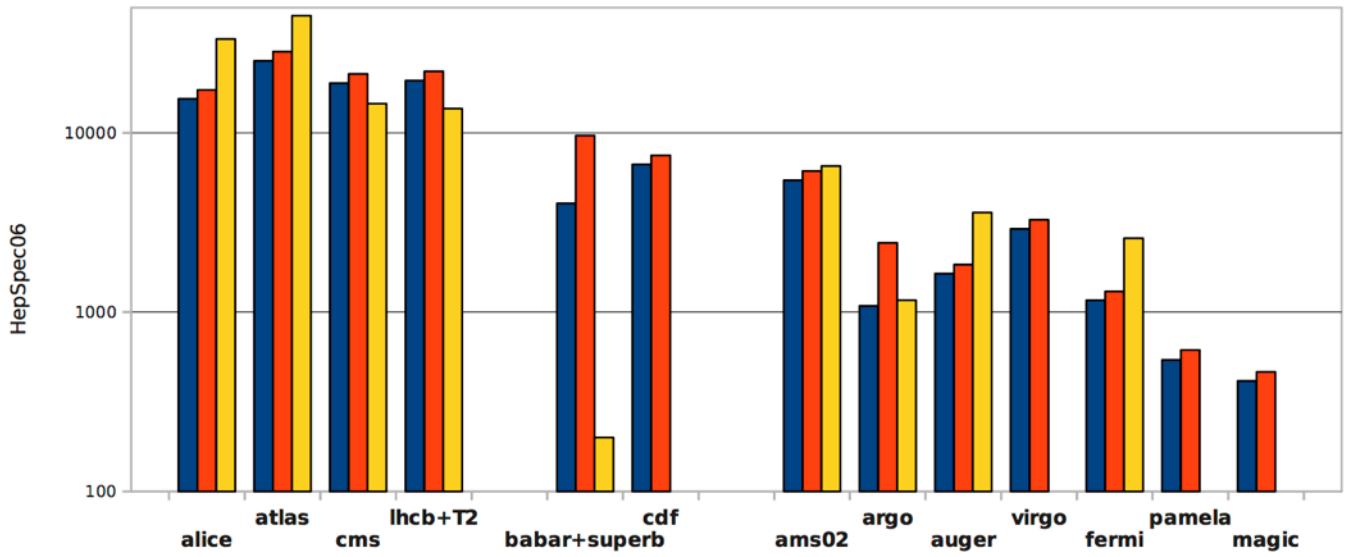
- 13k cores (130 kHEPSpec06), ALICE share is 18620 HS06 (about 1600 job slots)
  - LSF for management
  - Most WNs virtualized in a cloud-like architecture ("Worker Nodes On Demand", WNODes)
- 11PBn of disk, ALICE share is 1.7 PB (T1D0+T0D1) plus 3.7 PB of tapes
  - GPFS + TSM for management
  - Xrootd as a front-end protocol
- Staff of 23 (22 FTE) to babysit the whole centre
  - Plus 1 person dedicated to ALICE-specific operations (F. Noferini), not full time

- Storage 2012 was delayed by a tender problem
  - The winner delivered but the first shipped machines did not pass the preliminary tests (did not match tender requirements)
  - Could not fix the issue, order had to be cancelled and reassigned to runner-up
  - Lost 6 months (and some money)…
- Procurement undergoing to fulfill 2013 CPU pledges
  - CPU tenders for “blanket orders” nearly ready
  - Disk 2013 blanket order ready for 3.8PB overall
  - Tape tenders to be defined depending on technology (5TB or 7-9TB tapes + drives?)
  - Uplink upgrade to 40Gb/s

Farming

INFN CNAF

HEP-SPEC pledge HEP-SPEC attuali HEP-SPEC medi utilizzati
[dal 22/04/2013 al 21/05/2013]



2

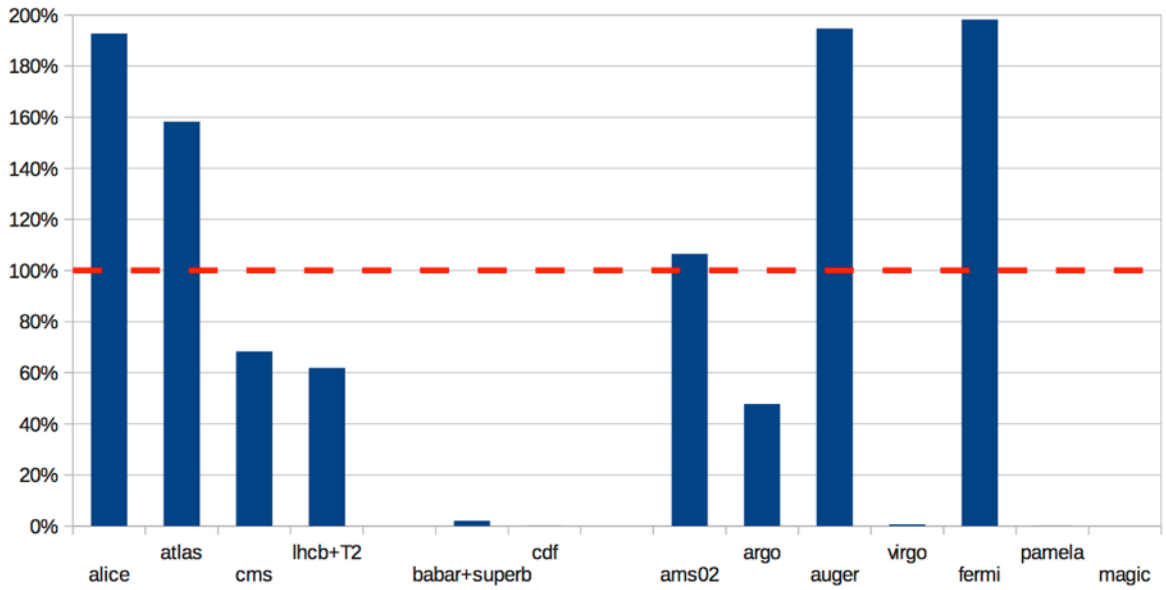
Stefano.Perazzini@cnaf.infn.it

Farming

INFN CNAF

wct_hep_day/assigned

[dal 22/04/2013 al 21/05/2013]



