

Prague Site Report

M. Adam, V. Říkal

D. Adamová, J. Chudoba, Z. Sobotka, J. Švec



Outline

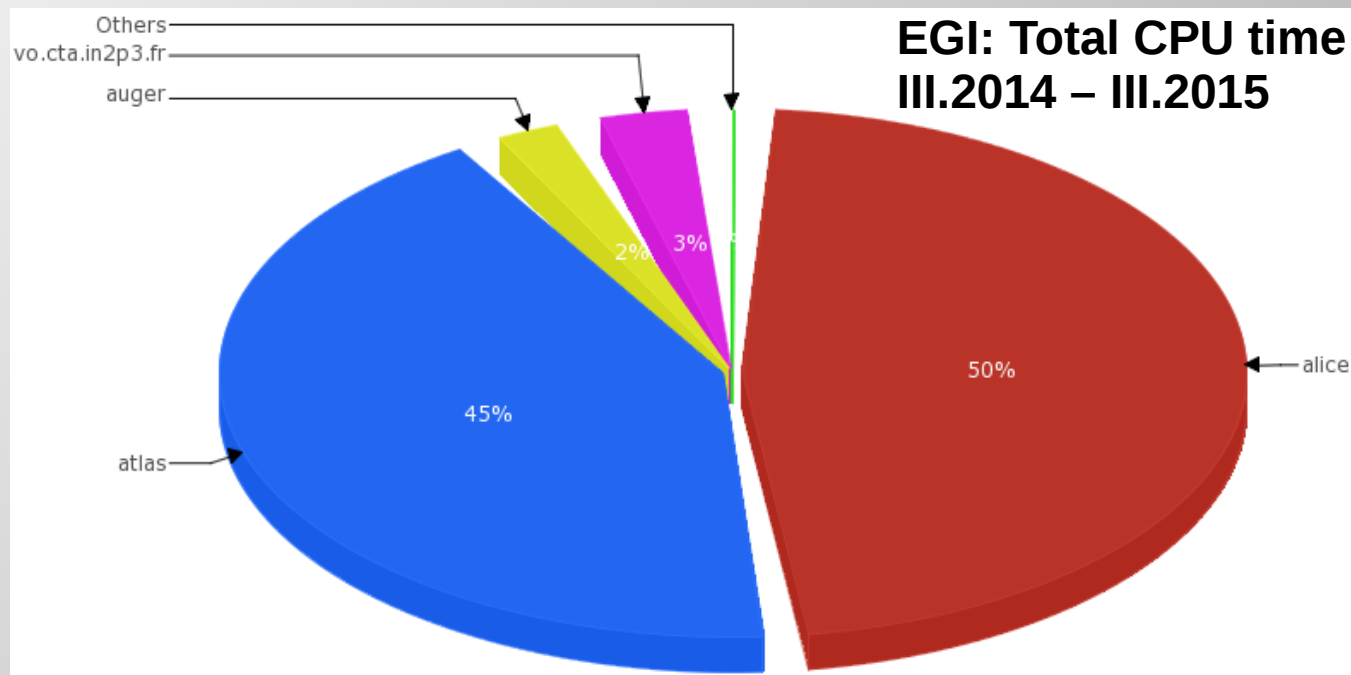
- Overview
- New Hardware
- Network
- Torque + maui mcore configuration
- ALICE
- Spacewalk
- XEN \Rightarrow KVM transition
- CFEngine \Rightarrow Puppet configuration engine
- Monitoring

Overview

- Distributed Tier-2 (some storages on external site)
- Up to 5000 cores published in the grid
- ~4 PB on disk servers (DPM, XrootD, NFS)
- Standard WLCG services (ARC CE installation)
- 3x10 Gbps external connection

Overview

- 1 batch system (torque + maui)
- 2 main WLCG VOs: ATLAS, ALICE
 - Other VOs: Auger, CTA
 - Non EGI: D0, Nova 25% overall fairshare



Main site

- Server room area: 62m²
- UPS:
 - 1X200 kVA + 2X100 kVA = 400 kVA
 - Diesel generator 350 kVA
- Air cooling: 108 kW, Water cooling: 176 kW
- Cluster LUNA
 - For Czech NGI distributed computer project

New Hardware

- Subcluster Aplex
 - Standard Asus servers in IBM iDataPlex rack
 - 30 servers with
 - 32 cores Xeon E5-2630 v3
 - 64GB RAM
 - 338 HS06 per server
 - 2 servers with
 - 54 cores Xeon E5-2695 v3
 - 128GB RAM
 - 595 HS06 per server
 - IBM vs. ASUS offer: 60% higher computing power



