Grid Operations in Germany

Kilian Schwarz

WooJin Park

Thilo Kalkbrenner

Christopher Jung

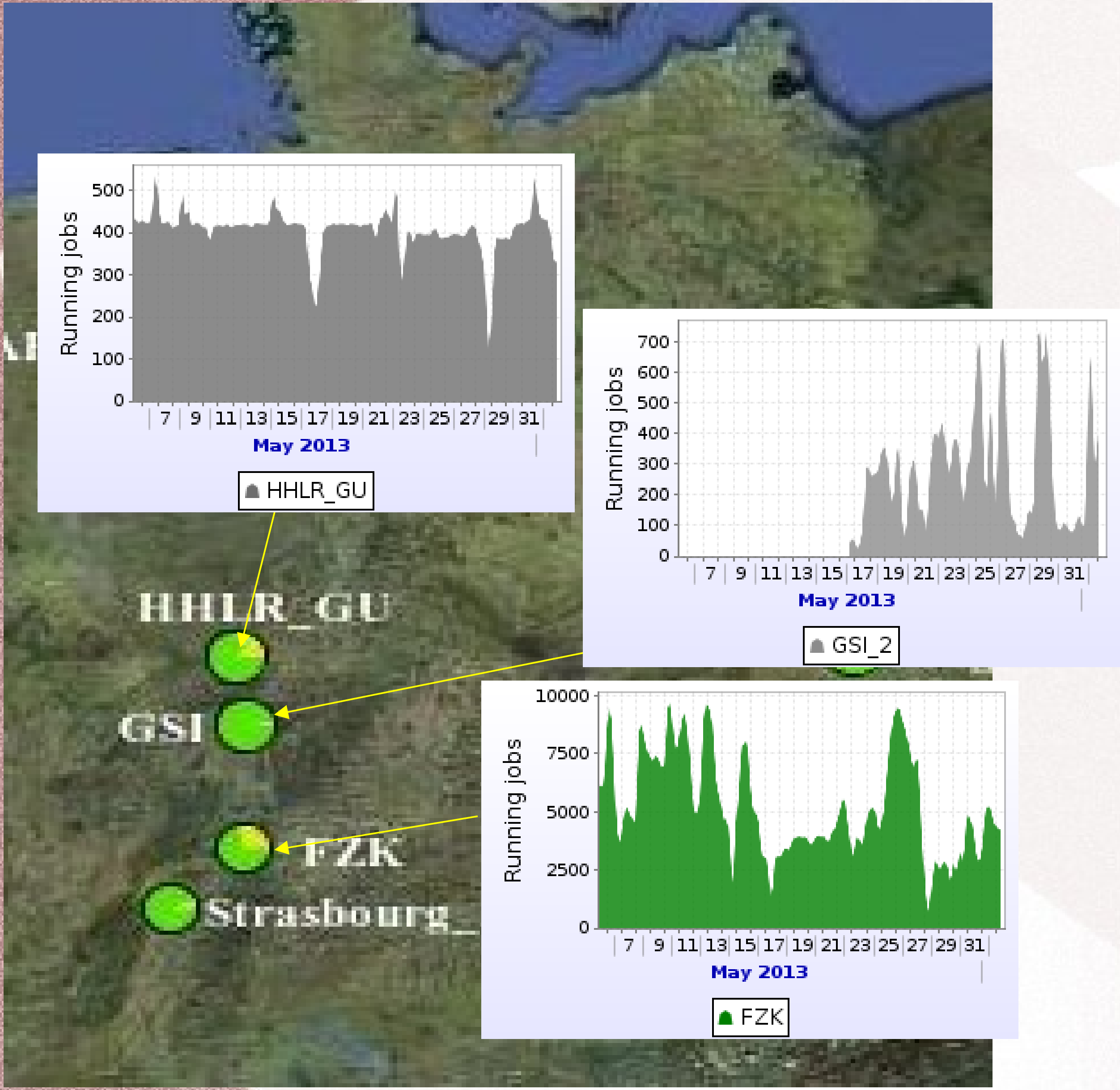
Table of contents

- Overview
- GridKa T1
- GSI T2
- HHLR-GU
- Summary

Table of contents

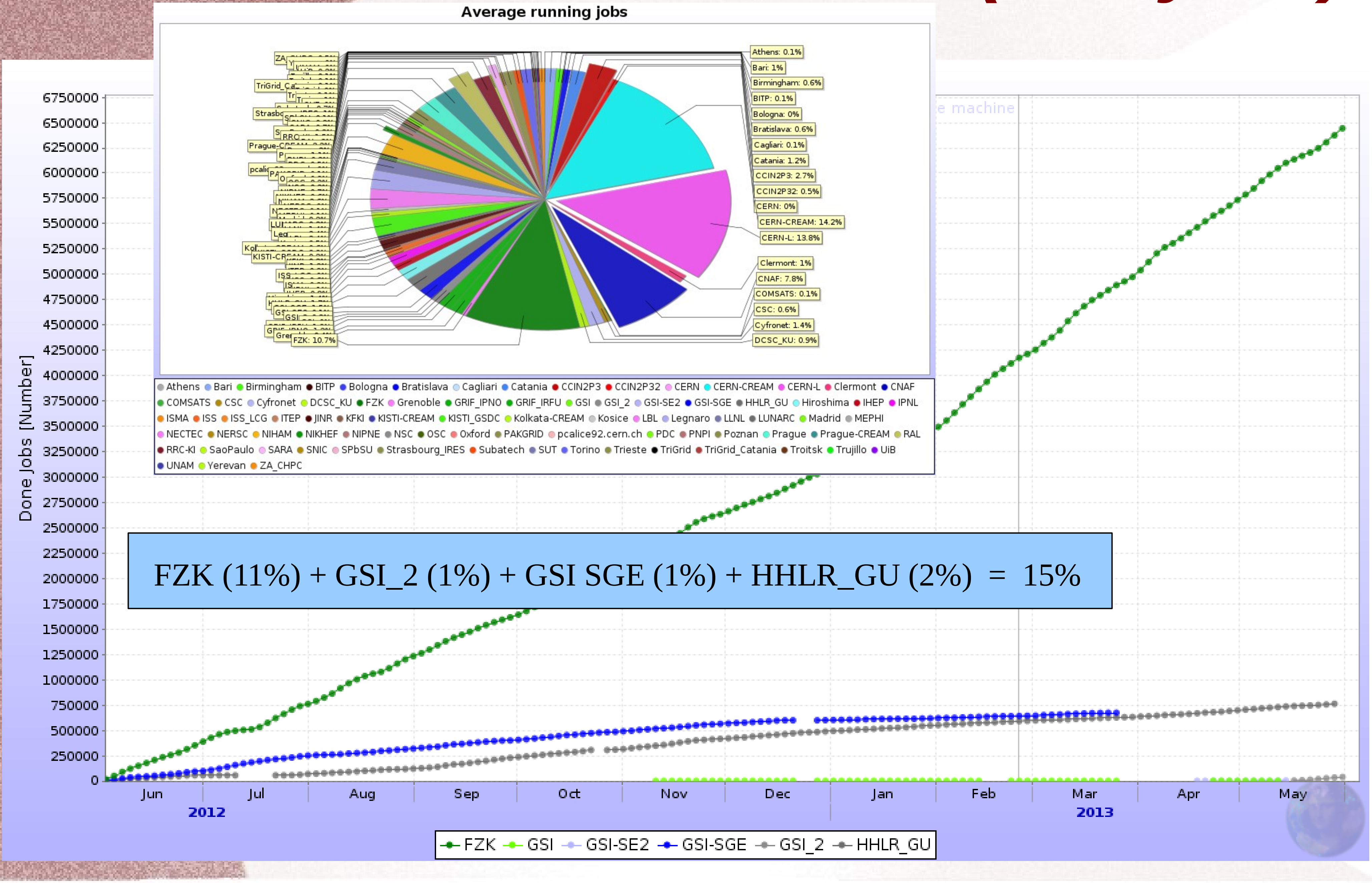
- Overview
- GridKa T1
- GSI T2
- HHLR-GU
- Summary

Map of German Grid sites



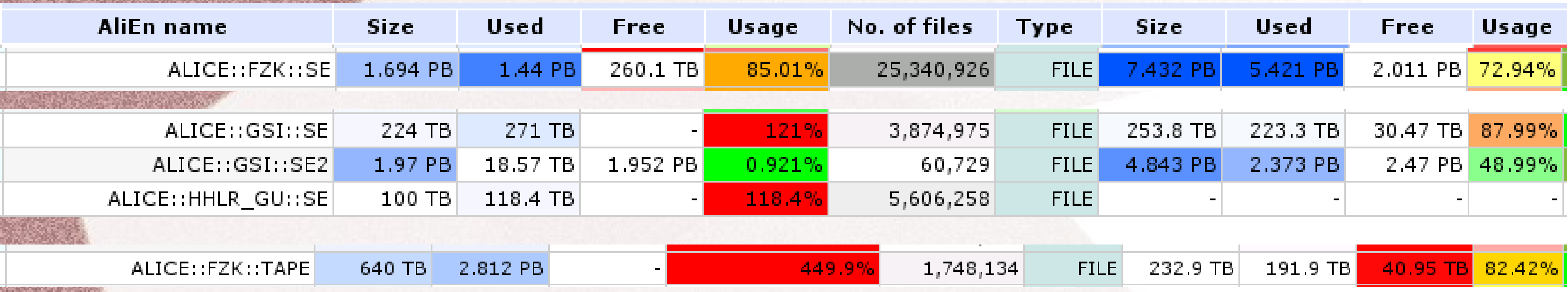
- T1: GridKa/FZK in Karlsruhe
- T2: GSI in Darmstadt
- HHLR_GU in Frankfurt

Job contribution (last year)



FZK (11%) + GSI_2 (1%) + GSI SGE (1%) + HHLR_GU (2%) = 15%

Storage contribution



AliEn name	Size	Used	Free	Usage	No. of files	Type	Size	Used	Free	Usage
ALICE::FZK::SE	1.694 PB	1.44 PB	260.1 TB	85.01%	25,340,926	FILE	7.432 PB	5.421 PB	2.011 PB	72.94%
ALICE::GSI::SE	224 TB	271 TB	-	121%	3,874,975	FILE	253.8 TB	223.3 TB	30.47 TB	87.99%
ALICE::GSI::SE2	1.97 PB	18.57 TB	1.952 PB	0.921%	60,729	FILE	4.843 PB	2.373 PB	2.47 PB	48.99%
ALICE::HHLR_GU::SE	100 TB	118.4 TB	-	118.4%	5,606,258	FILE	-	-	-	-
ALICE::FZK::TAPE	640 TB	2.812 PB	-	449.9%	1,748,134	FILE	232.9 TB	191.9 TB	40.95 TB	82.42%

Total size:

(GridKa: 2.3 PB Disk SE and 0.7 PB tape buffer – xroot infos not correct)

