Grid Operations in Germany

Kilian Schwarz

Christopher Jung

Guido Laubender

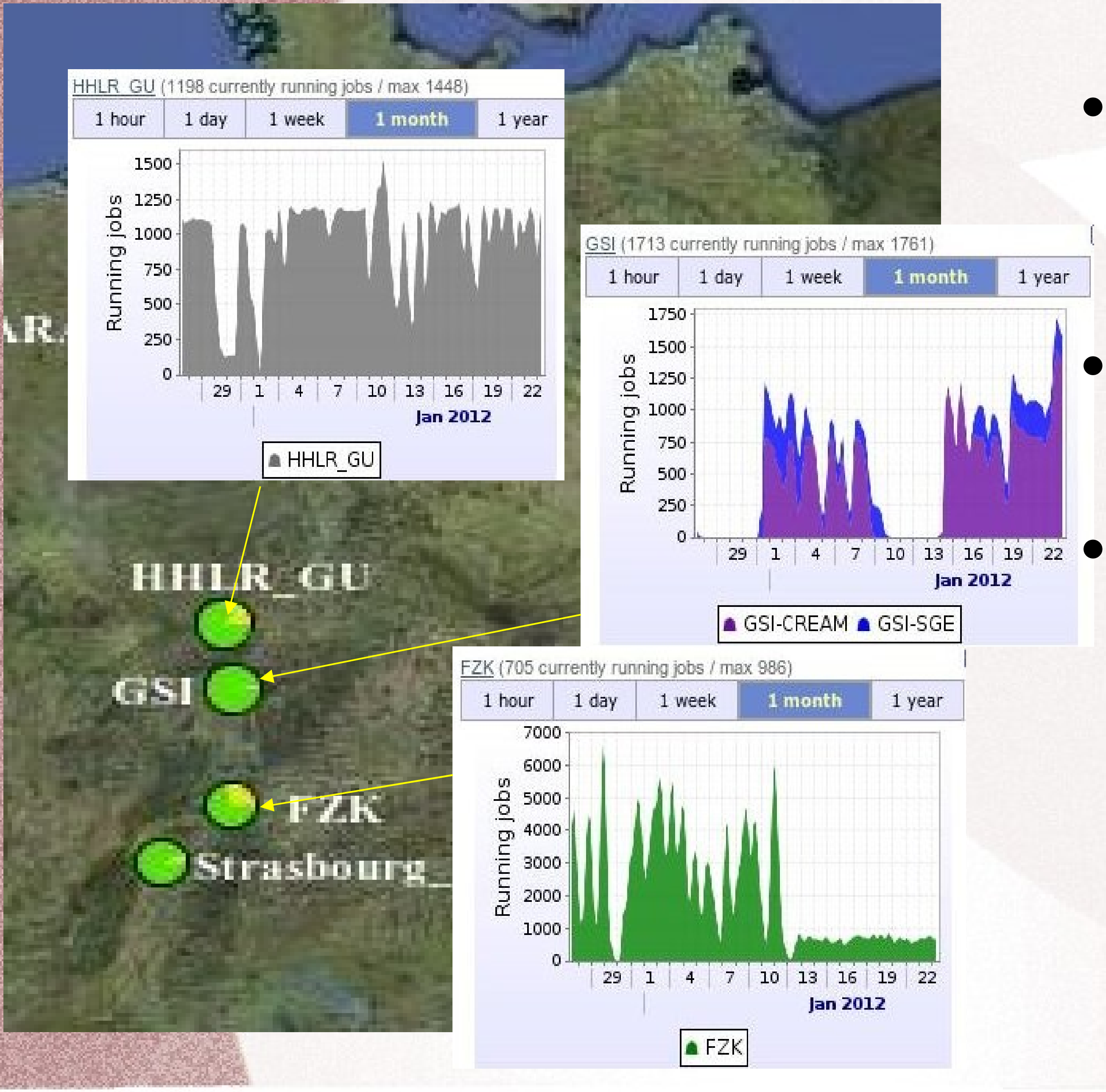
Table of contents

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

Table of contents

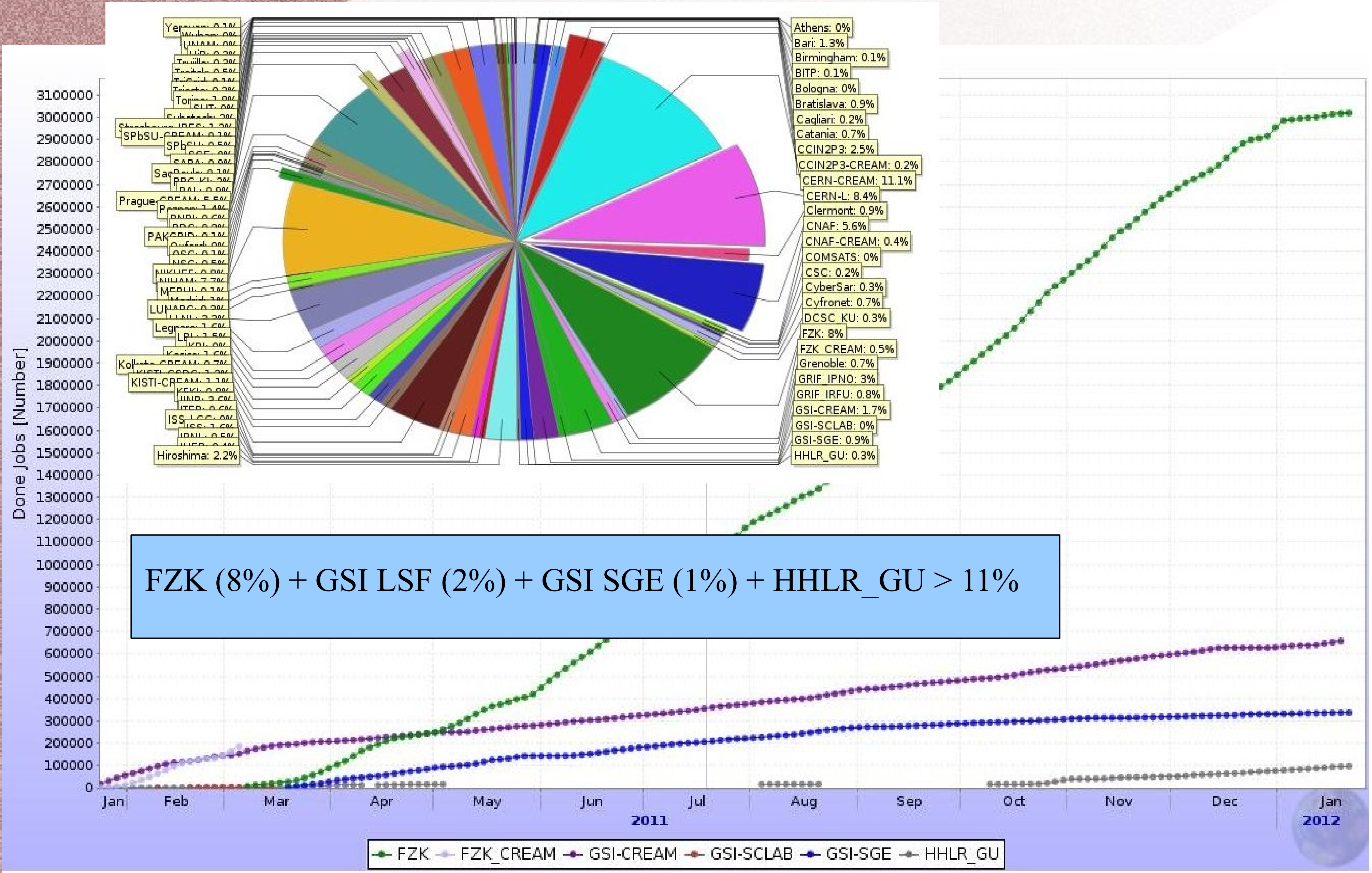
- Overview
- GridKa T1
- GSI T2
- HHLR-GU
- Summary

Map of German Grid sites



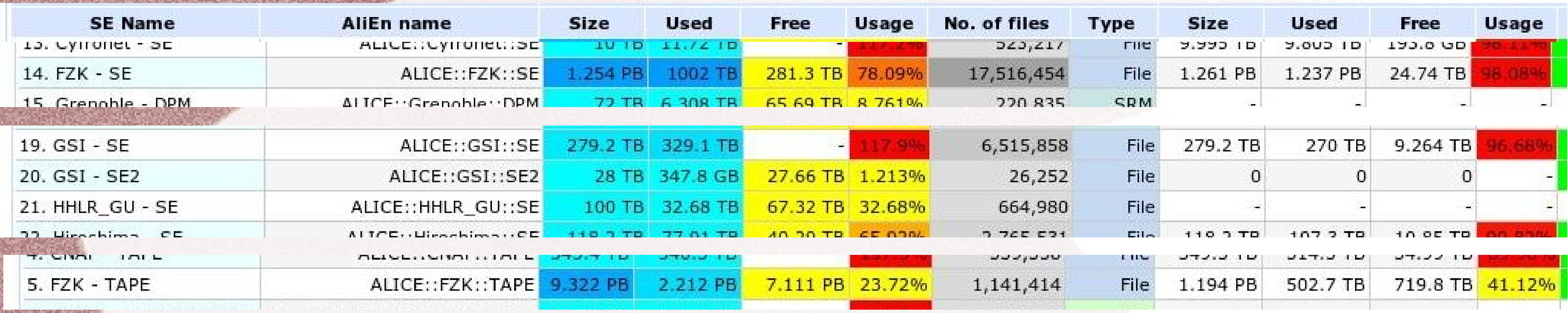
- T1: GridKa/FZK in Karlsruhe
- T2: GSI in Darmstadt
- HHLR_GU in Frankfurt

Job contribution (last year)



FZK (8%) + GSI LSF (2%) + GSI SGE (1%) + HHLR_GU > 11%

Storage contribution



SE Name	AliEn name	Size	Used	Free	Usage	No. of files	Type	Size	Used	Free	Usage
13. Cyfronet - SE	ALICE::Cyfronet::SE	10 TB	11.72 TB	-	117.2%	523,217	File	9.995 TB	9.803 TB	193.8 GB	98.11%
14. FZK - SE	ALICE::FZK::SE	1.254 PB	1002 TB	281.3 TB	78.09%	17,516,454	File	1.261 PB	1.237 PB	24.74 TB	98.08%
15. Grenoble - DPM	ALICE::Grenoble::DPM	72 TB	6.308 TB	65.69 TB	8.761%	220,835	SRM	-	-	-	-
19. GSI - SE	ALICE::GSI::SE	279.2 TB	329.1 TB	-	117.9%	6,515,858	File	279.2 TB	270 TB	9.264 TB	96.68%
20. GSI - SE2	ALICE::GSI::SE2	28 TB	347.8 GB	27.66 TB	1.213%	26,252	File	0	0	0	-
21. HHLR_GU - SE	ALICE::HHLR_GU::SE	100 TB	32.68 TB	67.32 TB	32.68%	664,980	File	-	-	-	-
22. Hiroshima - SE	ALICE::Hiroshima::SE	118.2 TB	77.01 TB	40.20 TB	65.02%	2,765,531	File	118.2 TB	107.3 TB	10.85 TB	[illegible]
4. CNAF - TAPE	ALICE::CNAF::TAPE	[illegible]	[illegible]		[illegible]	[illegible]	File	[illegible]	[illegible]	[illegible]	[illegible]
5. FZK - TAPE	ALICE::FZK::TAPE	9.322 PB	2.212 PB	7.111 PB	23.72%	1,141,414	File	1.194 PB	502.7 TB	719.8 TB	41.12%

Total size:

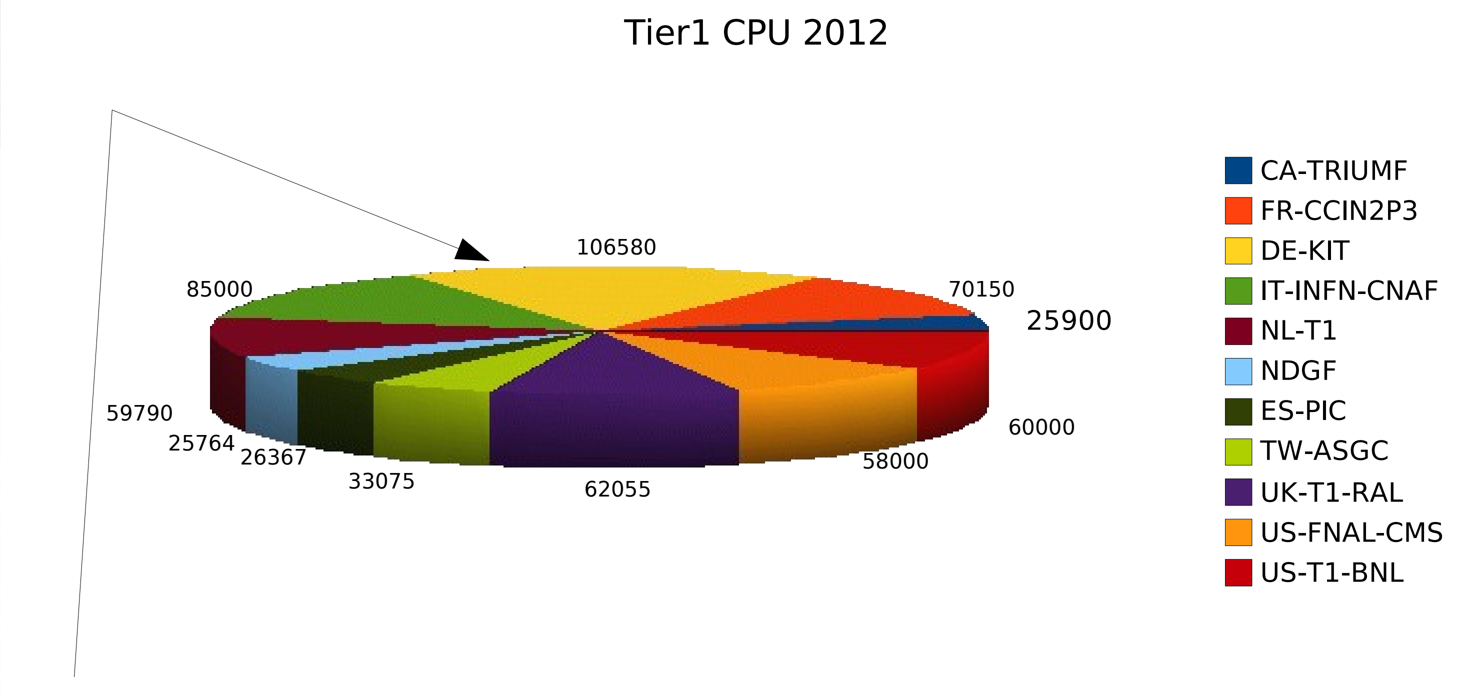
- 1.7 PB disk based SE (ALICE total: 13.2 PB)
- 1.2 PB disk buffer with Tape backend

Table of contents

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

WLCG Tier-1	CPU (HS06)	Disk	Tape
2012	553'000	67 PB	103 PB

Tier-1: GridKa

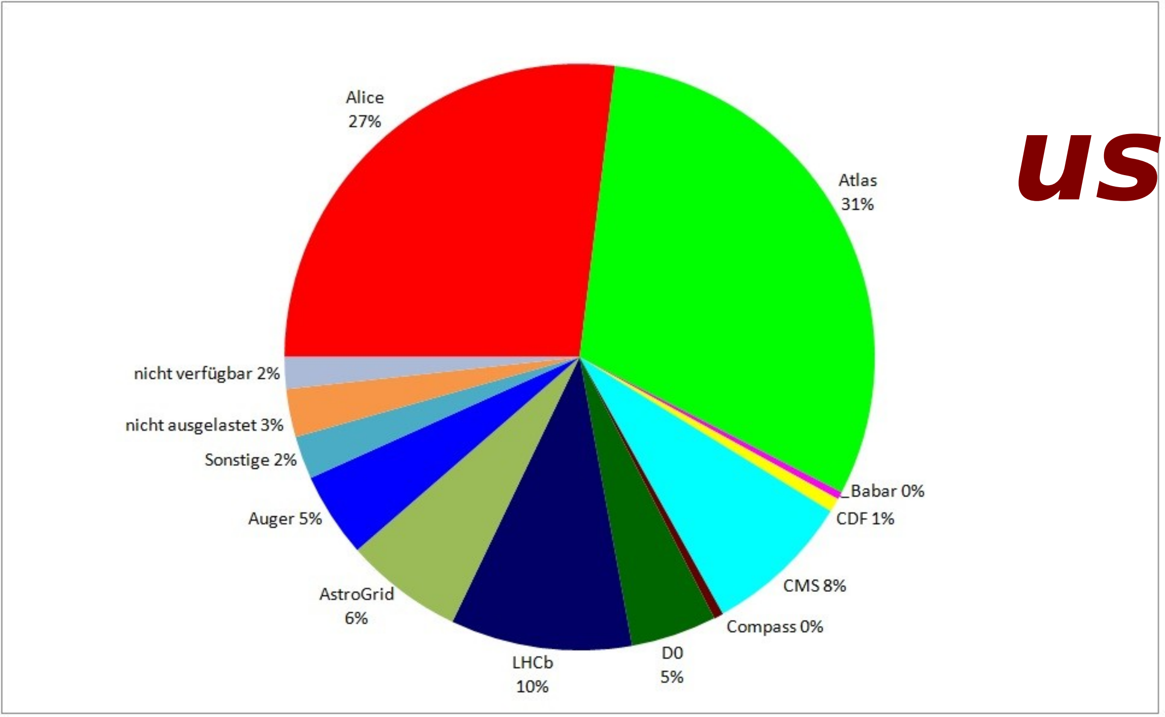


GridKa is the largest Tier1 in WLCG and provides about 15% of the total T1 recources

GridKa:	CPU (HS06)	%WLCG	Disk	%WLCG	Tape	% WLCG
ALICE :	40000	25%	2,7 PB	25%	5,2 PB	25%
ATLAS:	32400	12.5%	3,4 PB	12,5%	4,5 PB	12,5%
CMS:	24000	10%	2,2 PB	10%	5,1 PB	10%
LHCb:	19200	17%	1,6 PB	17%	1,6 PB	17%

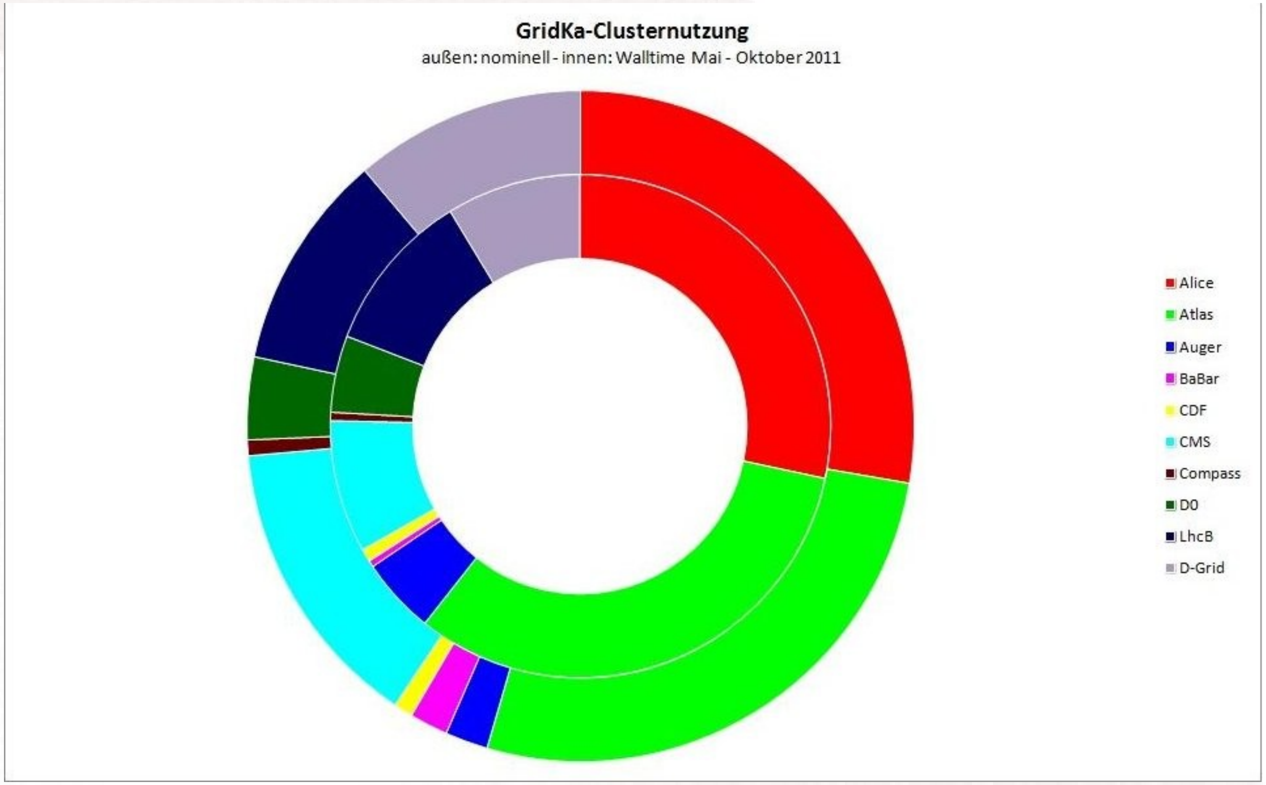
usage statistics

(last 6 months)



Centre is well used. 5% not available or non used.
Largest shares: LHC experiments.
(ALICE and ATLAS alone > 50%)

ALICE, ATLAS, LHCb, CDF, and D0
are using roughly their nominal share.

