Grid Operations in Germany

T1-T2 workshop 2017
Straßburg

Kilian Schwarz

Sören Fleischer

Raffaele Grosso

Jan Knedlik

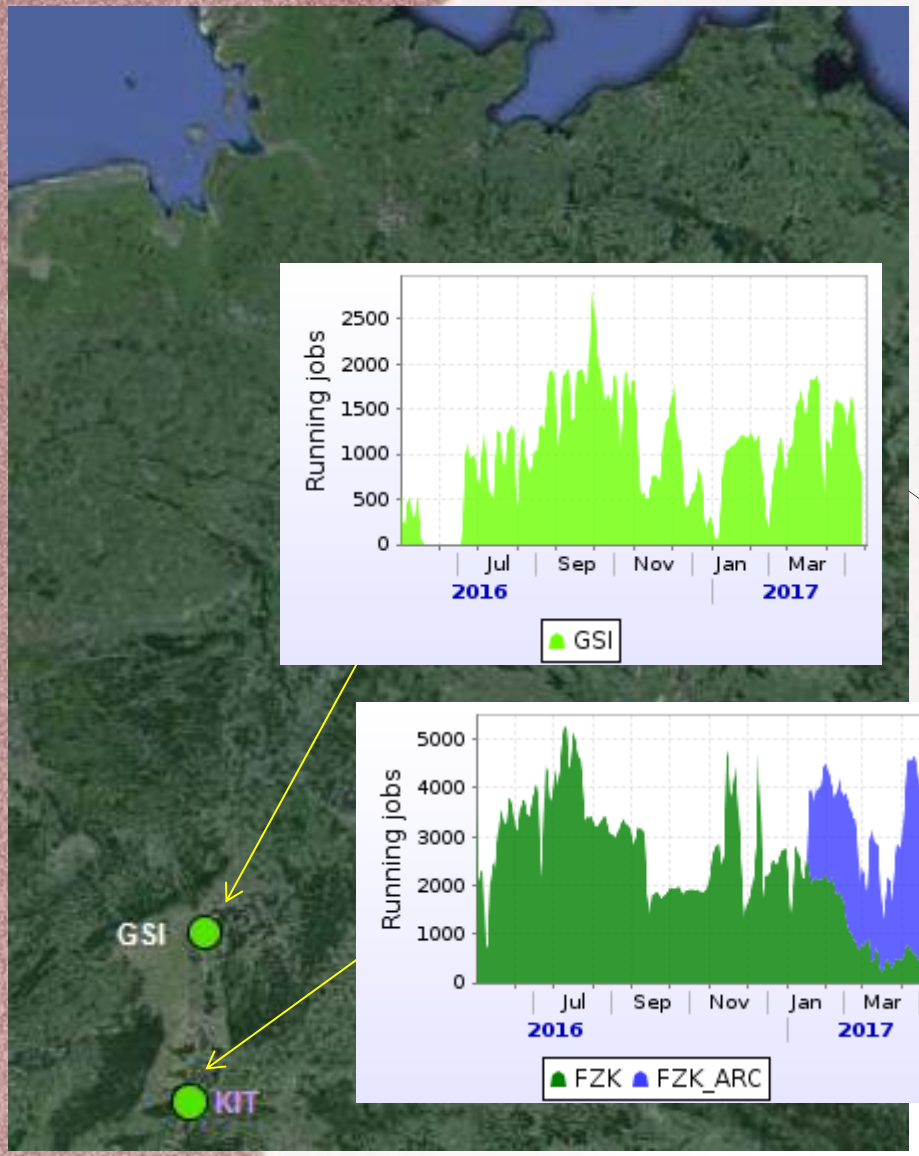
Max Fischer

Paul Kramp

Table of contents

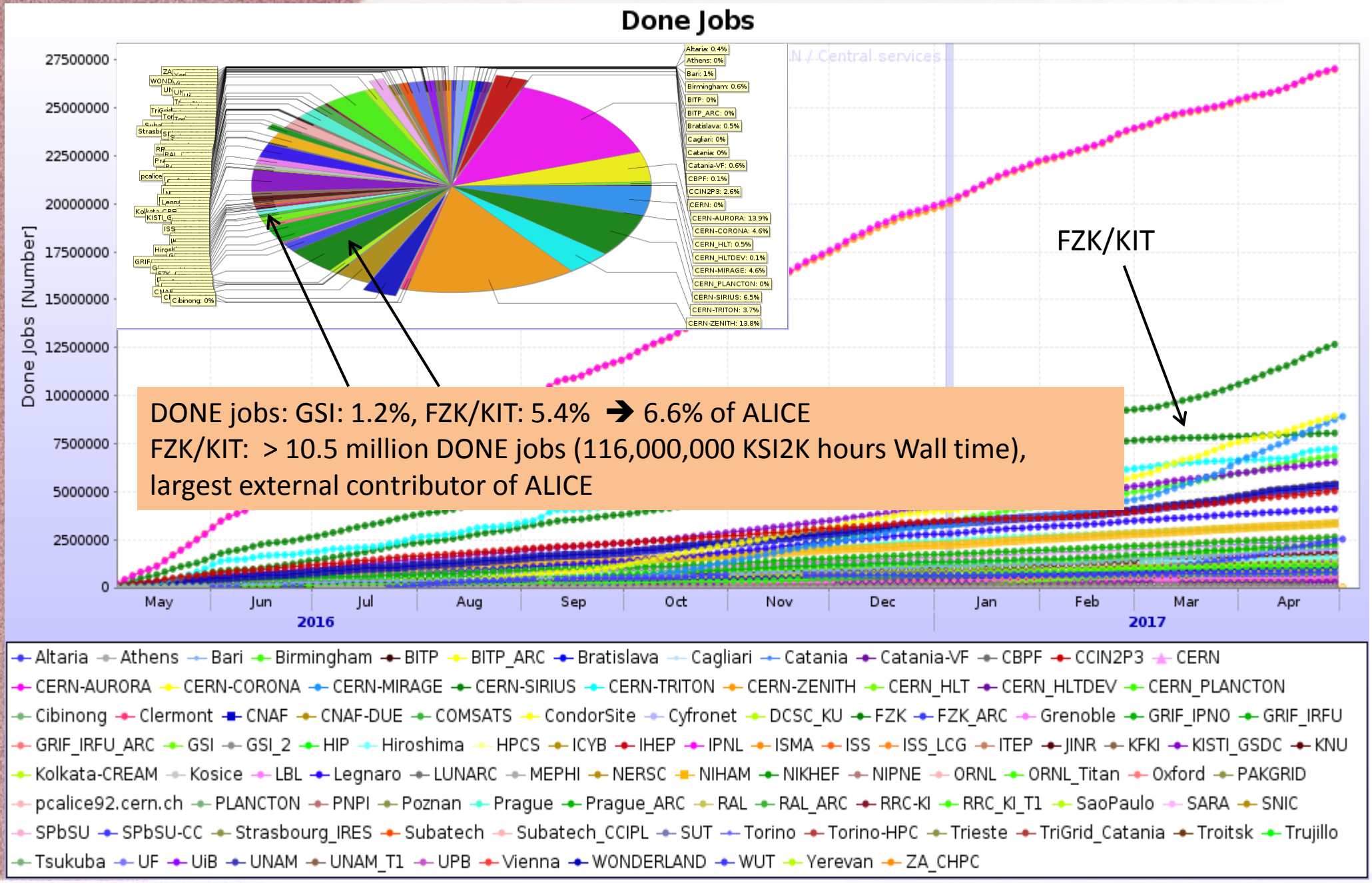
Table of contents

Map of German Grid sites



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

Job contribution (last year)



DONE jobs: GSI: 1.2%, FZK/KIT: 5.4% ➔ 6.6% of ALICE

FZK/KIT: > 10.5 million DONE jobs (116,000,000 KSI2K hours Wall time), largest external contributor of ALICE