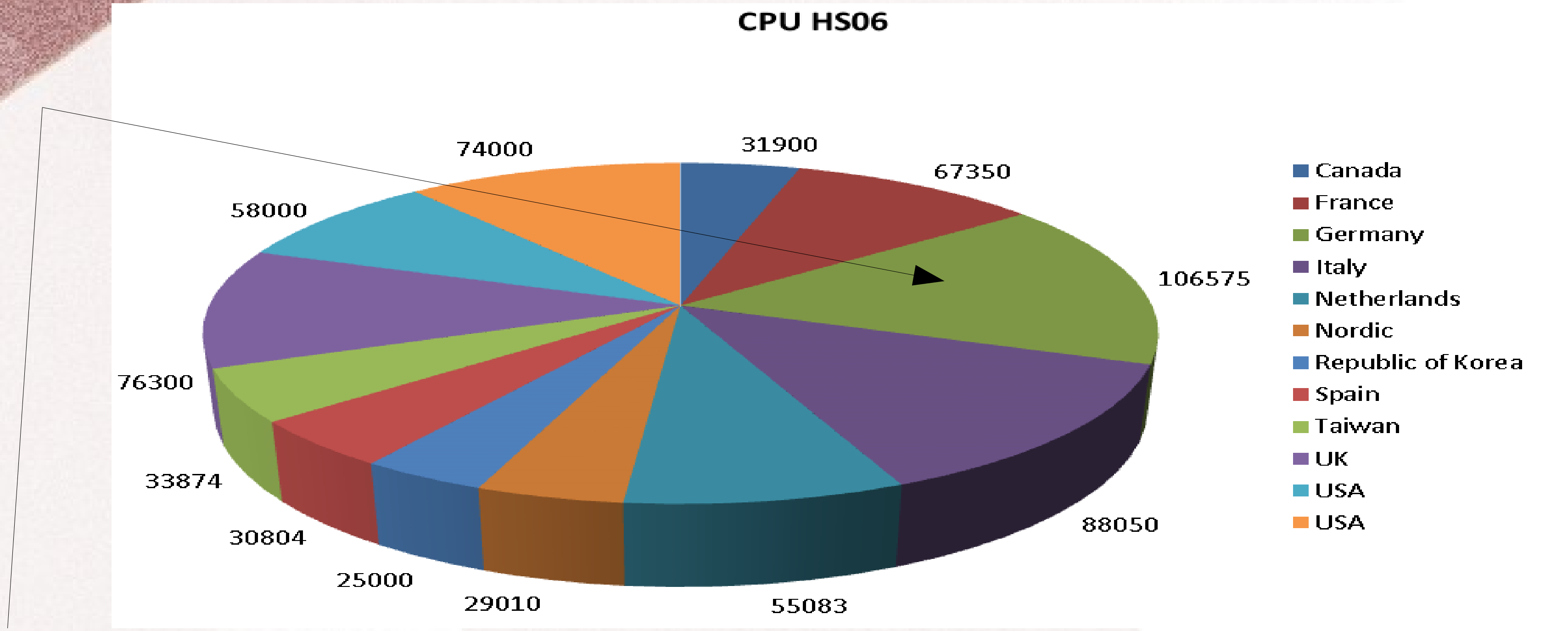
- 3.1 PB disk based SE (ALICE total: 32 PB)
- 700 TB disk buffer with Tape backend

Table of contents

- Overview
- GridKa T1
- GSI T2
- HHLR-GU
- Summary

Tier-1: GridKa

WLCG Tier-1	CPU (HS06)	Disk	Tape
2013	695'000	77.7 PB	106.8 PB

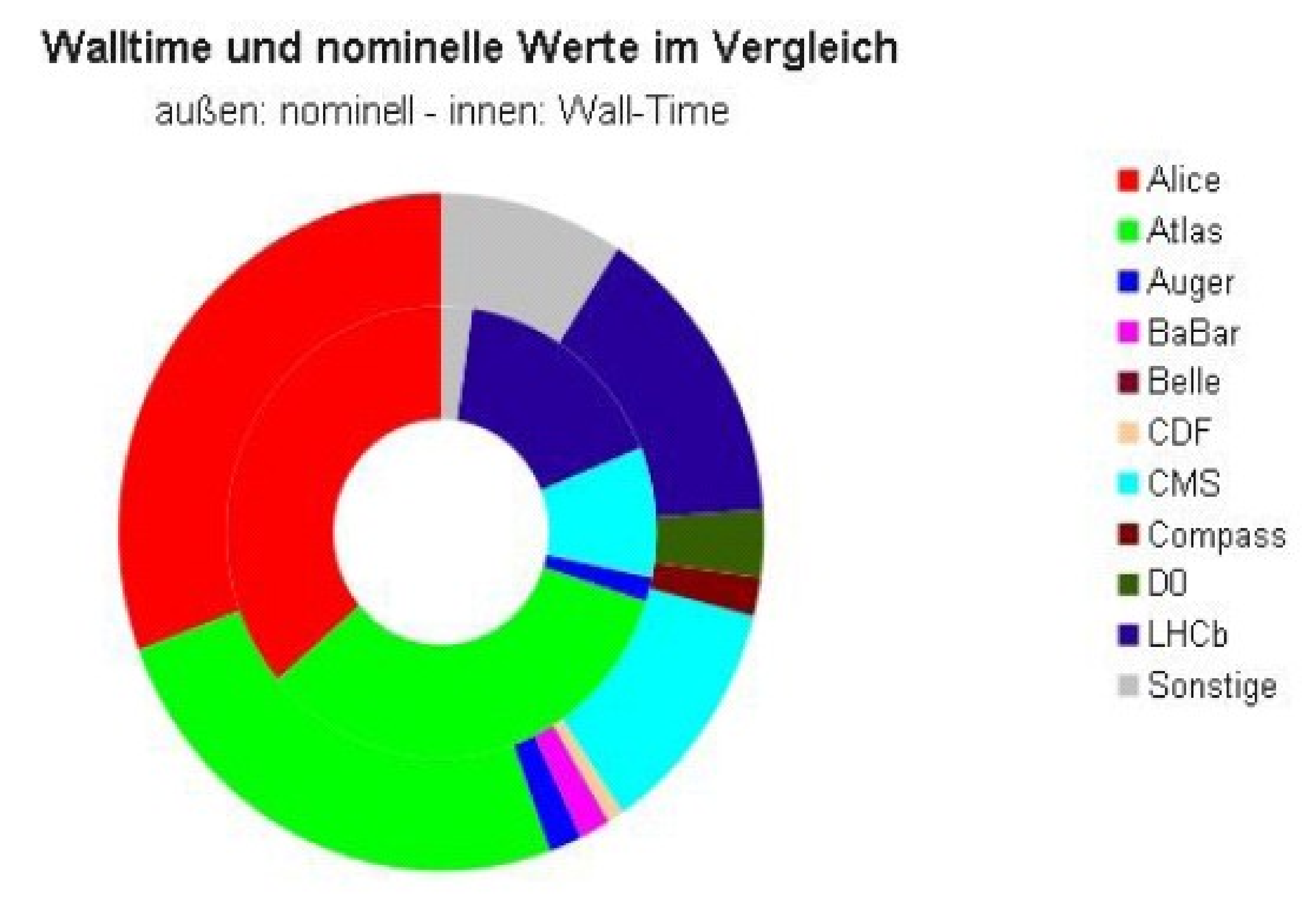


GridKa is the largest Tier1 in WLCG and provides about 15% of the total T1 recources

GridKa:	CPU (HS06)	%WLCG	Disk	%WLCG	Tape	% WLCG
ALICE :	30000	25%	2.7 PB	25%	5.2 PB	25%
ATLAS:	39875	12.5%	4.1 PB	12.5%	5.0 PB	12,5%
CMS:	16500	10%	2.6 PB	10%	5.0 PB	10%
LHCb:	18370	16.7%	1.4 PB	16.7%	1.1 PB	16.7%

usage statistics

(Jan-Apr 2013)



Centre is well used.

Largest shares: LHC experiments.

(ALICE and ATLAS alone > 50%)

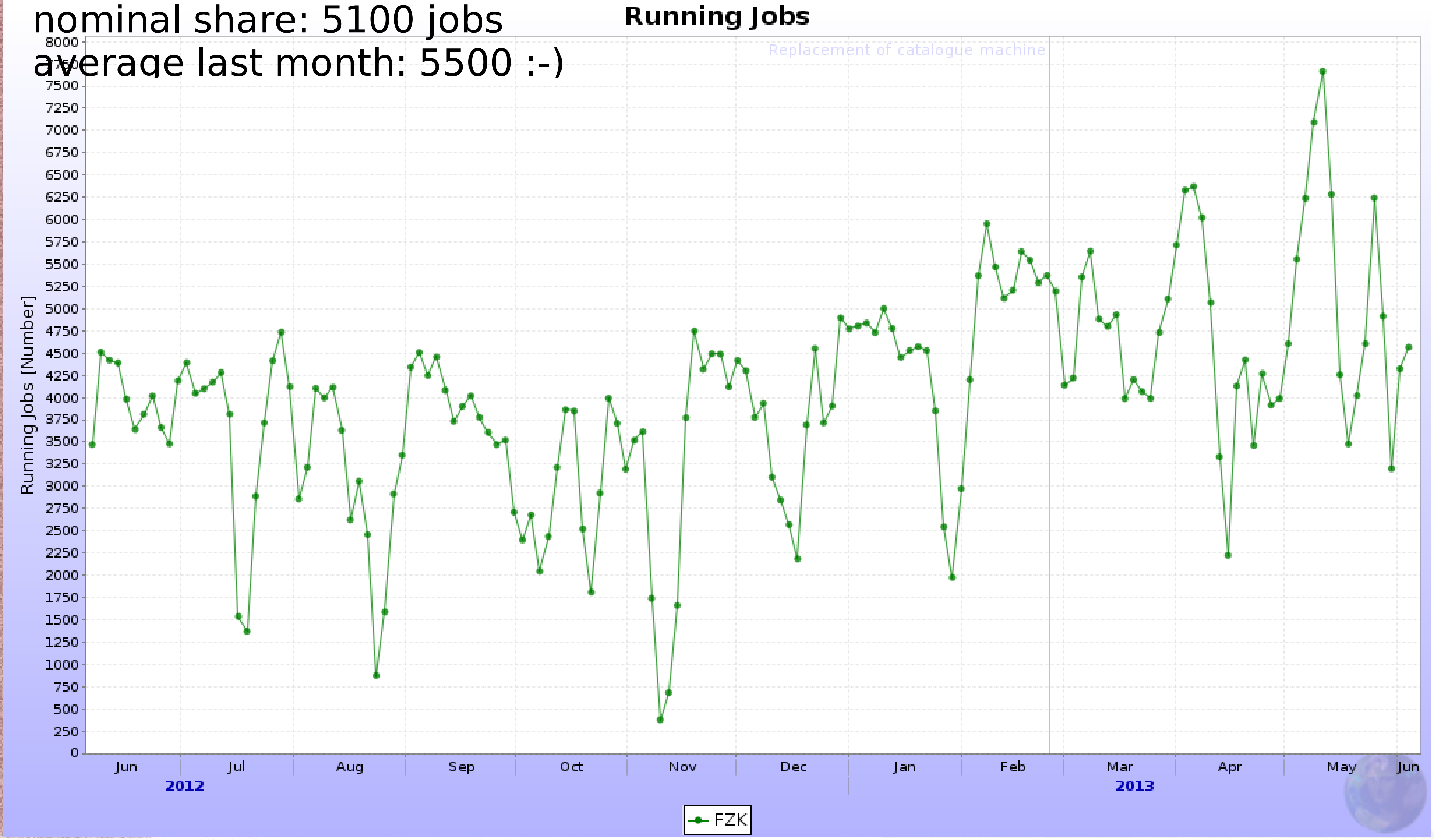
ALICE and ATLAS are using roughly their nominal share.