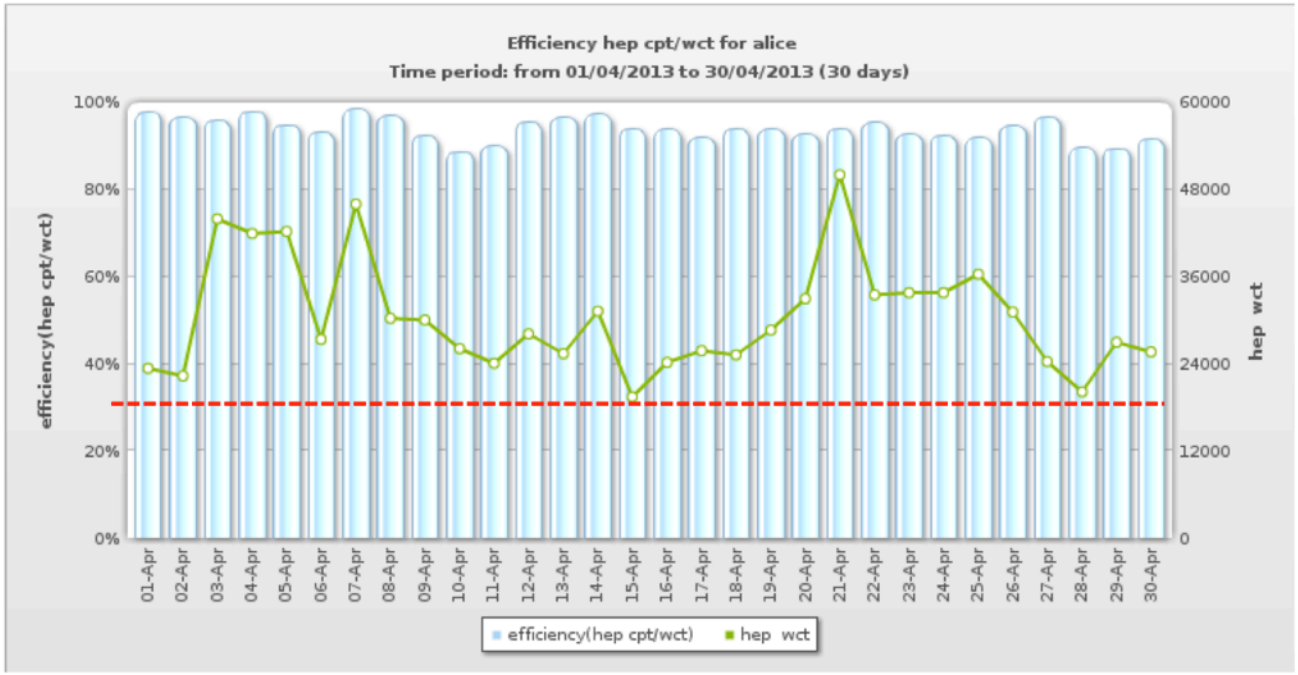
Stefano.Perazzini@cnaf.infn.it

Job Efficiency (HepSpec CPT/WCT)

INFN CNAF

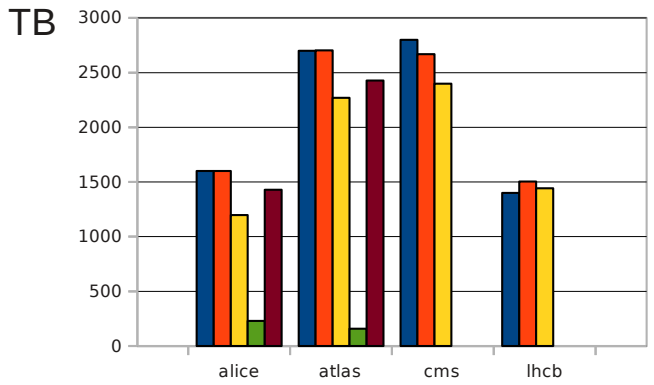
ALICE

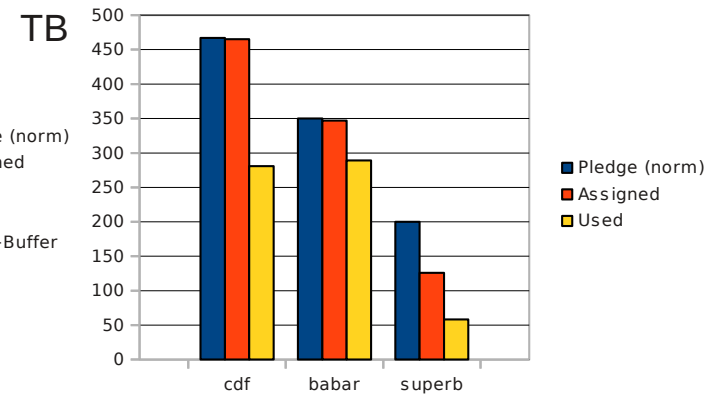
Pledge: 18620 HepSpec



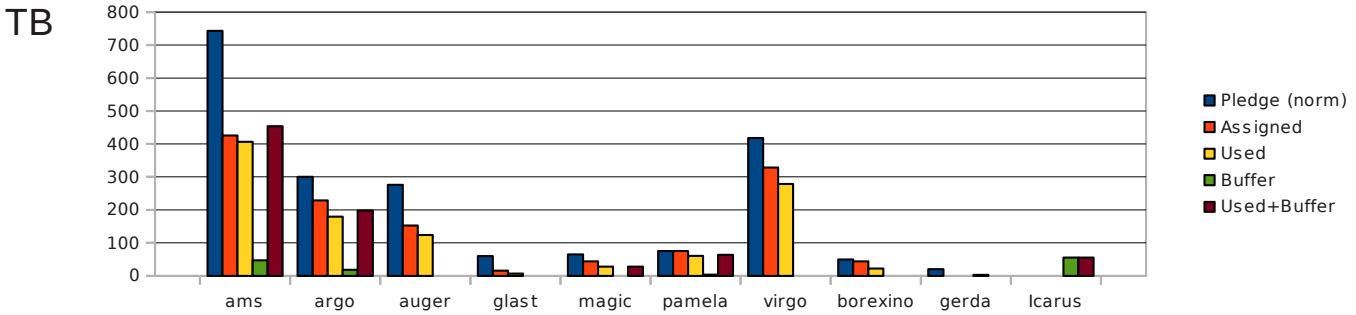
Stefano.Perazzini@cnaf.infn.it

Storage: DISCO (al 21/05/13)



