Grid Operations in Germany

T1-T2 workshop 2016

Bergen, Norway

Kilian Schwarz

Sören Fleischer

Raffaele Grosso

Christopher Jung

Table of contents

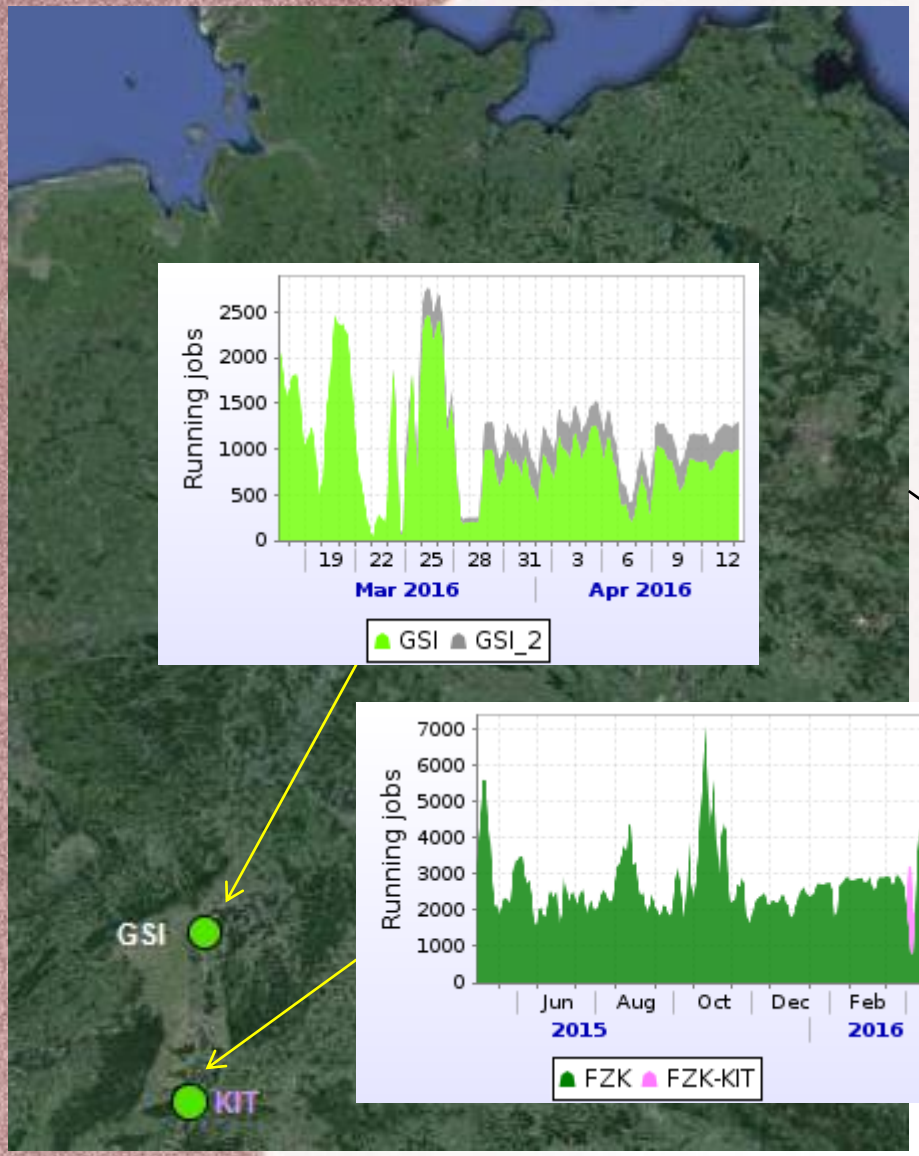
- Overview
- GridKa T1
- GSI T2
- UF
- Summary

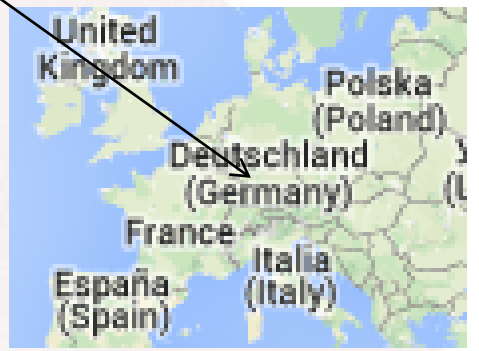
Table of contents

- Overview
- GridKa T1
- GSI T2
- UF
- Summary

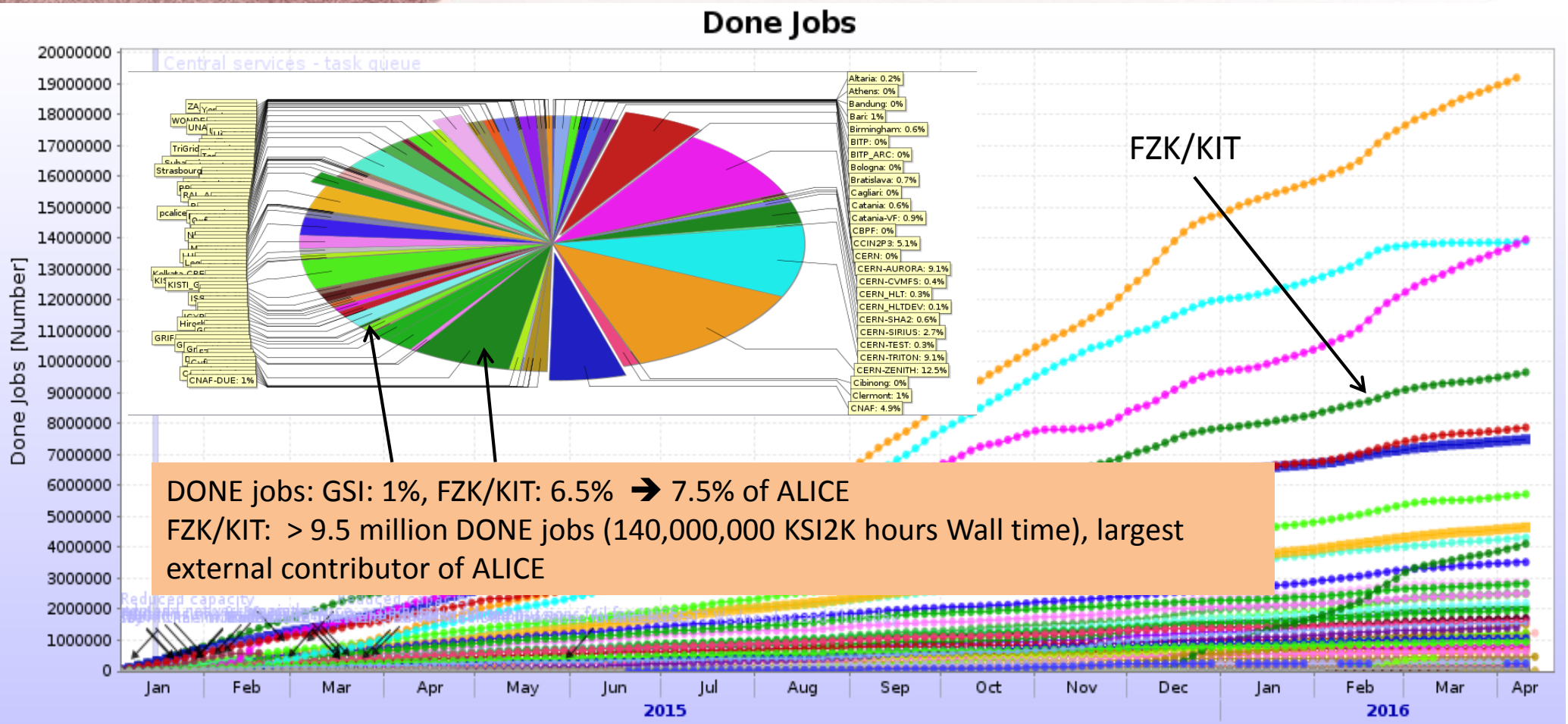
Map of German Grid sites

- T1: GridKa/FZK/KIT in Karlsruhe
- T2: GSI in Darmstadt





Job contribution (last year)