Storage contribution

AliEn name	Size	Used	Free	Usage	No. of files	Type	Size	Used	Free	Usage
ALICE::FZK::SE	4.5 PB	3.625 PB	895.7 TB	80.56%	81,741,328	FILE	7.563 PB	7.479 PB	86.3 TB	98.89%
ALICE::GSI::SE2	2.3 PB	1.514 PB	804.6 TB	65.84%	42,280,512	FILE	2.3 PB	1.345 PB	977.5 TB	58.5%
ALICE::FZK::TAPE	640 TB	5.033 PB	-	805.3%	3,392,308	FILE	640.4 TB	549.9 TB	90.48 TB	85.87%

Total size:

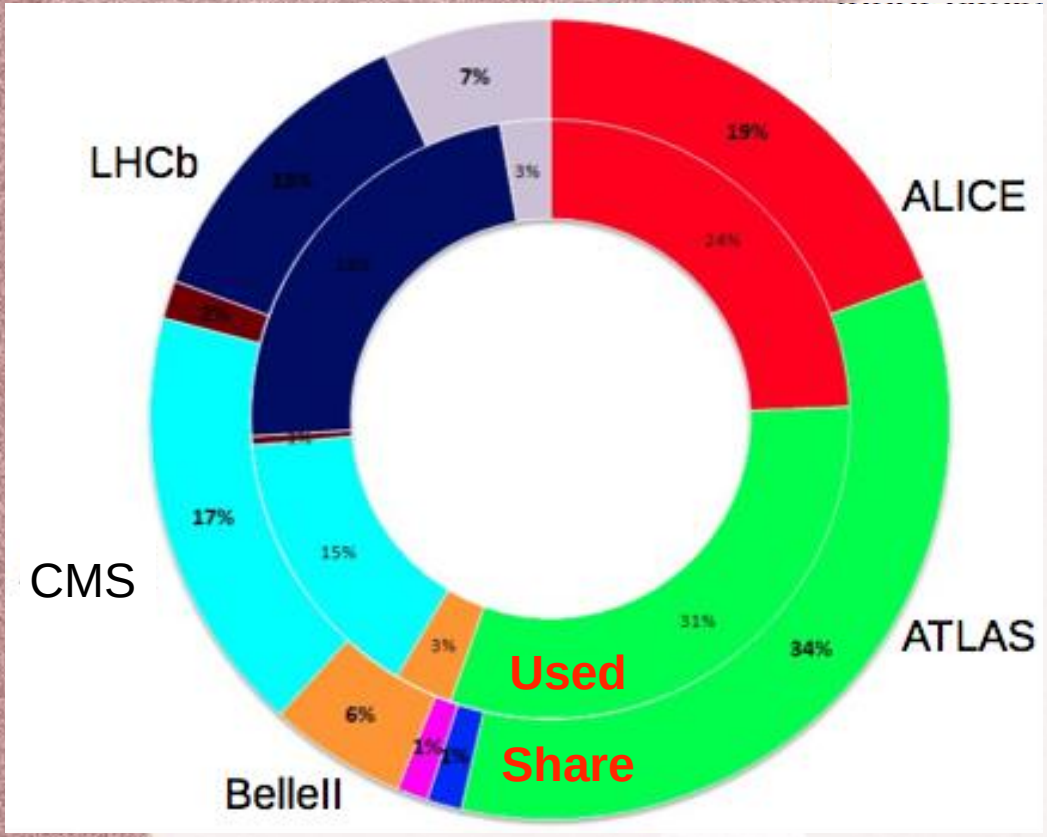
- GridKa: 4.5 PB Disk SE including 0.6 PB tape buffer
- GSI: 2.3 PB Disk SE

- FZK: xrootd still shows a wrong storage capacity
- GSI: via xrootd plugin – xrootd now displays Lustre quota for ALICE SE

- 6.8 PB disk based SE (14% of ALICE)
- 5.25 PB tape capacity
- 640 TB disk buffer with Tape backend

Table of contents

GridKa Usage

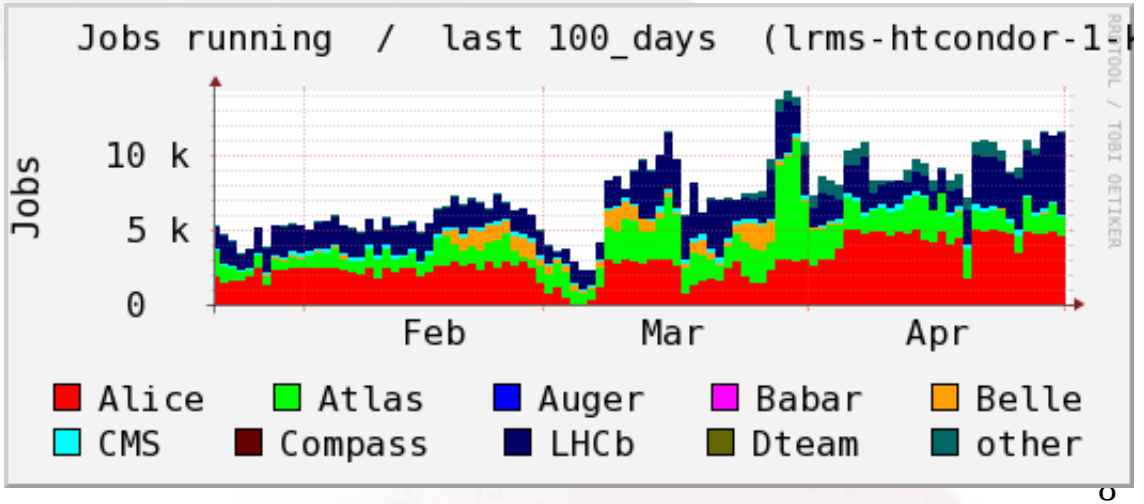


Usage May-Oct 2016

- Support of all 4 major LHC collaborations
- Biggest nominal share for ALICE and ATLAS
- ALICE and LHCb usage above nominal share
- all LHC experiments together use > 90%