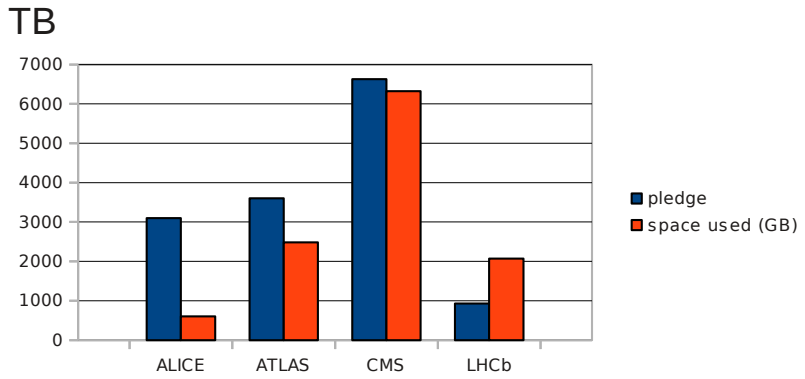


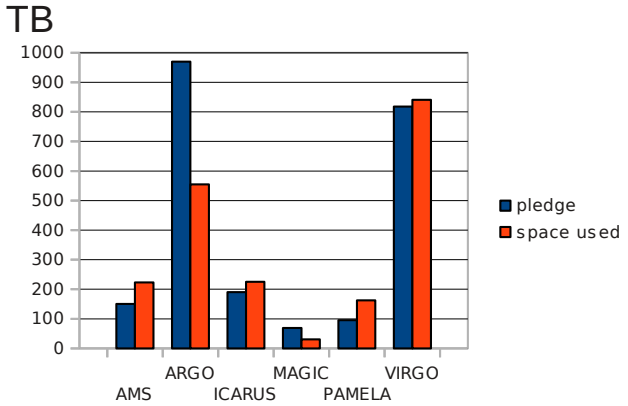
Nell'usato di AMS è considerato lo spazio prestato da SuperB in gpfs_superb/AMS 54 TB

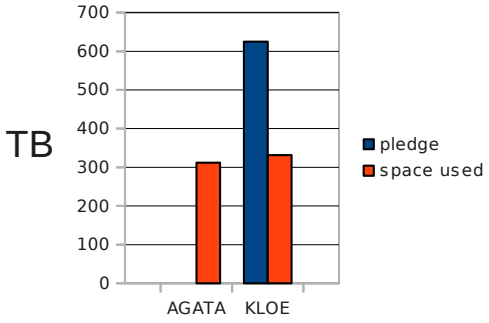


Stefano.Perazzini@cnaf.infn.it

Storage: TAPE (al 21/05/13)







Stefano.Perazzini@cnaf.infn.it

TIER 2 RESOURCES 2013

- As of April 2013
  - Includes “obsolescent” resources
  - Detailed info available upon request…
  - Procurement 2013: 202TB (Torino) + 648TB
  - No new CPU in 2013

	Bari	Catania	LNL-Padova	Torino	Cagliari	Totale
HS06	8984	3110	8264	7805	1960	30123
TB	450	258	357	634	70	1769
Full	95.4%	82.9%	96.0%	87.3%	98.5%	

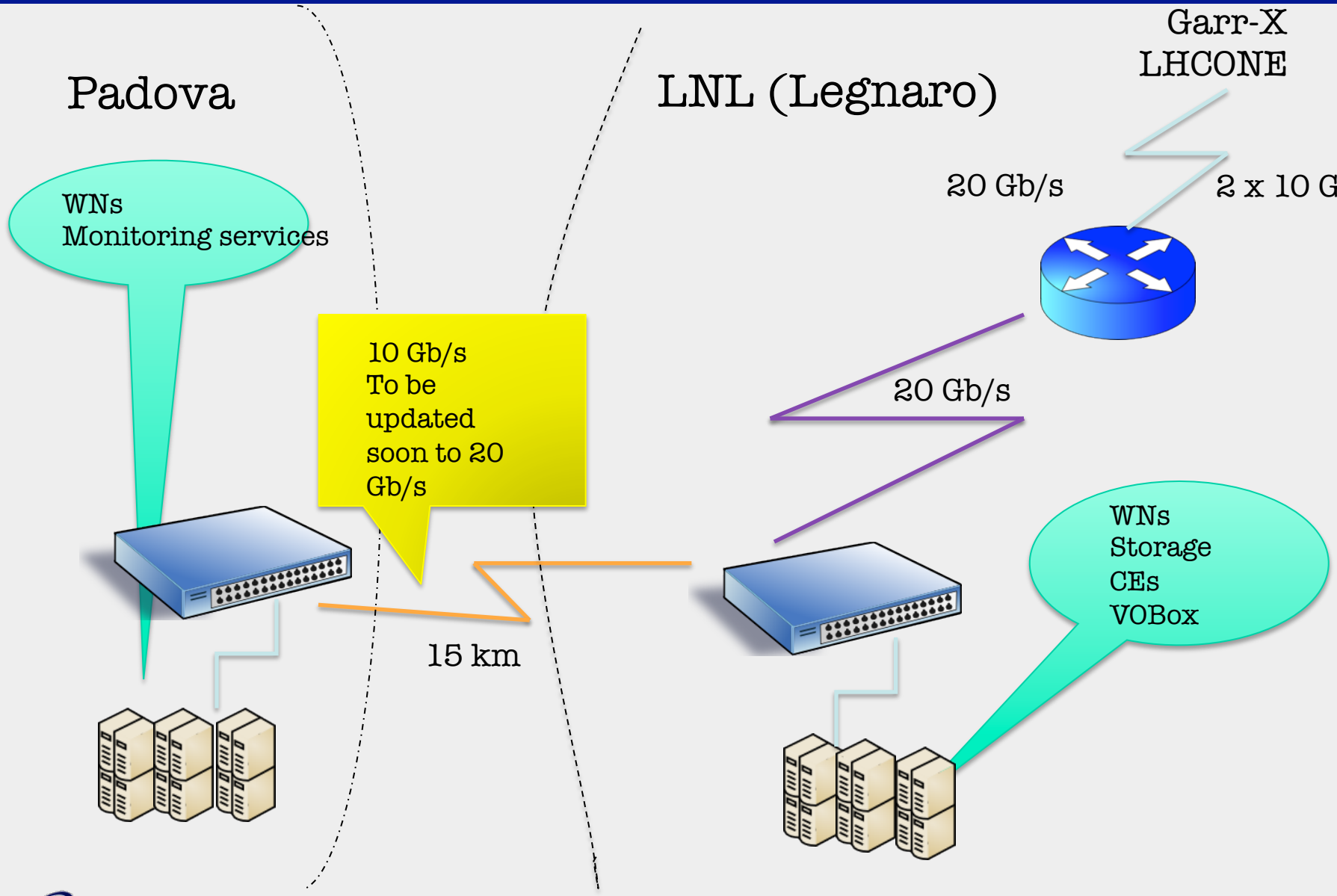
- Project to provide all Italian Tier-2s (and Tier-1) with 10Gb/s connections to LHCONE
  - Work done by GARR (The Italian NREN)
  - Dark fibers leased to GARR by providers, GARR owns both routing and transmissive devices
  - Most T2s easily upgradable to 20-40 Gb/s (should the need arise...)
- Deployment during 2012-2013
  - Bari, Legnaro (20Gb/s), Catania and CNAF (20Gb/s for LHCONE +LHCOPN out of 30Gb/s total) already done
  - Torino nearly there (we had to upgrade the core switch, router, optic fiber link,...)