DONE jobs: GSI: 1%, FZK/KIT: 6.5% ➔ 7.5% of ALICE

FZK/KIT: > 9.5 million DONE jobs (140,000,000 KSI2K hours Wall time), largest external contributor of ALICE

Altaria Athens Bandung Bari Birmingham BITP BITP_ARC Bologna Bratislava Cagliari Catania Catania-VF CBPF CCIN2P3
CERN CERN-AURORA CERN (Wigner) CERN (Meyrin) CERN-SIRIUS CERN-TEST CERN-TRITON CERN-ZENITH CERN_HLT CERN_HLTDEV
Cibinong Clermont CNAF CNAF-DUE COMSATS CondorCloud CondorSite CSC Cyfronet DCSC_KU FZK FZK-KIT Grenoble
GRIF_IPNO GRIF_IRFU GRIF_IRFU_ARC GSI GSI_2 HIP Hiroshima ICYB ICYB_ARC IHEP IPNL ISMA ISS ISS_LCG ITEP JINR
KFKI KISTI_GSDC KISTI_GSDC-T1 KNU Kolkata-CREAM Kosice LBL Legnaro LLNL LUNARC Madrid MEPHI NECTEC NERSC
NIHAM NIKHEF NIPNE ORNL OSC Oxford PAKGRID pcalice92.cern.ch PNPI Poznan Prague RAL RAL_ARC RRC-KI
RRC_KI_T1 SaoPaulo SARA SNIC SPbSU Strasbourg_IRES Subatech Subatech_CCIPL SUT Torino Trieste TriGrid_Catania
Troitsk Trujillo UF UiB UNAM UNAM_T1 Vienna WONDERLAND Wuhan WUT Yerevan ZA_CHPC

Storage contribution

AliEn name	Size	Used	Free	Usage	No. of files	Type	Size	Used	Free	Usage
ALICE::FZK::SE	1.901 PB	1.71 PB	196.6 TB	89.91%	39,836,201	FILE	3.998 PB	3.756 PB	247.8 TB	93.95%
ALICE::GSI::SE2	1000 TB	972.4 TB	27.58 TB	97.24%	33,839,231	FILE	14.55 PB	11.44 PB	3.106 PB	78.65%
ALICE::FZK::TAPE	640 TB	3.984 PB	-	637.4%	2,701,442	FILE	640.4 TB	574.1 TB	66.33 TB	89.64%

Total size:

- GridKa: 2.7 PB Disk SE including 0.6 PB tape buffer
- GSI: 1 PB Disk SE

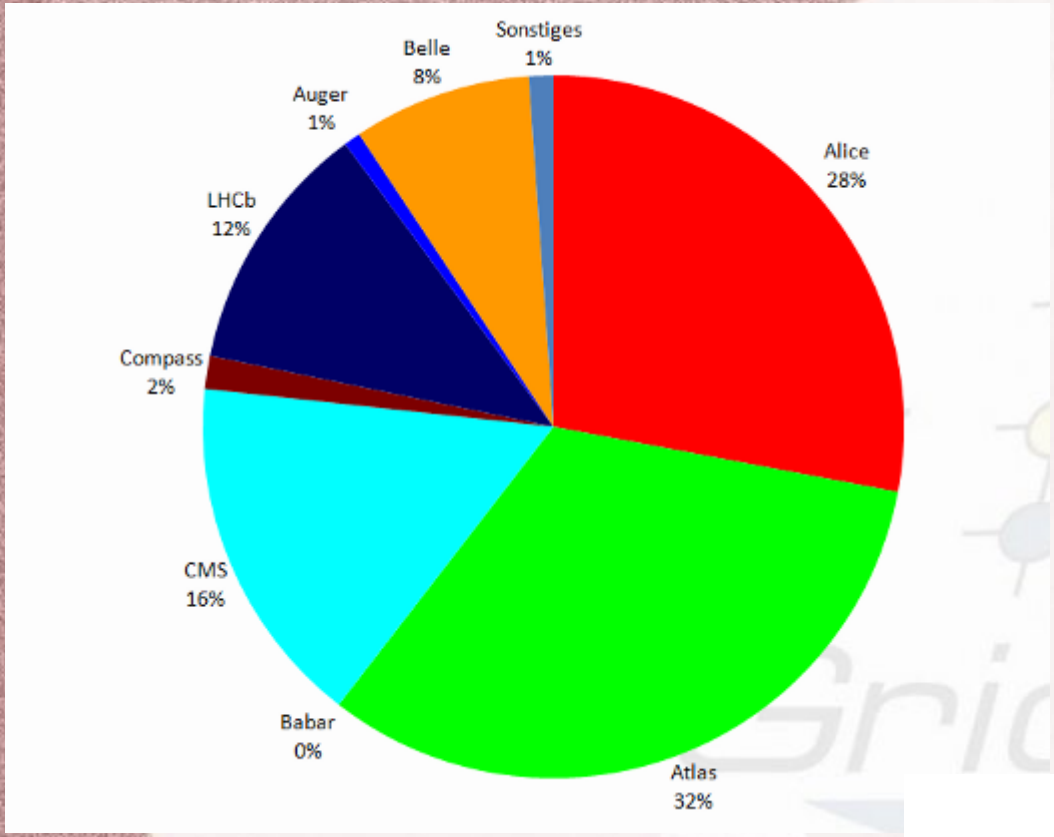
- GSI & FZK: xrootd shows twice the storage capacity
- GSI: an extra quota for the ALICE SE has been introduced in the Lustre file system. Can the displayed capacity be based on this quota ?

- 3.7 PB disk based SE
- 5.25 PB tape capacity
- 640 TB disk buffer with Tape backend

Table of contents

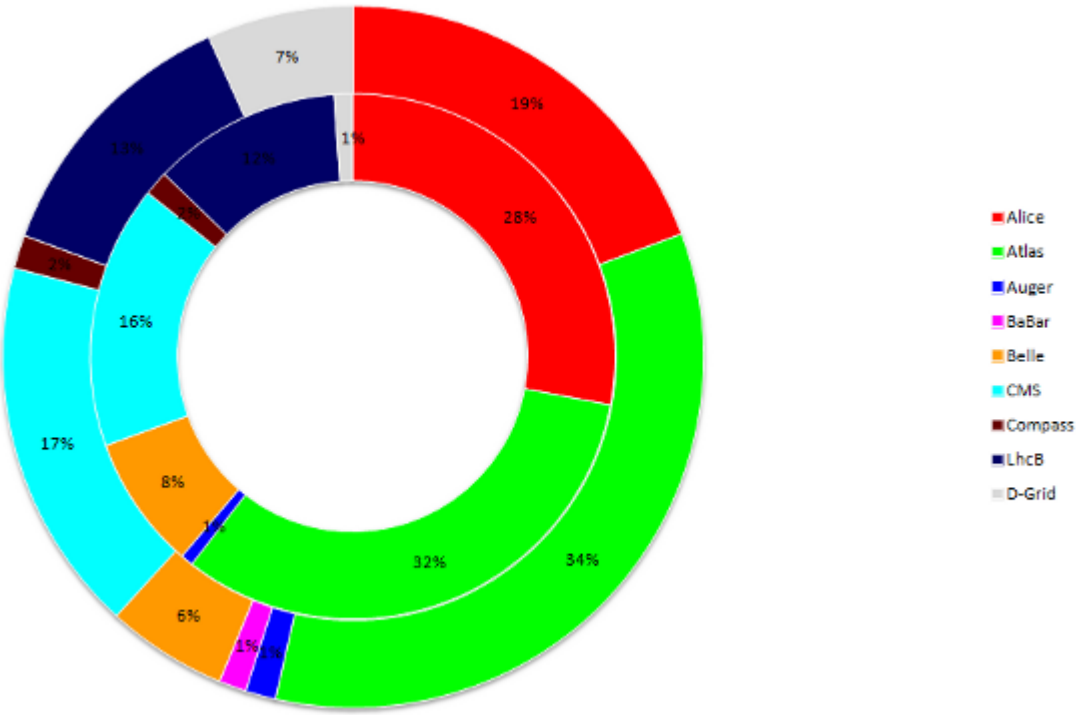
- Overview
- GridKa T1
- GSI T2
- UF
- Summary