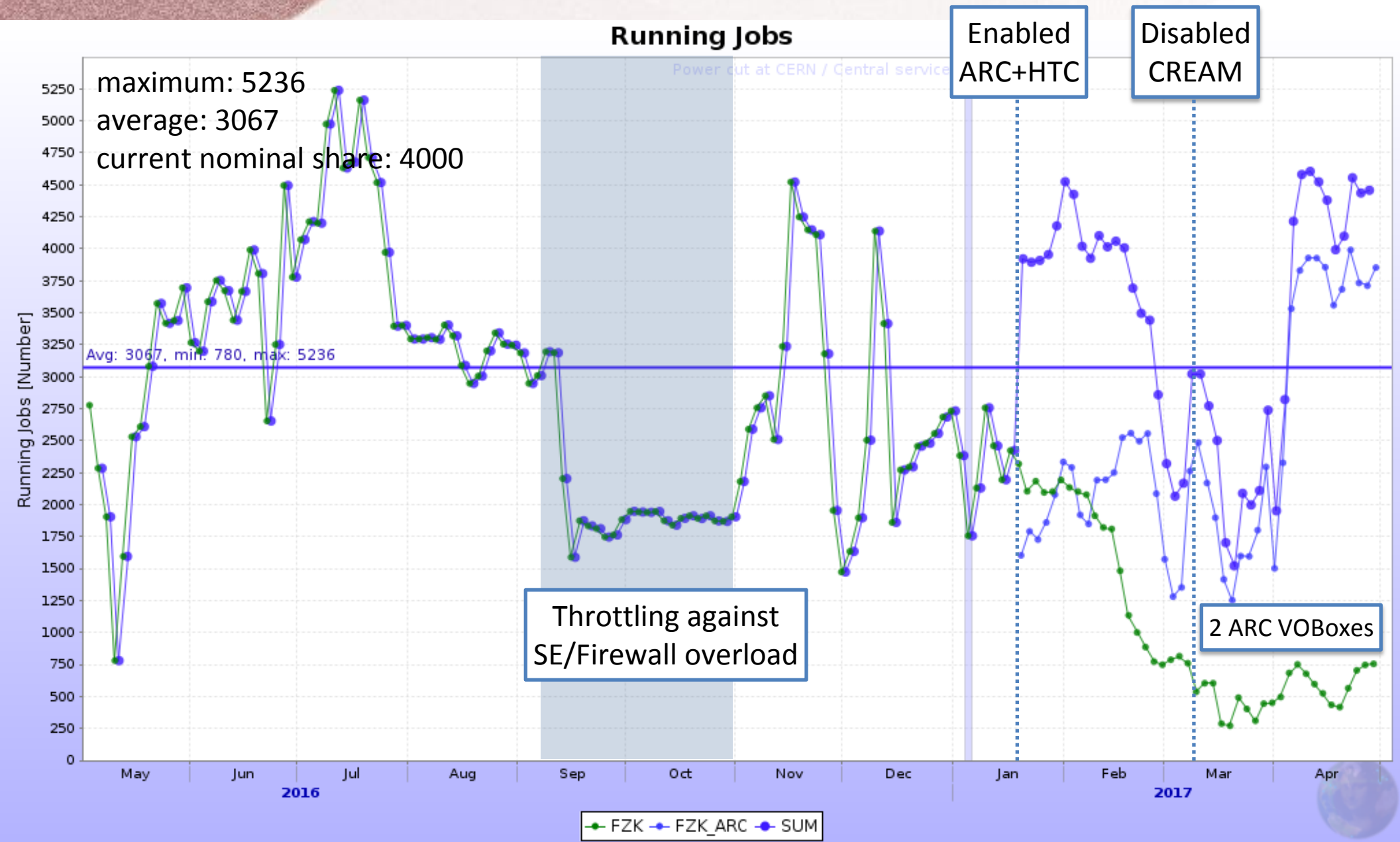
Batch System Migration

- Migrated from CREAM-CE+UGE to ARC-CE+HTCondor in Feb
- Overall satisfied, no major operational problems
- Stability/compatibility issues with ARC-CE, work is ongoing



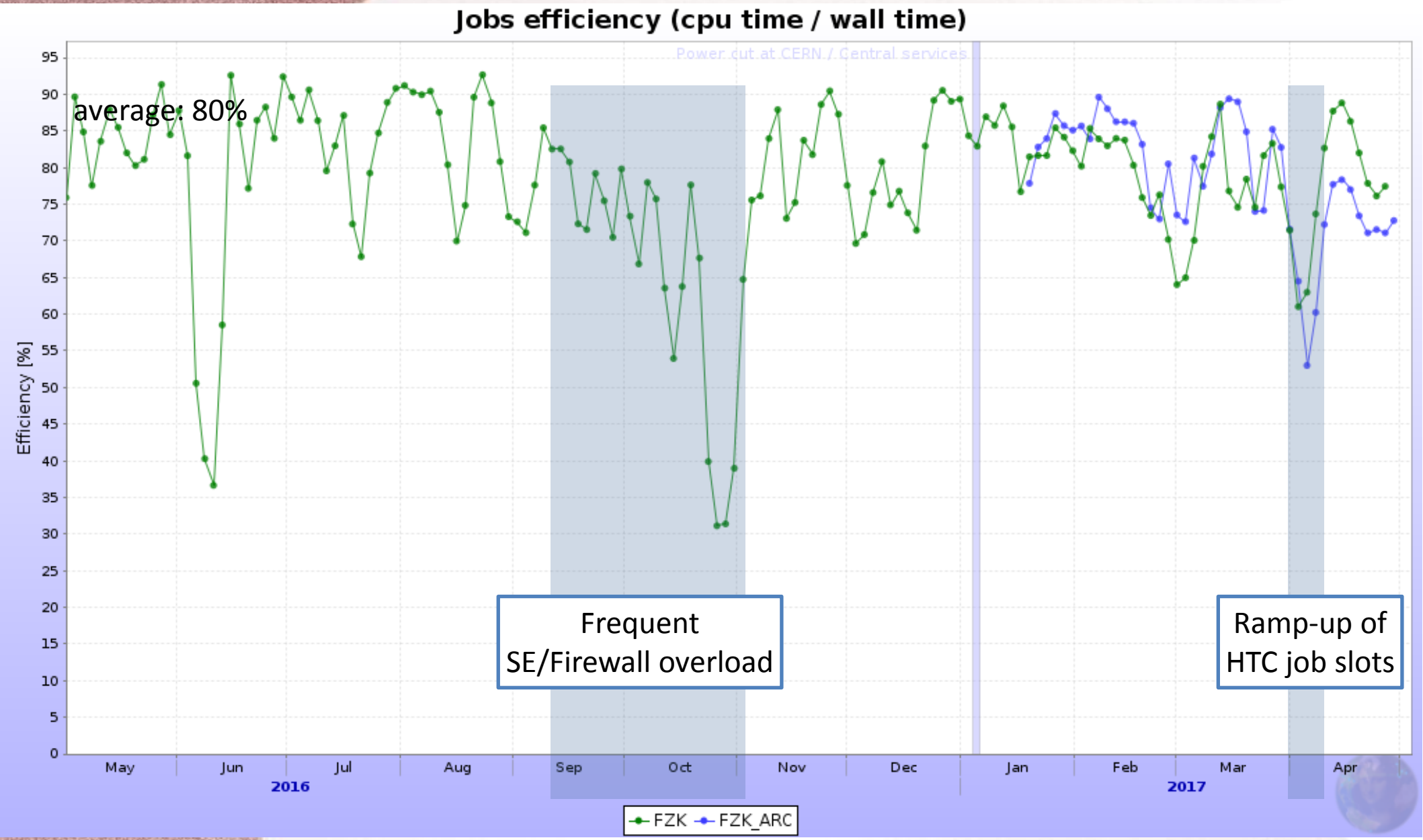
Jobs at GridKa

(over last year)