New Hardware

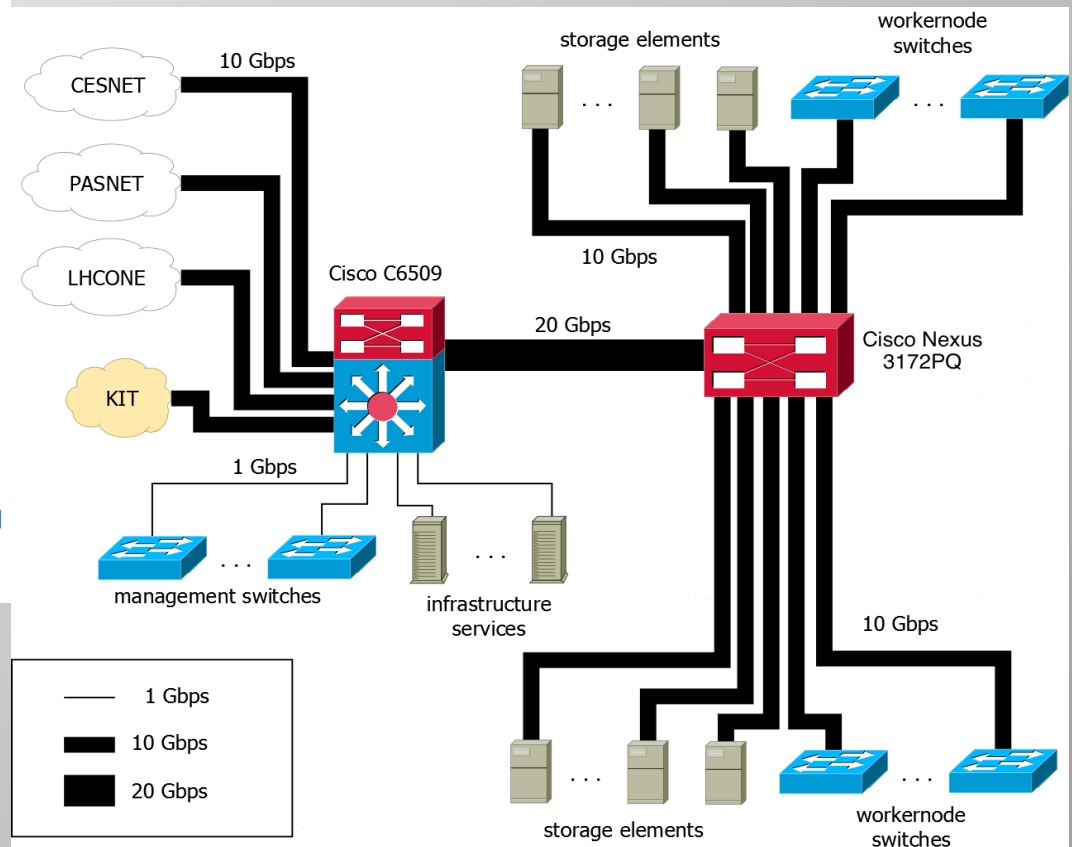
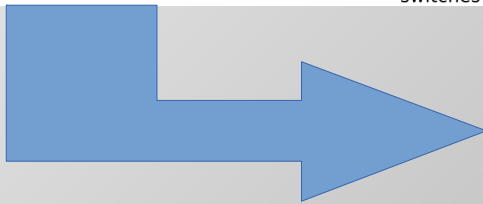
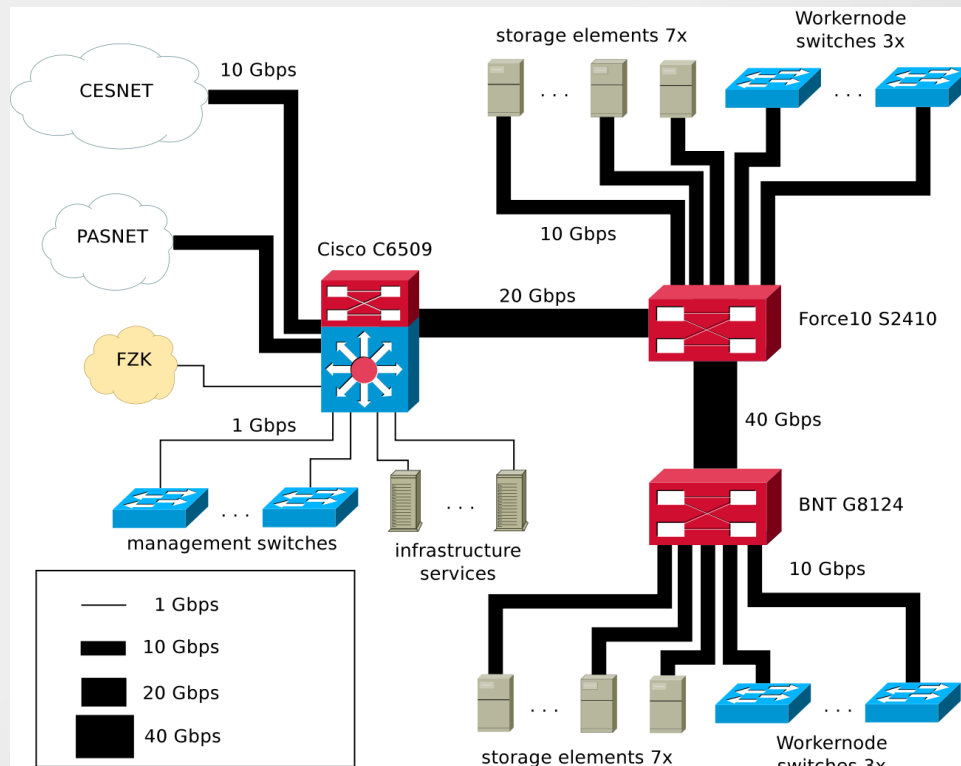
- ELI cluster
 - 0.5 MEuro (without VAT)
 - 768 + 192 TB disk space
 - 44 604 SPECfp_rate2006 base; > 30 kHS06
 - 40 kW
 - water cooled racks
 - worker nodes: 84 servers
 - IBM Next Scale nx360, Xeon E5-2600 v3, 128 GB RAM, 128 GB SSD, IB QSFP+
 - $84 \times 128 = 10\,752$ GB RAM
 - $84 \times 2 \times 8$ cores = 1 344 cores

New Hardware

- 2x DPMpool and XRrootD Storage element for ALICE
 - Server + 3 x MegaRAID SAS 2208
 - 24 core Intel(R) Xeon(R) CPU E5-2620 v2
 - 48GB RAM
 - ~400TB disk space
 - Controller 1: 4xFS (RAID6) + system
 - Controller 2,3: 4xFS (RAID6)
 - 1 FS: 44 HDDs ~ 30TB
 - 1 hotspare for all FSs