CATANIA UPGRADE AND TUNING

- Storage upgrade
  - added 110TBn (130TB RAW, 15% loss) last Dec 2012
  - **67TB** available of **358TB** total
- GPFS migrated from 3.2.0.21 to 3.4.0.15
  - enabled fast-enhanced-attributes
- VOBox
  - AliEn v2.19-197
- Installed and configured 4 xrootd servers
  - xrootd v3.2.4
  - network bandwidth 8 Gbps (4 x 2 Gbps)

Roberto.Barbera@ct.infn.it

THE DISTRIBUTED LNL-PD TIER2



Massimo.Sgaravatto@pd.infn.it

LNL-PD RESOURCES

- Storage
  - 386TB, basically full
    - 300 TB planned to be installed around end of the year
  - 1 redirector + 7 servers
  - Native xrootd (3.1.1) without "intermediate" layers
- Computing
  - 70 WNs, 1016 cores, 9752 HS06
  - Possibility to use CMS resources (110 WNs, 1384 cores, 15099 HS06) when not used
  - WNs running SL5 EMI2
    - Update to SL6 to be done by September
  - 6 CEs, used by all WNs
    - SL6, EMI2
  - 1 WLCG-VOBOX (SL6)

Massimo.Sgaravatto@pd.infn.it

LNL-PD STATUS

- Everything is running smoothly
- Only a few issues that had to be addressed
  - Problems with download of Torrent client from NAT-ed WNs
    - Because tcp_tw_recycle was enabled in alitorrent
  - Old WNs with small disks had to be excluded for ALICE after migration to Torrent
  - NAT network (1 Gbps) not far from being saturated
    - Update to 10 Gbps in progress
    - In the meantime 3 x 1Gbps NATs
  - Problems with jobs using too much memory
    - We have in place a script that monitors the memory usage in the WN and kills the processes using too much memory if needed (if the node is running out of mem)
    - Not fast enough for ALICE!
    - needed to also set a limit (4.5 GB) at batch system level