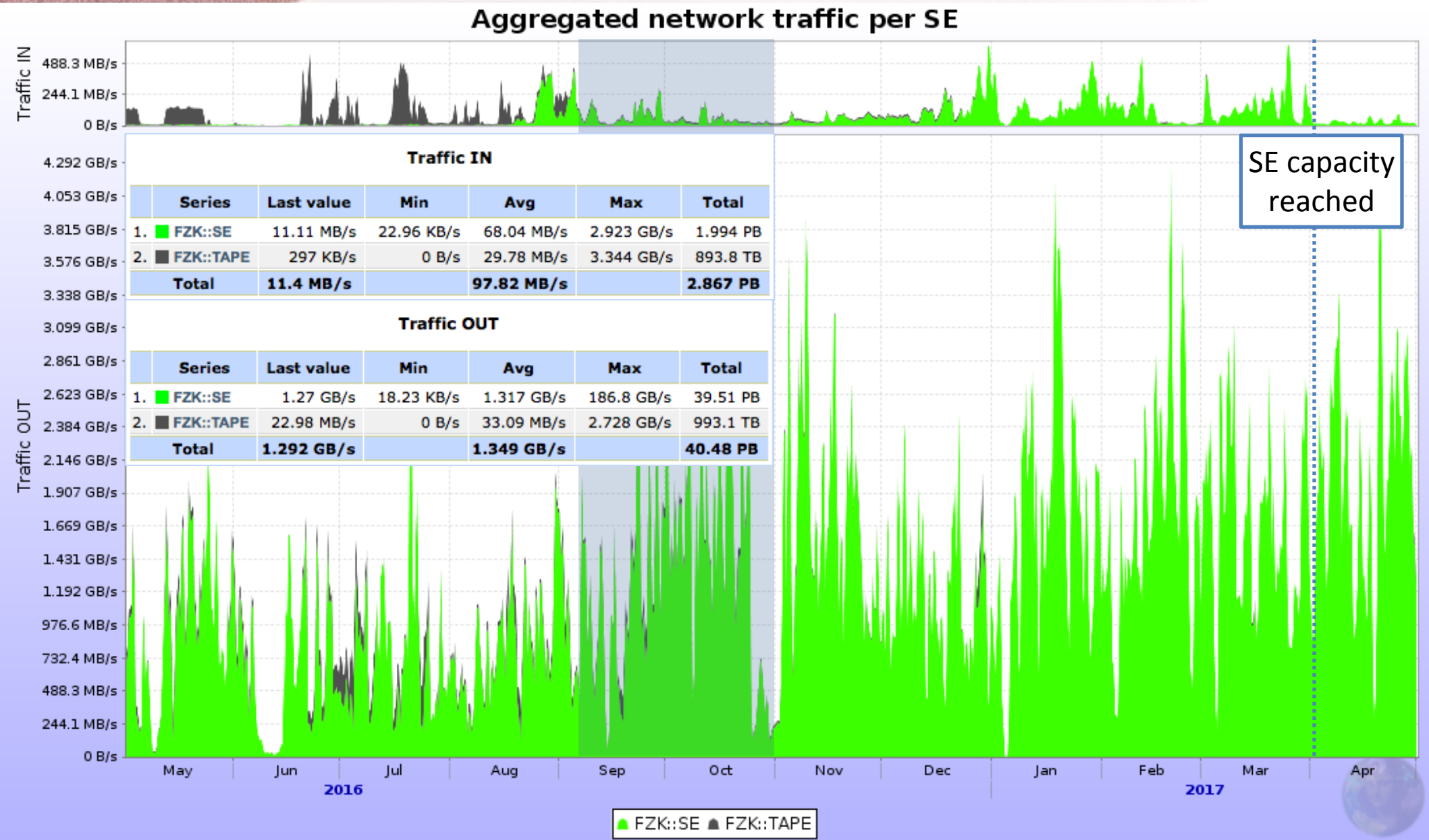
ALICE Job Efficiency@GridKa

Grid: 82%



XRootD usage@GridKa

- Heavy increase in disk SE usage since Aug 2016
- Tape performance degraded by congestion from VOs
- Disk pledges 2017 available within 1 month from now



XRootD@GridKa

In General

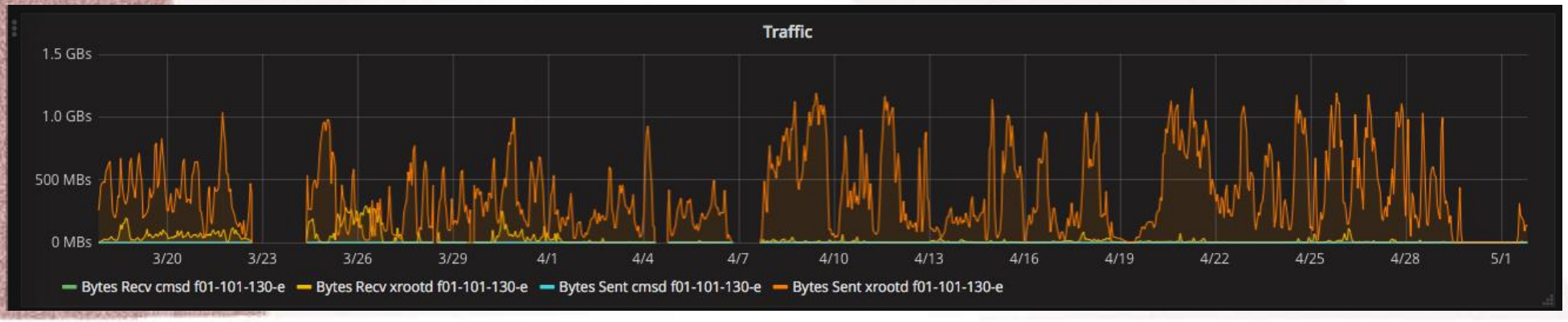
- Overall satisfied with xrootd, much less issues than VOs with other middleware
- Major hardware upgrades soon, software upgrades skipped if not critical
- Reduced ALICE dependency stack for compatibility with puppet and multi-VO
  - Replicated auth and extracted monitoring, removed configuration
  - Splitting monitoring stream to ALICE MonALISA and GridKa Grafana

ALICE::FZK::SE:

- Using multiple 2nd level redirectors after stability issues for *manager* on *server*
- Replacing entire storage system, new hardware operational since last week
  - Migrate data from existing 4.5PB to new 5.5 PB on single file system

ALICE::FZK::TAPE:

- Hit the Pledge end of 2016
- Adoption of HPSS storage system for production planned this year
- Existing tape library capacity being expanded to satisfy pledges



GridKa AOB

- **Funding and 2017 resource growth**
  - GridKa funding secured for next years
  - Disk: 5250TB -> 6150TB (includes tape buffer)
  - Tape: 5250TB -> 7725TB

- **IPv6 Adoption at GridKa**
  - Moving first services to IPv6, including new ALICE::FZK::SE servers
  - Building IPv6 firewall rules from scratch

- **Storage services**
  - Consolidating xrootd setup/configuration
  - Focus on backend performance and stability
  - No immediate plans to move to EOS
  - issue: 0 Byte files (should never happen on an SE <LB>)
    - 15829 / 68062075 files

- **ALICE representative**
  - Christopher Jung has left KIT and HEP
  - Max Fischer taken over since Jul 2017