Batch Submission

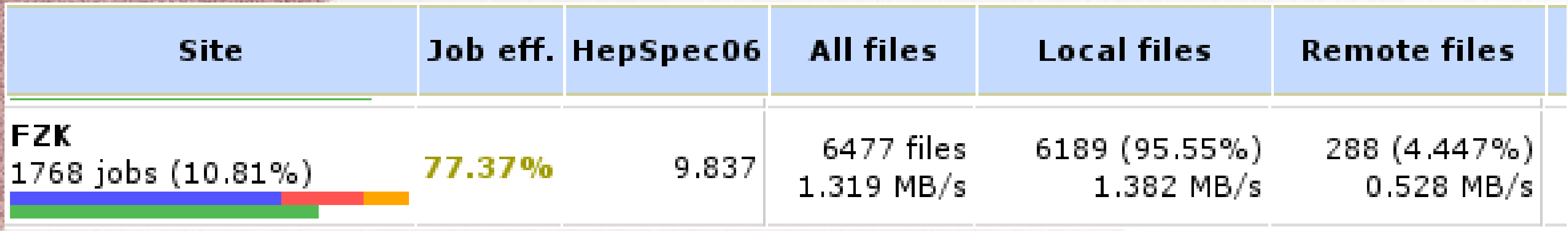
- Vobox migrated to WLCG VO-BOX
- switched from PBSPro to UniVa GridEngine
  - during PBS usage cluster had to be devided into 2 sub clusters.
  - GE is able to manage to whole cluster
  - Fair share values are computed daily. Current values for ALICE: about 30%
    - This computes to 5100 jobs
- Software distribution via Torrent is being tested
  - There is a communication problem among the Wns
- First Wns will be upgraded to SL6 this year

Jobs at GridKa within last year

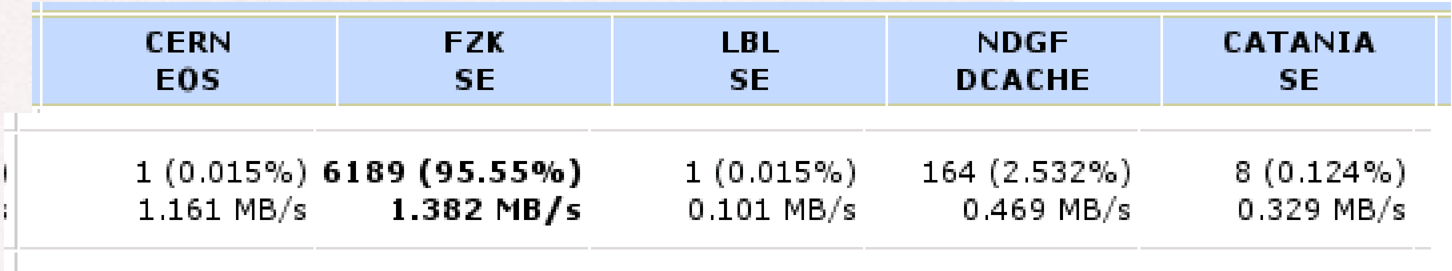
max. number of concurrent jobs: 9922 (ALICE record)
average number of jobs: 4009
nominal share: 5100 jobs
average last month: 5500 :-)



ALICE Job Efficiency



Site	Job eff.	HepSpec06	All files	Local files	Remote files
FZK 1768 jobs (10.81%)	77.37%	9.837	6477 files 1.319 MB/s	6189 (95.55%) 1.382 MB/s	288 (4.447%) 0.528 MB/s

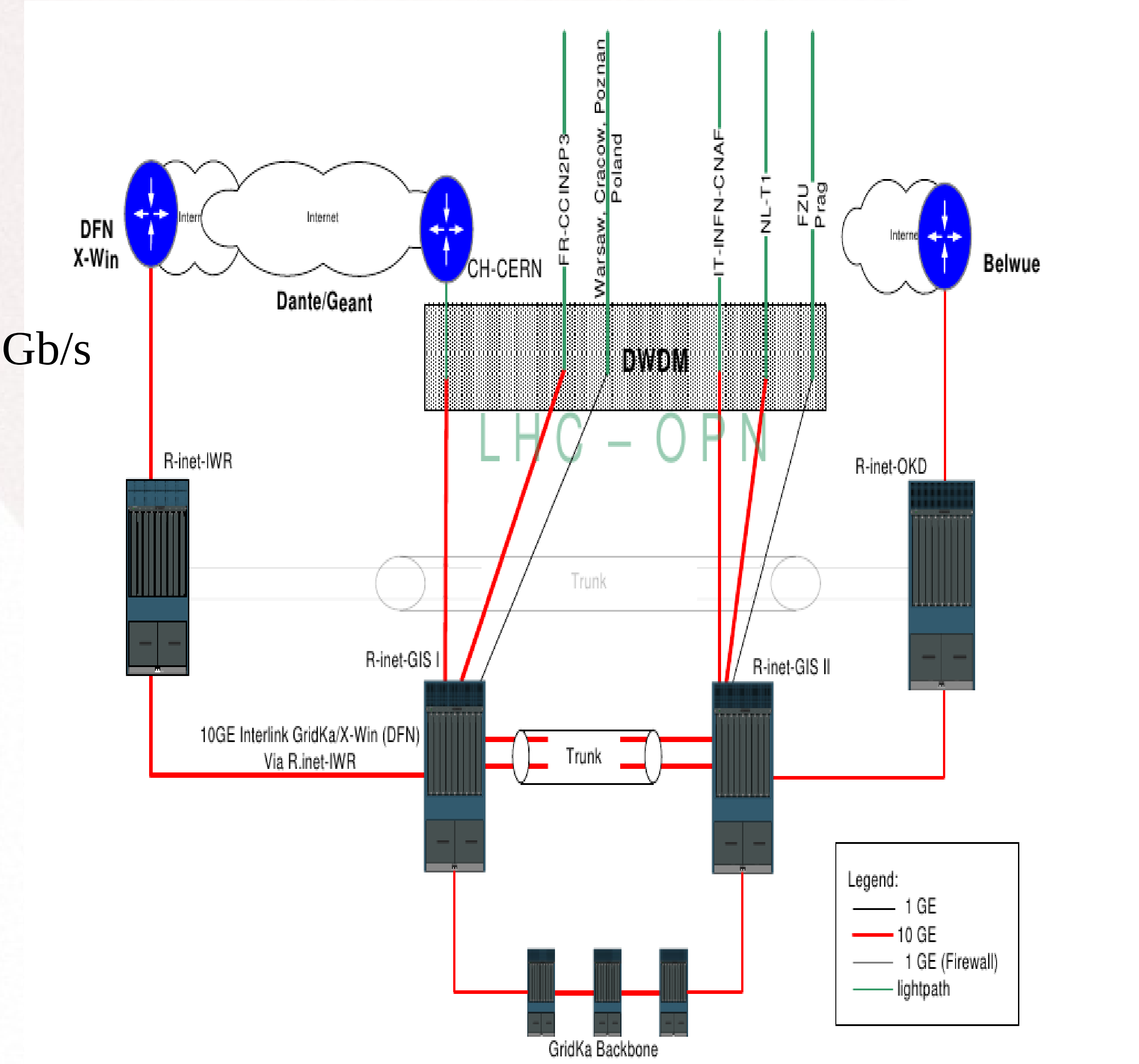


CERN EOS	FZK SE	LBL SE	NDGF DCACHE	CATANIA SE
1 (0.015%) 1.161 MB/s	**6189 (95.55%)** **1.382 MB/s**	1 (0.015%) 0.101 MB/s	164 (2.532%) 0.469 MB/s	8 (0.124%) 0.329 MB/s

- Firewall upgrade in calendar week 28
  - new FW should be able to cope with 100 Gb/s throughput
  - remote data storage will not affect site performance anymore

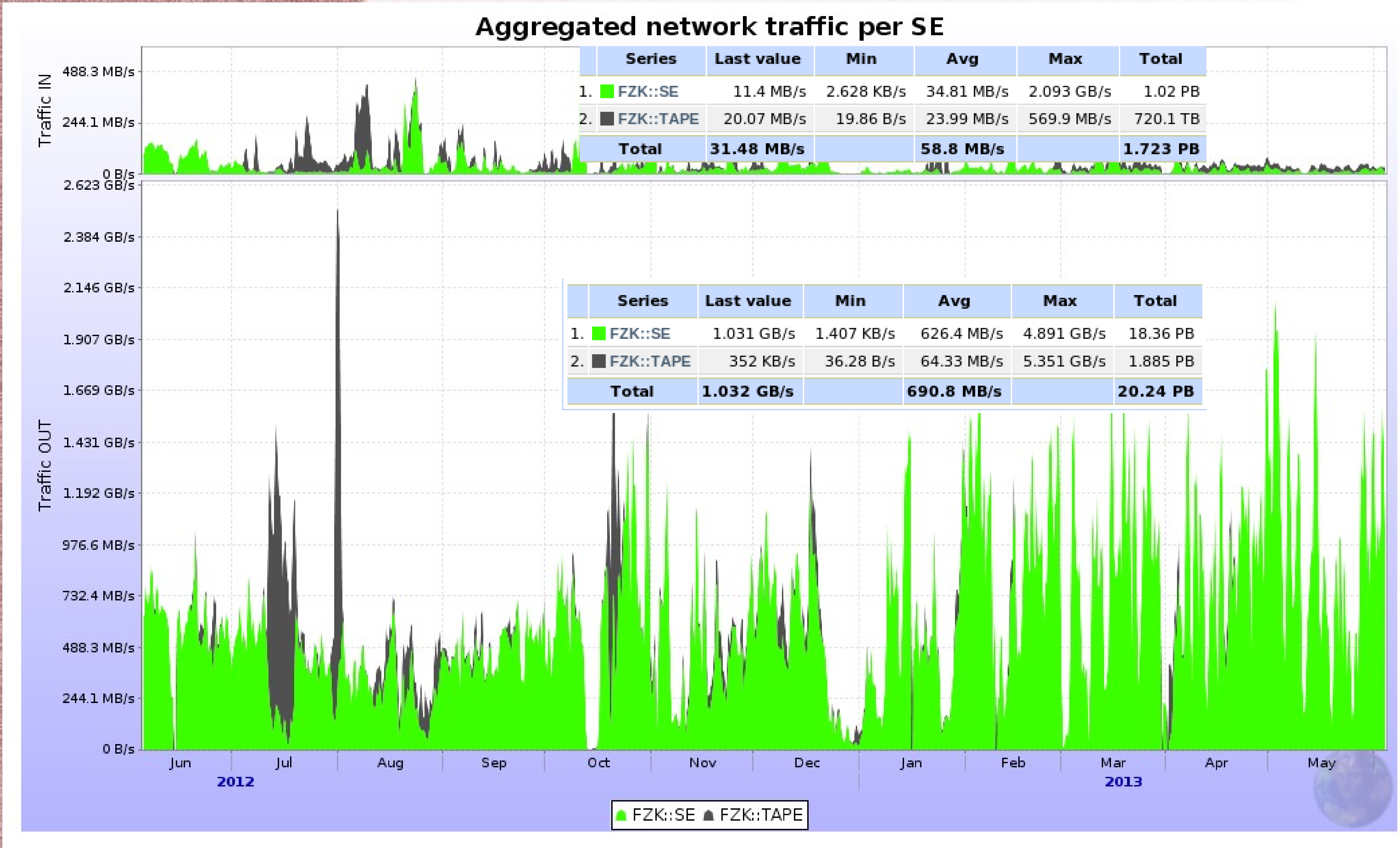
GridKa network setup

LHCOPN (CERN, CCIN2P3, CNAF, SARA, ..): 10 Gb/s
LHCONE: 10 Gb/s
German research network X-Win: 10 Gb/s
Prague (dedicated): 1 Gb/s
Poznan (dedicated): 1 Gb/s
==> 62 Gb/s total bandwidth



GridKa storage

- Xrootd based SE works well and is heavily used
- Reading from FZK::SE increased by a factor of 3 since last workshop