Massimo.Sgaravatto@pd.infn.it

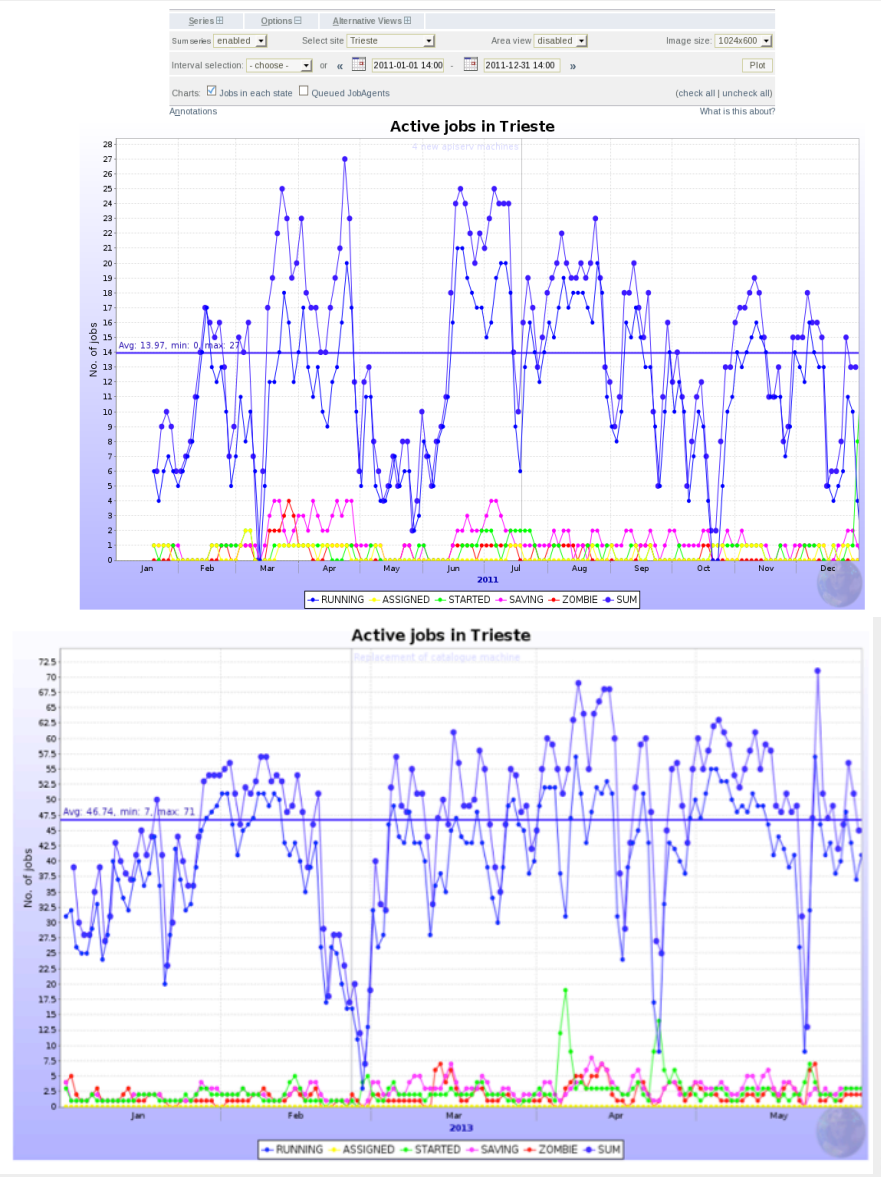
LNL-PD: OTHER ACTIVITIES

- PROOF
  - A prototype deployment of PoD done for CMS
    - Using LSF as batch system
    - Used for quasi interactive analysis
    - Being validated by the CMS local users
  - To be assessed if it can be useful also for local ALICE users
- CVMFS
  - Already deployed in all WNs and used by CMS and LHCb
  - 2 squid proxies, used also by CMS frontier
  - Using CVMFS v. 2.0.19 for the time being
  - Plan is to update to v. 2.1 along with SL6 update
    - Useful for us to share WN cache among multiple CVMFS repos
  - Migration to CVMFS for ALICE should then be quite straightforward

ALICE GRID IN TRIESTE

- 2010 – Alice-Grid set in Trieste (VOBOX and XROOTD)
- 2011 – Alice batch job run in Trieste (parasitic mode with 1 alice job / node with RAM > 8 GB, max job rss 4 GB in 2GB/core farm)
- 2012 – Optimization of xrootd server and grid queue priorities
- 2013 – Feasibility study of virtual infrastructure (OpenNebula) and Analysis Facility in Trieste (PROOF/LSF – XROOTD/GPFS): recycled 4 disused nodes (2 cpu x 2 core/cpu) totaling 16 cores – 140 HS06) with RAM and disks upgrade for testing purposes

- **Deployed resources for Alice: shared disk space 30 TB (GPFS) + 28 TB (XROOTD/GPFS), 24 CPU (80 cores) totaling 860 HEP-SPEC2006**

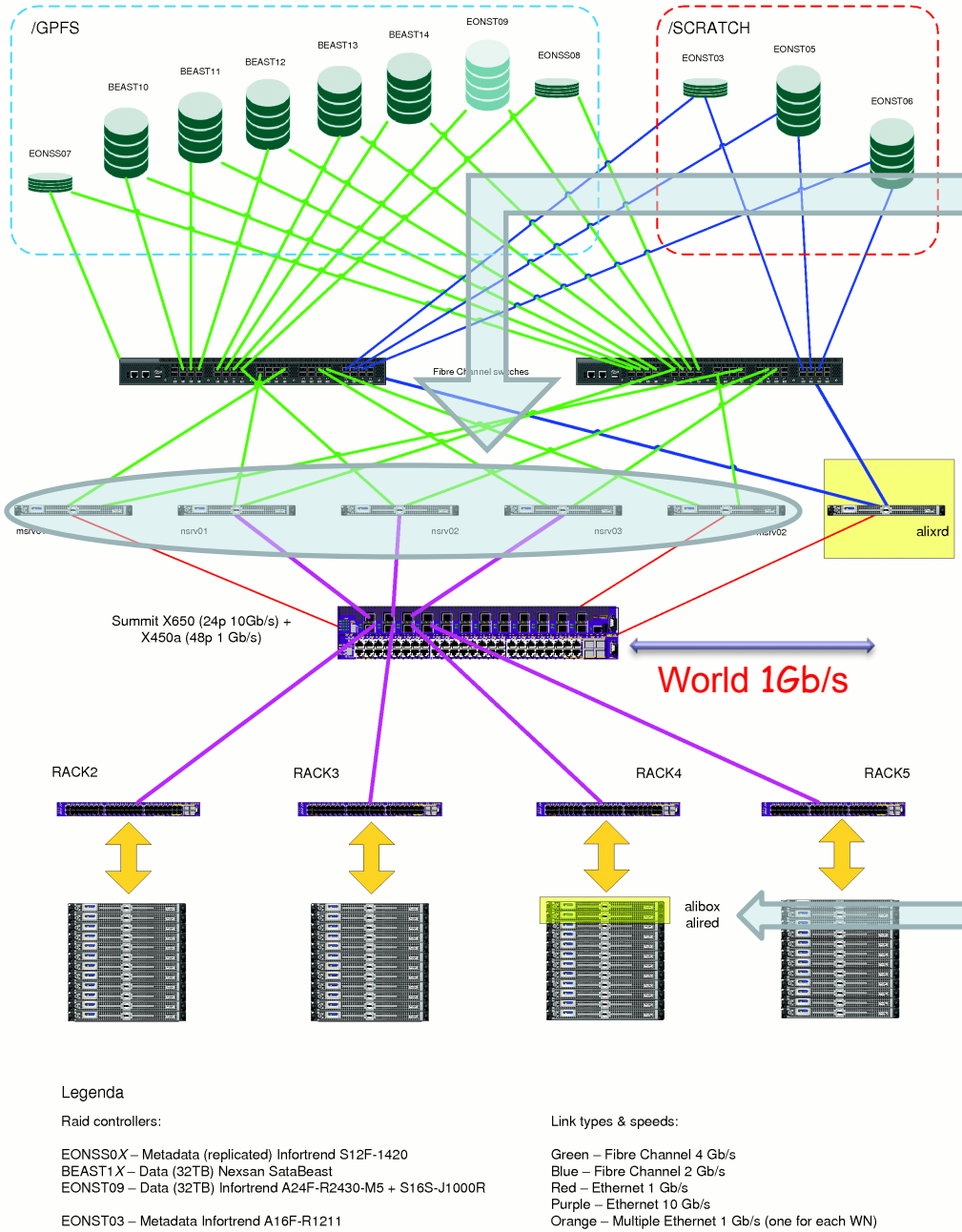


Stefano.Piano@ts.infn.it

THE TRIESTE SITE

- INFN – Trieste Computing Farm operating since 2003, shared by all experimental groups
- Funded by INFN Trieste, CCR, INFN-GRID, INFN groups and single experiments
- Architecture installation, configuration and upgrades: care of INFN - Trieste systems managers
- Present status: shared disk space 190 TB (GPFS), 106 CPU (382 cores) totaling 3300 HEP-SPEC2006
- Grid INFN-TS (LCG-CE + SRM) since 2002, embedded into INFN – Trieste Farm since 2008: local and grid jobs dynamically shared. Running VO's: cms, glast, theophys, lhcb, atlas ... and alice since 2011
- Alice hardware funded by GrIII, PRIN and Alice-Italy