GridKa Usage
(annual statistics 2015)



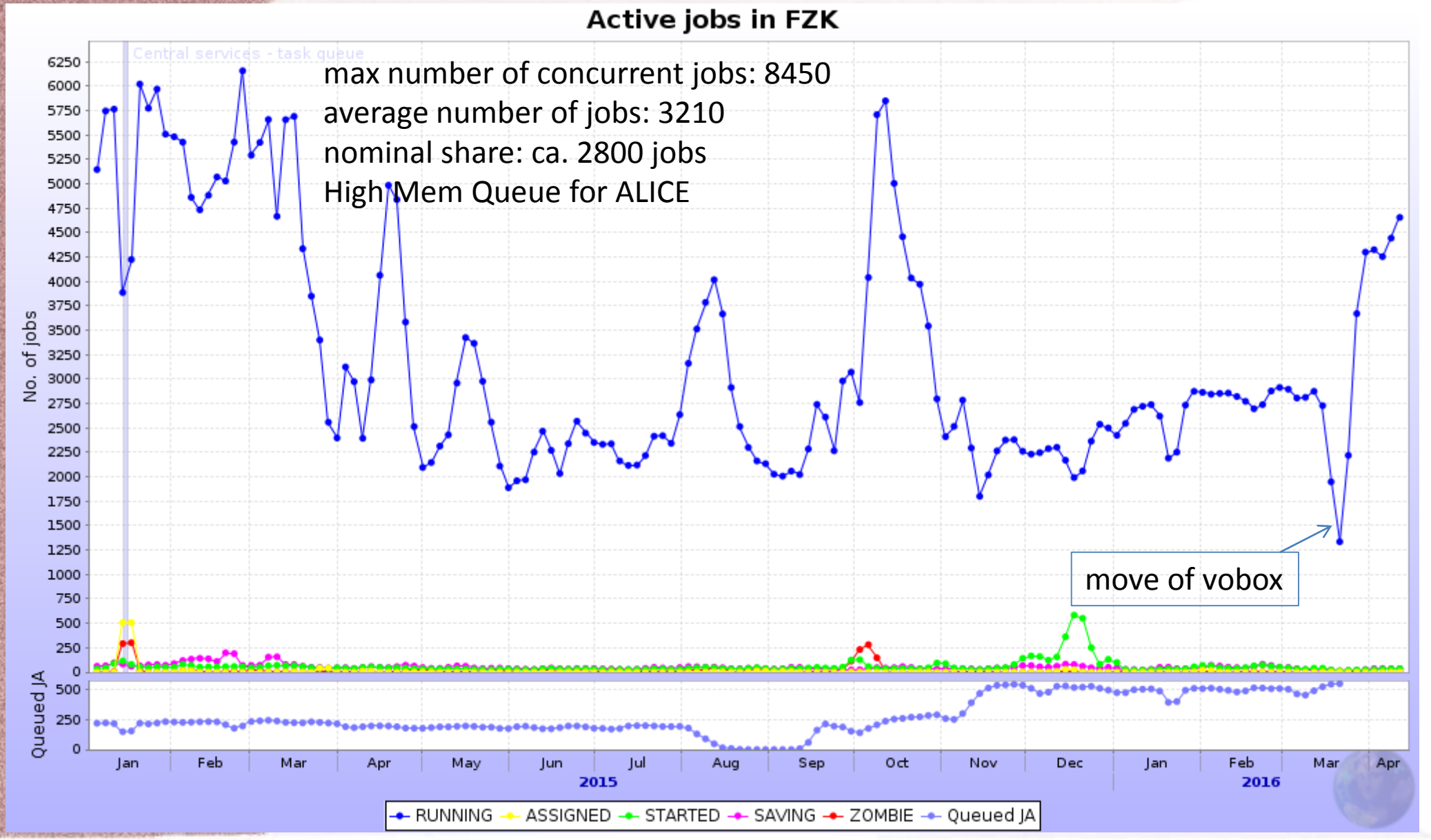
- The two largest users are ATLAS and ALICE
- The 4 LHC experiments together use about 90%

- ALICE and Belle use more than the nominal share
- ATLAS, CMS, and LHCb use about their nominal share

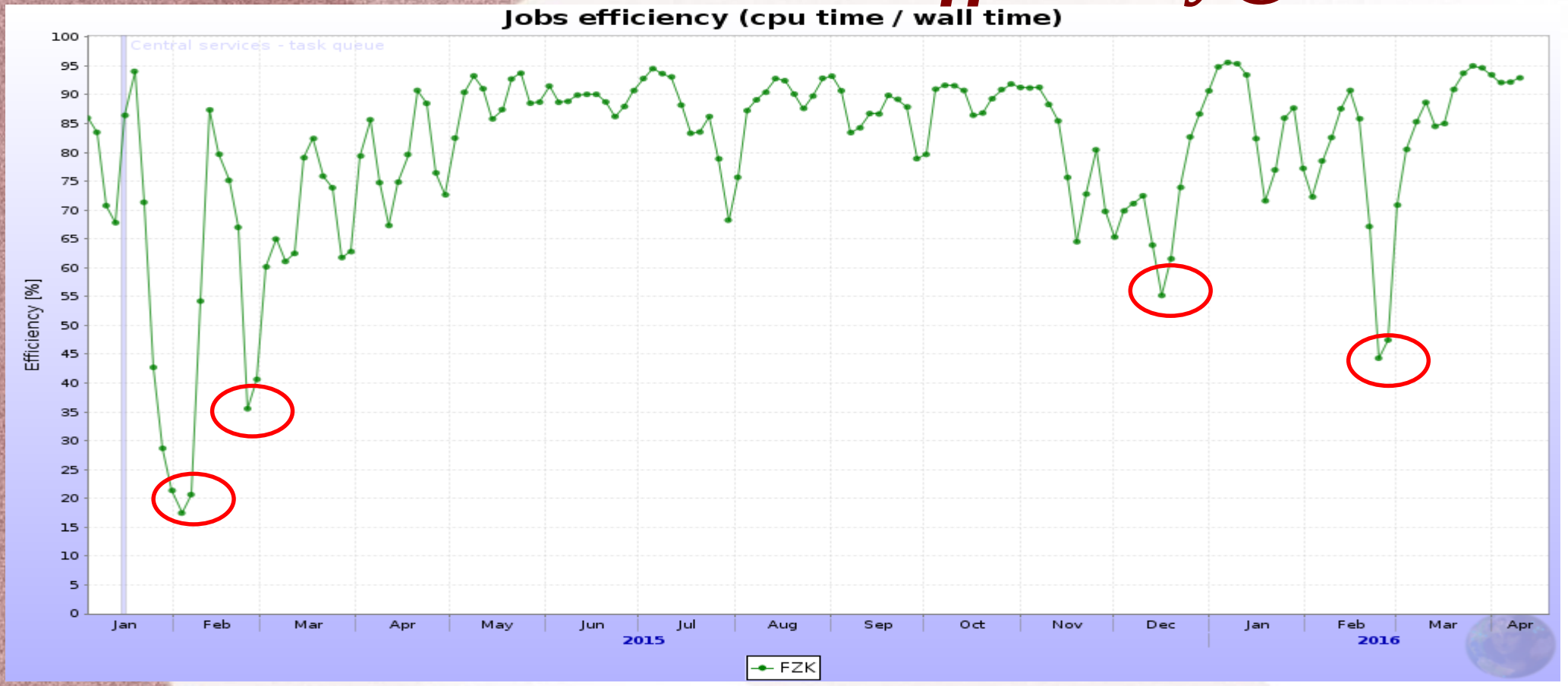


total used wall time: 106 500 284 hours

Jobs at GridKa within last year



ALICE Job Efficiency@GridKa



- average job efficiency: 80%

- 4 major efficiency drops:

- Jan/Feb 2015

frequent remote data reading from WNs.