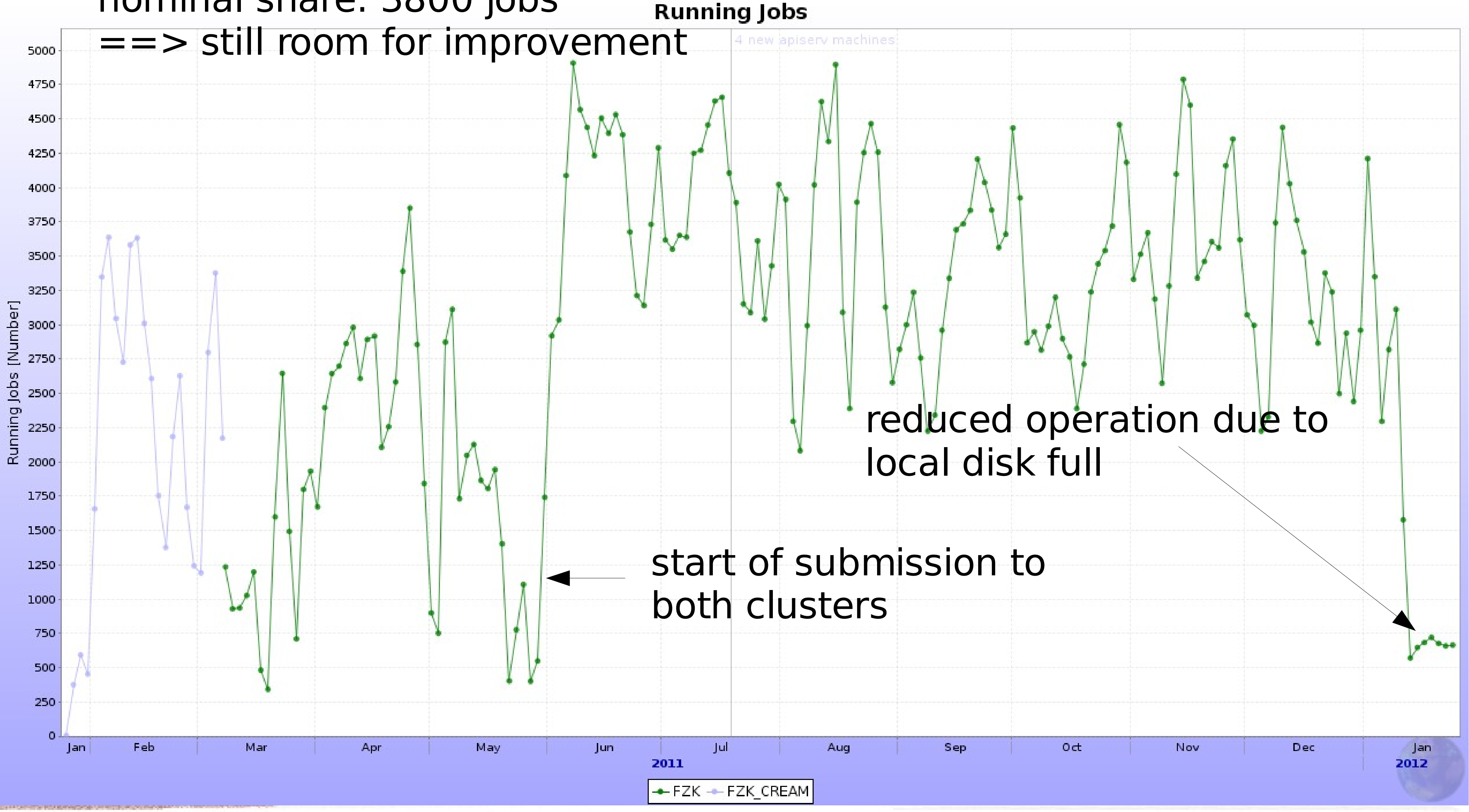


Batch Submission

- OS: SL5
- Used Batch System: PBSPro
- due to PBS problems in supporting large clusters division into 2 sub clusters a 8500 cores (ALICE nominal share: 30%) and 4200 cores (ALICE nominal share: 35%).
  - Fair share values are computed daily. Current values for ALICE: 24%(30%) and 34%(35%).
- Submission via CREAM CE to both clusters
- LDAP config: CE_LCGCE=(cream-1-fzk.gridka.de:8443/cream-pbs-aliceXL,cream-3-fzk.gridka.de:8443/cream-pbs-aliceXL,cream-5-kit.gridka.de:8443/cream-pbs-aliceXL),(cream-2-fzk.gridka.de:8443/cream-pbs-aliceXL,cream-4-kit.gridka.de:8443/cream-pbs-aliceXL)

Jobs at GridKa within last year

max. number of concurrent jobs: 9260
average number of jobs: 3000
average job number in last 6 months: 3200
nominal share: 3800 jobs
==> still room for improvement

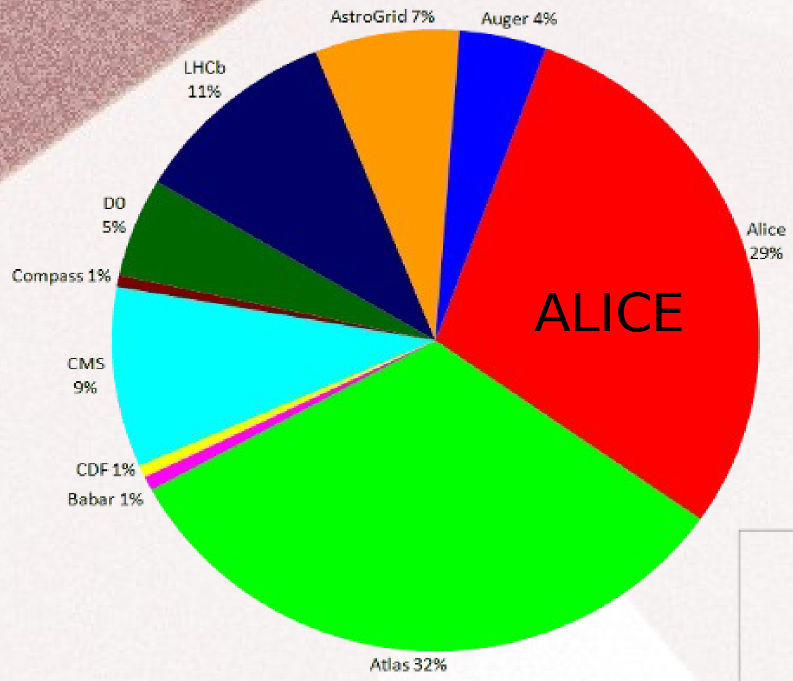


start of submission to both clusters

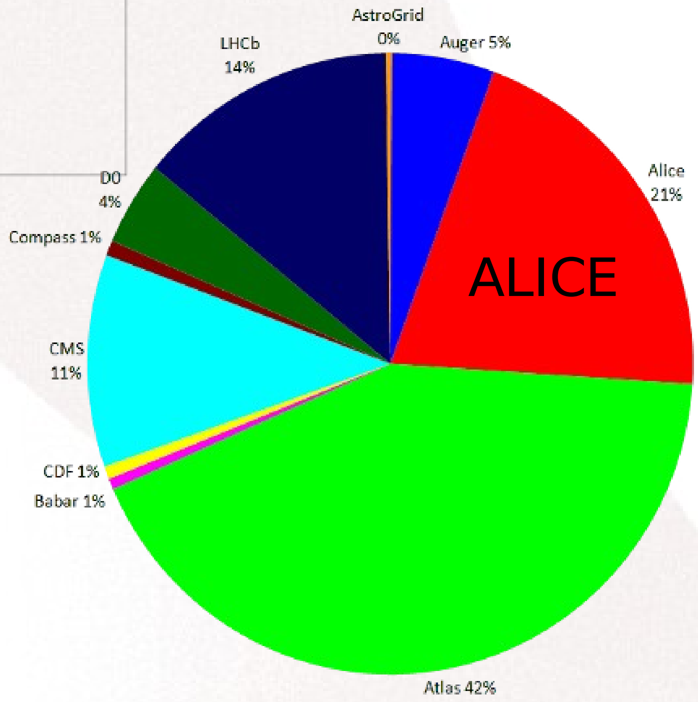
reduced operation due to local disk full

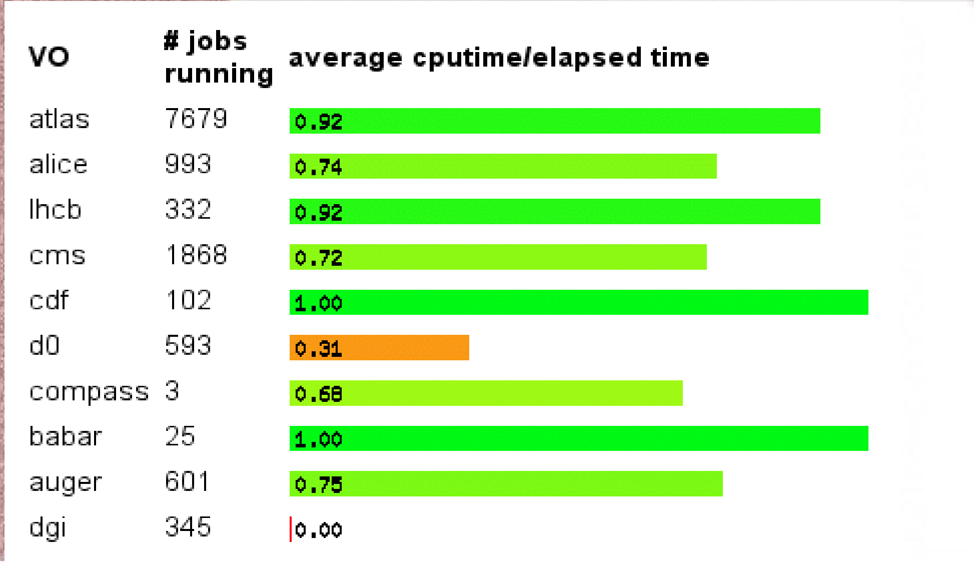
ALICE Job Efficiency

Wall Time 2011



CPU Time 2011

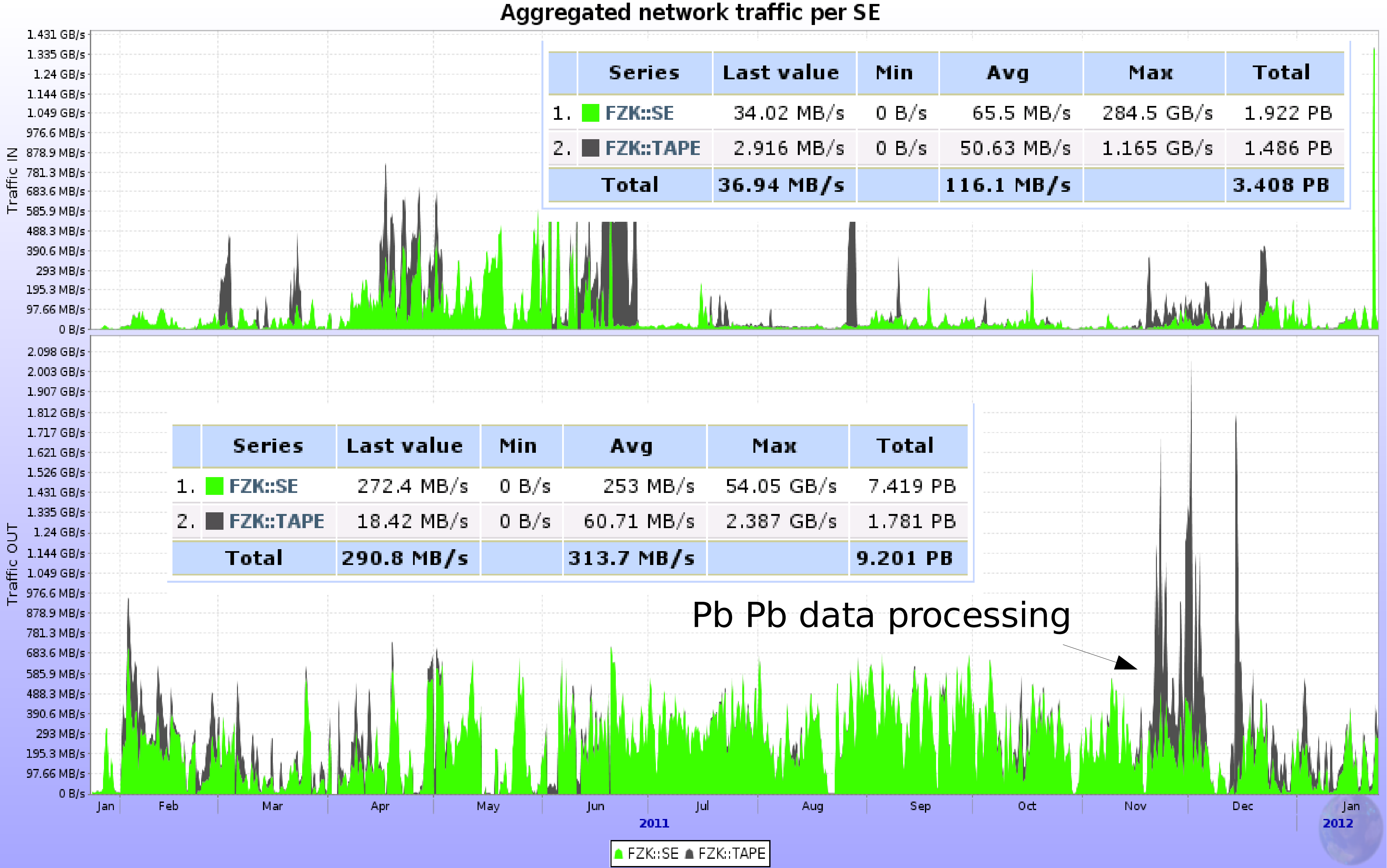




VO	# jobs running	average cputime/elapsed time
atlas	7679	0.92
alice	993	0.74
lhcb	332	0.92
cms	1868	0.72
cdf	102	1.00
d0	593	0.31
compass	3	0.68
babar	25	1.00
auger	601	0.75
dgi	345	0.00