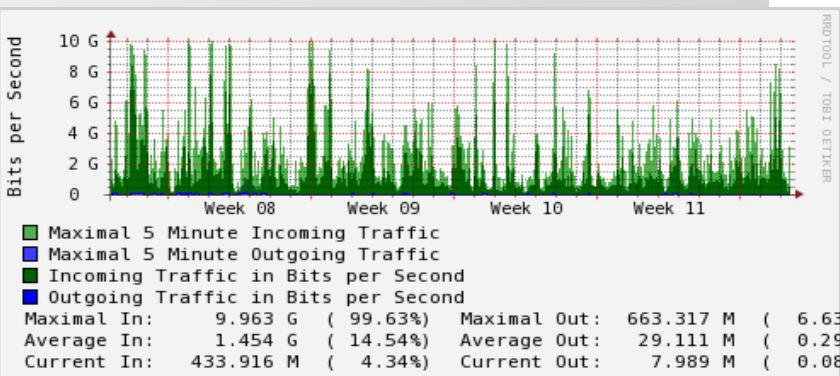
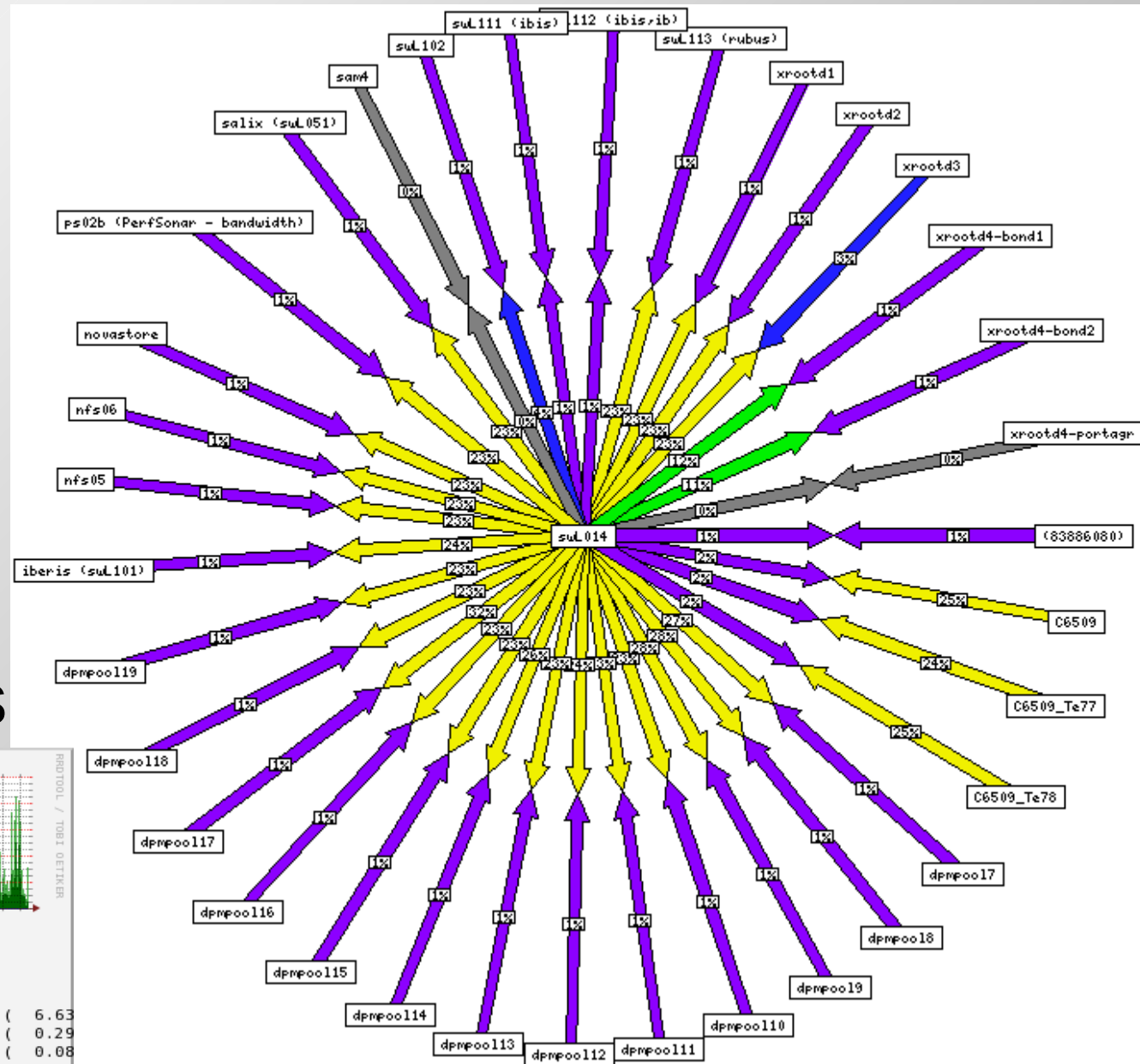
New Hardware

- Cisco Nexus 3172 PQ router (48x10Gbps)