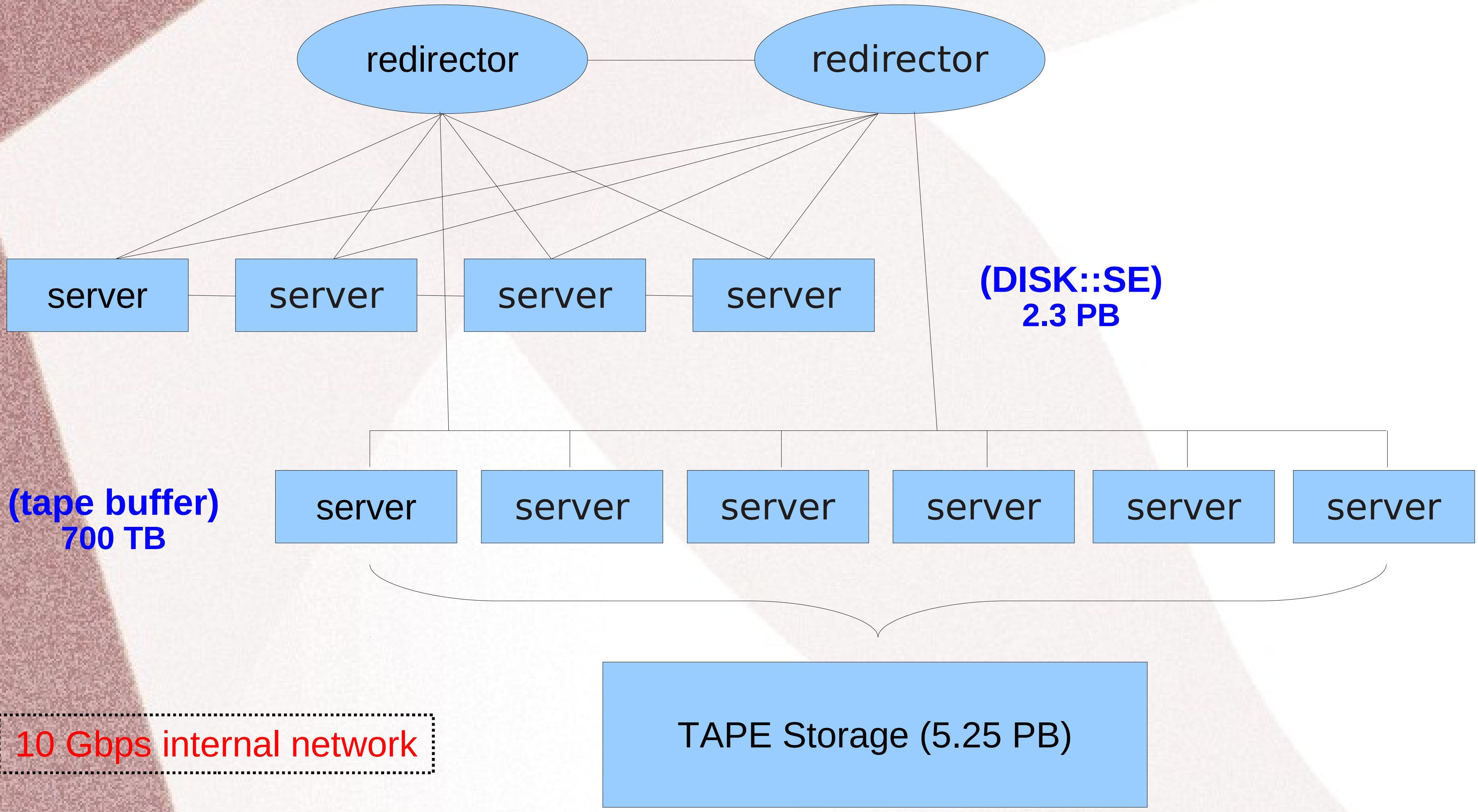


Aggregated network traffic per SE

	Series	Last value	Min	Avg	Max	Total
1.	🟩 FZK::SE	11.4 MB/s	2.628 KB/s	34.81 MB/s	2.093 GB/s	1.02 PB
2.	⬛ FZK::TAPE	20.07 MB/s	19.86 B/s	23.99 MB/s	569.9 MB/s	720.1 TB
	Total	**31.48 MB/s**		**58.8 MB/s**		**1.723 PB**

	Series	Last value	Min	Avg	Max	Total
1.	🟩 FZK::SE	1.031 GB/s	1.407 KB/s	626.4 MB/s	4.891 GB/s	18.36 PB
2.	⬛ FZK::TAPE	352 KB/s	36.28 B/s	64.33 MB/s	5.351 GB/s	1.885 PB
	Total	**1.032 GB/s**		**690.8 MB/s**		**20.24 PB**

XRootD Architecture at GridKa



10 Gbps internal network

Various points of interest

Xrootd:

- 4 data servers (and 2 redirectors) were replaced by more powerful machines
  - Disk space increased to 2.7 PB
    - But the borrowed 0.5 PB from 2011 still needs to be paid back
- Still trouble with writing via redirectors to the new servers
- Internal monitoring tool has been implemented

ALICE requested trade of tape to disk:

- This is not feasible at the moment due to limited budget
- No major hardware change planned within the next 2 years
  - Cold data sets could be moved to tape ...

ALICE Requirements at GridKa T1

GridKa (25% of the T1 requirements)

	CPU	**Disk**	**Tape**
	kHEPSPEC06	PB	PB
2015	27,5	2,625	5,3
2016	32,5	2,975	8,7
2017	40	3,825	12,2
2018	40	3,825	12,2
2019	40	3,825	12,2

Table of contents

- Overview
- GridKa T1
- GSI T2
- HHLR-GU
- Summary

Gesellschaft für Schwerionenforschung mbH (GSI)

employs about 1000 people

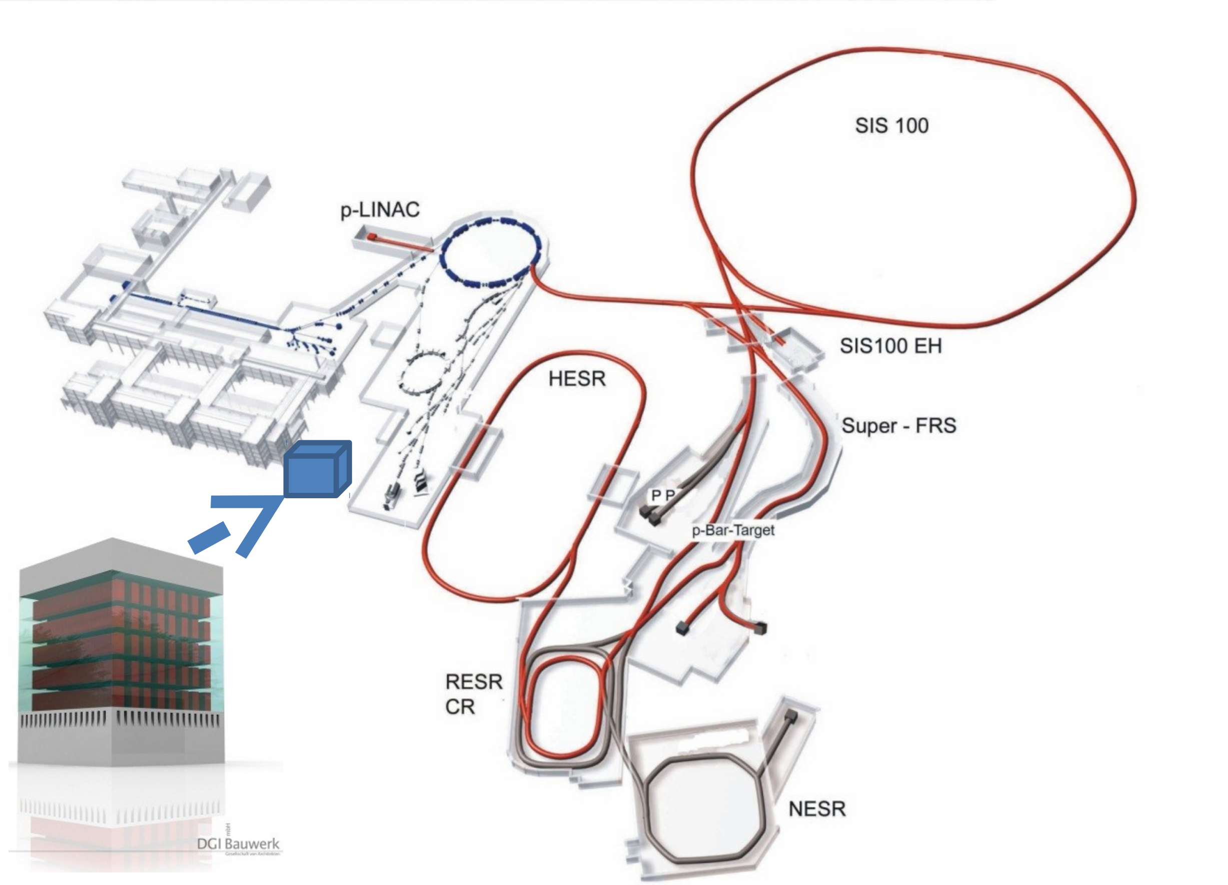
GSI: a national Research Centre for heavy ion research
FAIR: Facility for Ion and Antiproton Research ~2018

GSI computing today

ALICE T2/T3
HADES
~ 14000 cores,
~ 5.5 PB lustre
~ 9 PB archive capacity

FAIR computing 2018

CBM
PANDA
NuSTAR
APPA
LQCD
300000 cores
40 PB disk
40 PB archive



open source and community software
budget commodity hardware
support different communities
scarce manpower

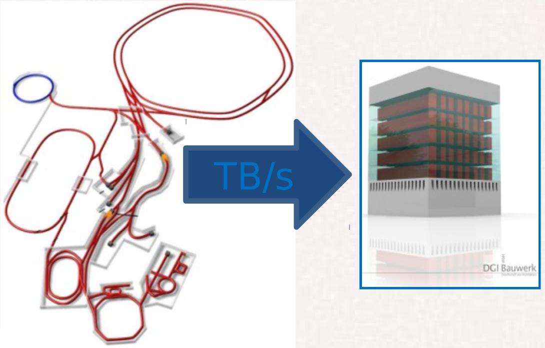
FAIR Computing: T0/T1 MAN & Grid/Cloud

GOETHE UNIVERSITÄT FRANKFURT AM MAIN

1Tb/s

1Tb/s

JOHANNES GUTENBERG UNIVERSITÄT MAINZ

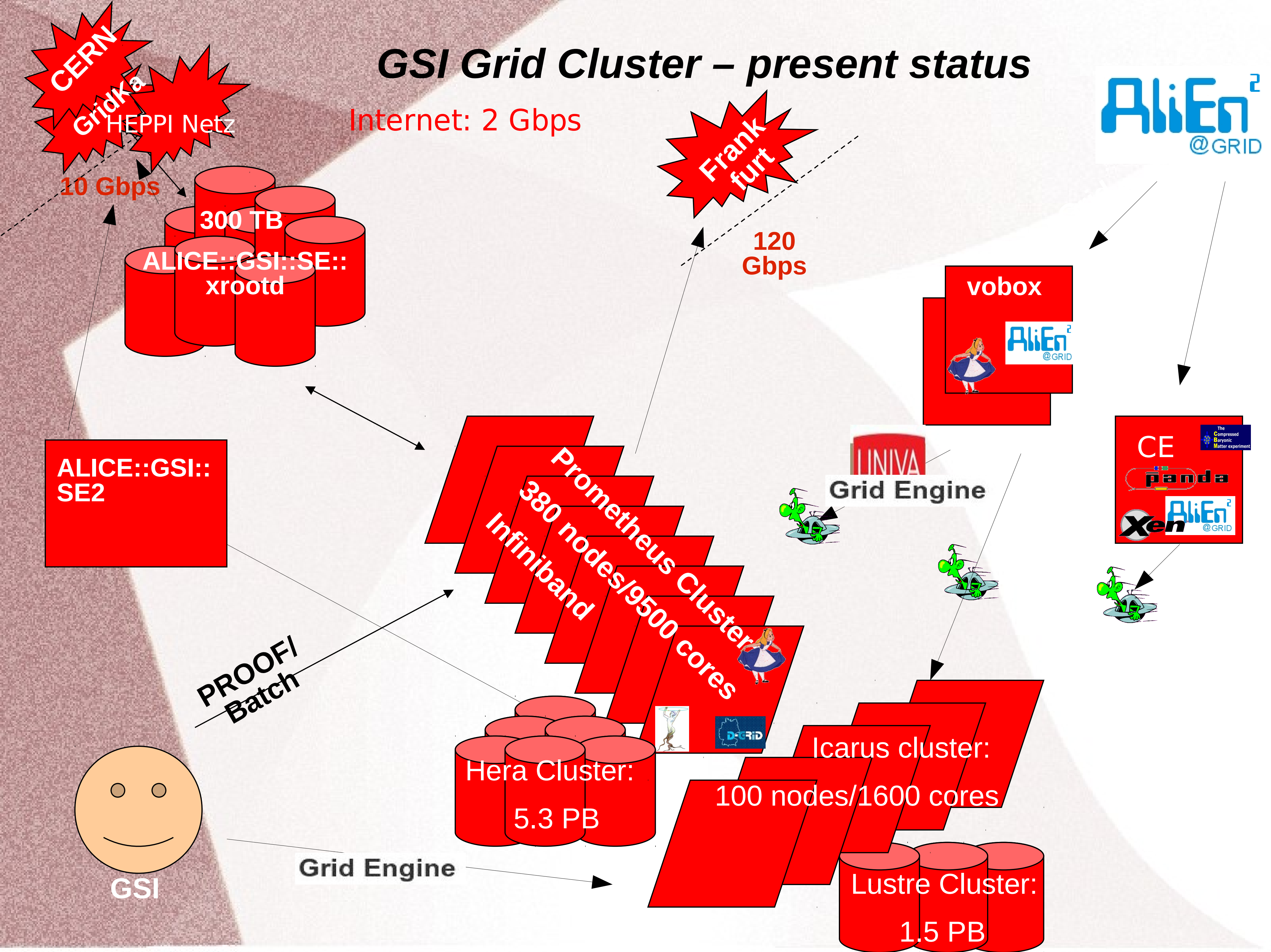


1Tb/s

1Tb/s

TECHNISCHE UNIVERSITÄT DARMSTADT

GSI Grid Cluster – present status



CERN

GridKa

HEPPI Netz

Internet: 2 Gbps

10 Gbps

Frank furt

120 Gbps

AliEn @GRID

300 TB

ALICE::GSI::SE:: xrootd

vobox

ALICE::GSI:: SE2

UNIVA Grid Engine

CE

Prometheus Cluster

380 nodes/9500 cores

Infiniband

PROOF/ Batch

Hera Cluster:

5.3 PB

Icarus cluster:

100 nodes/1600 cores

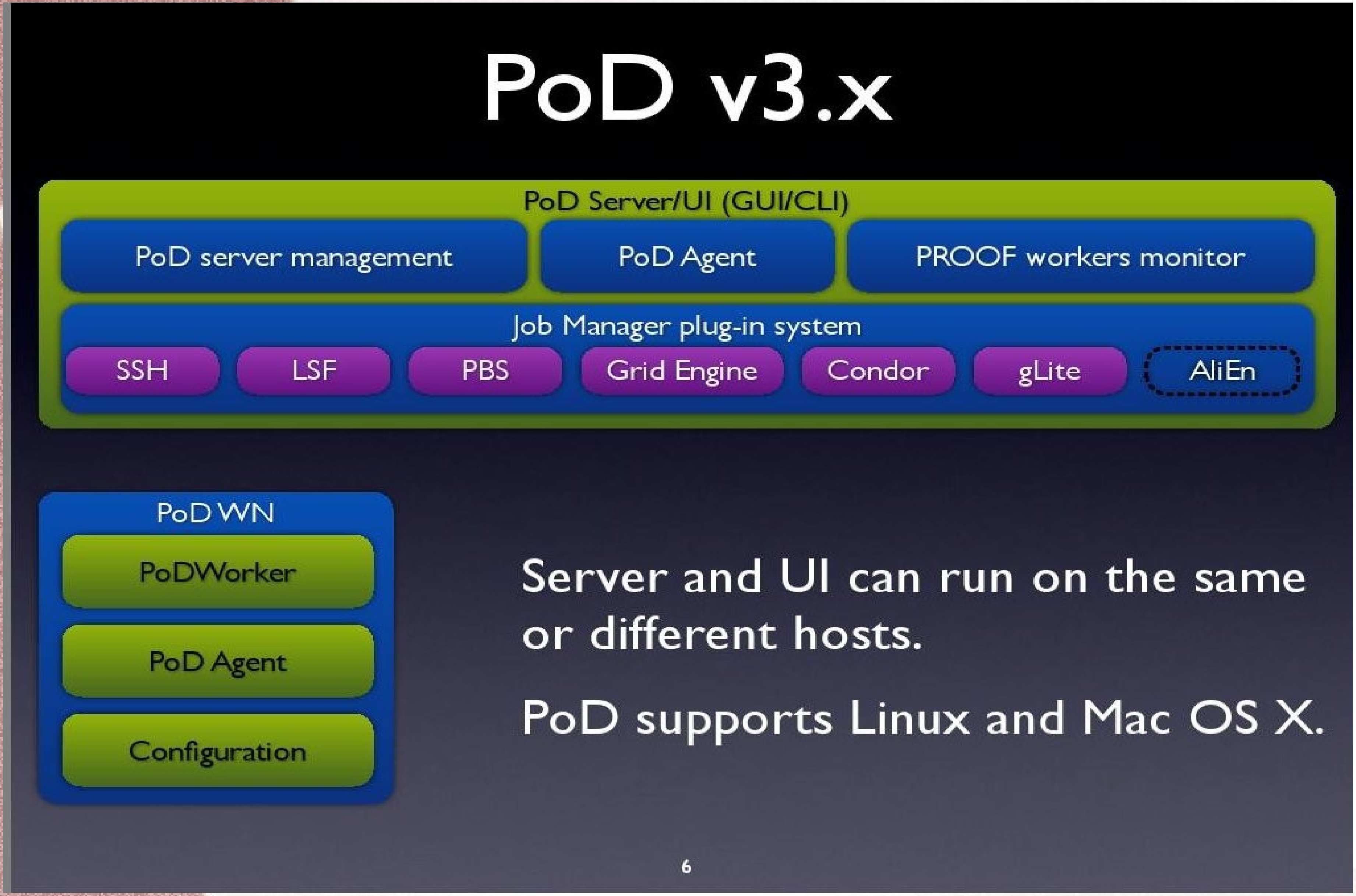
Lustre Cluster:

1.5 PB

GSI

Grid Engine

PROOF on Demand (PoD)



PoD v3.x

PoD Server/UI (GUI/CLI)

PoD server management

PoD Agent

PROOF workers monitor

Job Manager plug-in system

SSH

LSF

PBS

Grid Engine

Condor

gLite

AliEn

PoD WN

PoDWorker

PoD Agent

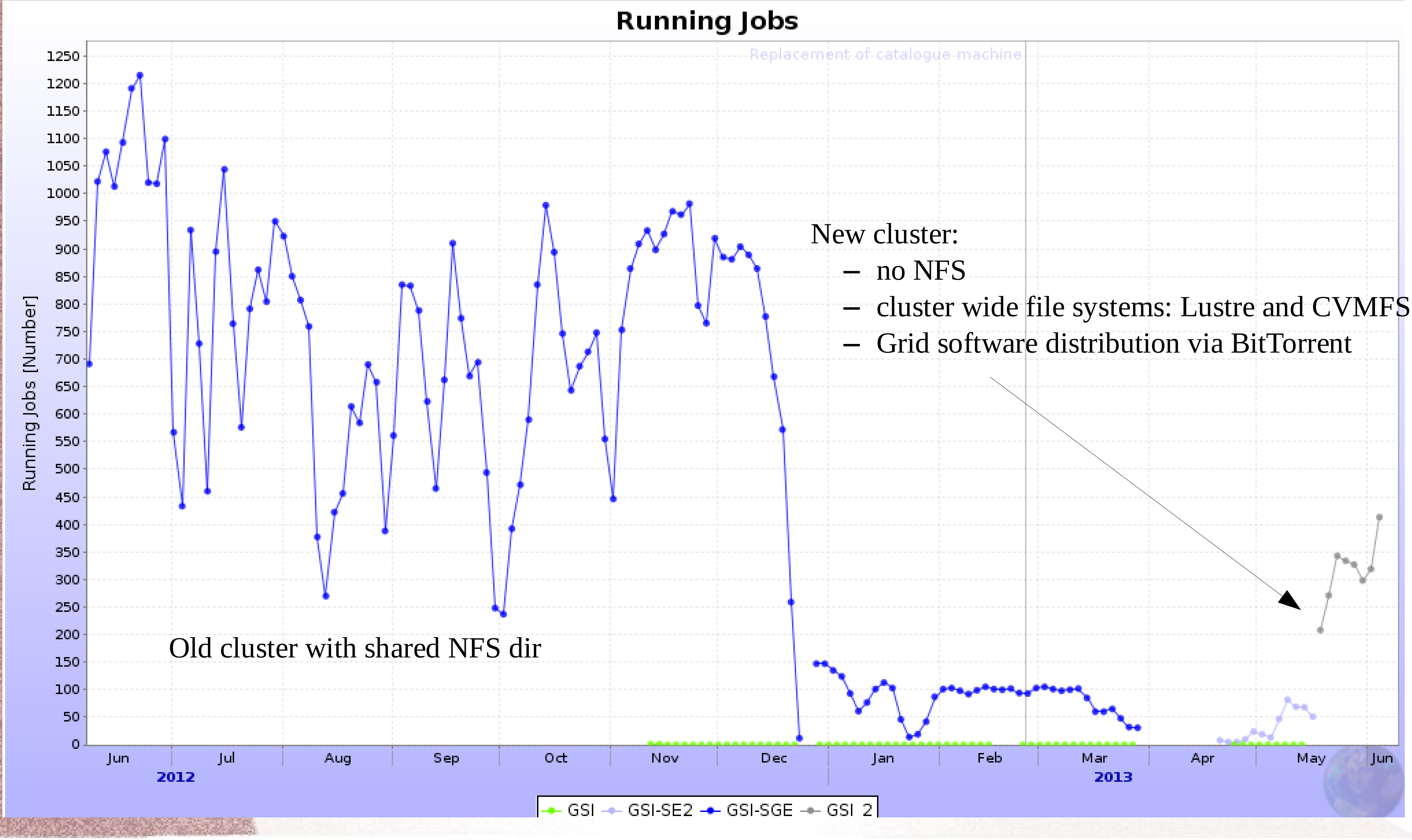
Configuration

Server and UI can run on the same or different hosts.

PoD supports Linux and Mac OS X.

jobs at GSI within last year

BitTorrent: too many pack/unpack activities on local disk lead to IO-Wait states on Wns.
Maybe this situation improves when switching to CVMFS



New cluster:

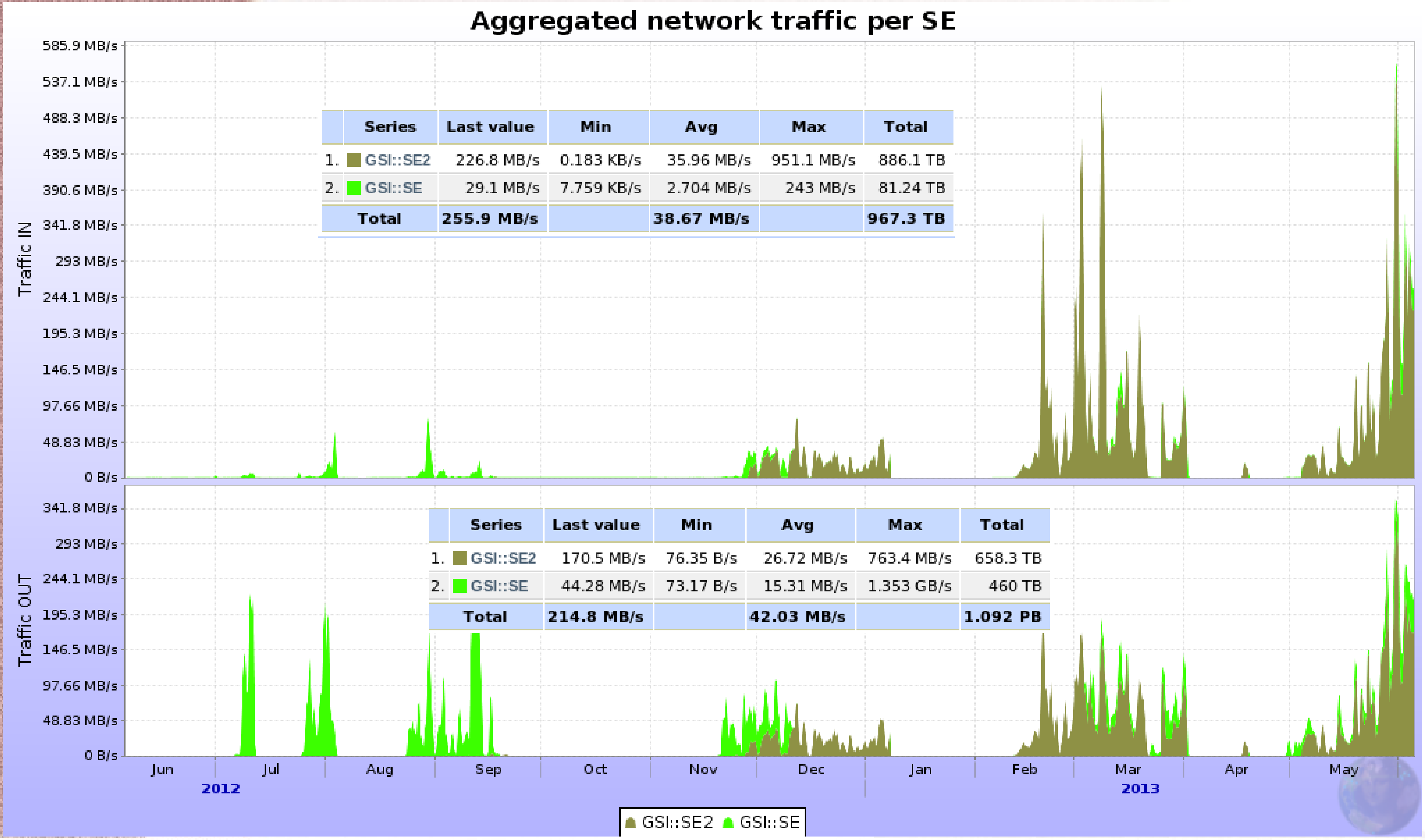
- no NFS
- cluster wide file systems: Lustre and CVMFS
- Grid software distribution via BitTorrent

Old cluster with shared NFS dir

GSI Storage Elements

– ALICE::GSI::SE is mainly used for read access

– ALICE::GSI::SE2 is mainly used for data transfer. The peak value shows that the 10 Gb link can be saturated. But most of the time t transfer speed is not satisfying. This still needs investigation.