xrootd SE works well and is heavily used

storage



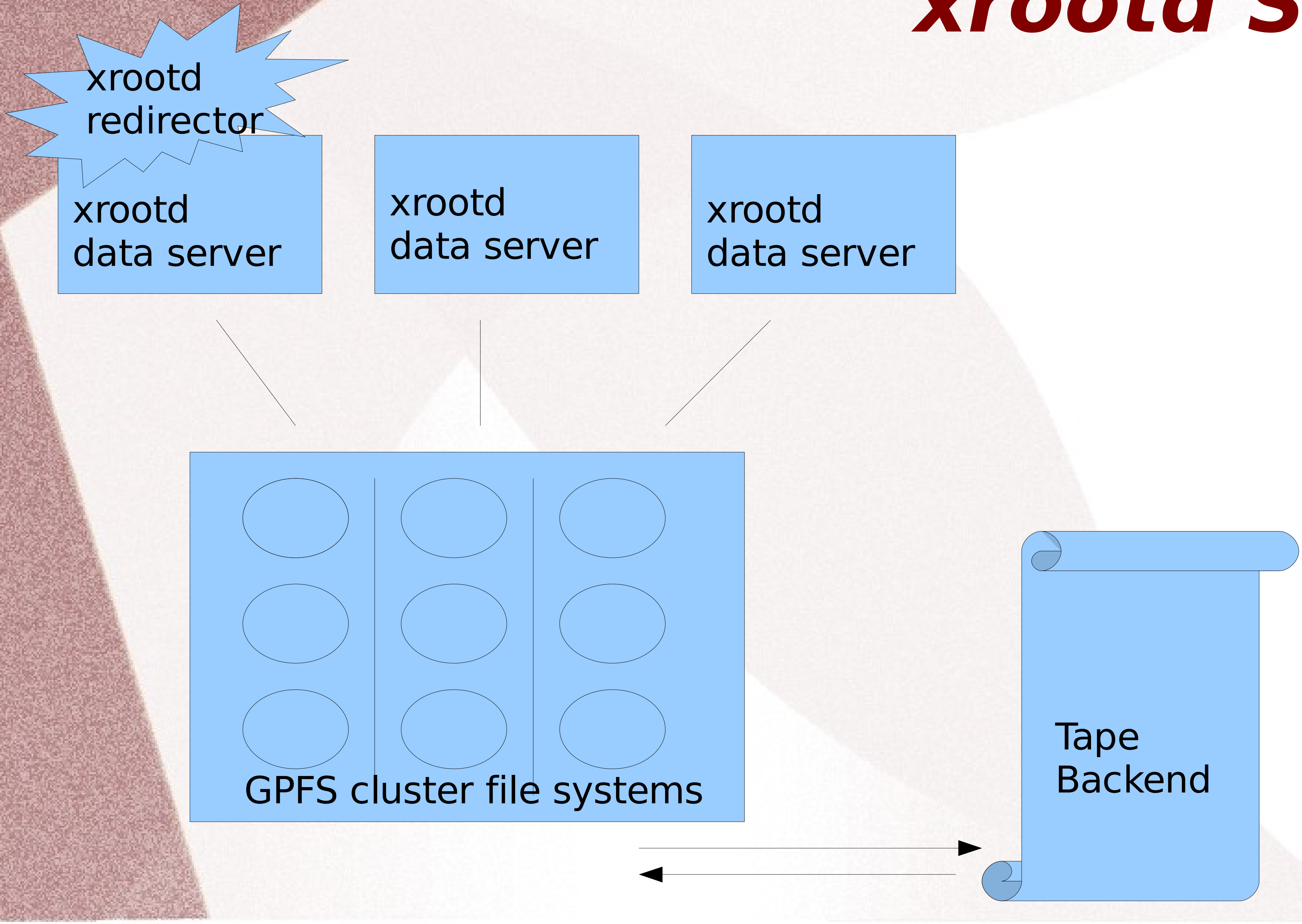
Aggregated network traffic per SE

	Series	Last value	Min	Avg	Max	Total
1.	🟩 FZK::SE	34.02 MB/s	0 B/s	65.5 MB/s	284.5 GB/s	1.922 PB
2.	⬛ FZK::TAPE	2.916 MB/s	0 B/s	50.63 MB/s	1.165 GB/s	1.486 PB
	Total	**36.94 MB/s**		**116.1 MB/s**		**3.408 PB**

	Series	Last value	Min	Avg	Max	Total
1.	🟩 FZK::SE	272.4 MB/s	0 B/s	253 MB/s	54.05 GB/s	7.419 PB
2.	⬛ FZK::TAPE	18.42 MB/s	0 B/s	60.71 MB/s	2.387 GB/s	1.781 PB
	Total	**290.8 MB/s**		**313.7 MB/s**		**9.201 PB**

Pb Pb data processing

architecture of xrootd SE



xrootd redirector

xrootd data server

xrootd data server

xrootd data server

GPFS cluster file systems

Tape Backend

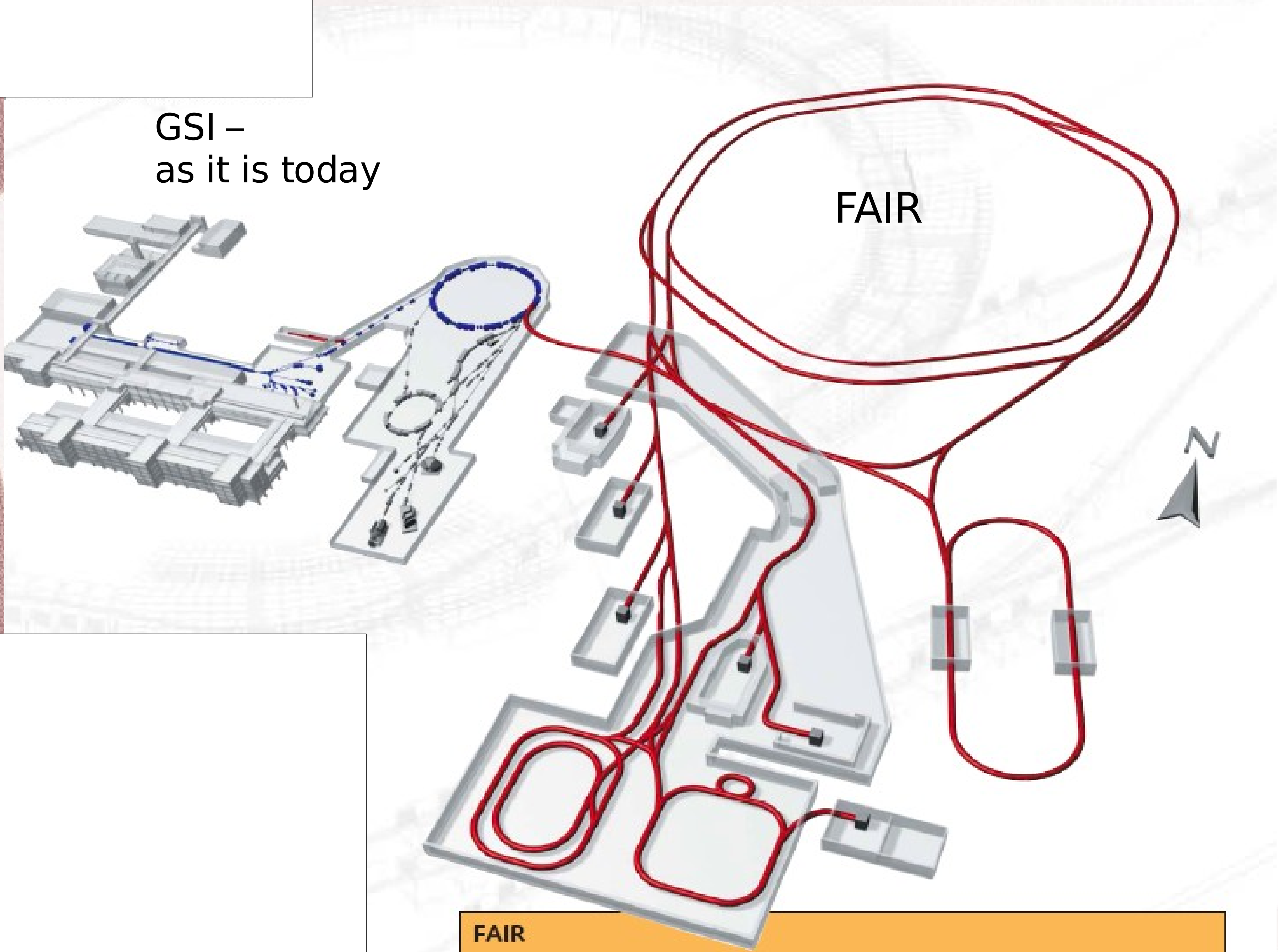
Table of contents

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

Gesellschaft für Schwerionenforschung mbH (GSI)