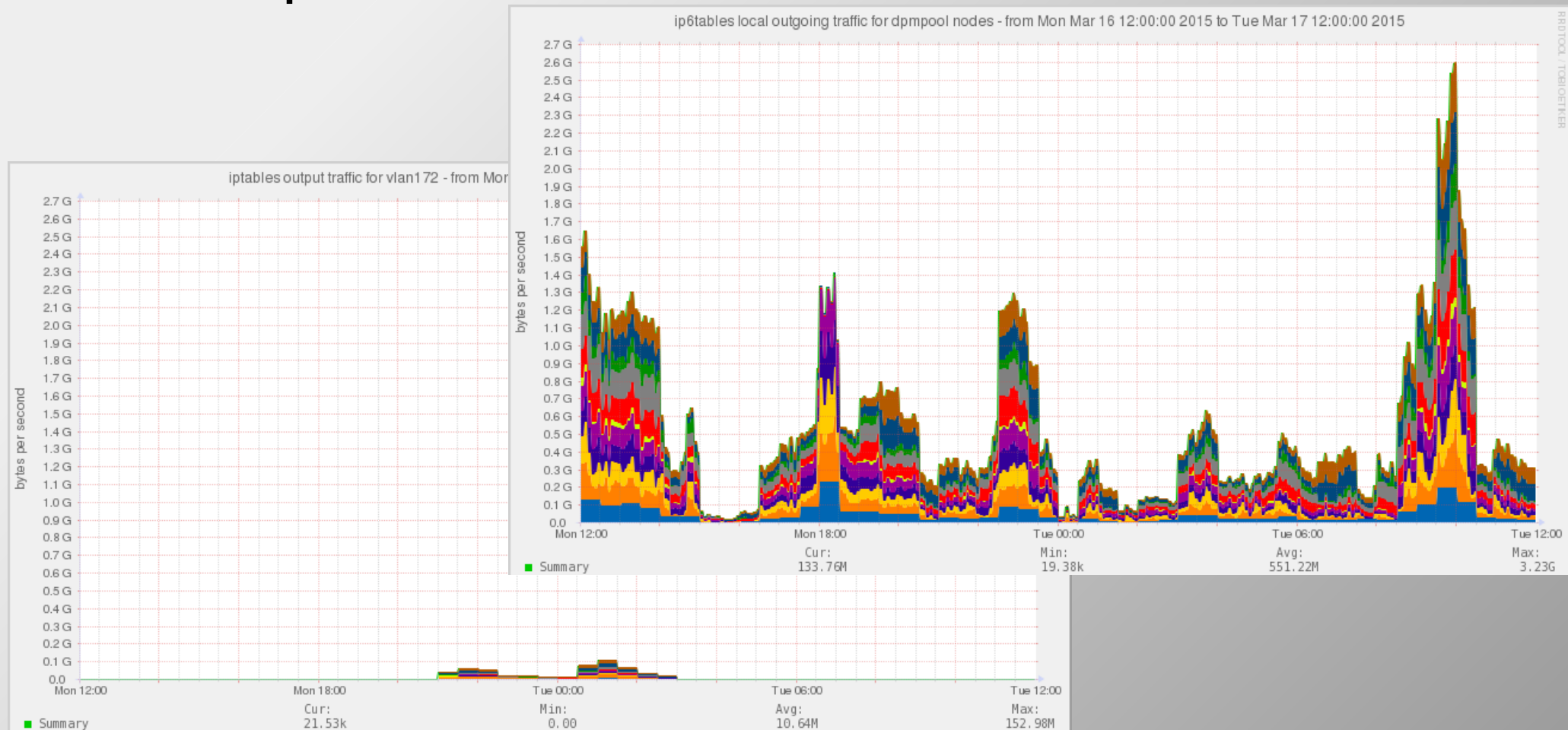
Local network

- Bottlenecks maybe between WN racks backbone
- Will be upgraded: 10 -> 2x10Gbps



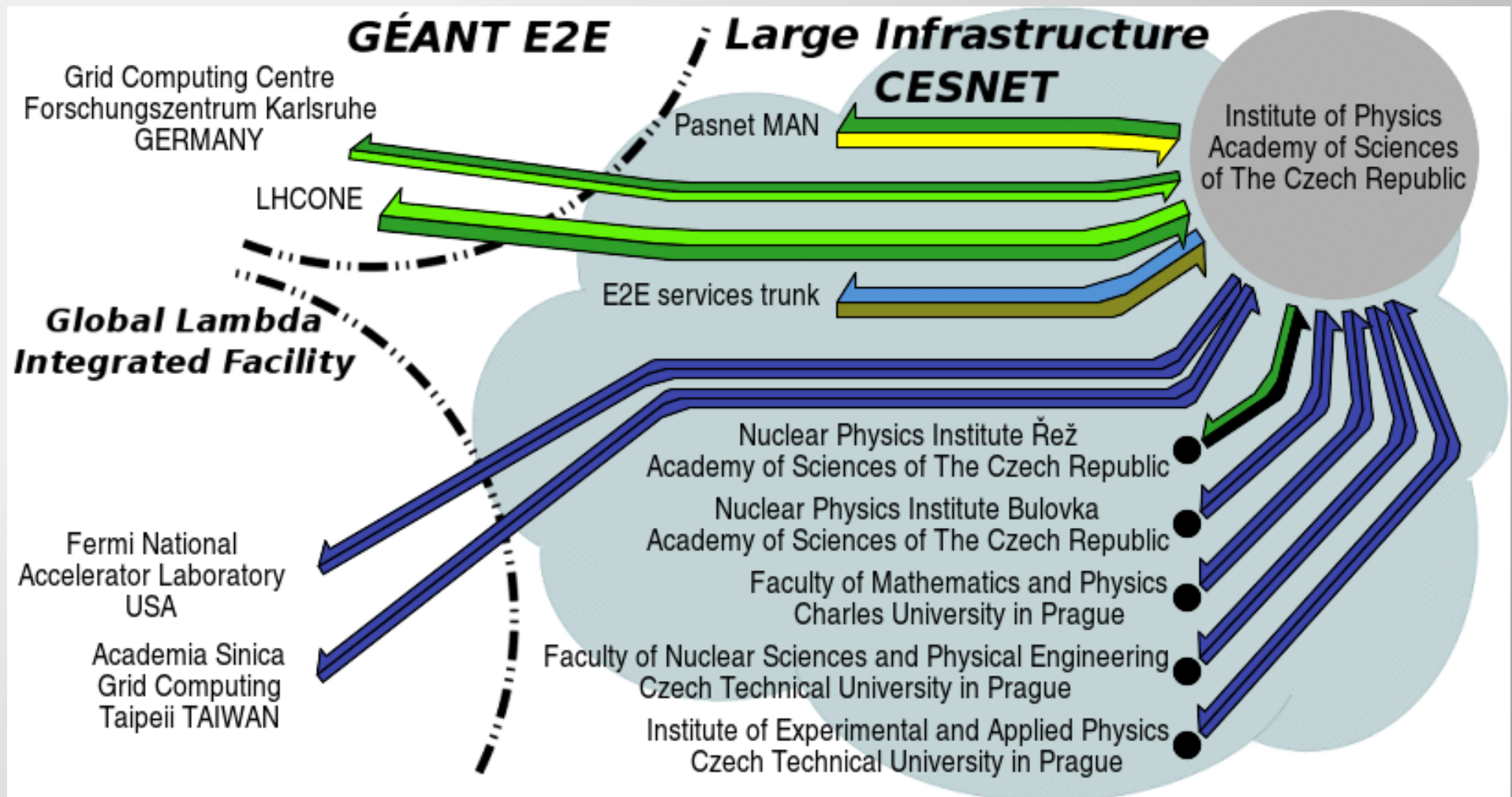
IPv6 usage

- DPM servers in dual-stack environment
- IPv6 preferred for communication with WNs