Storage ALICE::Trieste::SE



Legenda

Raid controllers:

EONSS0*X* – Metadata (replicated) Infortrend S12F-1420
BEAST1*X* – Data (32TB) Nexsan SataBeast
EONST09 – Data (32TB) Infortrend A24F-R2430-M5 + S16S-J1000R

EONST03 – Metadata Infortrend A16F-R1211
EONST05/6 – Data (15TB) Infortrend A24F-R2224

Link types & speeds:

Green – Fibre Channel 4 Gb/s
Blue – Fibre Channel 2 Gb/s
Red – Ethernet 1 Gb/s
Purple – Ethernet 10 Gb/s
Orange – Multiple Ethernet 1 Gb/s (one for each WN)

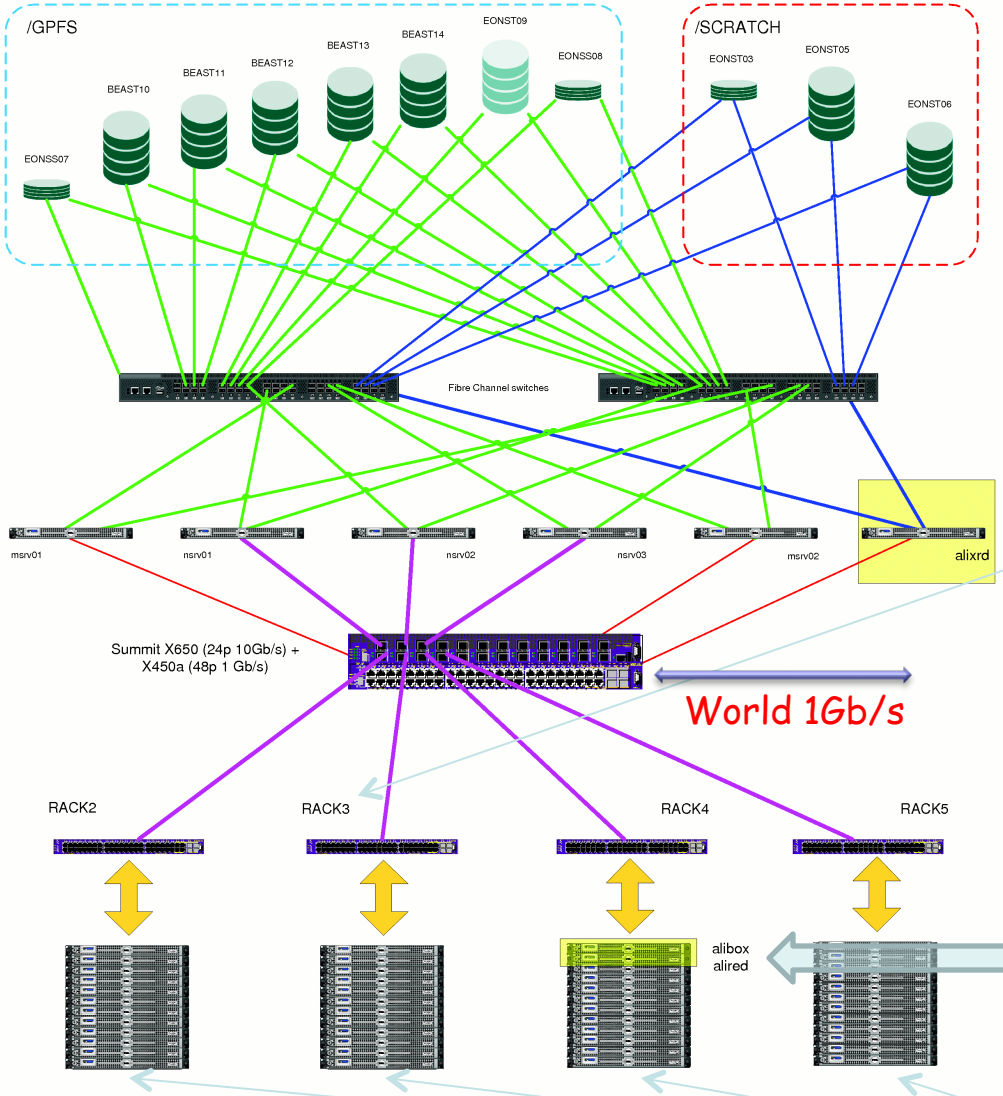
Local storage based on GPFS filesystem:
3 data server
-to disk FC 3x2x 4 Gb/s
-to WN 3x10 Gb/s
2 metada server

1 xrootd server
xrootd on GPFS:
Bandwidth:
- to disk FC 2x 2 Gb/s
- to WN Eth 1 Gb/s

1 xrootd redirector
xrootd on GPFS
Bandwidth:
- to disk Eth 1 Gb/s
- to WN Eth 1 Gb/s

Stefano.Piano@ts.infn.it

ALICE::Trieste::CREAM



Legenda

Raid controllers:

EONSS0*X* – Metadata (replicated) Infortrend S12F-1420
BEAST1*X* – Data (32TB) Nexsan SataBeast
EONST09 – Data (32TB) Infortrend A24F-R2430-M5 + S16S-J1000R

EONST03 – Metadata Infortrend A16F-R1211
EONST05/6 – Data (15TB) Infortrend A24F-R2224

Link types & speeds:

Green – Fibre Channel 4 Gb/s
Blue – Fibre Channel 2 Gb/s
Red – Ethernet 1 Gb/s
Purple – Ethernet 10 Gb/s
Orange – Multiple Ethernet 1 Gb/s (one for each WN)

Local job submission based on LSF Resource Manager
1 LSF Server:
Bandwidth:1 Gb/s

1 CREAM-CE:
Bandwidth: 1 Gb/s

1 VO-Box (alibox)
Bandwidth: 1 Gb/s

106 CPU's
Bandwidth:1 Gb/s

Stefano.Piano@ts.infn.it

Bari Farm - Computing Resures

- 18 Computer Racks
- ≃ 4000 CPU cores
- 1.6 PByte Global Storage
- 110 MB/s network bandwith for each WorkerNode
- 18k Jobs queue deep
- 125 kW power use up
- 24 Virtual Organizations
- Uman resources 4 FTE
  (Diacono Domenico, Gervasoni Riccardo, Franco Antonio, Donvito Giacinto, Spinoso Vincenzo)

About 520.000 running jobs per month !