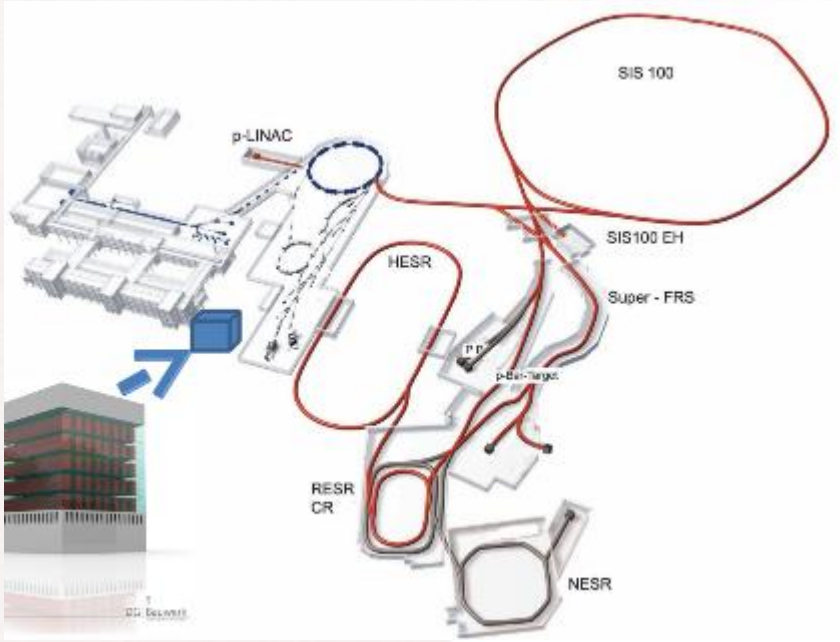
Table of contents

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

GSI computing 2016

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



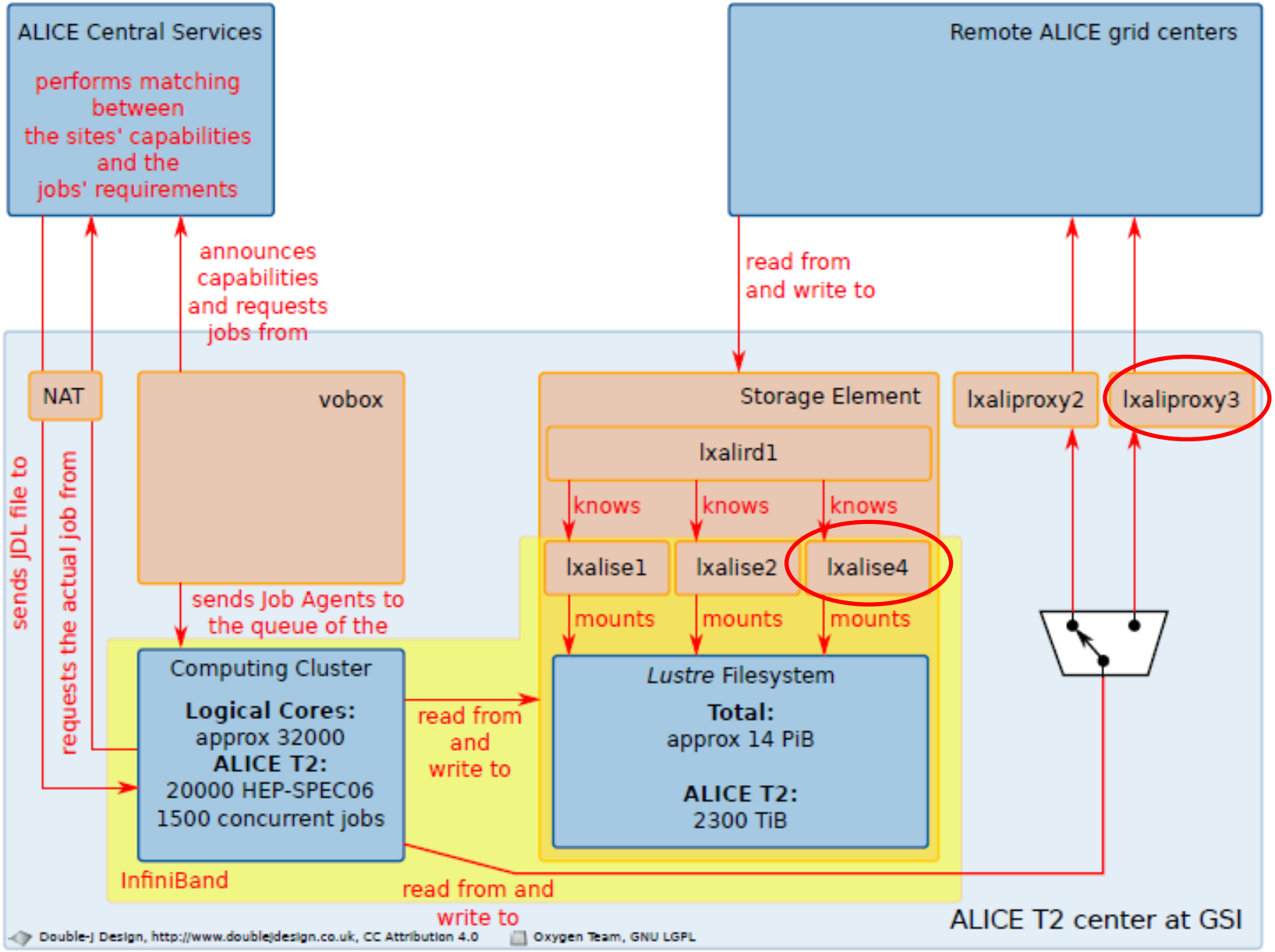
FAIR computing at nominal operating conditions

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

View of construction site

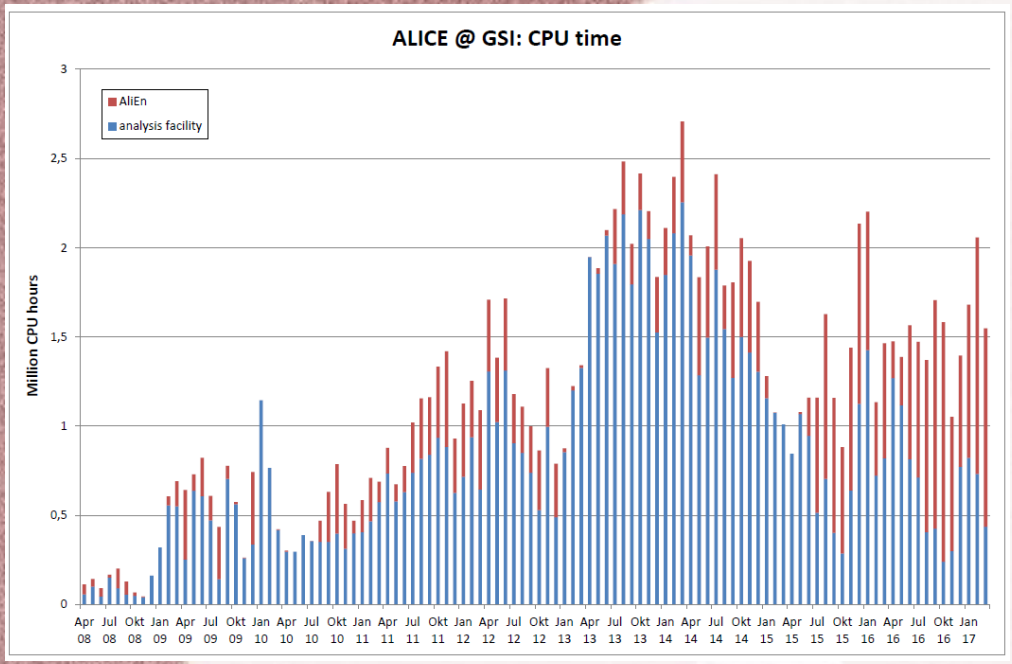
- Open source and community software
- budget commodity hardware
- support of different communities
- manpower scarce