External connections

- 10Gbps: LHCONe, PASNET, CESNET



Batch system

- Torque 2.4.16 & maui 3.2.6p21 on one server
 - Becomes unresponsive when submitting thousands of jobs at once
 - Planning to upgrade (5.x) or replace (HT Condor,...)
- Multicore
 - One queue: atlasmc
 - Maui backfilling cannot be effectively used for mc jobs
 - Depends on walltime
 - Using Jeff Templon's mcfloat script with some modifications
 - Dynamically assigning WNs to mcore queue

ALICE

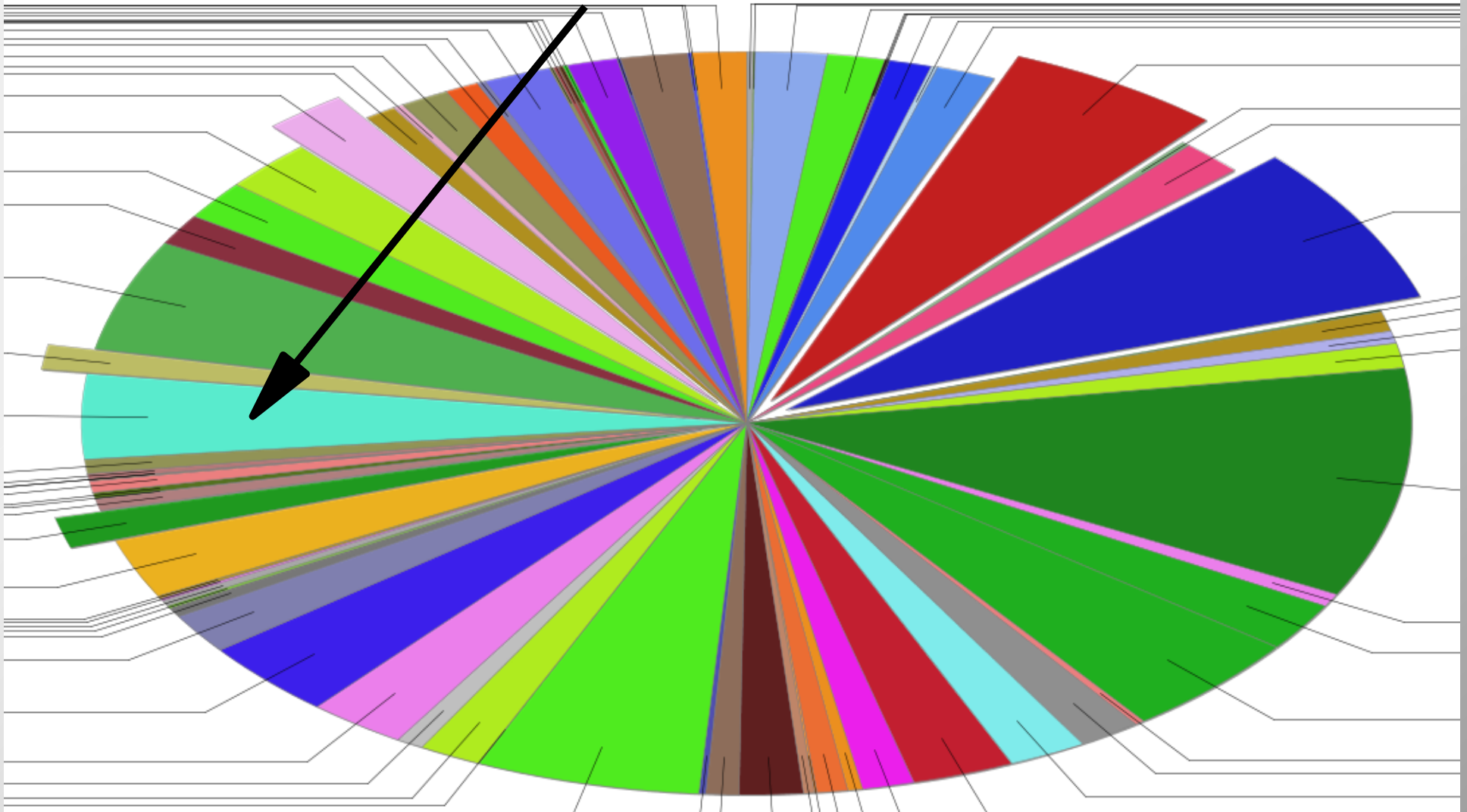
- The WLCG and ALICE pledges in 2014 of computing and storage resources were satisfied to more than 150%
 - Sharing HW between projects
 - Using servers out of warranty

	ALICE required	Delivered
CPU (kHS06)	5.7	8.8
Disk (PB)	0.4	1.1

ALICE

avg. running jobs Tier-1 & Tier-2 (last year)