Communication takes place via Firewall (NAT)

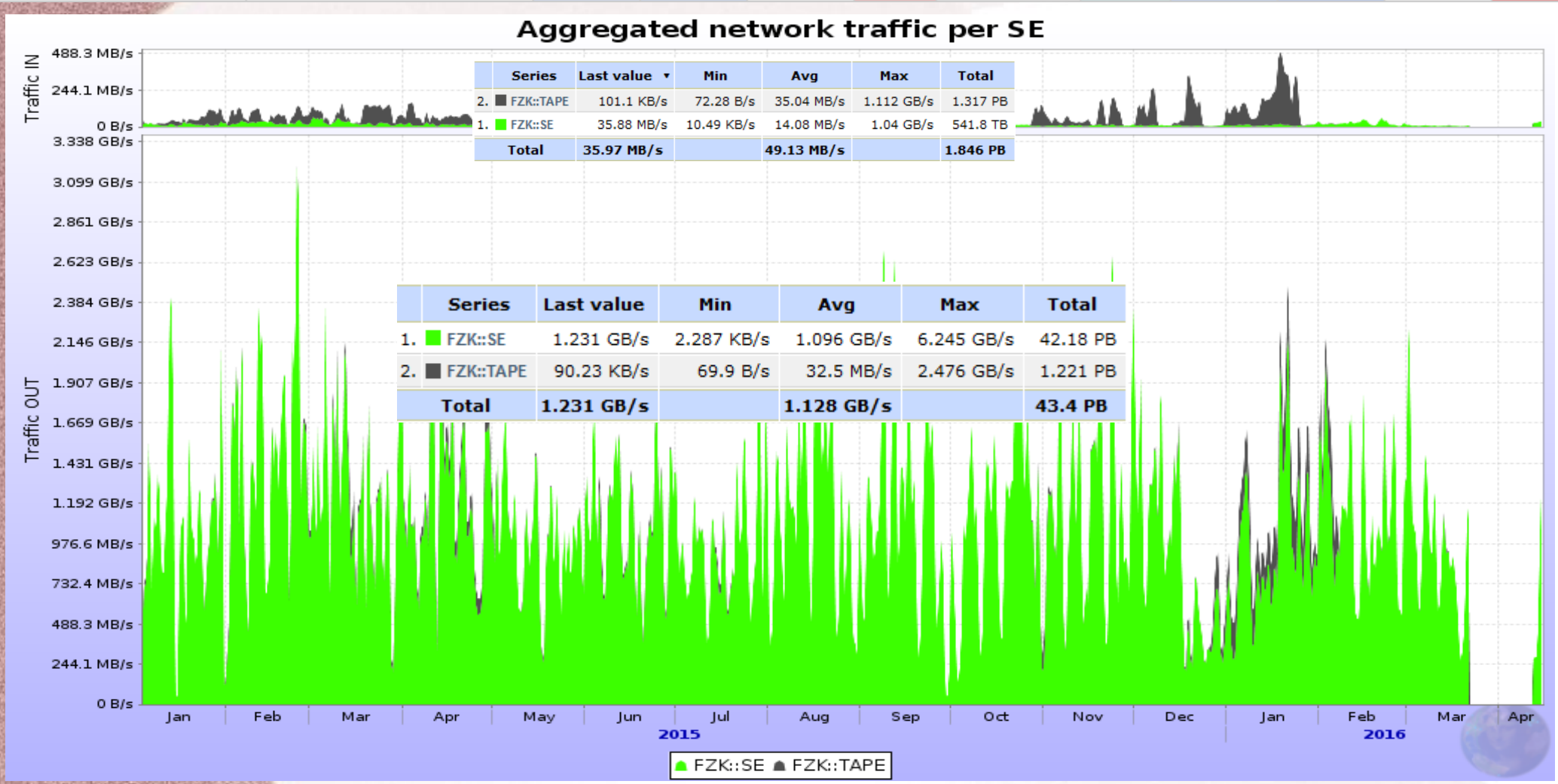
(Firewall upgrade from 2.5 to 10 Gb took place in spring 2015)

- later efficiency drops (Dec 2015 & Feb 2016) affected also other sites (raw data calibration)

xrootd based SEs work fine and are used intensively.

- Disk-based xrootd-SE has a size of 2.1 PB (+ 0.6 PB tape buffer)
- 43 PB have been read from ALICE::FZK::SE in 2015 (average reading rate: ~1.1 GB/s)
  - (3/2 of what has been reported in Torino)
- incoming data went mainly to tape (1.3 PB)

	AliEn SE	Statistics						Xrootd info			
SE Name	AliEn name	Size	Used	Free	Usage	No. of files	Type	Size	Used	Free	Usage
13. FZK - SE	ALICE::FZK::SE	1.901 PB	1.713 PB	193.2 TB	90.08%	40,005,709	FILE	3.998 PB	3.765 PB	239.3 TB	94.16%



ALICE::FZK::TAPE	640 TB	3.984 PB	-	637.4%	2,701,442	FILE	640.4 TB	574.1 TB	66.33 TB	89.64%

XRootD Issues at GridKa

ALICE::FZK::SE:

- xrootd SE and duplicated name space - in MonaLisa still the correct size is not published (currently reported size is 3.9 PB for disk SE, correct should be 2.1 PB)
- ALICE::FZK::SE (v4.0.4) and ALICE::FZK::TAPE (v4.1.3) both still not running on the newest xrootd version (upgrade is planned with next full-time experiment representative on site)

- ALICE::FZK::Tape
  - using FRM (File Residency Manager) feature of xrootd
    - migrating to tape and staging from tape has been fixed.
  - file consistency check alien FC ⇔ FZK::Tape done
    - inconsistencies found by W. Park
    - New transfer on-going: about 70% done, so far all files correctly copied
- Internal monitoring : individual server monitoring page has been setup:
  - http://web-kit.gridka.de/monitoring/xrootd.php
  - number of active files, daemon status, internal reading/writing test and network throughputs, internal/external data access are displayed.
  - will be connected to the icinga alarm system

Various points of interest

- Resources at GridKa:
  - ALICE has requested (pledged number in WLCG REBUS) 4.01 PB of disk space in 2015
    - pledges will arrive April 2016
    - 2016 pledges will arrive end of 2016
    - WLCG has been informed.
- IPv6
  - GridKa internal tests ongoing (currently with ATLAS services)
  - GridKa maintains a IPv6 testbed in the context of the HEPiX-IPv6 Working Group.
  - xrootd software can in principle be operated in dual stack mode

more points of interest

change of local experiment support at GridKa

- WooJin Park: left to Korea (thanks a lot for his great work)
- Christopher Jung: interim experiment representative, in parallel to his LSDMA activities
- Max Fischer: will take over summer 2016

GridKa Downtimes

- tape backend: In order to migrate the core management services of the tape backend to different hardware a downtime (2 days) is needed. **Are there no-go dates for ALICE ? Current proposed date: April, 26 & 27**
- maintenance of cooling water pipes: May 9-11, 2016

Table of contents

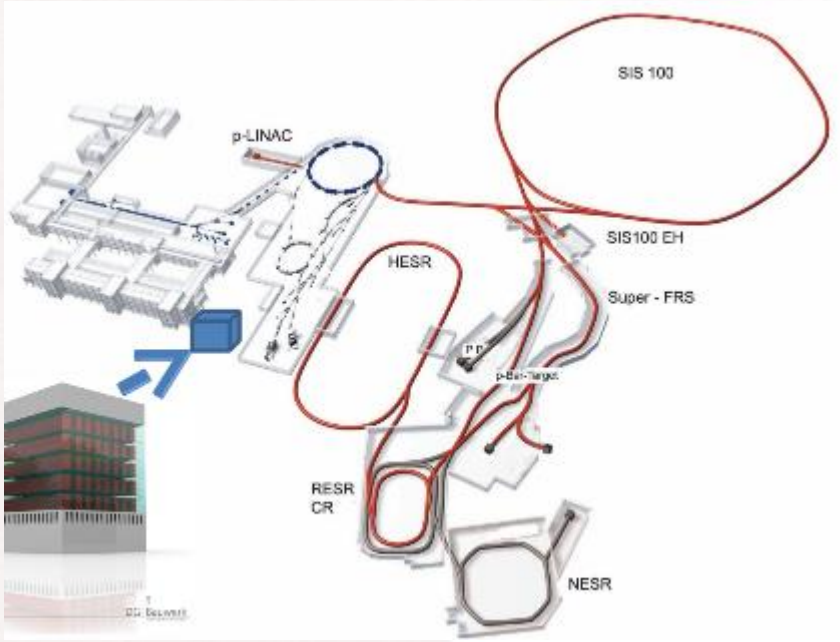
- Overview
- GridKa T1
- GSI T2
- UF
- Summary