employs about 1000 people

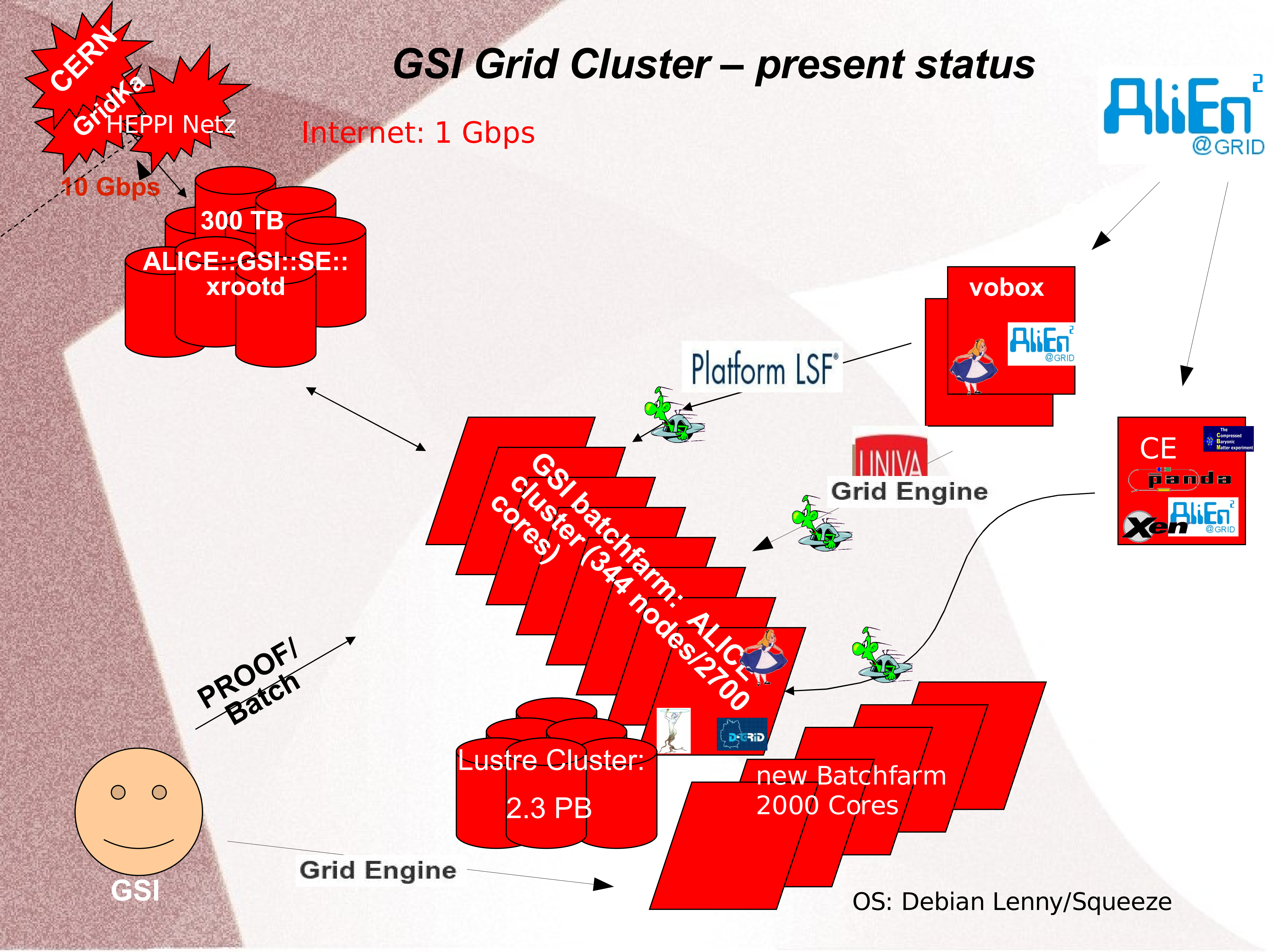
FAIR – *Facility for Antiproton and Ion Research*



GSI –
as it is today

FAIR

GSI Grid Cluster – present status



CERN

GridKa

HEPPI Netz

Internet: 1 Gbps

10 Gbps

300 TB

ALICE::GSI::SE::
xrootd

Platform LSF

vobox

UNIVA
Grid Engine

CE

GSI batchfarm: ALICE cluster (344 nodes/2700 cores)

PROOF/
Batch

Lustre Cluster:
2.3 PB

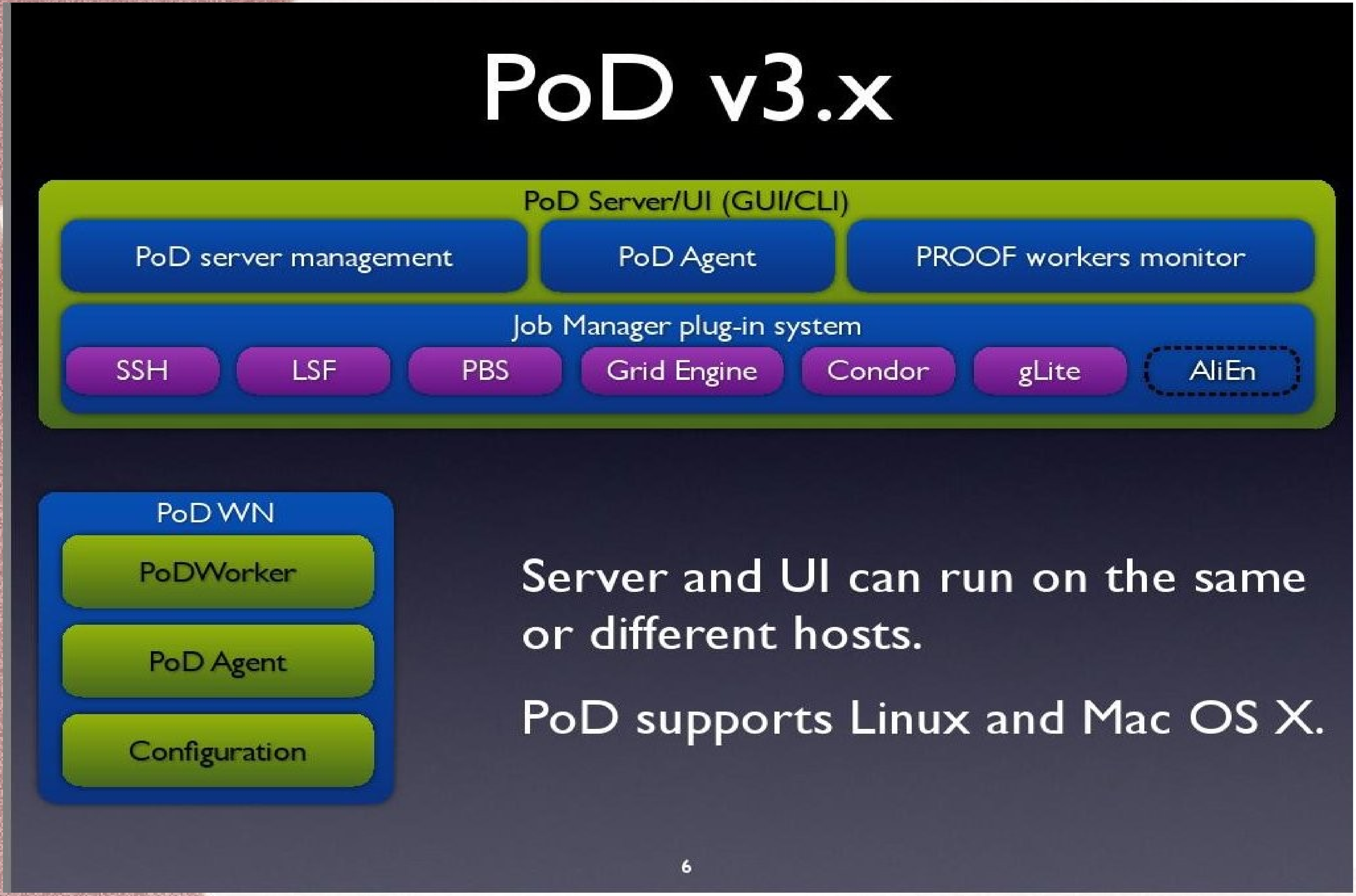
new Batchfarm
2000 Cores

GSI

Grid Engine

OS: Debian Lenny/Squeeze

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

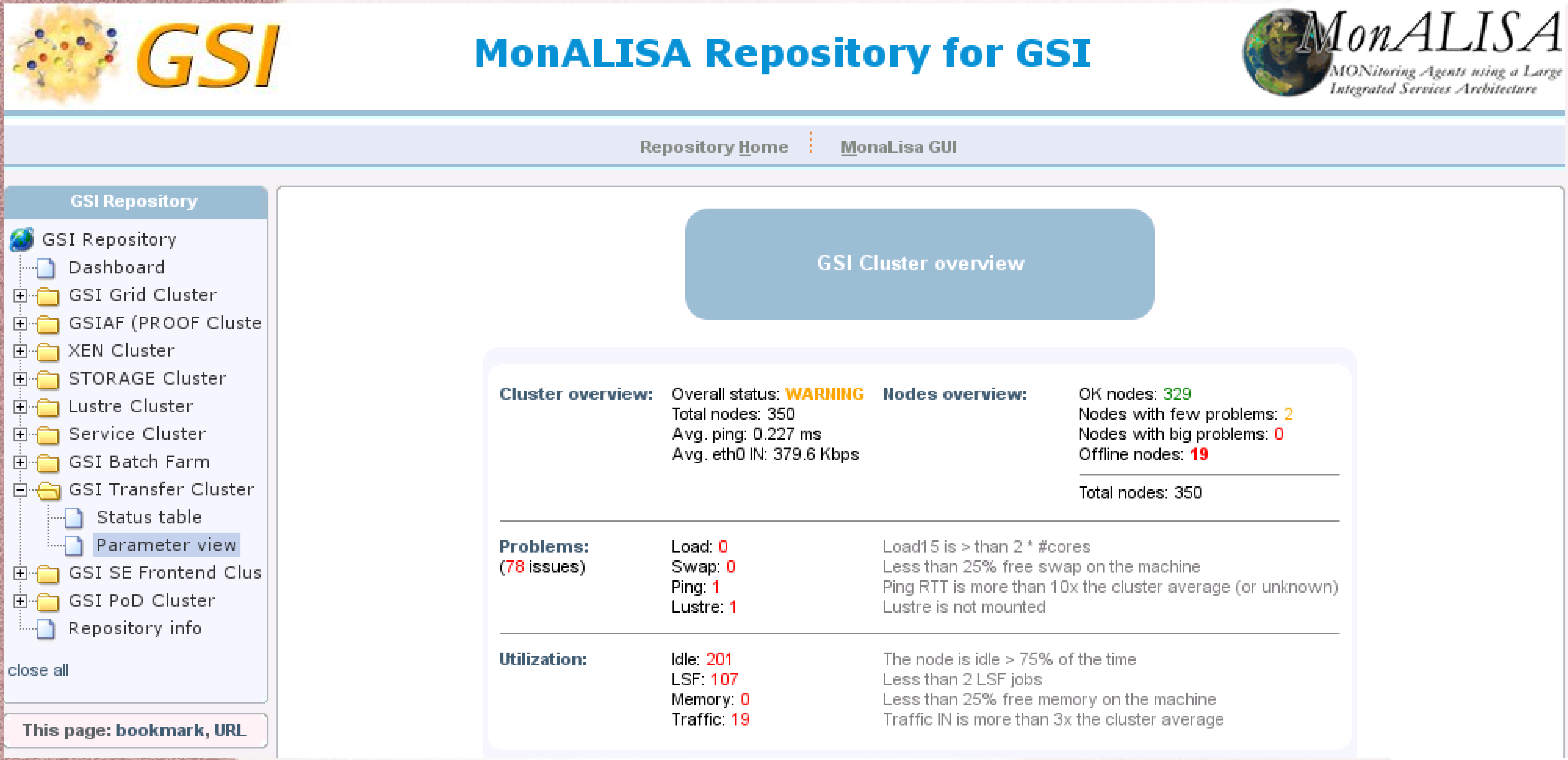
PoD Agent

Configuration

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

PoD supports Linux and Mac OS X.