	Series	Last value	Min	Avg	Max	Total
1.	🟫 GSI::SE2	226.8 MB/s	0.183 KB/s	35.96 MB/s	951.1 MB/s	886.1 TB
2.	🟩 GSI::SE	29.1 MB/s	7.759 KB/s	2.704 MB/s	243 MB/s	81.24 TB
	Total	255.9 MB/s		38.67 MB/s		967.3 TB

	Series	Last value	Min	Avg	Max	Total
1.	🟫 GSI::SE2	170.5 MB/s	76.35 B/s	26.72 MB/s	763.4 MB/s	658.3 TB
2.	🟩 GSI::SE	44.28 MB/s	73.17 B/s	15.31 MB/s	1.353 GB/s	460 TB
	Total	214.8 MB/s		42.03 MB/s		1.092 PB

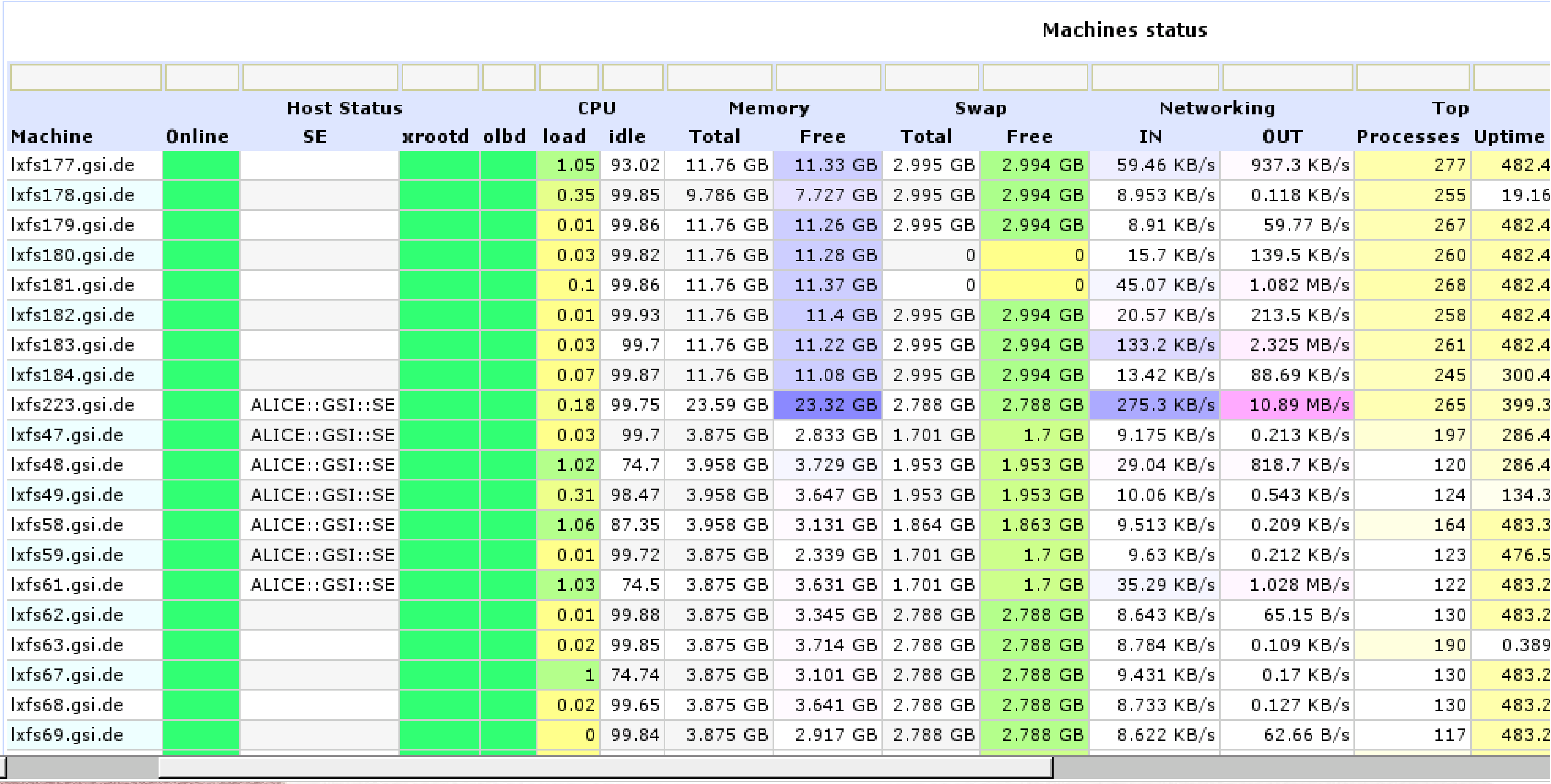
ALICE::GSI::SE - architecture

36 file server and 1 redirector providing 300 TB disk space

file servers come into age and start refusing service

disks are full ...

Storage Cluster



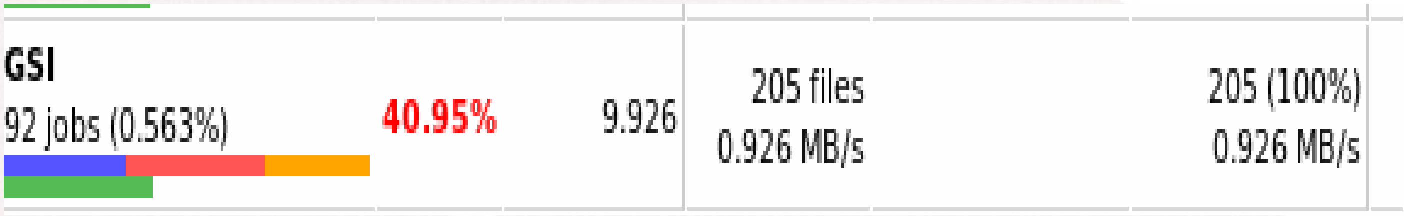
Machines status

		Host Status			CPU		Memory		Swap		Networking		Top	
Machine	Online	SE	xrootd	olbd	load	idle	Total	Free	Total	Free	IN	OUT	Processes	Uptime
lxfs177.gsi.de	🟩		🟩	🟩	1.05	93.02	11.76 GB	11.33 GB	2.995 GB	2.994 GB	59.46 KB/s	937.3 KB/s	277	482.4
lxfs178.gsi.de	🟩		🟩	🟩	0.35	99.85	9.786 GB	7.727 GB	2.995 GB	2.994 GB	8.953 KB/s	0.118 KB/s	255	19.16
lxfs179.gsi.de	🟩		🟩	🟩	0.01	99.86	11.76 GB	11.26 GB	2.995 GB	2.994 GB	8.91 KB/s	59.77 B/s	267	482.4
lxfs180.gsi.de	🟩		🟩	🟩	0.03	99.82	11.76 GB	11.28 GB	0	0	15.7 KB/s	139.5 KB/s	260	482.4
lxfs181.gsi.de	🟩		🟩	🟩	0.1	99.86	11.76 GB	11.37 GB	0	0	45.07 KB/s	1.082 MB/s	268	482.4
lxfs182.gsi.de	🟩		🟩	🟩	0.01	99.93	11.76 GB	11.4 GB	2.995 GB	2.994 GB	20.57 KB/s	213.5 KB/s	258	482.4
lxfs183.gsi.de	🟩		🟩	🟩	0.03	99.7	11.76 GB	11.22 GB	2.995 GB	2.994 GB	133.2 KB/s	2.325 MB/s	261	482.4
lxfs184.gsi.de	🟩		🟩	🟩	0.07	99.87	11.76 GB	11.08 GB	2.995 GB	2.994 GB	13.42 KB/s	88.69 KB/s	245	300.4
lxfs223.gsi.de	🟩	ALICE::GSI::SE	🟩	🟩	0.18	99.75	23.59 GB	23.32 GB	2.788 GB	2.788 GB	275.3 KB/s	10.89 MB/s	265	399.3
lxfs47.gsi.de	🟩	ALICE::GSI::SE	🟩	🟩	0.03	99.7	3.875 GB	2.833 GB	1.701 GB	1.7 GB	9.175 KB/s	0.213 KB/s	197	286.4
lxfs48.gsi.de	🟩	ALICE::GSI::SE	🟩	🟩	1.02	74.7	3.958 GB	3.729 GB	1.953 GB	1.953 GB	29.04 KB/s	818.7 KB/s	120	286.4
lxfs49.gsi.de	🟩	ALICE::GSI::SE	🟩	🟩	0.31	98.47	3.958 GB	3.647 GB	1.953 GB	1.953 GB	10.06 KB/s	0.543 KB/s	124	134.3
lxfs58.gsi.de	🟩	ALICE::GSI::SE	🟩	🟩	1.06	87.35	3.958 GB	3.131 GB	1.864 GB	1.863 GB	9.513 KB/s	0.209 KB/s	164	483.3
lxfs59.gsi.de	🟩	ALICE::GSI::SE	🟩	🟩	0.01	99.72	3.875 GB	2.339 GB	1.701 GB	1.7 GB	9.63 KB/s	0.212 KB/s	123	476.5
lxfs61.gsi.de	🟩	ALICE::GSI::SE	🟩	🟩	1.03	74.5	3.875 GB	3.631 GB	1.701 GB	1.7 GB	35.29 KB/s	1.028 MB/s	122	483.2
lxfs62.gsi.de	🟩		🟩	🟩	0.01	99.88	3.875 GB	3.345 GB	2.788 GB	2.788 GB	8.643 KB/s	65.15 B/s	130	483.2
lxfs63.gsi.de	🟩		🟩	🟩	0.02	99.85	3.875 GB	3.714 GB	2.788 GB	2.788 GB	8.784 KB/s	0.109 KB/s	190	0.389
lxfs67.gsi.de	🟩		🟩	🟩	1	74.74	3.875 GB	3.101 GB	2.788 GB	2.788 GB	9.431 KB/s	0.17 KB/s	130	483.2
lxfs68.gsi.de	🟩		🟩	🟩	0.02	99.65	3.875 GB	3.641 GB	2.788 GB	2.788 GB	8.733 KB/s	0.127 KB/s	130	483.2
lxfs69.gsi.de	🟩		🟩	🟩	0	99.84	3.875 GB	2.917 GB	2.788 GB	2.788 GB	8.622 KB/s	62.66 B/s	117	483.2

ALICE::GSI::SE2 architecture

- a box with 2 10 Gb network interfaces
- on one end: Lustre mounted via LNET routers with 10 Gb
- on the other end: connected to LHCONE with 10 Gb
- xroot running on top of Lustre
- via symlinks files are stored also with LFN name ==> usable in a transparent way for local users

GSI – job efficiency



GSI 92 jobs (0.563%)	40.95%	9.926	205 files 0.926 MB/s	205 (100%) 0.926 MB/s

- 41% Job Efficiency is clearly not sufficient !!!
  - Auto SE discovery does not seem to work still
  - 100% of all data are stored at other sites
    - files are stored at CERN::EOS (6%) and NDGF::DCACHE(75%)
    - this needs to be investigated !!!