GSI: a national Research Centre for heavy ion research
FAIR: Facility for Ion and Antiproton Research

GSI computing 2015

ALICE T2/NAF
HADES
LQCD (#1 in Nov‘ 14 Green 500)
~30000 cores
~ 25 PB lustre
~ 9 PB archive capacity

Green IT Cube Computing Centre
- Construction finished
- Inauguration Meeting 22.1.2016



FAIR computing at nominal operating conditions

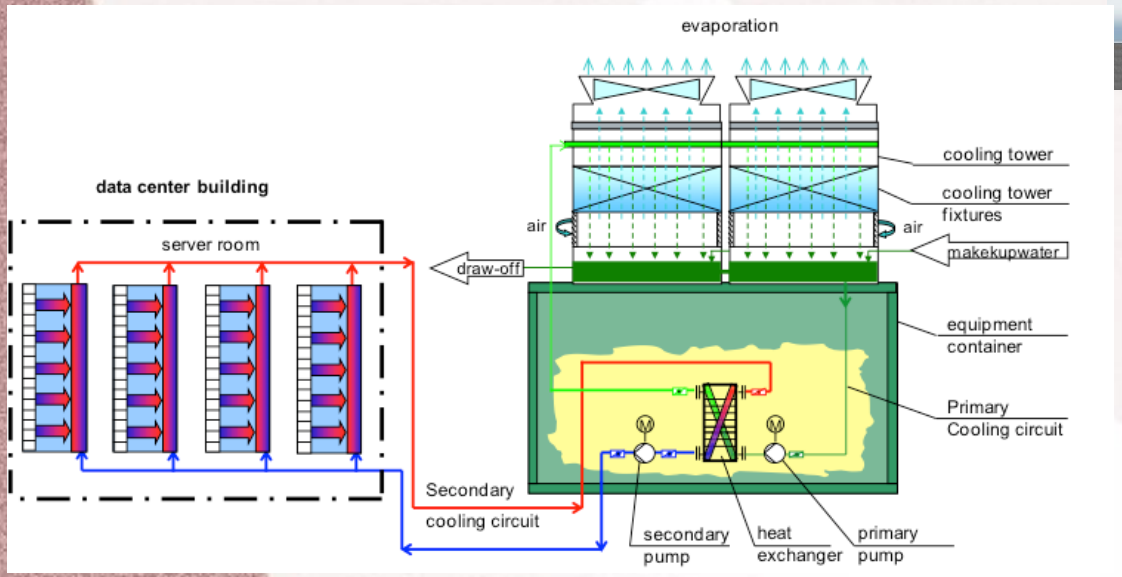
CBM
PANDA
NuSTAR
APPA
LQCD
300000 cores
40 PB disk/y
30 PB tape/y

View of construction site

- Open source and community software
- budget commodity hardware
- support of different communities
- manpower scarce

FAIR Data Center

A common data center
for FAIR (Green IT Cube)



6 floors, 4.645 sqm
room for 768 19” racks (2,2m)
4 MW cooling (baseline)
Max cooling power 12 MW
Fully redundant (N+1)
PUE <1.07

Green IT Cube

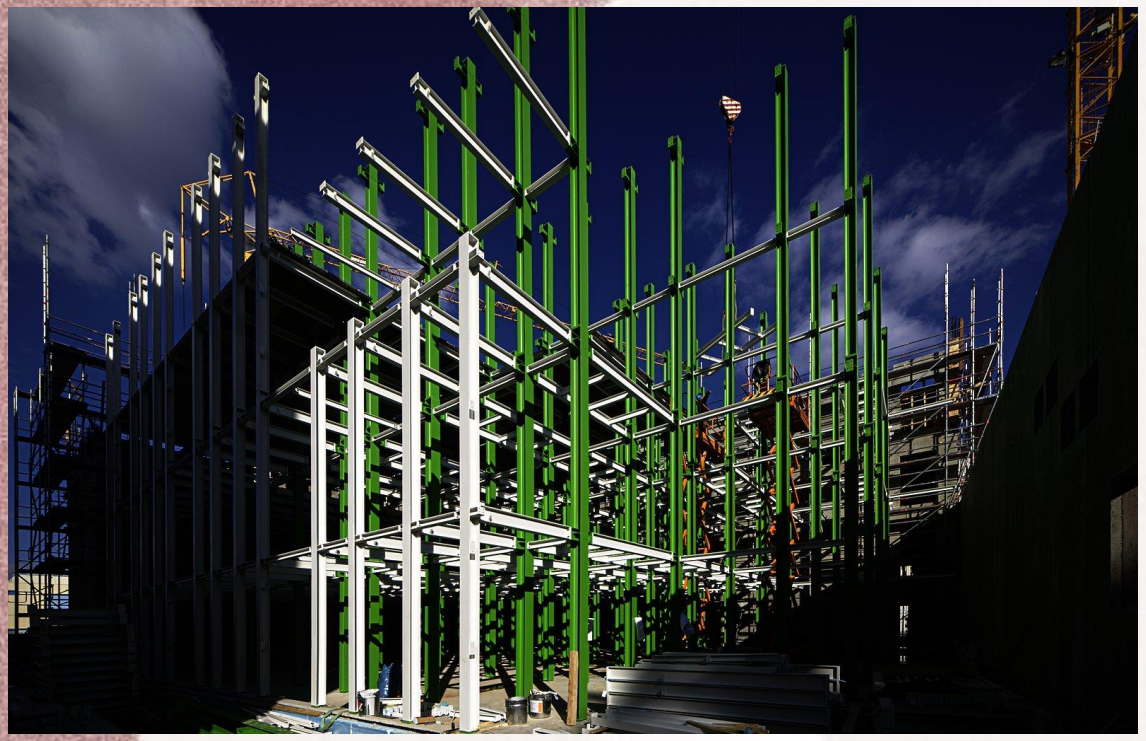
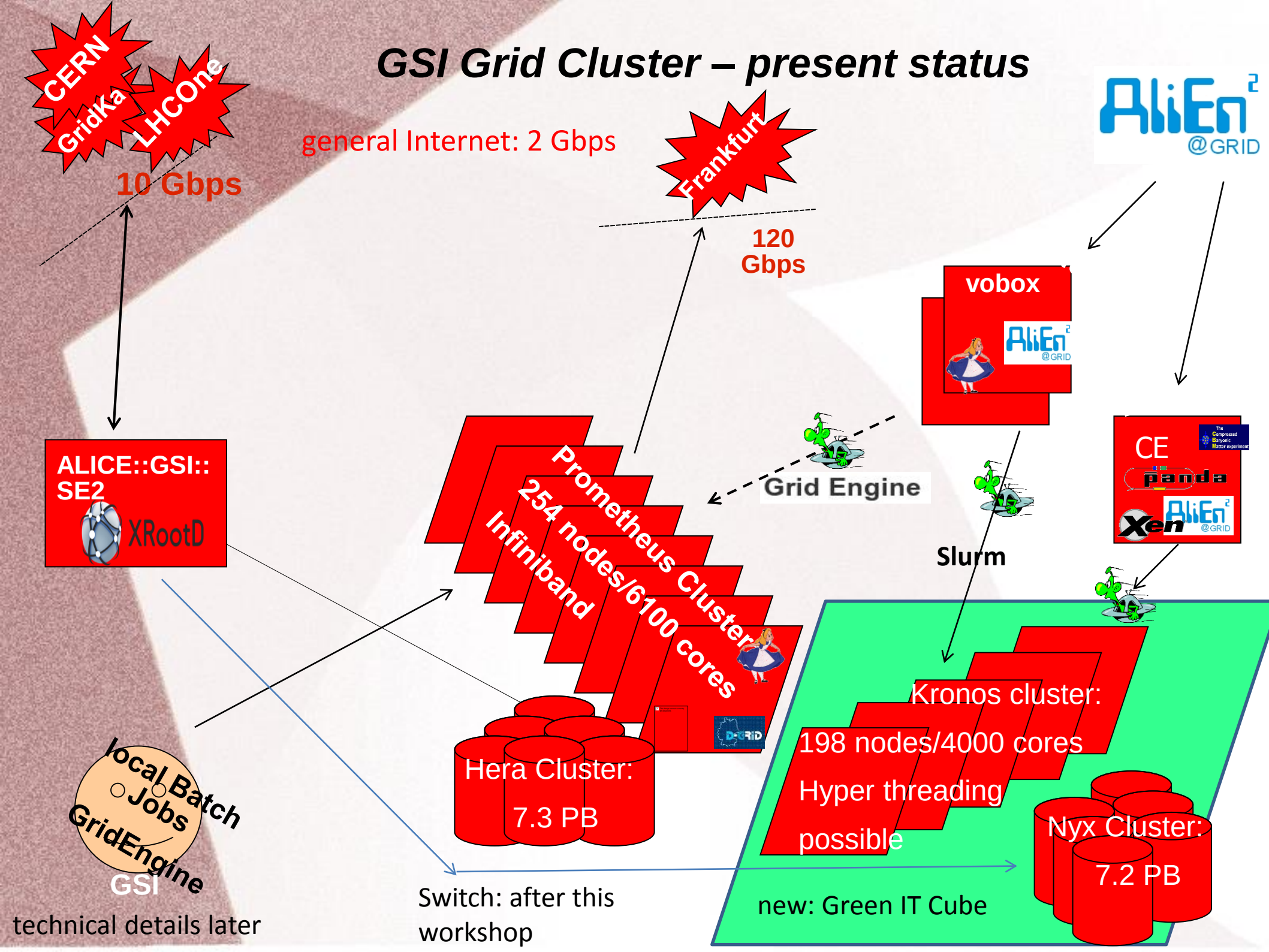