Farm monitoring via MonaLisa



MonALISA Repository for GSI

GSI Cluster overview

Cluster overview:	Overall status: WARNING Total nodes: 350 Avg. ping: 0.227 ms Avg. eth0 IN: 379.6 Kbps	**Nodes overview:**	OK nodes: 329 Nodes with few problems: 2 Nodes with big problems: 0 Offline nodes: **19** Total nodes: 350
Problems: (78 issues)	Load: 0 Swap: 0 Ping: 1 Lustre: 1	Load15 is > than 2 * #cores Less than 25% free swap on the machine Ping RTT is more than 10x the cluster average (or unknown) Lustre is not mounted	
Utilization:	Idle: 201 LSF: 107 Memory: 0 Traffic: 19	The node is idle > 75% of the time Less than 2 LSF jobs Less than 25% free memory on the machine Traffic IN is more than 3x the cluster average	

GSI SCLAB: Grid site in a Cloud

GSI Cloud:

- Debian Lenny as host OS
- KVM as virtual machine hypervisor
- libvirt (virtualisation API) as abstraction layer above
- OpenNebula toolkit for building the cloud
- 16 physical boxes ==> 100 virtual machines in parallel

AliEn Grid site:

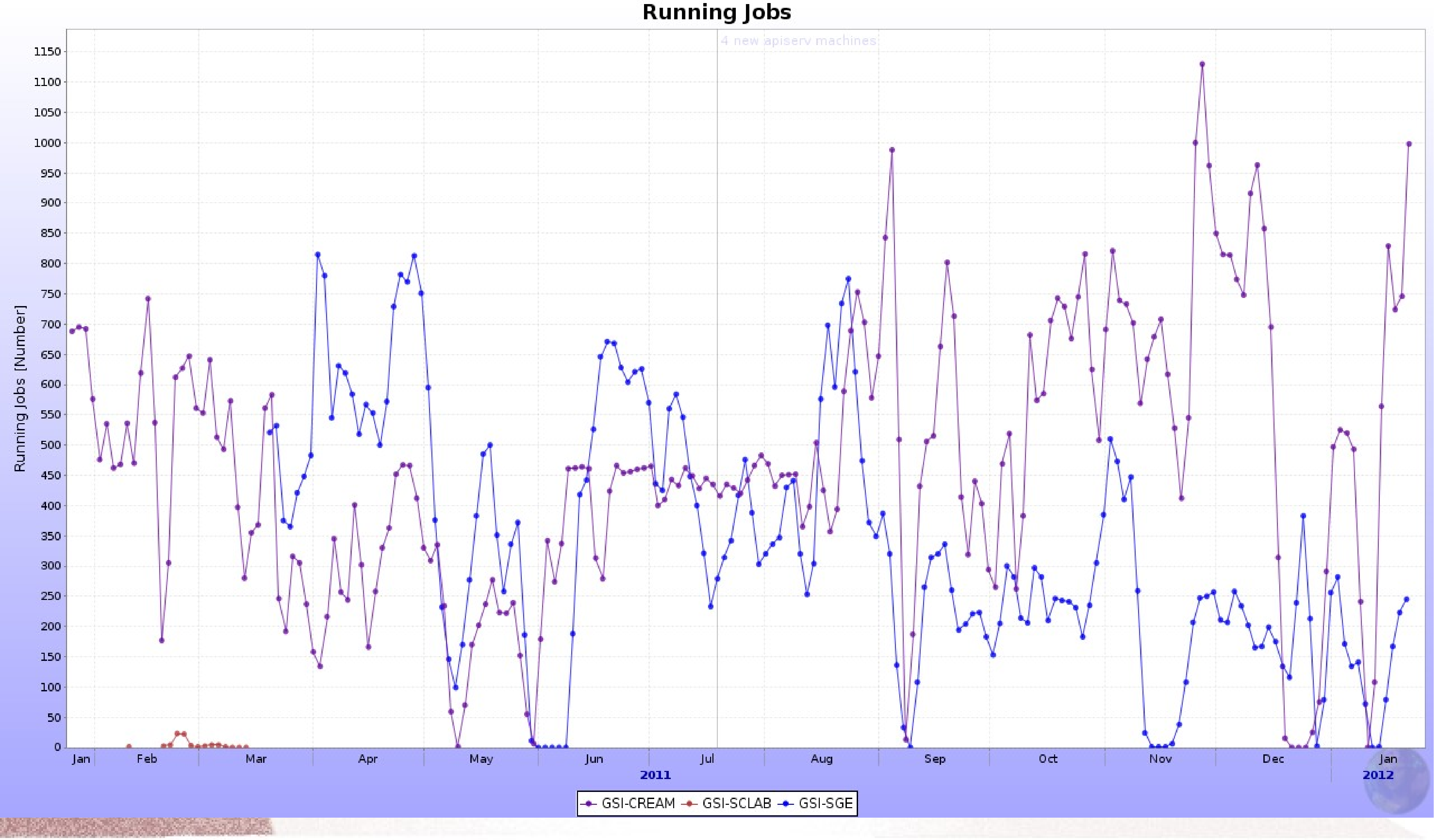
- all jobs run on virtual SL5 machines
- no shared directories
- software packages are installed and distributed using AliEn PackMan and BitTorrent

Grid site in a cloud:

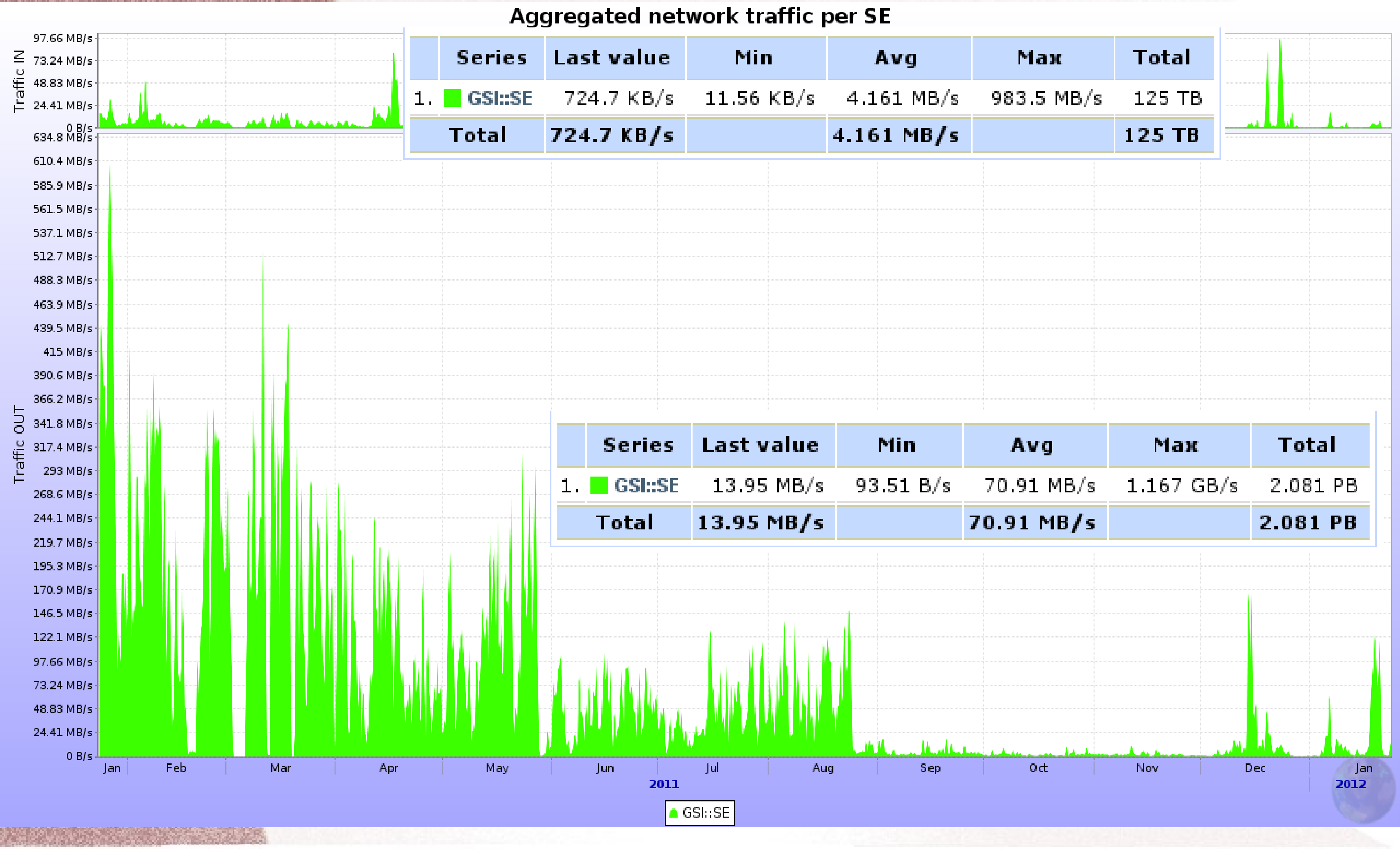
- prepare to be able to startup an AliEn Grid site in any available Cloud

jobs at GSI within last year

average: 800 concurrent jobs



ALICE::GSI::SE



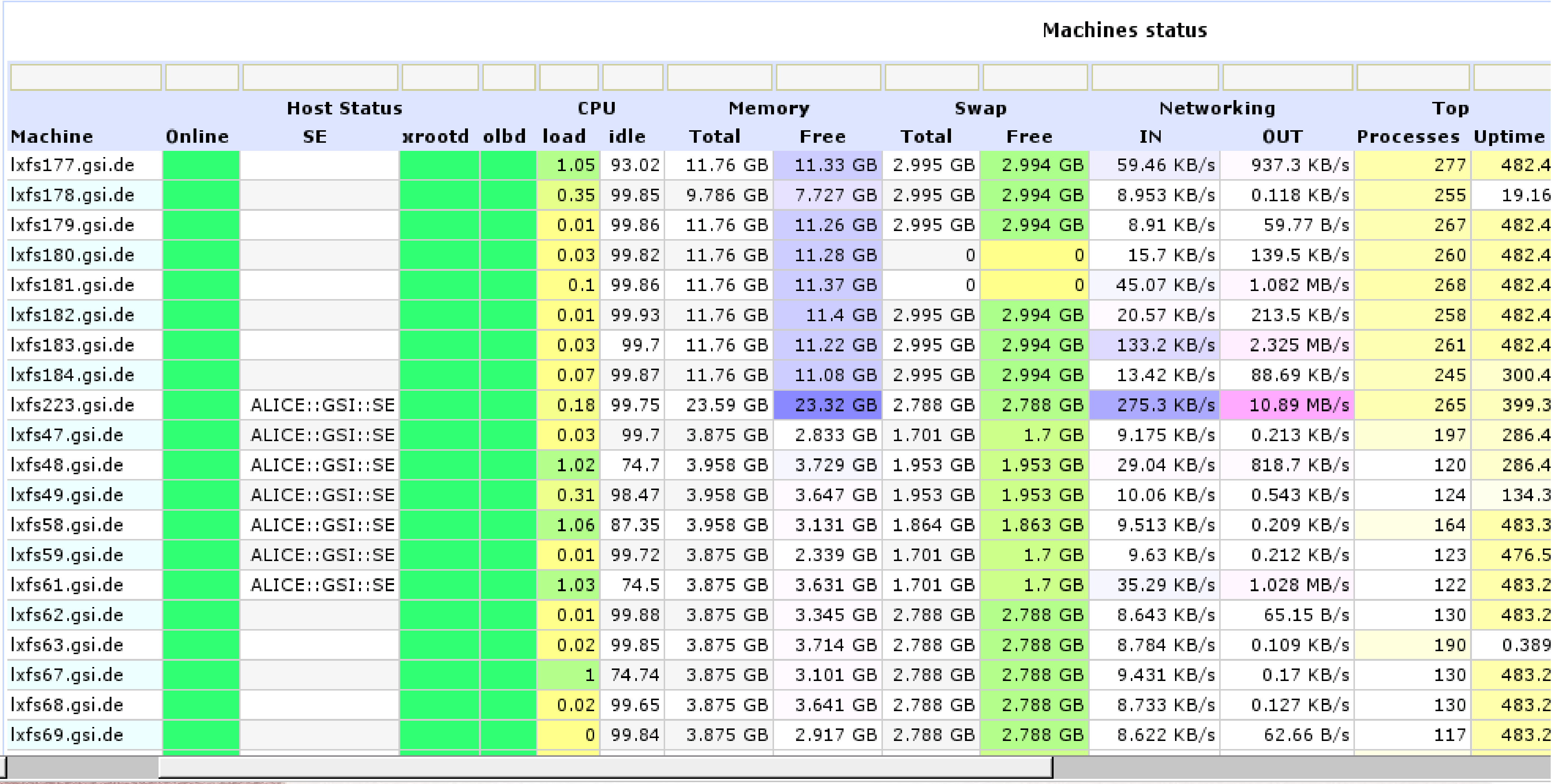
Aggregated network traffic per SE

	Series	Last value	Min	Avg	Max	Total
1.	🟩 GSI::SE	724.7 KB/s	11.56 KB/s	4.161 MB/s	983.5 MB/s	125 TB
	Total	**724.7 KB/s**		**4.161 MB/s**		**125 TB**