GSI: next activities

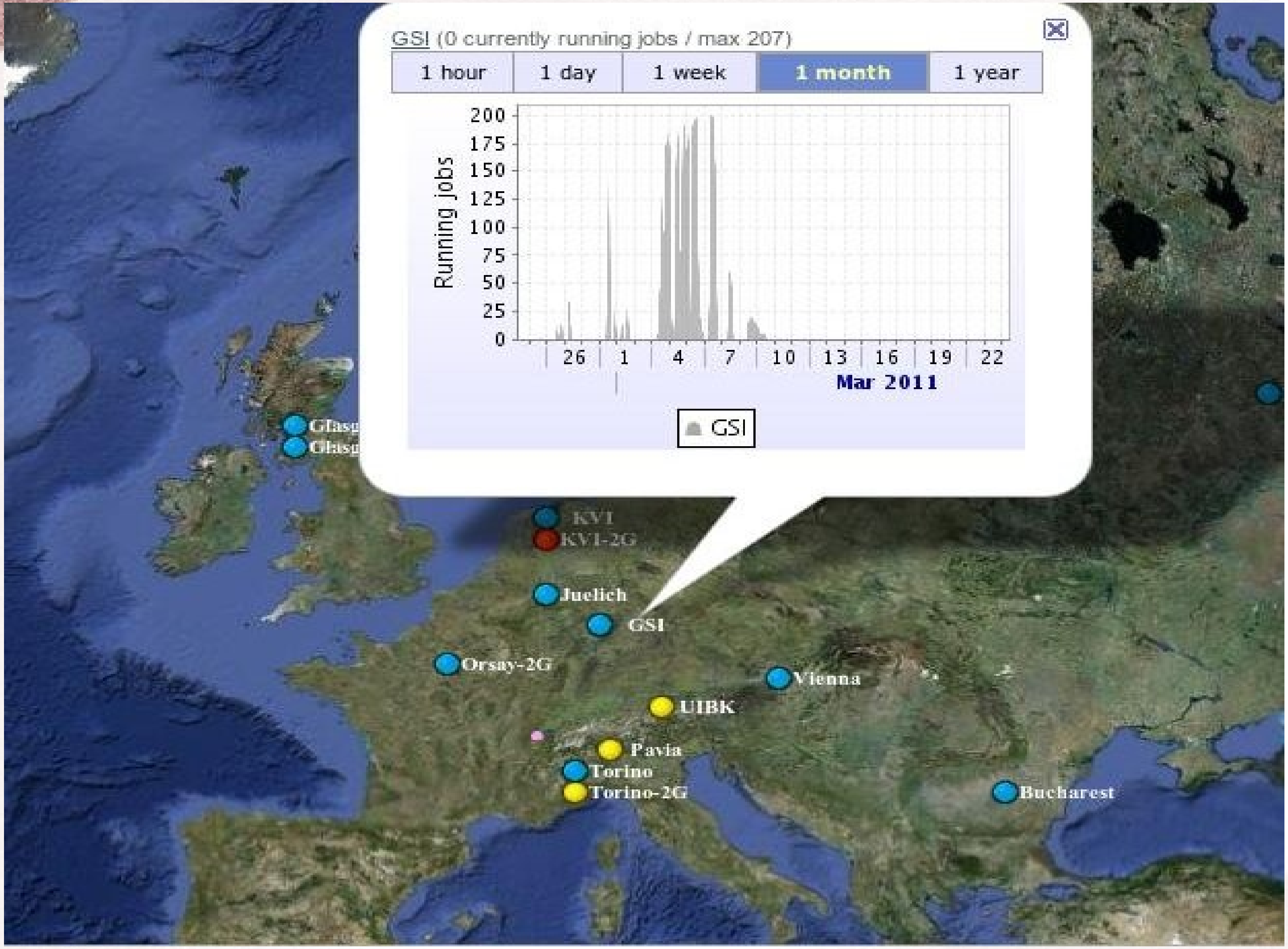
- Fix pending problems
- Try CVMFS for software distribution
- Include Prometheus Cluster in ALICE Grid
- No plans for IPv6 this year

ALICE requirements at GSI T2

GSI (20% of the T2 requirements) 1 core ~ 12 HEPSPEC06

	CPU	**Disk**	**no of cores**
	kHEPSPEC06	PB	
2015	40	3,22	3333
2016	44	4,1	3667
2017	48	4,96	4000
2018	48	4,96	4000
2019	48	4,96	4000

LHC Computing – Prototype for FAIR

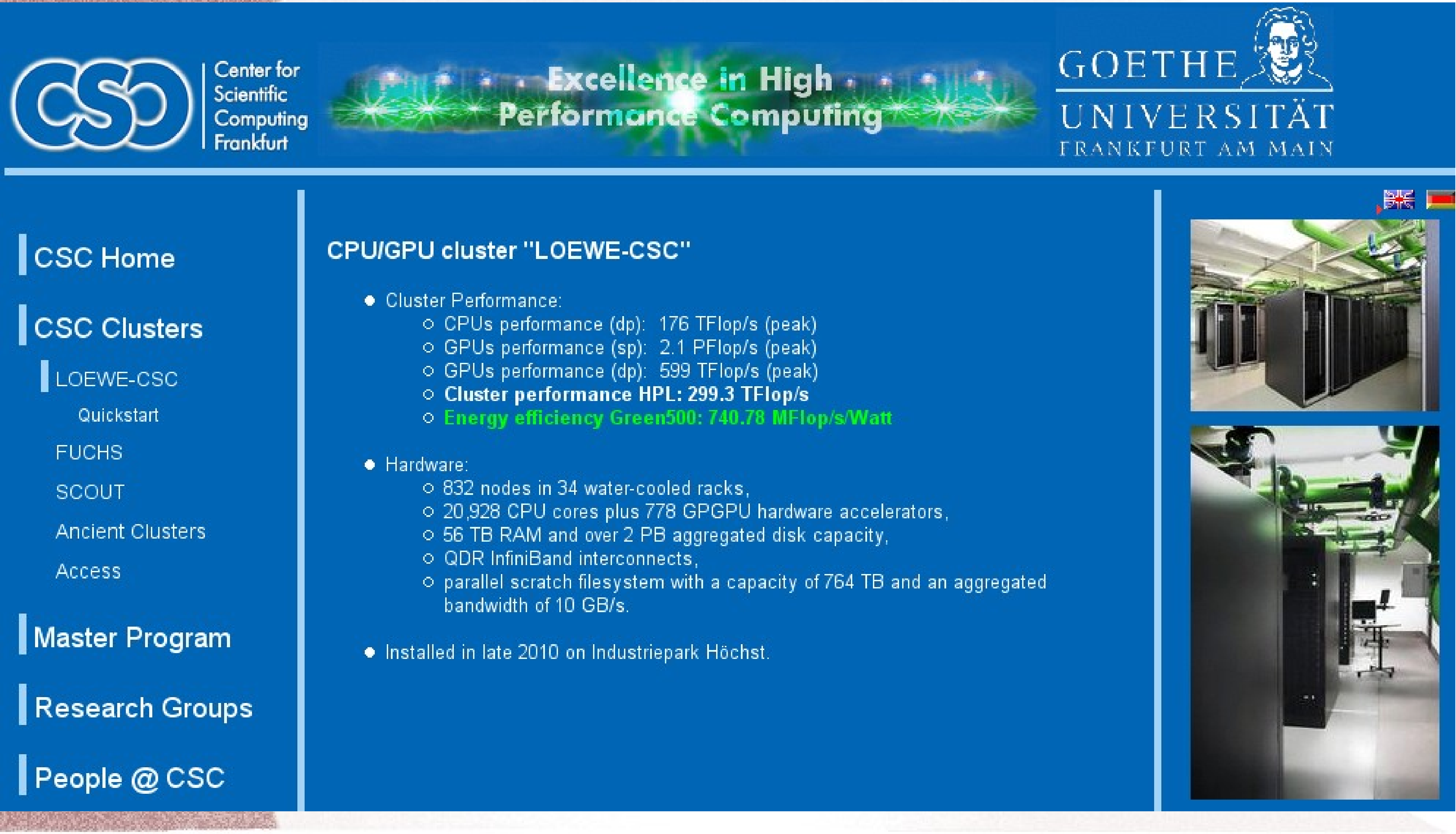


PandaGrid – up since 2004

Table of contents

- Overview
- GridKa T1
- GSI T2
- HHLR-GU
- Summary

(HHLR_GU) Hessisches Hochleistungsrechenzentrum Goethe Universität

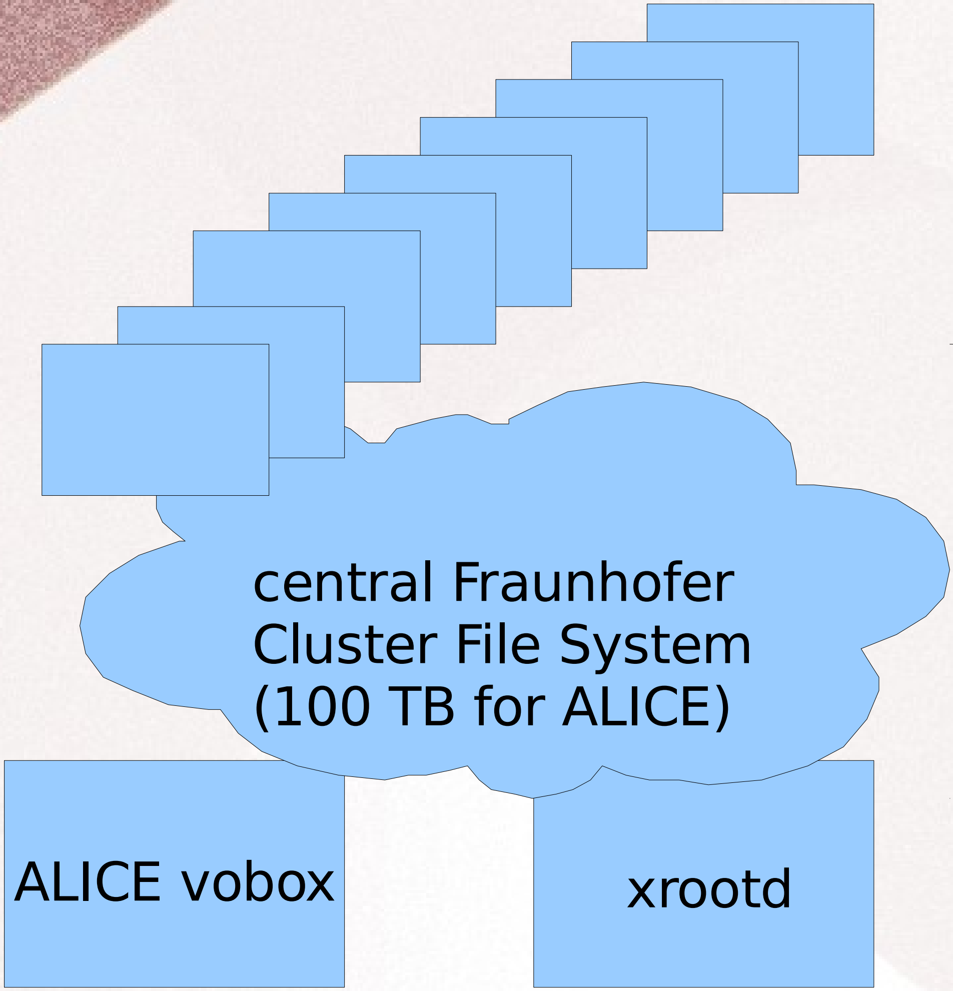

CSC
Center for Scientific Computing Frankfurt
Excellence in High Performance Computing
GOETHE UNIVERSITÄT FRANKFURT AM MAIN
CSC Home
CSC Clusters
LOEWE-CSC
Quickstart
FUCHS
SCOUT
Ancient Clusters
Access
Master Program
Research Groups
People @ CSC

CPU/GPU cluster "LOEWE-CSC"

- Cluster Performance:
  - CPUs performance (dp): 176 TFlop/s (peak)
  - GPUs performance (sp): 2.1 PFlop/s (peak)
  - GPUs performance (dp): 599 TFlop/s (peak)
  - **Cluster performance HPL: 299.3 TFlop/s**
  - **Energy efficiency Green500: 740.78 MFlop/s/Watt**
- Hardware:
  - 832 nodes in 34 water-cooled racks,
  - 20,928 CPU cores plus 778 GPGPU hardware accelerators,
  - 56 TB RAM and over 2 PB aggregated disk capacity,
  - QDR InfiniBand interconnects,
  - parallel scratch filesystem with a capacity of 764 TB and an aggregated bandwidth of 10 GB/s.
- Installed in late 2010 on Industriepark Höchst.