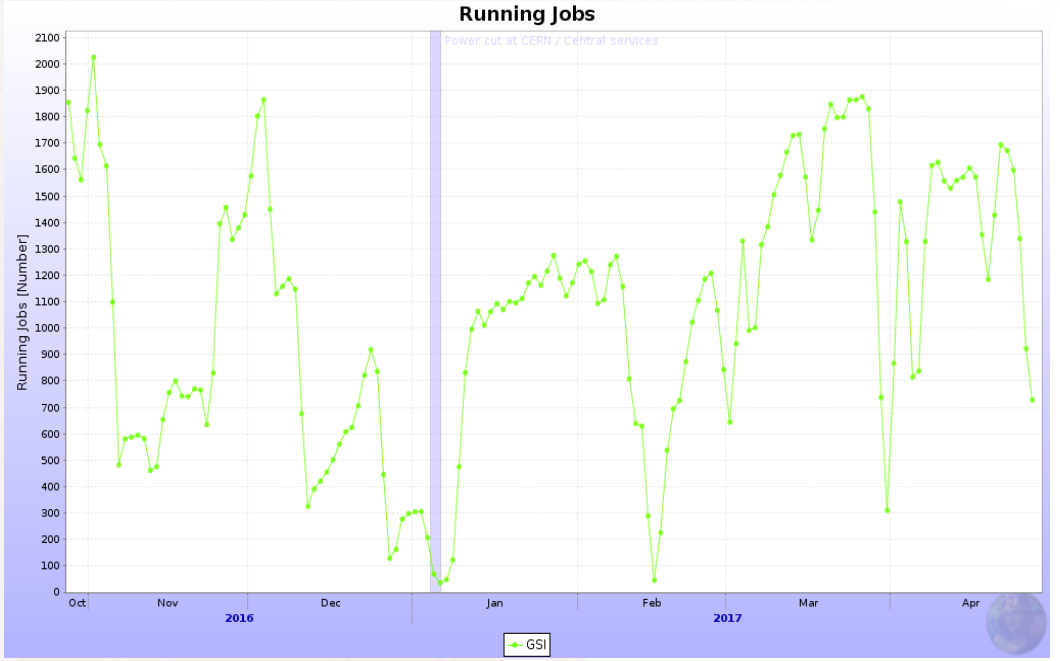
ALICE T2 Centre at GSI



redundant proxy servers

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



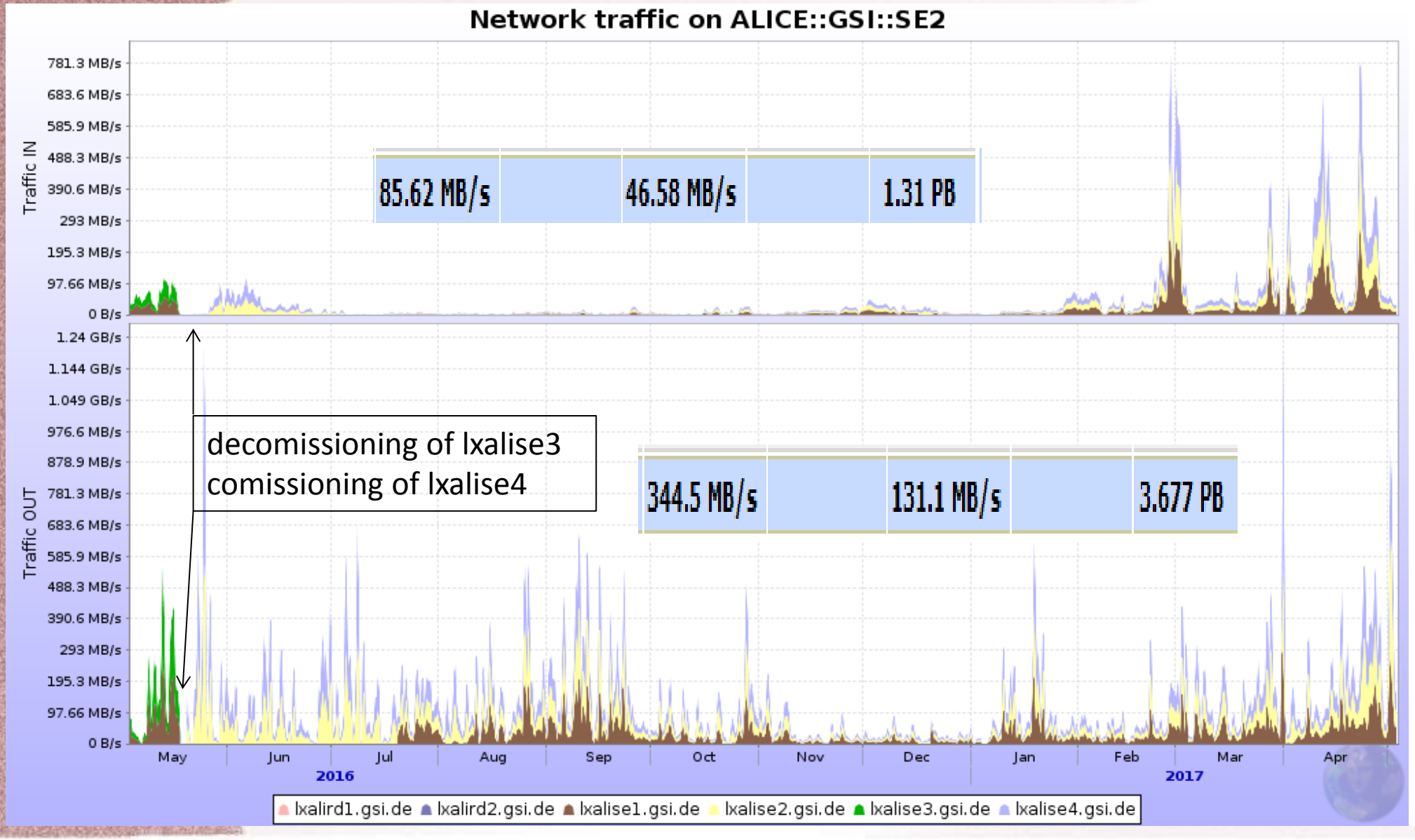


Green IT Cube

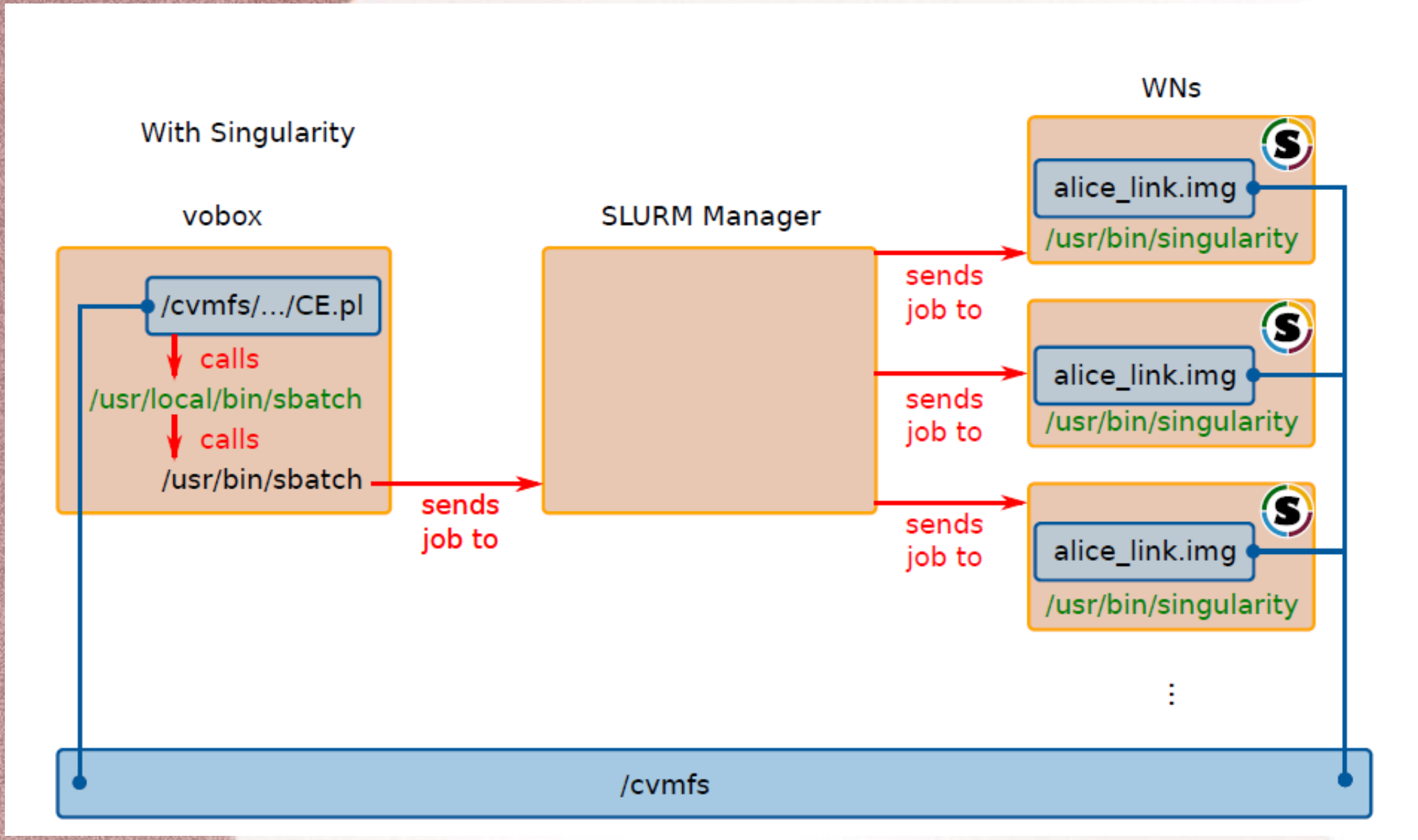
- T2: 7% of ALICE T2 (2017: CPU: 20kHS06 (5%, pledged), 26kHS06 (7%, delivered), disk: 2.3 PB)
  - about 1/3 Grid (red), 2/3 NAF (blue)
- on average: 1100 jobs
- pledges 2016/2017 fulfilled
- fulfilling pledges 2018/2019 planned