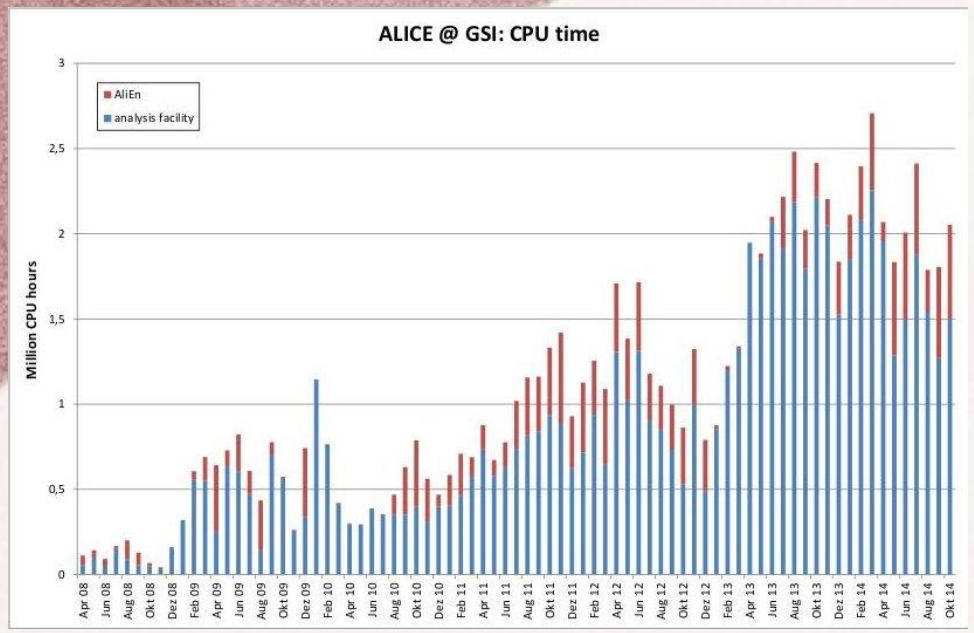
Foto: C. Wolbert

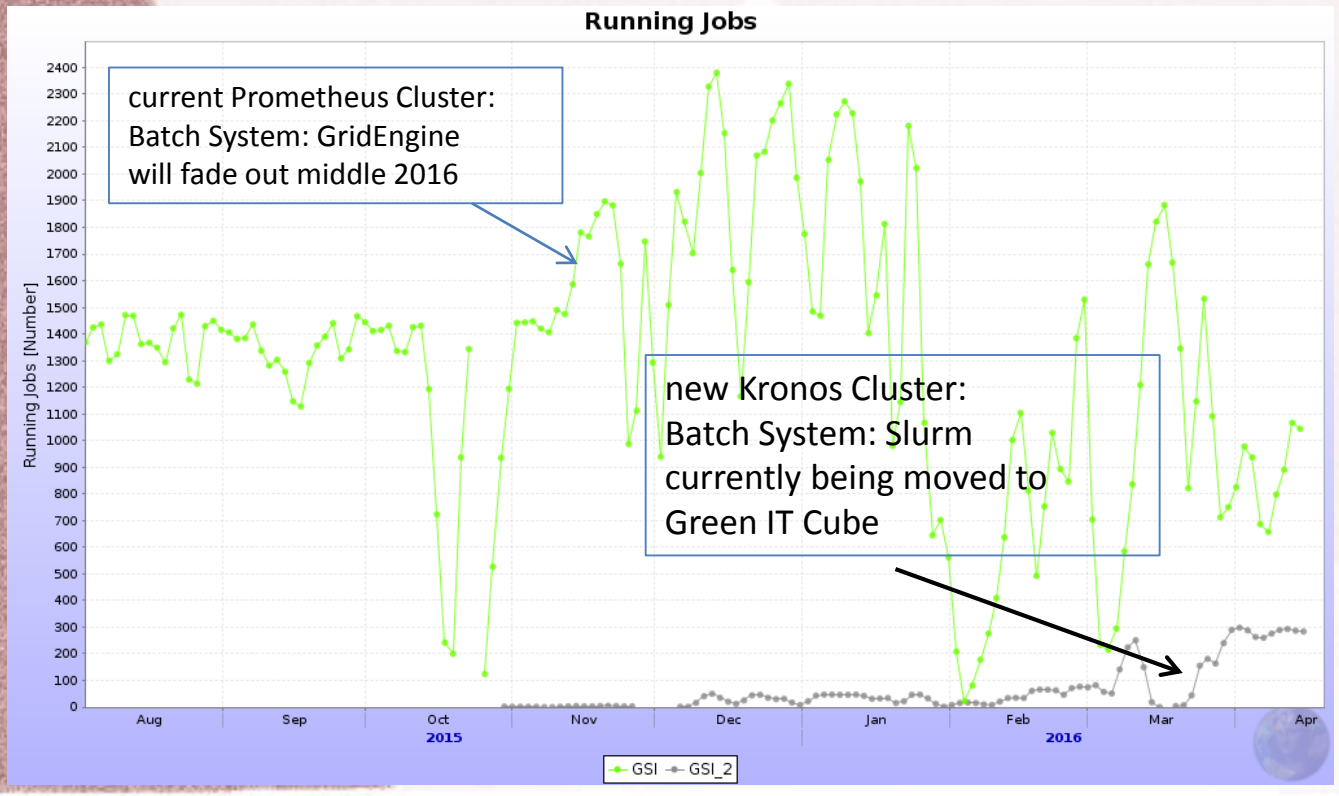
- Construction finished (started in December 2014)
- Inauguration ceremony January 22, 2016
- move of GSI compute and storage clusters to GC up to Easter 2016