	Series	Last value	Min	Avg	Max	Total
1.	🟩 GSI::SE	13.95 MB/s	93.51 B/s	70.91 MB/s	1.167 GB/s	2.081 PB
	Total	**13.95 MB/s**		**70.91 MB/s**		**2.081 PB**

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

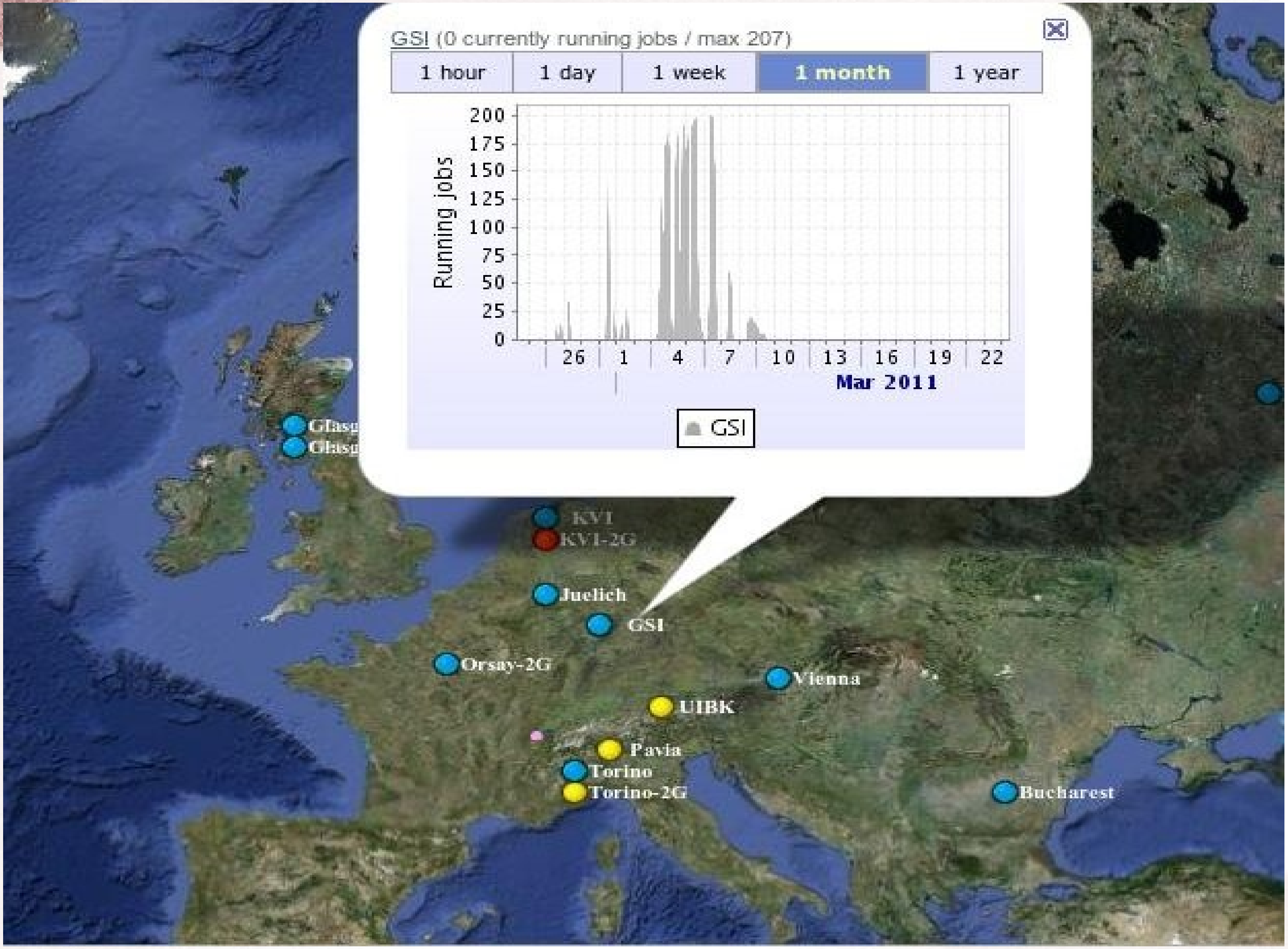
Machines status

		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

GSI: next activities

- include new SGE cluster (2000 cores) in the Grid
- setup new SE on top of Lustre file system with xrd-dm plugin
  - Lustre has currently 270 TB free space and this needs to be shared with local users
  - no quotas enabled

LHC Computing – Prototype for FAIR

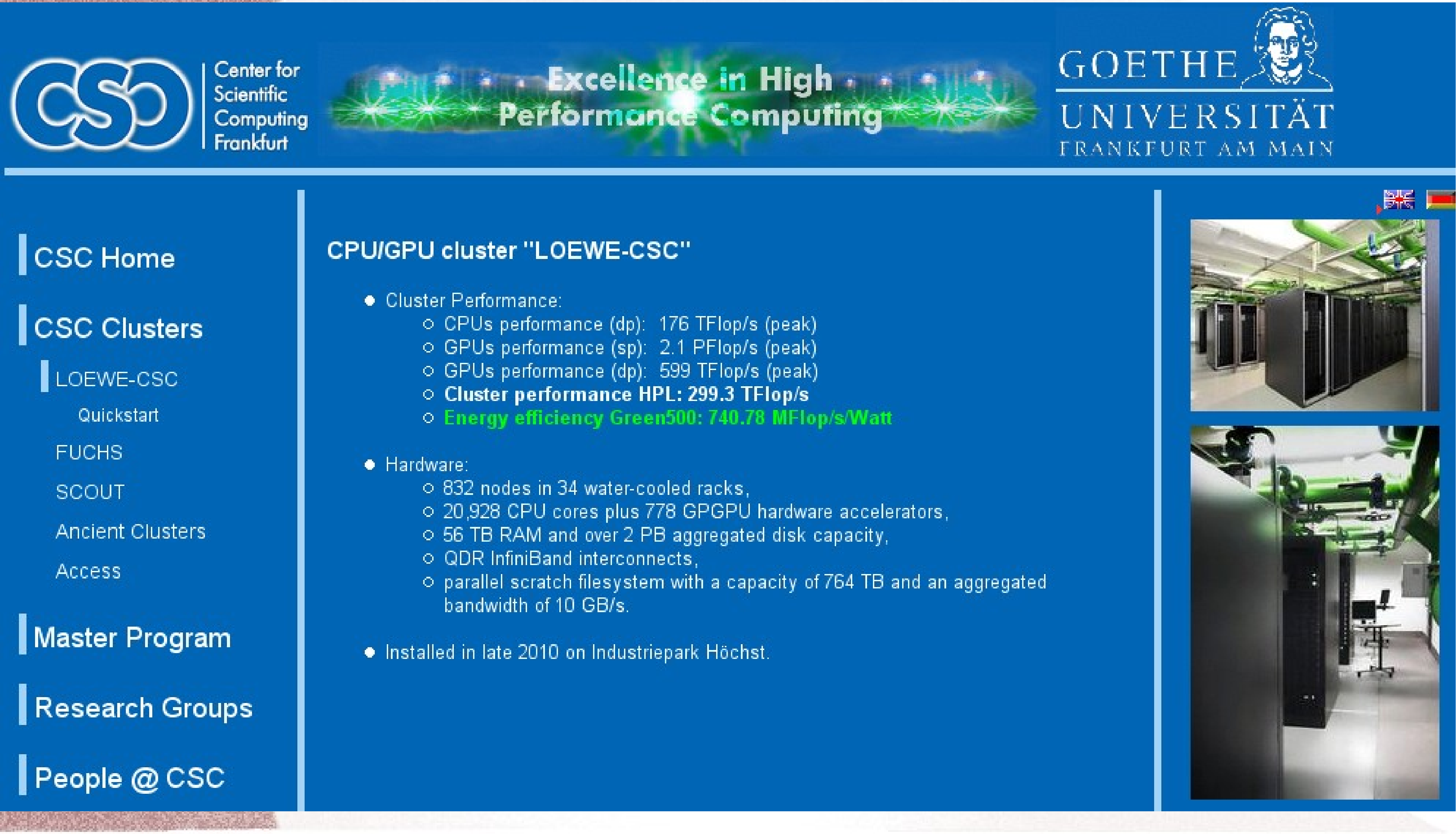


PandaGrid – up since 2004

Table of contents

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

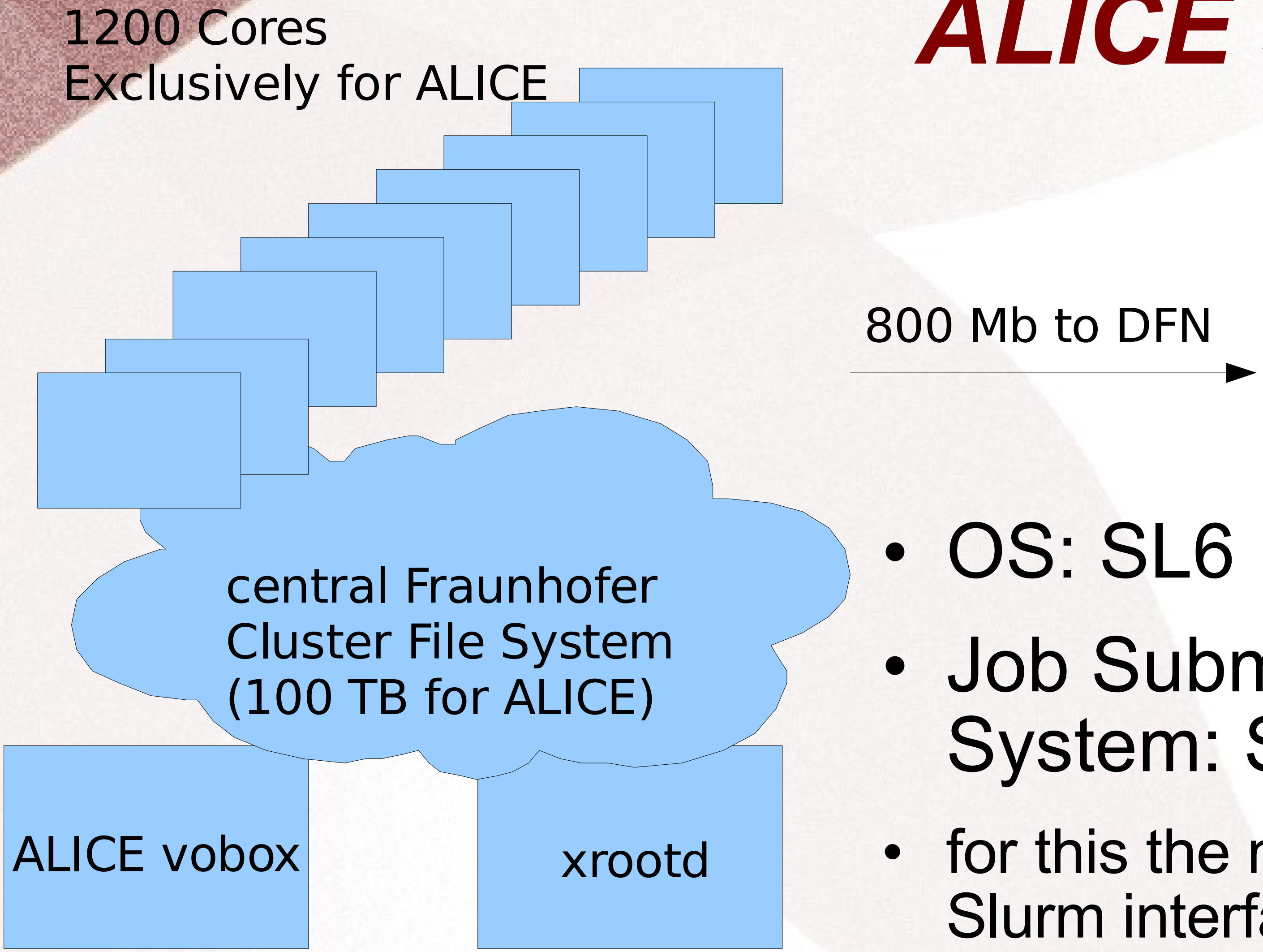
(HHLR_GU) Hessisches Hochleistungsrechenzentrum Goethe Universität


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.