GSI Storage Element

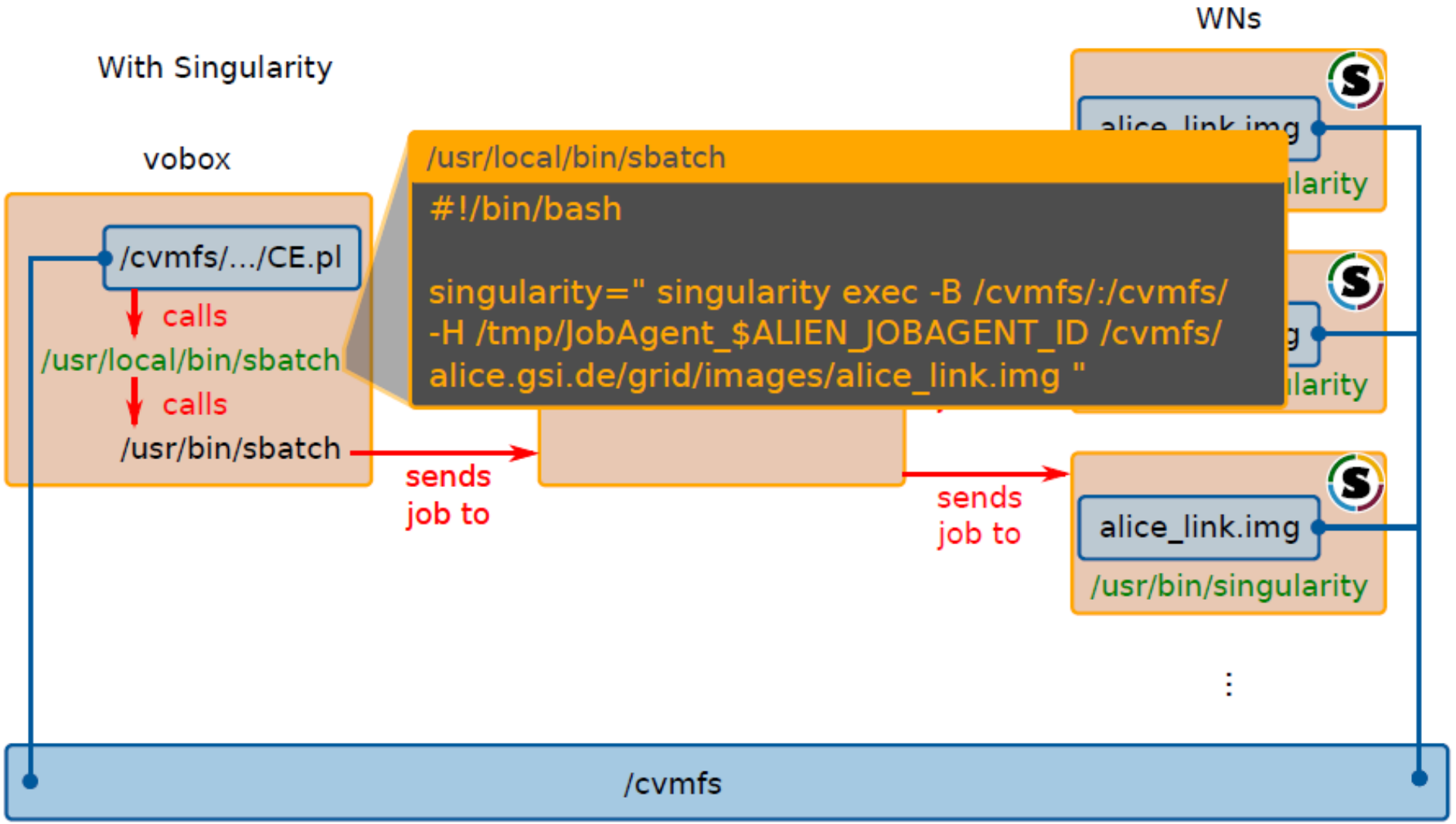
- *ALICE::GSI::SE2 works well and is used extensively*
- *in 2016 about 1.3 PB have been written, 3.6 PB read*
  - *(significant increase in writing since 2017)*



GSI ALICE T2 – Singularity

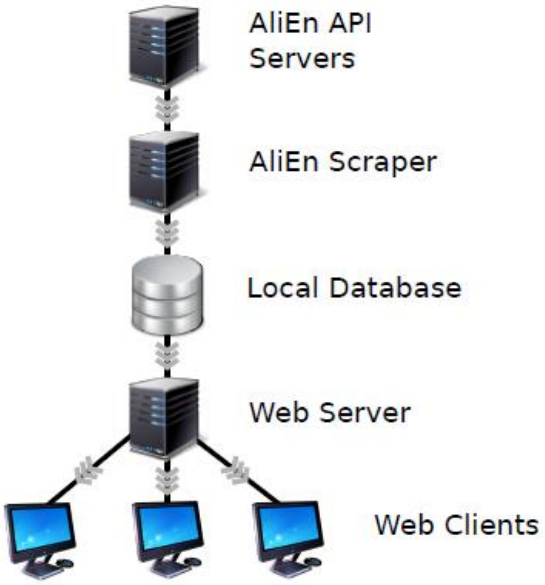


GSI ALICE T2 – Singularity



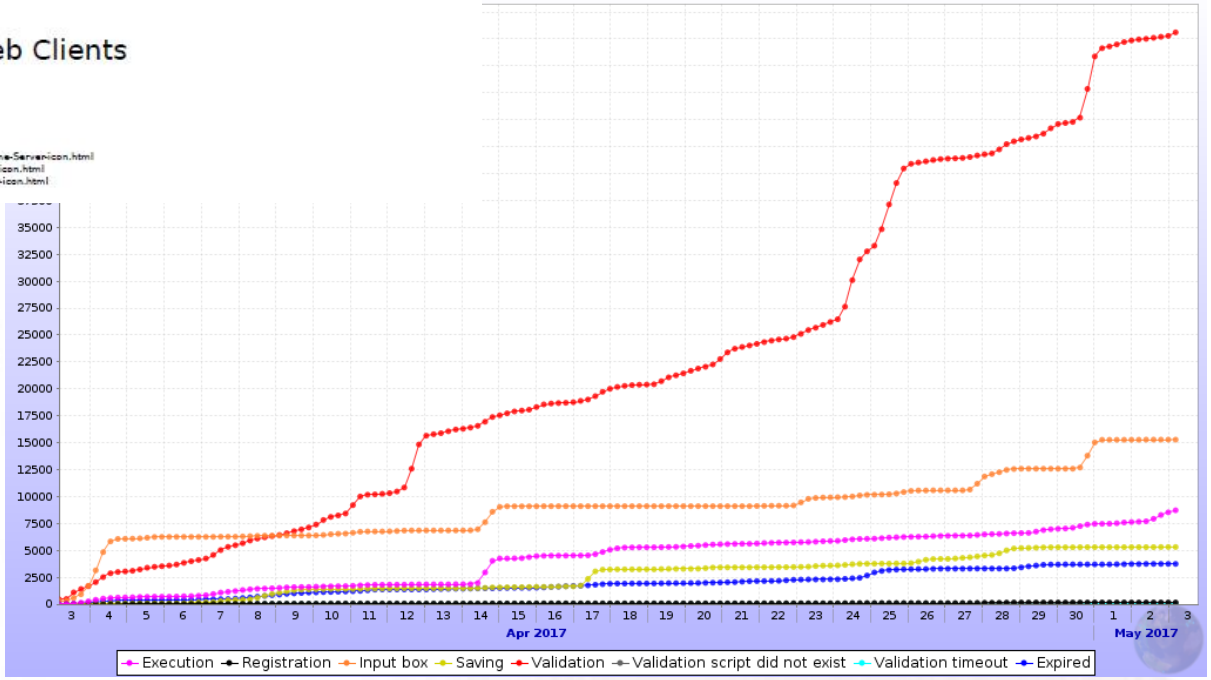
GSI ALICE T2 – Scraper

Ansatz: Periodic querying of the AliEn interface about JobIDs which have visited GSI and writing the information to a local database



http://www.iconarchive.com/show/vista-hardware-devices-icons-by-icons-land/Home-Server-icon.html
http://www.iconarchive.com/show/vista-2-icons-by-gakuseisean/Misc-Database-3-icon.html
http://www.iconarchive.com/show/3d-bluefx-desktop-icons-by-wallpaperfx/Monitor-icon.html

Error jobs in GSI



GSI ALICE T2 – Scraper

Example: Job categories 2017-03-14

Distribution of JobIDs per category:

```
MariaDB [alien]> call p_results_kpis2("2017-03-14 00:00:00","2017-03-14 23:59:59");
+----------+-----------------+---------------+-----------------------------------+
| category | number_absolute | number_percent | category_description              |
+----------+-----------------+---------------+-----------------------------------+
|        0 |               6 |          0.03 | Did not finish (yet)              |
|        1 |              10 |          0.05 | Failed elsewhere                  |
|        4 |             538 |          2.68 | Failed here                       |
|        5 |             181 |          0.90 | Failed generally                  |
|        6 |             162 |          0.81 | Failed here, succeeded elsewhere  |
|        7 |              21 |          0.10 | Failed generally, succeeded elsewhere |
|        8 |           17353 |         86.50 | Succeeded here                    |
|        9 |            1630 |          8.12 | Failed elsewhere, succeeded here  |
|       11 |               1 |          0.00 | Failed elsewhere, succeeded generally |
|       12 |             151 |          0.75 | Failed and succeeded here         |
|       13 |               9 |          0.04 | Failed generally, succeeded here  |
+----------+-----------------+---------------+-----------------------------------+
11 rows in set (0.53 sec)

+---------------+
| total_job_ids |
+---------------+
|         20062 |
+---------------+
1 row in set (0.53 sec)
```

could that be an ansatz for all sites ?
If managed on central service machines KPIs for each sites could easily be extracted ...