D. Di Bari - A. Franco

2

giovedì 13 dicembre 12

Antonio.Franco@ba.infn.it

BARI STATUS AND ISSUES

- Storage full since Feb 2012!
- Bad power cut in March killed several machines, including the xrootd redirector that was reinstalled on a spare machine.
- Currently running about 500 jobs.
- The farm is not yet in full production because of UPS issues not yet fully solved
  - One UPS needs to be replaced and the system to be upgraded

Antonio.Franco@ba.infn.it

TORINO STATUS AND ISSUES

- Storage
  - Current status: Lustre+xrootd for data access, GlusterFS for Cloud backend
  - Migrating all storage to GlusterFS while commissioning additional 200TB
- CPU
  - Most resources and services managed in an IaaS cloud infrastructure (the Clumsy Infrastructure?)
  - Some (variable) fraction dedicated to the Torino Virtual Analysis Facility
  - Migration to SL6 complete
- Remarks
  - Willing to test CVMFS: configured for all nodes, squid proxy ready
  - Since many people mentioned it, we're using puppet as well
  - The chiller that cools the Tier-2 is having problems
  - We will likely have to add a second one, if we can secure funding

THE TORINO CLOUD: TWO CLUSTERS

VMs providing **critical services**:

- in- & out-bound connectivity
- public & private IP
- live migration
- no special I/O requirements

VMs providing **computing workforce**:

- example: Grid WNs
- private IP only
- high storage I/O performance

THE TORINO CLOUD: TWO CLUSTERS

- Server-class hardware
- Shared image repository
- Resiliency-optimized FS for shared system disks
- Currently 4 hosts

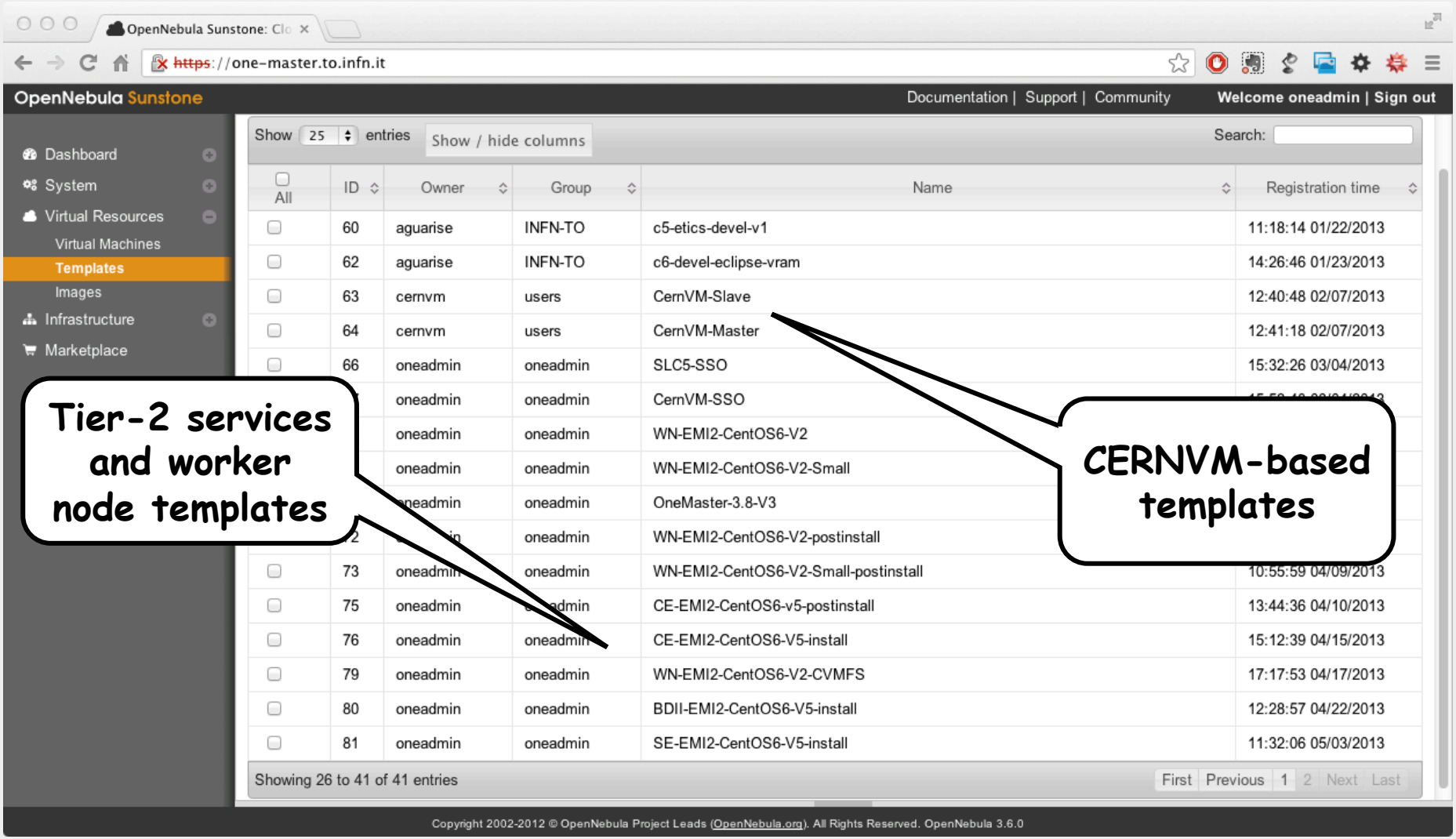
- Working-class hardware ☺
- Cached image repository
- Access to perfomance-optimized FS for data needs
- Currently 35 hosts