GSI Grid Cluster – present status



GSI Darmstadt: 1/3 Grid, 2/3 NAF

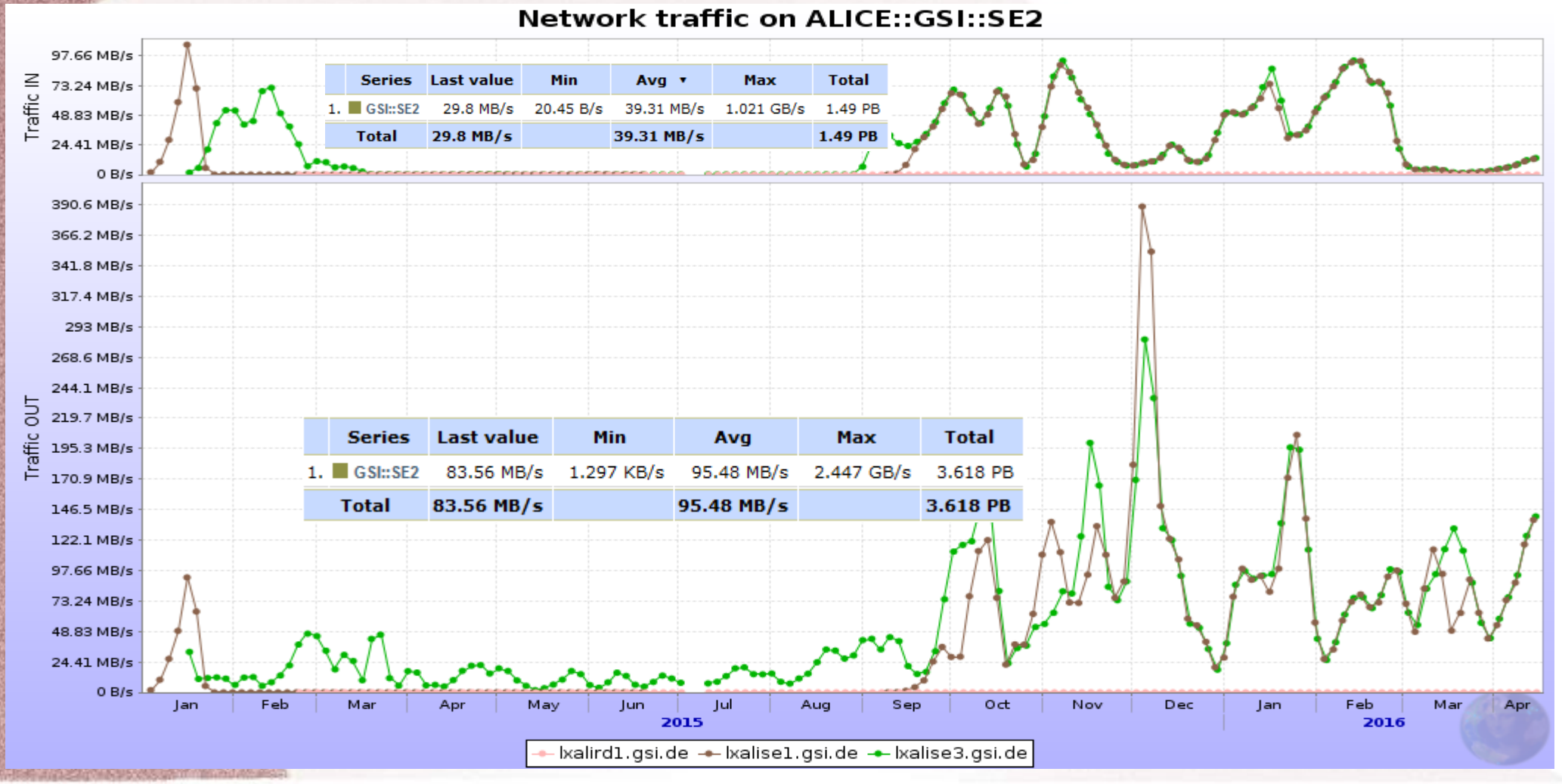




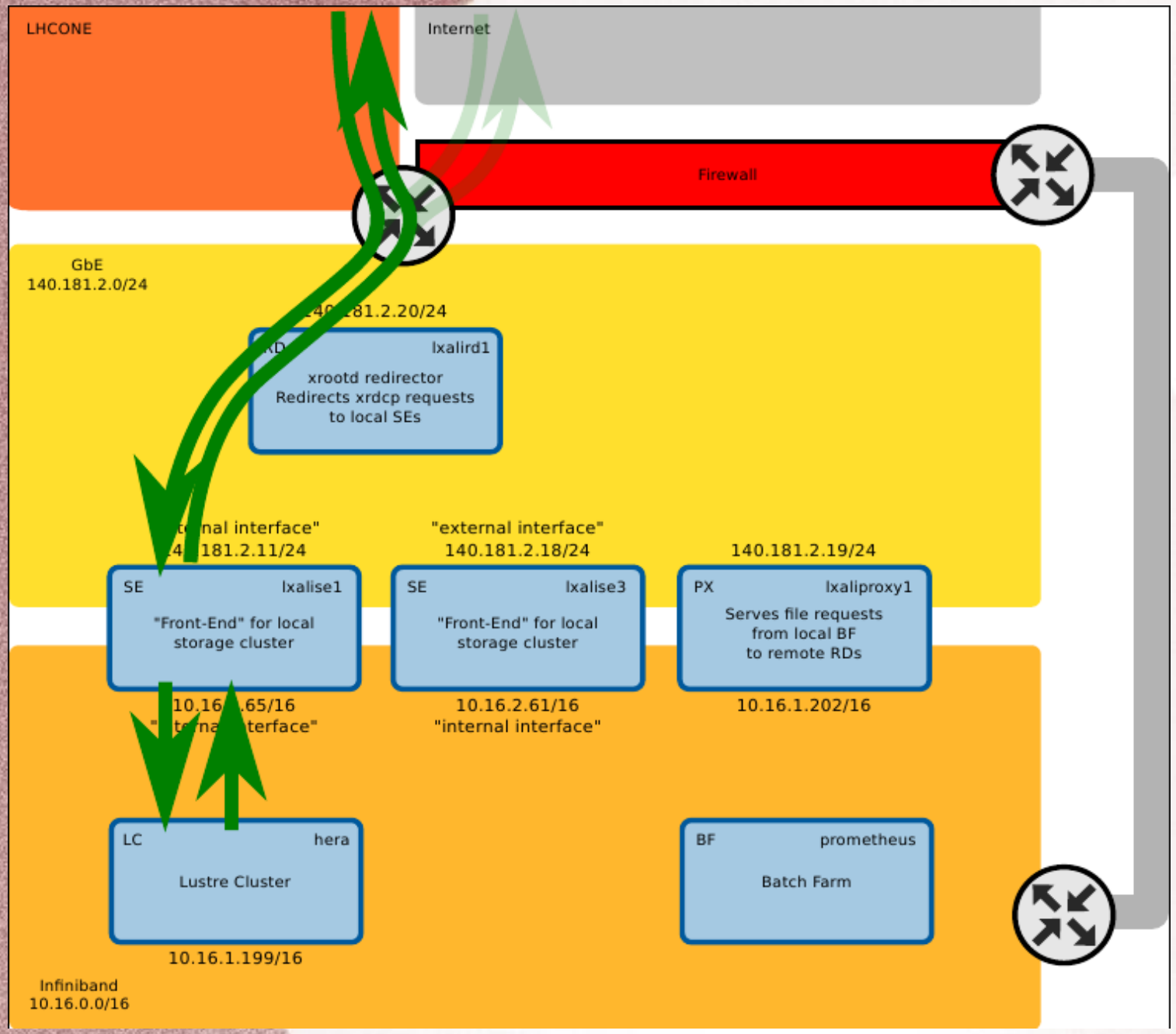
- Current state:
  - Prometheus (Grid Engine)
  - SE on „old“ Lustre /hera
- Future state:
  - Kronos (Slurm)
  - SE on „new“ Lustre /nyx
- ALICE T2 infrastructure is currently being moved into new cluster in GC
- average: 1400 T2 Jobs
  - max (3300 jobs)
- about 1/3 Grid (red), 2/3 NAF (blue)
- Intensive usage of the local GSI cluster through ALICE NAF (ca. 30% of the cluster capacity)