HHLR_GU ALICE setup

1200 cores (now 600) exclusively for ALICE



central Fraunhofer Cluster File System (100 TB for ALICE)

ALICE vobox

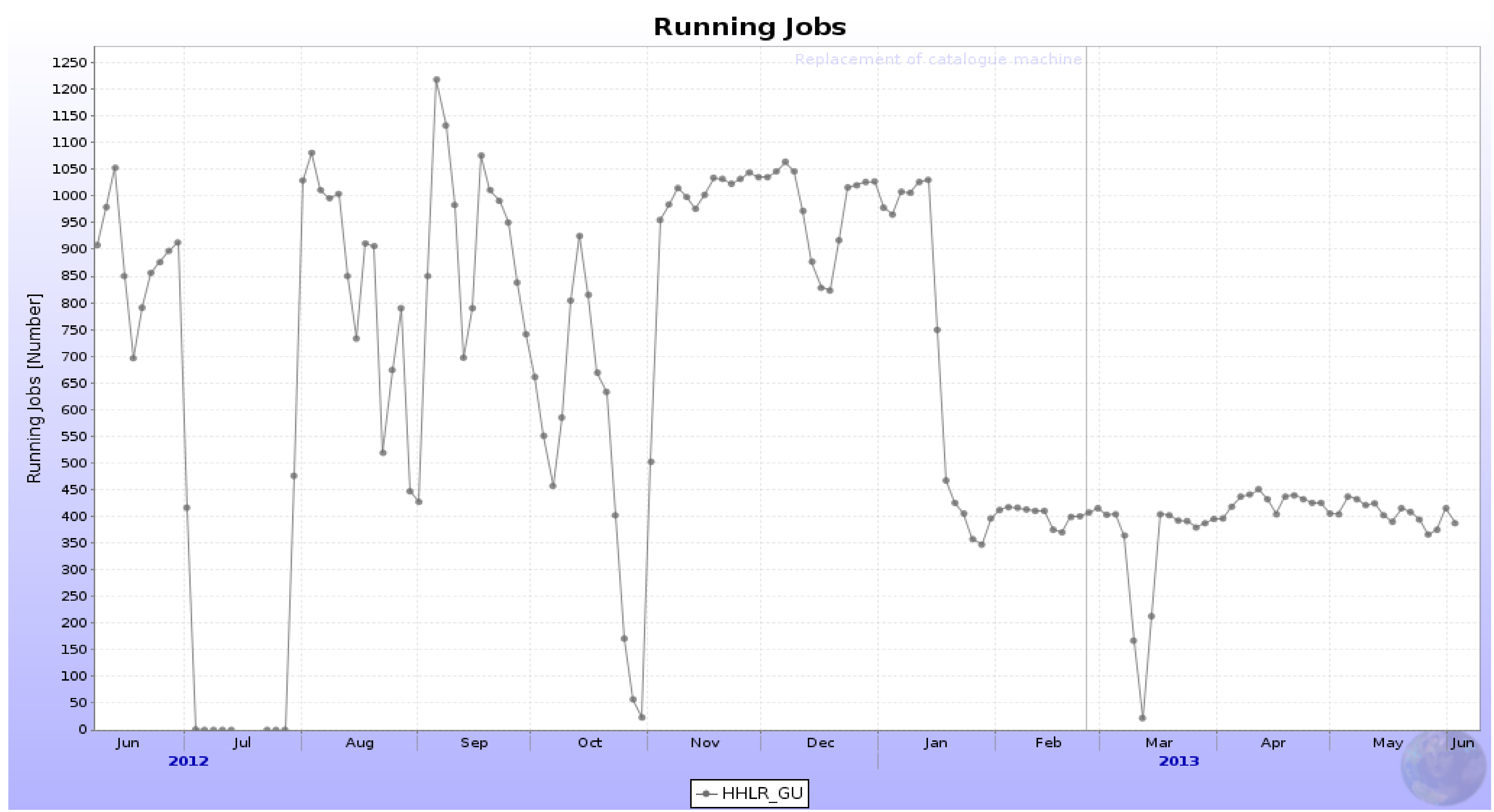
xrootd

800 Mb to DFN

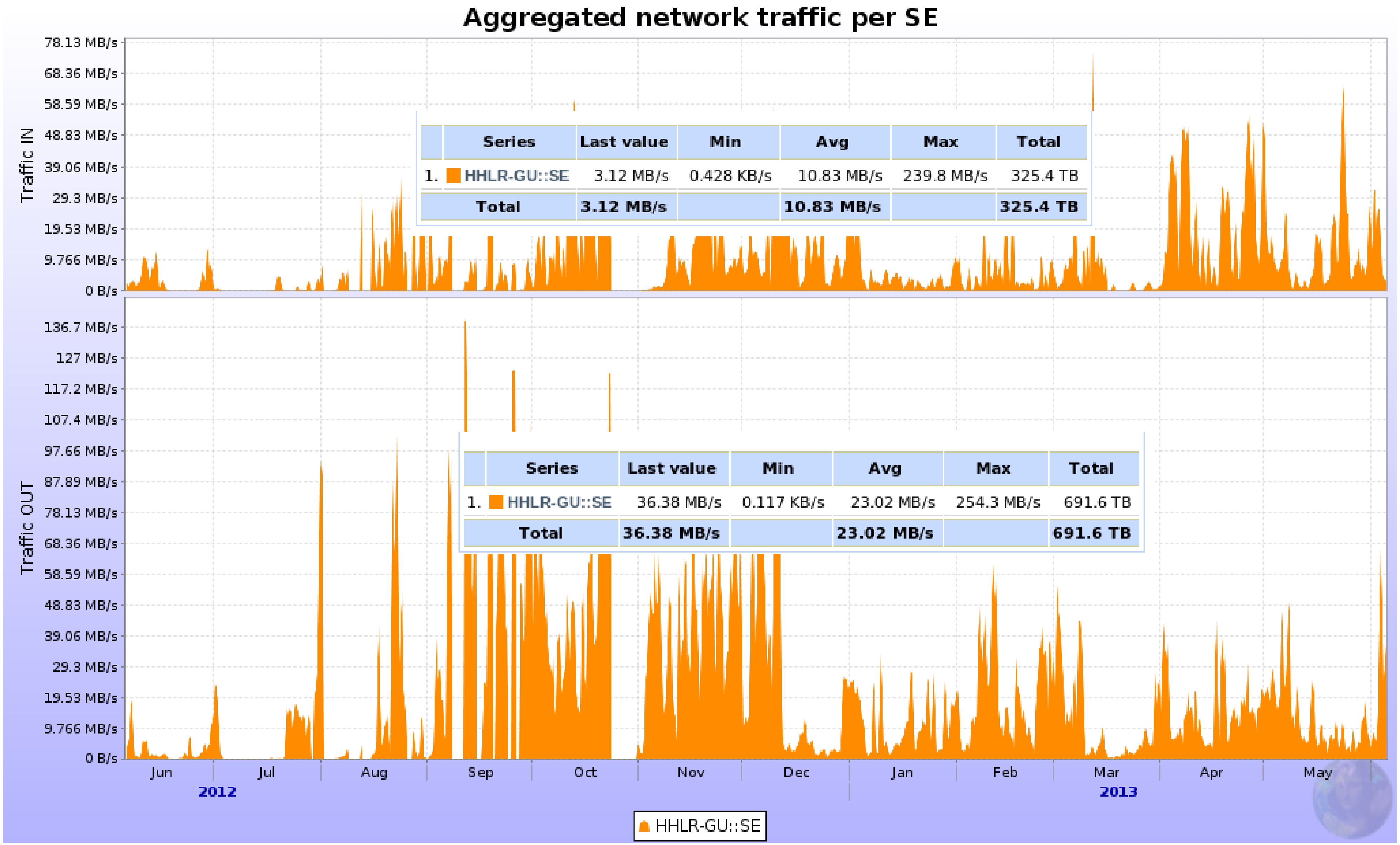
- Job Submission System: Slurm
- for this the native AliEn Slurm interface has been reactivated (A. Montiel Gonzalez)

- Almost continuous operation now
- Beginning of 2013 reduced job number for ALICE
- Files older than 3 months may be removed from fhgfs at some point
- No plans for IPv6

Jobs at Loewe CSC



storage at Loewe CSC



	Series	Last value	Min	Avg	Max	Total
1.	🟧 HHLR-GU::SE	3.12 MB/s	0.428 KB/s	10.83 MB/s	239.8 MB/s	325.4 TB
	Total	**3.12 MB/s**		**10.83 MB/s**		**325.4 TB**

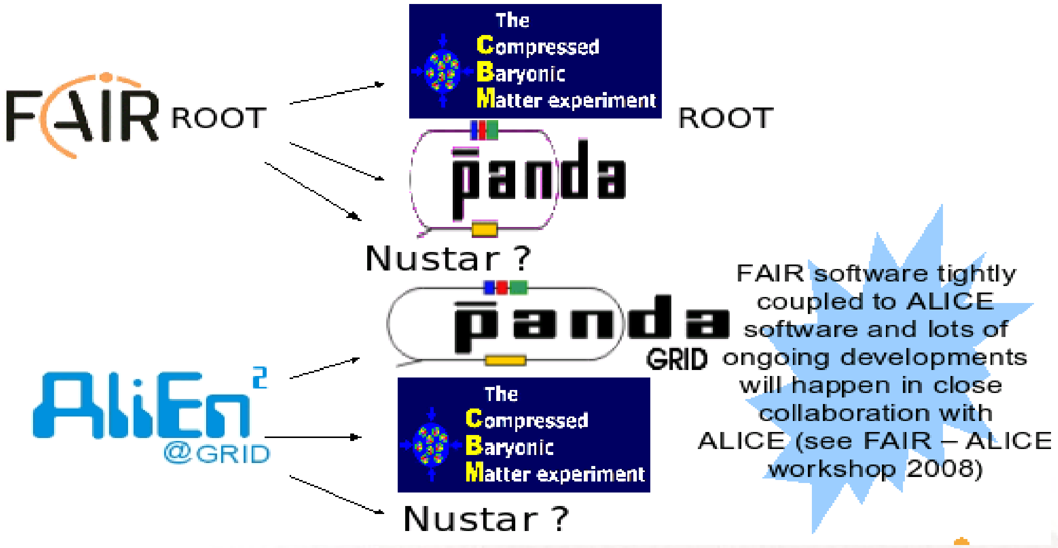
	Series	Last value	Min	Avg	Max	Total
1.	🟧 HHLR-GU::SE	36.38 MB/s	0.117 KB/s	23.02 MB/s	254.3 MB/s	691.6 TB
	Total	**36.38 MB/s**		**23.02 MB/s**		**691.6 TB**

Table of contents

- Overview
- GridKa T1
- GSI T2
- HHLR-GU
- Summary

Summary

- German sites provide a valuable contribution to ALICE Grid
- new developments are on the way
- FAIR will play an increasing role (funding, network architecture, software development and more ...)


FAIR ROOT
The Compressed Baryonic Matter experiment
ROOT
panda
Nustar ?
panda GRID
AliEn2 @GRID
The Compressed Baryonic Matter experiment
Nustar ?

FAIR software tightly coupled to ALICE software and lots of ongoing developments will happen in close collaboration with ALICE (see FAIR – ALICE workshop 2008)