GSI ALICE T2 – XRootD Plugins

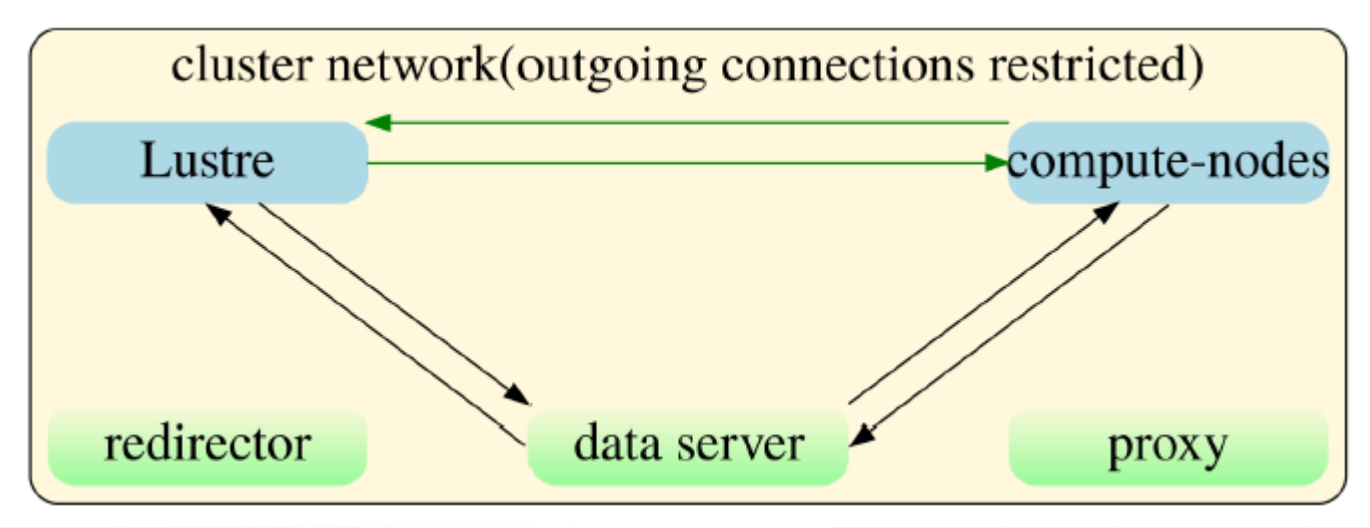
ALICE::GSI::SE2	2.3 PB	1.514 PB	804.6 TB	65.84%	42,280,512	FILE	2.3 PB	1.345 PB	977.5 TB	58.5%

XrdLustreOssWrapper - Plug-in:
Solution: An XRootD data server plug-in
Changes the data server's implementation
Calls the Lustre-api for quota statistics of the current user instead
Configure the server with
ofs.osslib= /path/to/XrdLustreOssWrapper.so
➔ in production use at GSI

XRootD Client Plug-in – XrdProxyPrefix:
Prefixes the client's destination with the proxy address
root://externalSE//file.root -> root://proxy.example.com:1094//root://externalSE//file.root
Is transparent to any application using the new XRootD client, changes the underlying implementation
Uses the XRootD v4.X client plug-in API (XrdCl::FilePlugIn,XrdCl::FileSystemPlugIn etc.)
➔ Currently Proxy Prefix is done by AliEn

GSI ALICE T2 – XRootD Plugins

XRootD Client Plug-in – XrdOpenLocal:
Clients should open a file directly from Lustre
if the GSI
SE is the destination of a call
Solution: Another XRootD client plug-in
"XrdOpenLocal"



Currently being reimplemented as **Server (Redirector) Plugin.**
Client will still need new XRootD Client, though

GSI ALICE T2 – alice-xrootd.deb package

- alice-xrootd.deb package
  - ALICE's xrootd binaries compiled for for Debian Jessie x86_64
  - not site-specific
  - separate unit files for xrootd, cmsd, apmon
  - default config file
  - designed for ease of use with configuration management tools such as Chef
  - runs successfully at GSI (since a week ago)

GSI: AOB

- plans for IPv6:
  - GSI is preparing for IPv6
    - security features, planning, routing, tests
  - no campus-wide use yet in 2017
- EOS
  - will it work on top of Lustre ?
- issues:
  - 0 Byte files on SE, needs further investigation
  - delete campaign: many delete errors, PFNs on storage element need to be checked

Table of contents

UF

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

It is managed by Andres Gomez

No news from this site

Table of contents

ALICE Computing Germany

		Tier 1 (GridKa)	Tier 2 (GSI)	NAF (GSI)
2016	CPU[(1)]	40	16	32
	Disk (PB)	5.2	1.8	3.6
	Tape (PB)	3.9	-	-
2017	CPU[(1)]	61	26	52
	Disk (PB)	6.3	2,3	4,6
	Tape (PB)	7.6	-	-
2018	CPU[(1)]	76.5	30	60
	Disk (PB)	8.0	2,9	5,8
	Tape (PB)	10.2	-	-

[(1)] CPU in kHEPSPEC06

NAF = National Analysis Facility

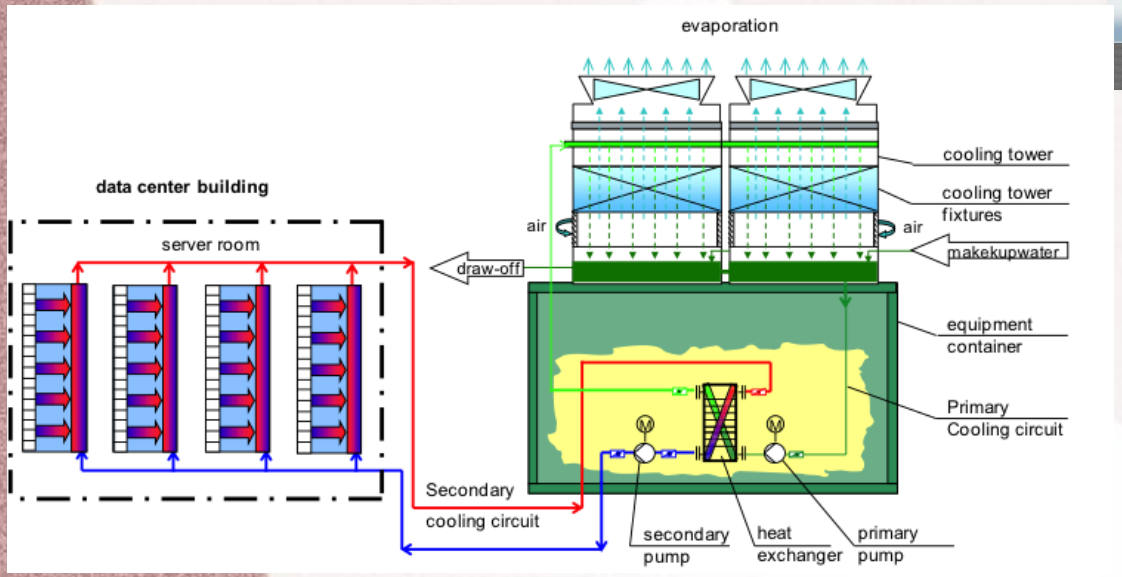
Summary & Outlook

- 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 ...
  - 5 year anniversary ☺

- http://www.springer.com/physics/particle+and+nuclear+physics/journal/41781 (Computing and Software for Big Science)

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