ALICE Tier1 – Tier2 Workshop

ITALY STATUS AND PLANS

University of Tsukuba, Japan – March 5-7, 2014

INFN

OUTLINE OF THE TALK

- Resources and operations
- Networking status
- Resource evolution
- The Tier-1 at INFN-CNAF
- Status of Tier-2 sites
- The STOA-LHC project and Virtual Analysis Facilities

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

OPERATIONS ORGANIZATION

- ALICE-IT Computing coordination: Domenico Elia, Stefano Bagnasco deputy
- New Tier-2 Operations Coordinator: Stefano Piano
  - Monthly phone conference for coordination and performance monitoring
  - Yearly face-to-face workshop, last one in Trieste in december
- Monthly Tier-1 Management Board at CNAF
  - Not much local activity from us, compared to ATLAS and CMS

NETWORKING: LHCONE, IPV6 ETC.

- All Tier-2s connected to LHCONE with at least 10Gbps links
  - Through GARR-X (The Italian NREN)
  - Dark fibers leased to GARR by providers, GARR directly owns both routing and transmissive devices
  - Most of T2s easily upgradable to 20-40 Gbps (should the need arise). Padova/Legnaro already at 20Gbps.
  - Tier-1 at 20 Gbps (LHCONE + LHCOPN) + 10 Gbps (General purpose)
  - Trieste coming soon
- IPv6 readiness
  - Not much activity so far
  - All INFN sites will plan and act coordinately at some point (no exact estimate yet)

STATUS AT THE END OF 2013

- **Tier-1:**

CPU: 18620 HS06 / DISK: 1.7 PB / TAPE: 3.7 PB

- **Tier-2 (+ Cagliari):**

CPU: 30120 HS06 / DISK: 1.9 PB

	Bari	Catania	LNL-Padova	Torino	Cagliari	Totale
HS06	8984	3110	8264	7805	1960	30123
TB	452	358	357	634	70	1871
Full	53%	87%	98%	61%	98%	

Procurement 2013 (not included above):

- No new CPU
- Storage: additional ~250 TB each for Bari, Catania and LNL-Padova

Domenico.Elia@ba.infn.it

EXPECTED IN 2014

- **Tier-1:**

Will match INFN share of the T1 requests at the last RRB

CPU: 20900 HS06 / DISK: 1.9 PB / TAPE: 2.5 PB

- **Tier-2 (+ Cagliari):**

Same as for T1, INFN pledges according to requirements

CPU: 37050 HS06 / DISK: 2.5 PB

To be replaced:

	Bari	Catania	LNL-Padova	Torino	Cagliari	Totale
HS06	7416	2035	1168	981	0	11600
TB	92	0	32	72	0	196

Procurement 2014:

- CPU: 18500 HS06 (including 11600 HS06 replacements)
- Storage: 280 TB (including 200 TB replacements)

Domenico.Elia@ba.infn.it

PERPECTIVES FOR 2015

As from RRB October 2013:

Resource	Site(s)	2014 request	2015 request
CPU (kHS06)	T0	135	175
	T1	110	120
	T2	190	200
Disk (PB)	T0	8.3	11.5
	T1	10.1	17.8
	T2	12.8	22.1
Tape (PB)	T0	12.0	16.2
	T1	6.0	10.2

Table 11 Summary of ALICE resource requests for 2014 and 2015. The 2014 requests are unchanged since Spring 2013. CERN CAF resources are included in T0. The T1 disk requirements include 2.35 PB of disk buffers for the tape systems.

According to INFN share, that would imply:

	increase wrt 2014
T1	
CPU: 22800 HS06	+9%
DISK: 3.4 PB	+79%
TAPE: 4.2 PB	+68%
T2s	
CPU: 39000 HS06	+5%
DISK: 4.3 PB	+72%

Domenico.Elia@ba.infn.it

THE TIER 1 AT CNAF

- 17k cores overall (about 180 kHEPSpec06)
- ALICE share is 18620 HS06 (about 1600 job slots)
  - LSF for queue management
  - Few WNs virtualized in a cloud-like architecture ("Worker Nodes On Demand", WNODes) but not used by ALICE
- 14PB of disk
- ALICE share is 1.6 PB (T1D0+T0D1) plus tapes for 3.7 PB (700 TB used by ALICE)
  - GPFS + TSM for management
  - Xrootd as a front-end protocol
- Staff of 20 (19 FTE) to babysit the whole centre
  - Plus 1 person dedicated to ALICE-specific operations (F. Noferini), not full time

XROOTD TAPE AT CNAF

- A plugin to manage recall via xrootd was developed.
  - http://www.bo.infn.it/alice/cnaf/XrdxFtsOfs.cc
  - It returns a WAIT to the client in case the file is not available in the buffer (at the same time it triggers the recall).
- Recently added: additional feature adding the md5sum as an extended attribute when a file is written, in order to have it always available even if the file is not in the buffer.
  - For TAPE instance only
  - We did this because last year we found a relevant fraction of file with a wrong md5sum and we were not able to understand why this happened. To fix the problem we recalled all the files: we want to avoid to do that in the future.