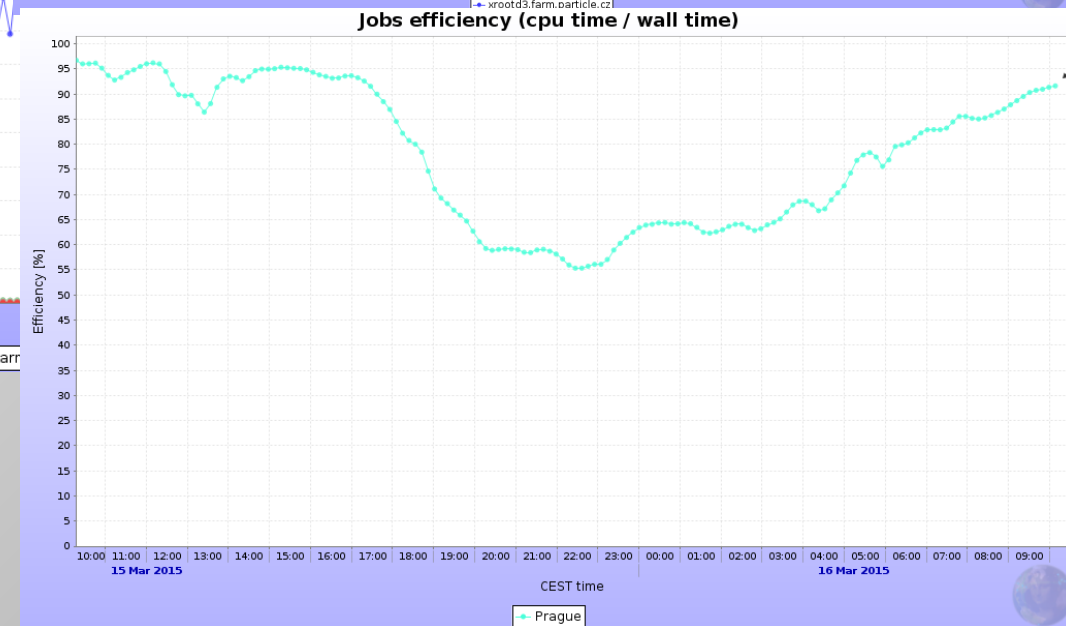
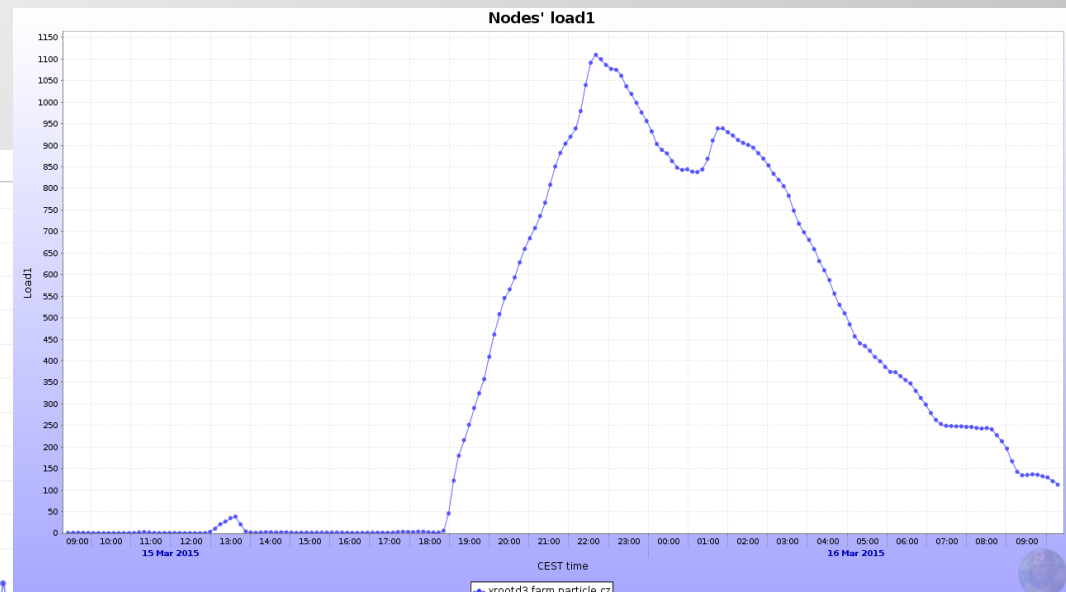
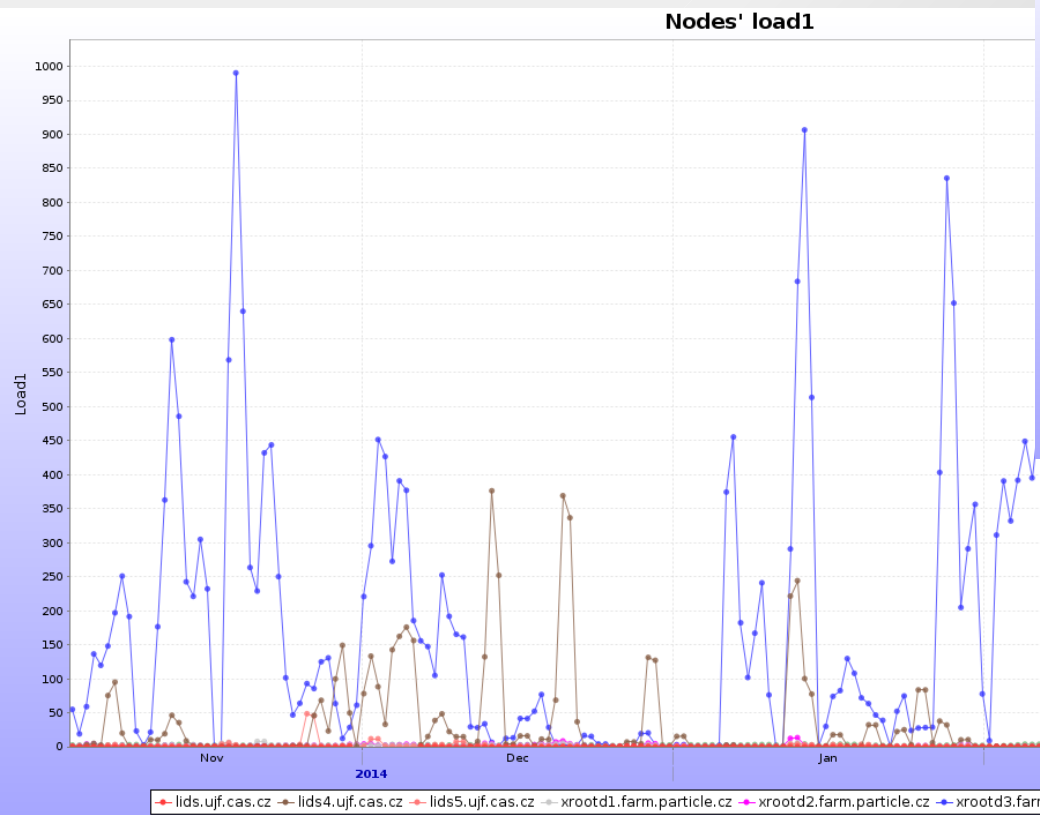
Prague (1542, 3.7%)