GSI Storage Element

- ALICE::GSI::SE2 works well and is used extensively
- in order to further increase the reliability several improvements have been introduced
- in 2015 about 1.5 PB have been written, 3.5 PB read
  - (reading increased by 0.5 PB)
- SE is currently being moved to Green IT Cube (move of data servers, file system will be based on new Lustre cluster (content needs to be migrated), new proxy server)



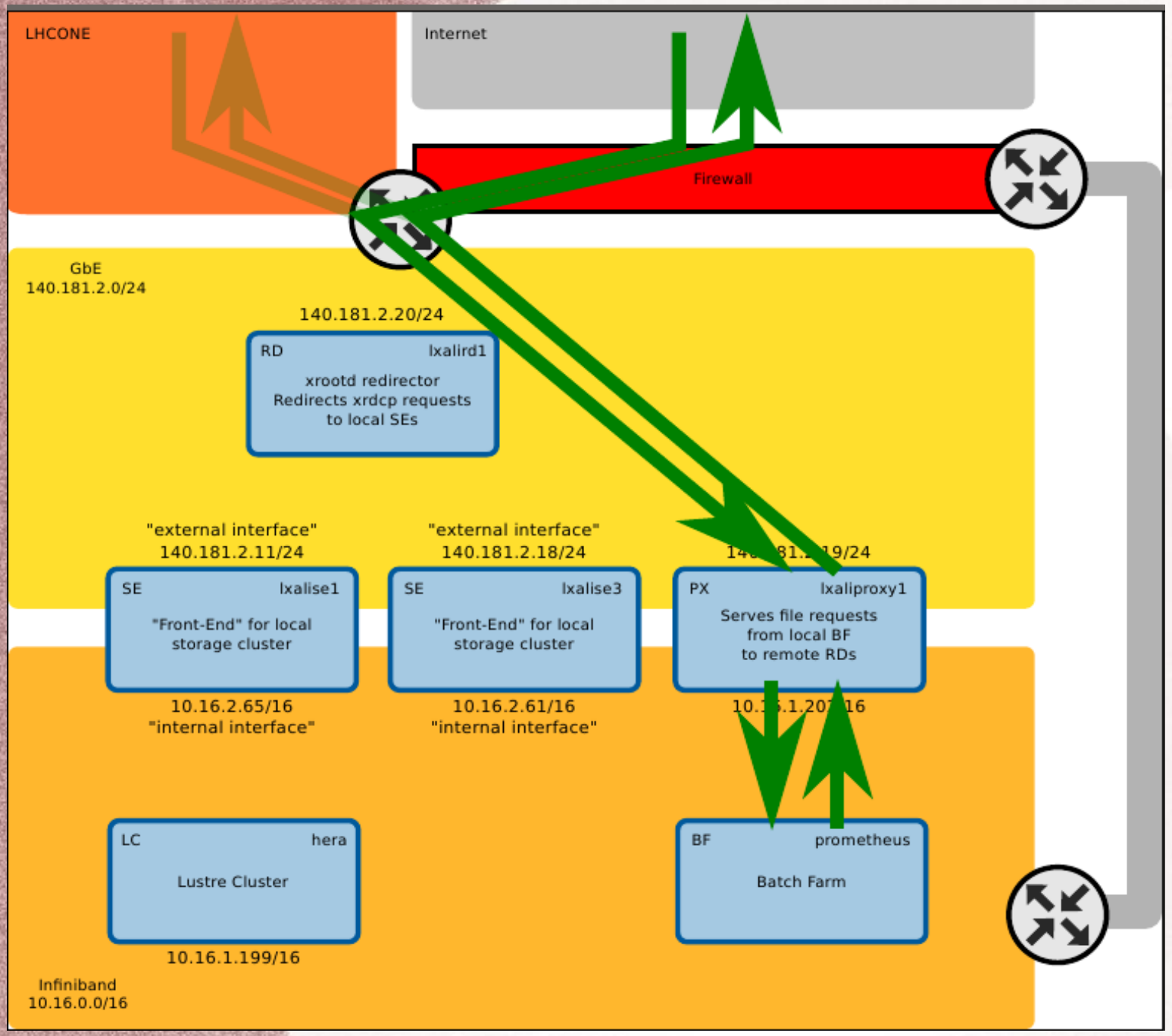
ALICE T2 – storage setup



xrootd redirector

- points external clients to external SE interfaces
- points internal clients to internal SE interfaces

ALICE T2 – storage setup



xrootd proxy (upper limit seems to be around 2000 jobs) tunnels traffic from/to clients within the HPC environment to/from Storage Elements outside GSI

xrootd v4 client plugin is available and would be able to point clients directly to Proxy URL.

➔ application software does not need to be proxy-aware

same plugin could also provide direct access to Lustre file system and does not have to read via xrootd data server

ALICE T2: improved infrastructure

- all ALICE T2 services have been implemented with
  - auto configuration (Chef cookbooks to deploy new machines and to keep existing ones consistent)
  - monitoring and auto restart as well as information service (Monit)
  - monitoring (Ganglia, MonaLisa)
- GridEngine and Slurm interfaces of AliEn have been patched: getNumberRunning, getNumberQueueing now provides correct number of JobAgents

```
# Make a symlink from /usr/lib/x86_64-linux-gnu/libcrypto.so.1.0.0 (actual file) to /usr/lib64/libcrypto.so (link)
link "/usr/lib64/libcrypto.so" do
  to "/usr/lib/x86_64-linux-gnu/libcrypto.so.1.0.0"
  link_type :symbolic
end

# xrootd manual package "installation"
remote_directory "xrootd" do
  path node["xrootd"]["installdir"]  # Path on target machine
  mode "0755"
  owner node["users"]["storage-user"]["user"]
  group node["users"]["storage-user"]["group"]
  files_owner node["users"]["storage-user"]["user"]
  files_group node["users"]["storage-user"]["group"]
  files_mode "0755"
  action :create
end
```

GSI: next activities

- Continue the move of the ALICE T2 infrastructure to new cluster and new environment (Green IT Cube). After the move is finished a storage downtime will be announced in order to do final data migration. After that (a few weeks from now), production will be ramped up again.
- new manpower will be hired
  - ALICE T2 will profit from this, too
- plans for IPv6:
  - first IT services at GSI support IPv6
  - campus-wide use is not yet planned

Table of contents

- Overview
- GridKa T1
- GSI T2
- UF
- Summary

UF

UF is an ALICE Grid research site at University of Frankfurt.

It is managed by Andres Gomez (see his talk on Wednesday) and is used to investigate

“Intrusion Prevention and Detection in Grid Computing”

The new security framework contains among others

- execution of job payloads in a sandboxed context
- process behavior monitoring to detect intrusions

Table of contents

- Overview
- GridKa T1
- GSI T2
- Summary

•*ALICE Computing Germany*

		Tier 1 (GridKa)	Tier 2 (GSI)	NAF (GSI)
2016	CPU(1)	42	17	34
	Disk (PB)	4,6	1,9	3,8
	Tape (PB)	5,0	-	-
2017(*)	CPU(1)	55	20	40
	Disk (PB)	5,5	2,3	4,6
	Tape (PB)	7,1	-	-
2018(*)	CPU(1)	71	23	46
	Disk (PB)	7,0	2,9	5,8
	Tape (PB)	10,0	-	-

(1) CPU in kHEPSPEC06
(*) 2017 and 2018 requests to be approved by CERN RRB
NAF = National Analysis Facility

Summary & Outlook

- German sites provide a valuable contribution to ALICE Grid
  - Thanks to the centres and to the local teams
- new developments are on the way
- Pledges can be fullfilled
- FAIR will play an increasing role (funding, network architecture, software development and more ...)
- annual T1T2 meetings started 2012 at KIT. We propose to have the 5th anniversary meeting jointly hosted by Strasbourg & KIT in 2017