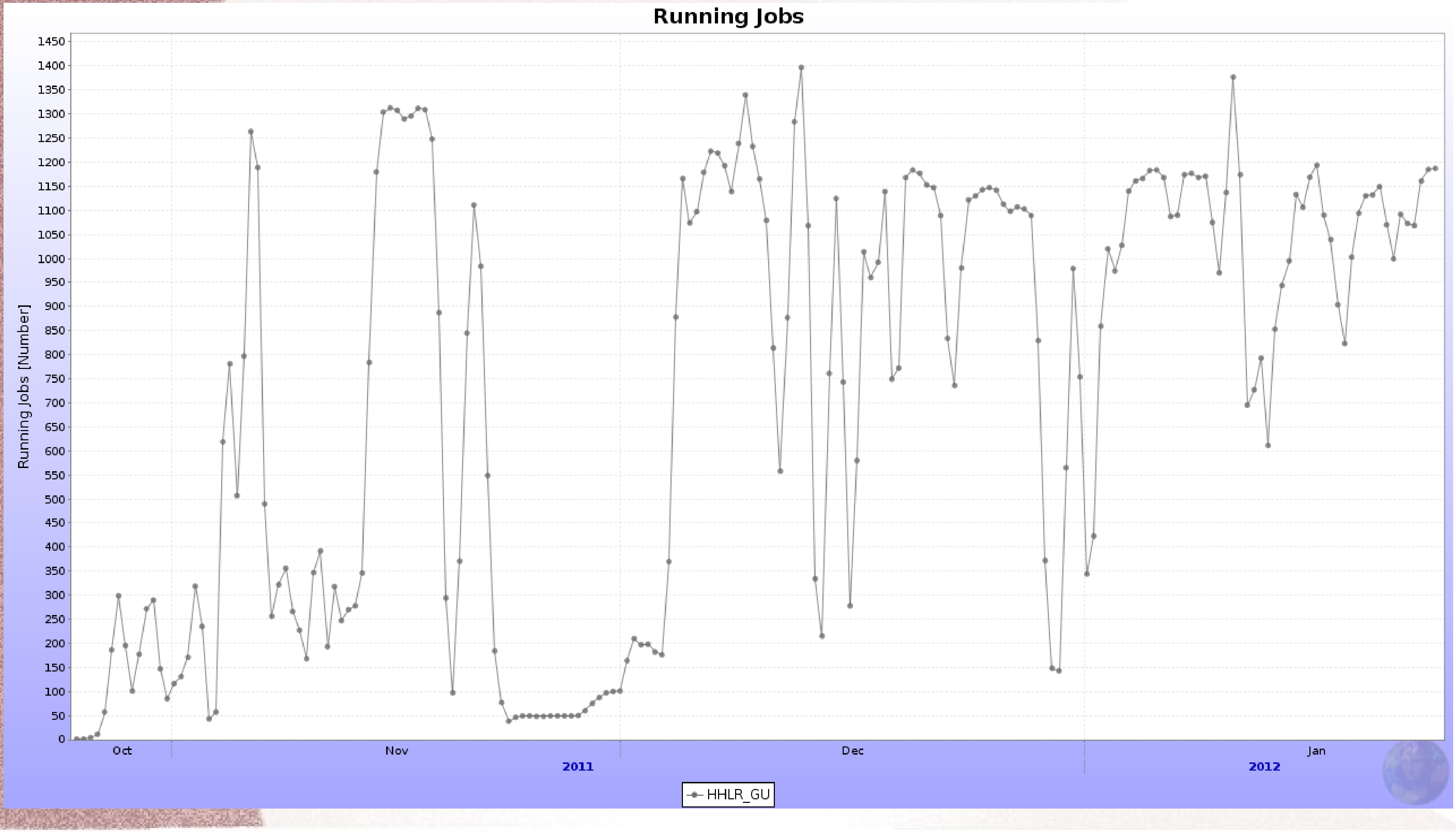
HHLR_GU ALICE setup



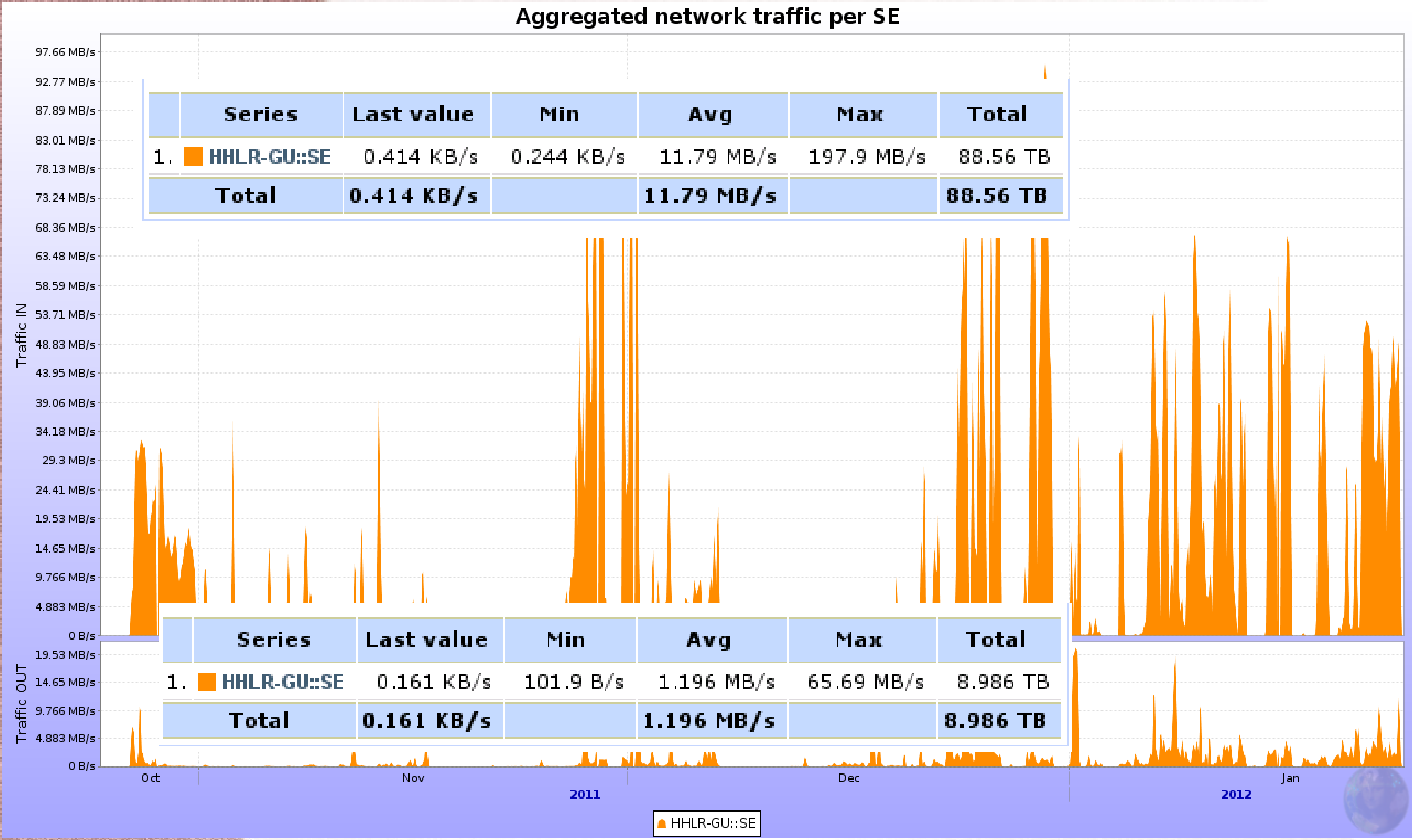
- OS: SL6
- Job Submission System: Slurm
- for this the native AliEn Slurm interface has been reactivated (A. Montiel Gonzalez)

Jobs at Loewe CSC

- continuous operation since October 2011
- average job #: 720
- max job #: 2400



storage at Loewe CSC



Aggregated network traffic per SE

	Series	Last value	Min	Avg	Max	Total
1.	🟧 HHLR-GU::SE	0.414 KB/s	0.244 KB/s	11.79 MB/s	197.9 MB/s	88.56 TB
	Total	**0.414 KB/s**		**11.79 MB/s**		**88.56 TB**

	Series	Last value	Min	Avg	Max	Total
1.	🟧 HHLR-GU::SE	0.161 KB/s	101.9 B/s	1.196 MB/s	65.69 MB/s	8.986 TB
	Total	**0.161 KB/s**		**1.196 MB/s**		**8.986 TB**

HHLR-GU: next steps

- increase network bandwidth. At some point Loewe CSC will be part of the federated FAIR T0 cloud ==> high bandwidth at least to GSI. But intermediate solutions may be needed
- create distributed file system based on local disk of Wns. Expected technology to be used: EOS

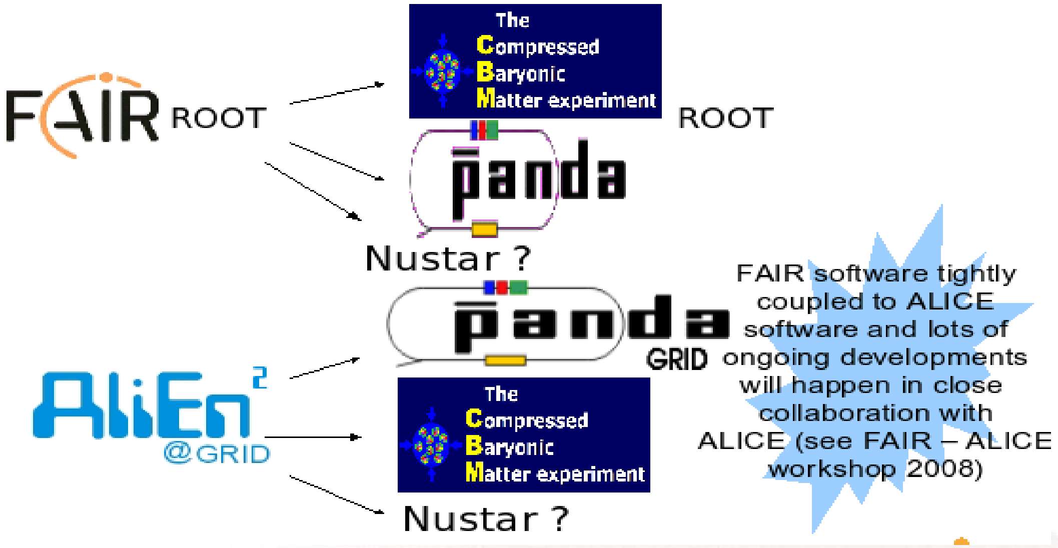
  This file system will be included in ALICE Grid.

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 ?
AliEn2 @GRID
panda 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)