ALICE








- Problems with the ALICE jobs processing efficiency (CPU/wall) for about a year
 - Average efficiency is $\sim 75\%$ for the last 6 months, which is lower than at Tier-2 sites of a similar rank (cf. Legnaro $\sim 87\%$)
- No recognized reason, we only have some hints like
 - The correlation between the load peaks on two of the XRootD servers and the efficiency drops.
 - Efficiency drops show up following higher network traffic on the links between storages and worker nodes

ALICE



ALICE

- Other problems
 - Saturated network between storage servers and worker nodes
 - Some of the XRootD servers equipped with only 1Gb/s ethernet cards

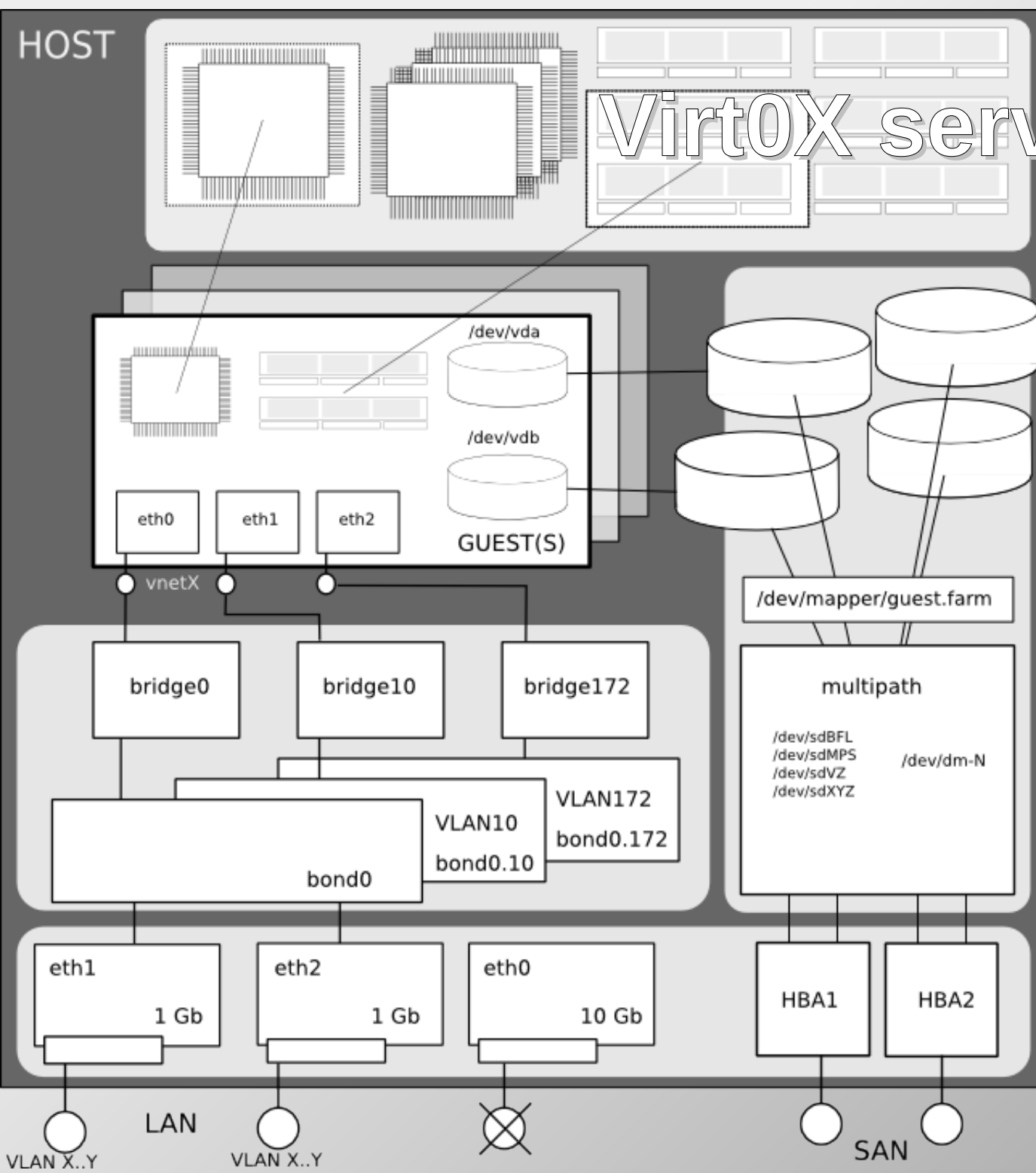
Traffic OUT						
	Series	Last value	Min	Avg	Max	Total
1.	 lids.ujf.cas.cz	4.44 MB/s	70.55 B/s	20.39 MB/s	118.2 MB/s	51.02 TB
2.	 lids4.ujf.cas.cz	3.625 MB/s	54.71 B/s	15.65 MB/s	111 MB/s	39.19 TB
3.	 lids5.ujf.cas.cz	1.397 MB/s	54.09 B/s	35.62 MB/s	118.3 MB/s	89.21 TB
4.	 xrootd1.farm.particle.cz	3.793 MB/s	0.257 KB/s	58.69 MB/s	606.9 MB/s	147 TB
5.	 xrootd2.farm.particle.cz	4.056 MB/s	0.202 KB/s	63.48 MB/s	679.3 MB/s	159 TB
6.	 xrootd3.farm.particle.cz	5.645 MB/s	5.828 KB/s	89.86 MB/s	641.9 MB/s	225 TB
7.	 xrootd4.farm.particle.cz	98.9 MB/s	0.991 KB/s	94.24 MB/s	1.544 GB/s	236 TB
Total		121.9 MB/s		377.9 MB/s		946.4 TB