THE INGREDIENTS

- Cloud management Toolkit: **OpenNebula**
  - OpenStack, now widely adopted in new projects, was too embrionic when we started
  - ...and arguably* OpenNebula is better suited at Data Center Nebulization
  - Currently using version 3.6, will migrate to 3.8/4.0 soon
  - Templates based on few very simple images plus full contextualization via scripts and puppet (looking into CloudInit)
- Backend storage: **GlusterFS**
  - Flexible enough to cater to different needs with a single tool
- VM network management: **OpenWRT**
  - OpenVSwitch was not integrated in OpenNebula when we started

* See e.g. blog.opennebula.org/?p=4042

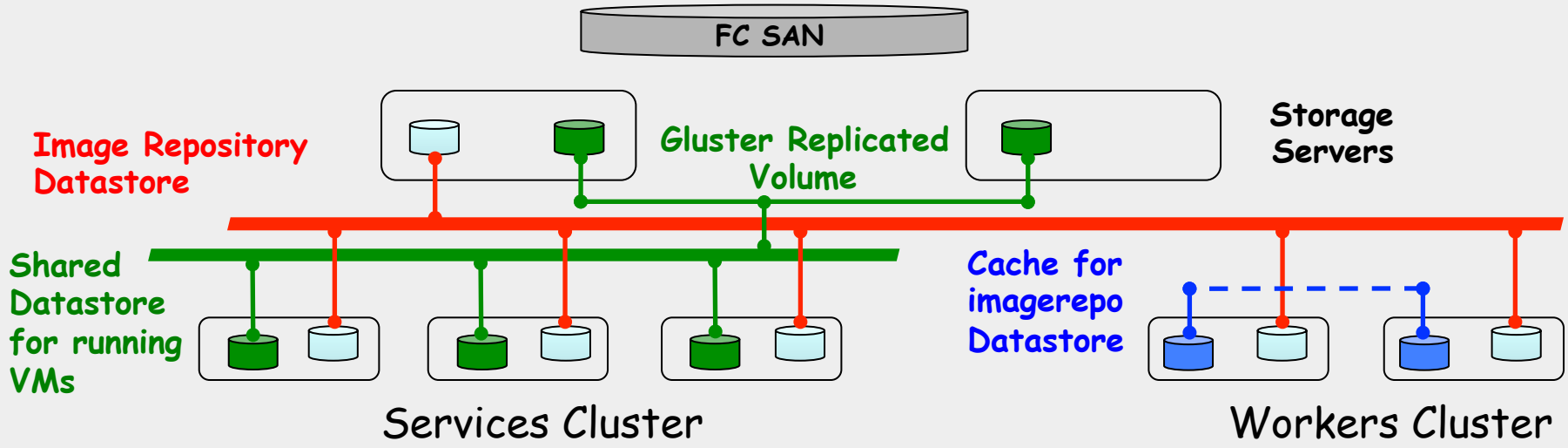
SUNSTONE DASHBOARD



MULTIPURPOSE STORAGE: GLUSTERFS

Two storage servers with 10Gbps interface provide some of the LUNs through GlusterFS

- All the virtual machines run on RAW or QCOW file images
- Services System Datastore is **shared** to allow live migration
- Workers System Datastore is **local** to the hypervisors to increase I/O capacity. Images repository is locally **cached** on each hypervisor to reduce startup time.
  - An ad-hoc script synchronizes the local copies using a custom "torrent-like" tool (scpWave + rsync) when new versions of the images are saved



- Variable number of workers
  - It's a Virtual Analysis Facility
  - See D. Berzano's presentation at last Offline Week
  - 50TB dedicated storage

ALICE PROOF Clusters

What is this about?

Cluster list

	Cluster					ROOT	Aggregated disk space			AF xrootd		xrootd
Name	Online	Status	Proof master	Workers	Users	Version	Total	Free	Used	Running	Latest	Version
1. CAF		Stable	alice-caf.cern.ch	116	1	v5-34-05	162 TB	12.9 TB	149.1 TB	1.0.50	1.0.50	20100510-1509_dbg
2. CAF_TEST				-	-		-	-	-			
3. JRAF		Stable	jraf.jinr.ru	48	0	v5-34-05	14.13 TB	9.912 TB	4.215 TB	1.0.50	1.0.50	20100510-1509_dbg
4. KIAF		Stable	kiaf.sdfarm.kr	96	0	v5-34-05	171.9 TB	5.393 TB	166.5 TB	1.0.50	1.0.50	20100510-1509_dbg
5. LAF				-	-		-	-	-			
6. SAF		Maintenance sin...	nansafmaster.in2p3.fr	48	0	v5-34-05	12.07 TB	1.473 TB	10.6 TB	1.0.50	1.0.50	20100510-1509_dbg
7. SKAF				-	-		0	0	0			v3.3.1
8. SKAF_TEST				-	-		-	-	-			
9. TAF		Virtually stable!	pmaster.to.infn.it	85	0	v5-34-05	49.1 TB	1.545 TB	47.56 TB			v3.0.4_dbg
Total				**393**	**1**		**409.2 TB**	**31.23 TB**	**378 TB**			

OUTLOOK 1

- Funding for resources in the next two years will be (mostly) coming from specific projects aimed at southern regions ("PON")
  - Two separate projects will provide CPU (RECAS) and Storage (PEGASUS) resources
  - ...but only in Napoli, Cosenza, **Bari** and **Catania**
  - INFN will only replace obsolescent resources in other sites for 2014 and 2015
  - Infrastructure-only projects
- This has implications:
  - Constraints in resource distribution
  - Manpower for management will be an issue due to recruitment limitations in Italy and the ending of EU Grid Projects

OUTLOOK 2

- Some manpower coming from a National Research Project ("PRIN") approved for 2013-2015
  - STOA-LHC: Optimization of data access, network and interactive data analysis for LHC experiments
- Most ALICE-relevant activities focused on resource federation (xrootd and clouds) and interactive analysis (PROOF)
  - Activities on interactive analysis on cloud infrastructures (Torino, Trieste), optimization of data access (Bari), development of a Science Gateway for ALICE (Catania)
- Most funding is for manpower
  - Funds made available last March
  - Boils down to o(1) postdoc or equivalent per site
  - Some sites managed to get the person (e.g. Torino, welcome Sara!), in others the process is ongoing (e.g. Legnaro, Catania)

THANKS!

- Questions?