Francesco.Noferini@bo.infn.it

TIER 2 SITE: GENERAL REMARKS

- Recent in-depth review of INFN Tier-2 operations requested by INFN management
  - Passed without any relevant issues
- All sites can allow for more resources coming in without big infrastructural investments
  - Bari and Catania will become rather large in the next future, but funding for infrastructural upgrades is secured (see next slide)
  - All sites working to expand support to more VOs beyond LHC ones, to allow for resource optimization
- Manpower is tight (but this is hardly news...)

TIER 2 SITES: RESOURCES

- CPU and storage usually bought through common tenders
  - Good collaboration with CMS sites
- Resources on average balanced across four sites
  - Past exceptions: electrical power limitations, procurement constraints (e.g. full storage enclosures)
  - Current and future exceptions: Bari and Catania special funding (RECAS project) will put most of the resources there for 2014 and a sizeable contribution for 2015.
  - For 2014 INFN will only fund replacement of obsolescent resources elsewhere.

TIER 2 SITES : UPDATES

- CVMFS deployed and working everywhere
  - Bari and Torino used for a comparison between CVMFS and Torrent
  - A bit more communication would have been appreciated ☺
- XROOTD upgrade to at least 3.2.1 everywhere
  - Except Trieste (planned soon, waiting for new hardware)
- Migration EMI-2 to EMI-3 ongoing
  - Will meet deadlines

TIER2 SITE: SITE SPECIFIC REMARKS

Bari:

- Resource sharing with CMS was not always optimal (fixed now)
- An UPS issue reduced the number of available slots in 2013

Torino:

- A BES-III Tier-2 being setup alongside the WLCG one
  - Different virtual WNs, but same physical machines that can be used by ALICE when not in use by BES
- AliEn submission seen as “too little aggressive”, other VOs difficult to quench by only setting priorities
- (nearly) all storage migrated from Lustre to GlusterFS
  - 51 TBs of storage currently dedicated to VAF: can we do something about this? (see later)

TIER2 SITE: SITE SPECIFIC REMARKS

Catania:

- Storage is currently offline for some refurbishing
- Electrical power funding issue solved

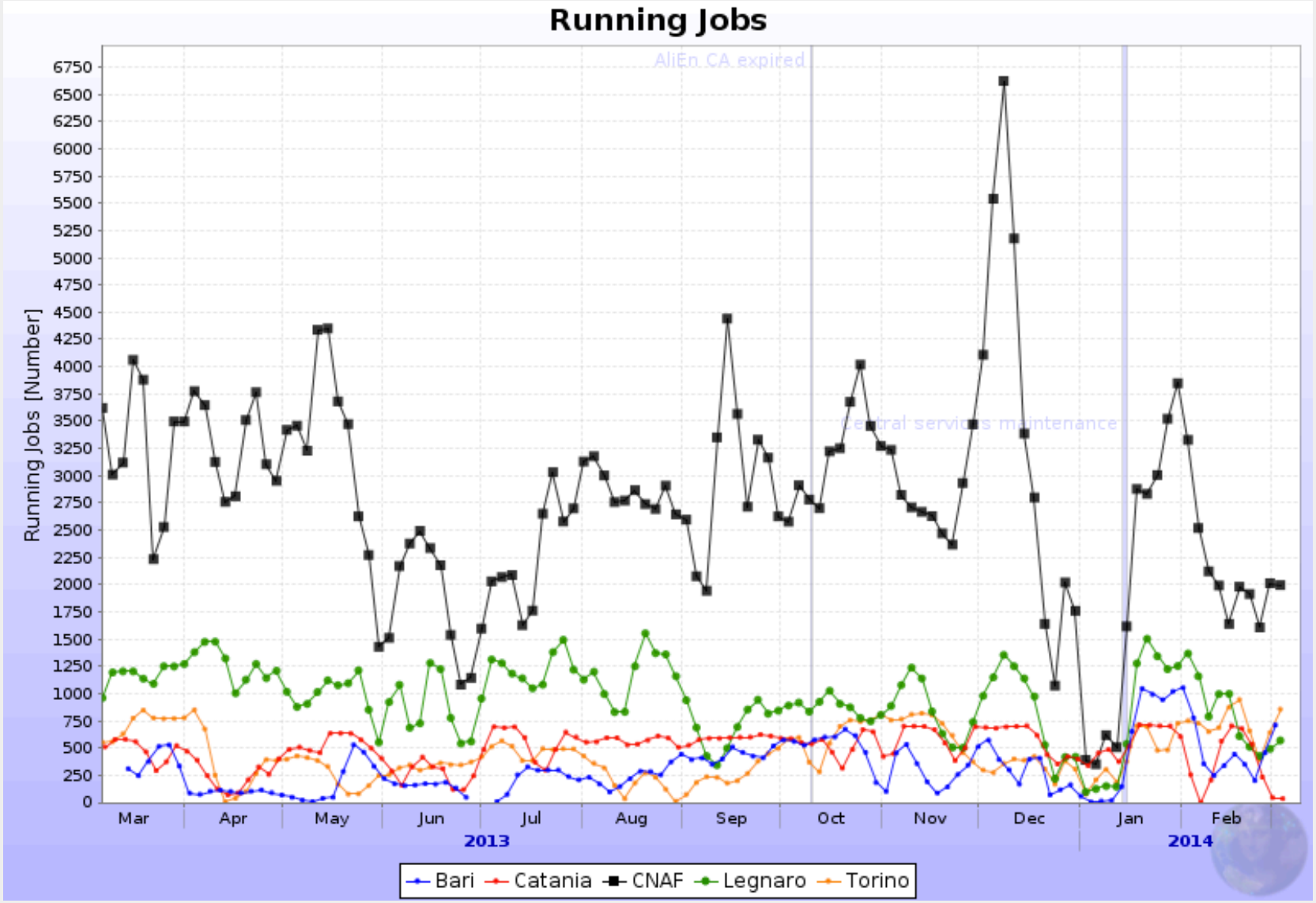
Padova/Legnaro:

- The split across two physical sites (INFN site at Padova University and INFN's Legnaro National Laboratories) proved to be a good way to optimize scarce manpower
- Outstanding availability
- An OpenStack-based Cloud infrastructure being set-up (not integrated with the Tier-2, at least in the beginning)

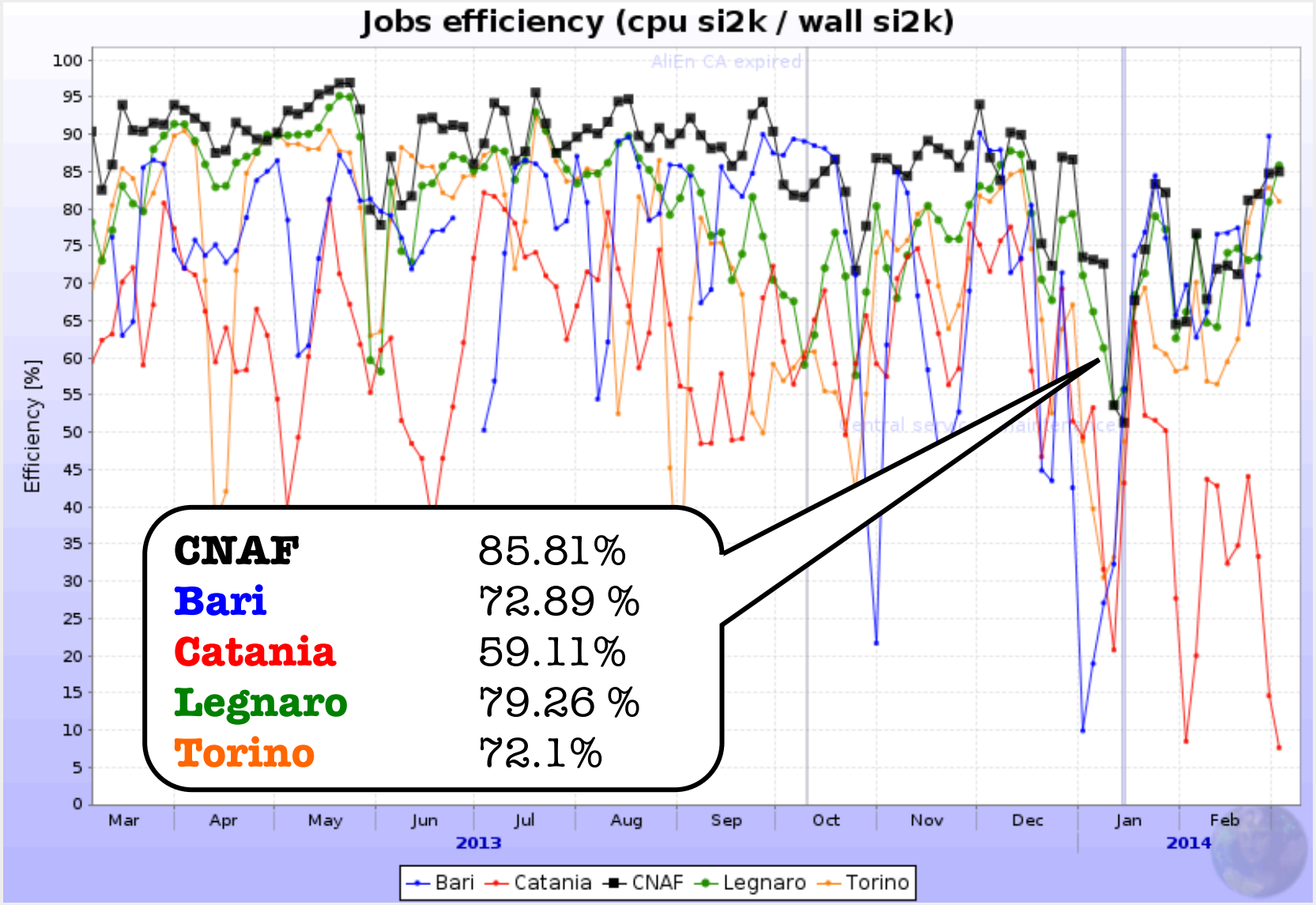
Trieste:

- XROOTD to be decoupled from GPFS (pending arrival of new hardare, ~2 months)

MANDATORY PLOT 1: RUNNING PROFILE



MANDATORY PLOT2: CPU EFFICIENCY



CNAF	85.81%
Bari	72.89 %
Catania	59.11%
Legnaro	79.26 %
Torino	72.1%

THE STOA LHC PROJECT

- Three-year 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)
- Post-doc positions in Torino, Bari, Legnaro and Trieste
  - Recruitment completed, work started for good
  - Kick-off meeting in Padova in January, coordination meeting at CERN by end of March.

The actual implementation:

- A working prototype of computing resources integration between batch and interactive analysis via cloud computing technologies through dynamic "elastic provisioning" of Virtual Analysis Facilities
  - Based on Dario Berzano's doctoral work in Torino and at CERN
  - Already in production in Torino since two years
  - See also Dario Berzano and Sara Vallero presentations at CHEP2013
  - On the OpenNebula middleware stack
- Bari, Legnaro, Trieste sites are replicating the Turin infrastructure and integrate into existing local cluster
  - On the OpenStack middleware stack
  - **Bari**: existing PRISMA cloud testbed
  - **Padova/Legnaro**: upcoming "Paduan Area Scientific Cloud"
  - **Trieste**: prototype deployment ready

The virtual analysis facility

PROOF	PoD	µCernVM	Elastiq
HTCondor	CVM online	CernVM-FS	authn/authz

- **What:** a cluster of µCernVMs with HTCondor
  → *One **head node** plus a scalable number of **workers***
- **How:** contextualization configured on the Web
  → *Simple web interface: http://cernvm-online.cern.ch*
- **Who:** so easy that can even be created by end users
  → *You can have your **personal analysis facility***
- **When:** scales up/down automatically
  → *Optimal usage of resources: fundamental **when you pay** for them!*

Dario.Berzano@cern.ch - TAF: deployment perspectives in other sites 12

Dario.Berzano@to.infn.it

Elastiq queue monitor

Python app to monitor HTCondor and scale up or down

Jobs waiting too long will trigger a **scale up**

Demo!

Supports **minimum and maximum quota** of VMs

You deploy only the master node: minimum quota immediately launches VMs automatically

Supports **any HTCondor cluster** (also **hybrid**: physical+virtual) and has a **trivial configuration**

Code available at **http://bit.ly/elastiq**

Dario.Berzano@to.infn.it

Outlook and issues:

- The scalability and elasticity of the computing resources via dynamic provisioning is limited by the size of the computing site:
  - The Grid Tier-2s are anelastic!
  - Scheduling VMs is still an issue (but lots of work planned by INFN, not only in this context)
  - Federate different sites
  - Elasticity among different clouds (“Cloudbursting”)

- Data access is still an issue
  - Avoid dedicated storage (but need to collect data anyway)
  - Minimize catalogue queries (Try TDataSetManagerAliEn?)
  - Minimize cross-site data transfer
  - Look into storage federation along with cloud federation
  - Investigate data distribution strategy

THANKS!

- Questions?