Manual PXE \Rightarrow Spacewalk

- Spacewalk
 - PXE + cobbler kickstart management
 - Systems inventory (HW and SW info)
 - Systems installation and software updates
 - Collection and distribution of custom software packages into groups
 - Provisioning (kickstart) of our systems
- Our current status
 - virtual guest, 6 cores, 16GB RAM, 60GB + 160GB HDD
 - 200+ WNs, 3 virtualization servers, 16 virtual guests, 4 dpm pools + dpmhead, 3 UIs

XEN \Rightarrow KVM

- 3 virtualisation servers
- Migration of 13 virtual servers
 - New installation (11 servers)
 - With OS and services update
 - Virtual disk image transfer (2 servers)
- Current status
 - 21 virtual guests (argus, cream, ldap, uiX, www,...)
 - Live migration of virtual guests tested



- 32 cores
- 128GB RAM
- 3 NIC:
 - 2x1Gb
 - 1x10Gb
 - 10xVLANs
- 2xSAN HBA

CFE \Rightarrow Puppet

- CFEngine
 - 2.2.10, since 2007, gentoo virtual server, svn
- Puppet
 - 3.7.4 (current), SL6 virtual server, Git repository
 - separation of code and configuration
 - 32 github/cern modules (cvmfs, ganglia, lcgdm-*, nagios, munin-node,...)
 - configuring 271 servers
 - workernodes, dpmpools, dpmhead, virt0X hosts, nfs servers, user interfaces

Puppet GUI

Background Tasks

4 pending tasks

Nodes

0 Unresponsive

0 Failed

0 Pending

3 Changed

213 Unchanged

55 Unreported

271 All

Add node

Radiator View

Group

Add group

Class

Add class

Nodes

Daily run status

Number and status of runs during the last 30 days:



Nodes

Export nodes as CSV

Node	Latest report	Resources				
		Total	Failed	Pending	Changed	Unchanged
Total		83761	0	0	219	83542
✓ ib23.farm.particle.cz	2015-03-19 08:20 UTC	394	0	0	0	394
✓ malva04.farm.particle.cz	2015-03-19 08:20 UTC	388	0	0	0	388
✓ rubuk01.farm.particle.cz	2015-03-19 08:19 UTC	388	0	0	0	388
✓ ibis04.farm.particle.cz	2015-03-19 08:19 UTC	394	0	0	0	394
✓ ibis37.farm.particle.cz	2015-03-19 08:19 UTC	394	0	0	0	394
✓ ibis54.farm.particle.cz	2015-03-19 08:19 UTC					
✓ ibis13.farm.particle.cz	2015-03-19 08:19 UTC					
✓ ibis19.farm.particle.cz	2015-03-19 08:19 UTC					
✓ ibis64.farm.particle.cz	2015-03-19 08:19 UTC					
✓ iberis17.farm.particle.cz	2015-03-19 08:19 UTC					
✓ iberis22.farm.particle.cz	2015-03-19 08:19 UTC					
✓ iberis27.farm.particle.cz	2015-03-19 08:19 UTC					
✓ malva10.farm.particle.cz	2015-03-19 08:19 UTC					
✓ rubus16.farm.particle.cz	2015-03-19 08:19 UTC					
✓ iberis35.farm.particle.cz	2015-03-19 08:19 UTC					
✓ rubus22.farm.particle.cz	2015-03-19 08:19 UTC					
✓ ib19.farm.particle.cz	2015-03-19 08:19 UTC					
✓ iberis04.farm.particle.cz	2015-03-19 08:19 UTC					
✓ dpmpool18.farm.particle.cz	2015-03-19 08:18 UTC					
✓ iberis32.farm.particle.cz	2015-03-19 08:18 UTC					

« Previous 1 2 3 4 5 6 7 8 9 ... 13 14 Next » Per page: 20 100 all

JVM Heap
bytes

141M



Nodes
in the population

216



Resources
in the population

99,351



Resource duplication
% of resources stored

88.1%



Catalog duplication
% of catalogs encountered

64.6%



Command Queue
depth

0



Command Processing
sec/command

0.163



Command Processing
commands/sec

0.142



Processed
since startup

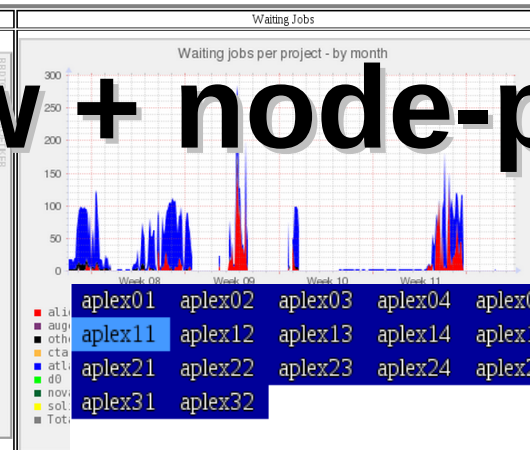
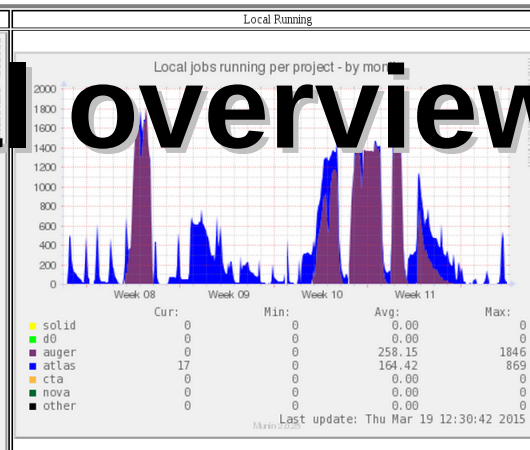
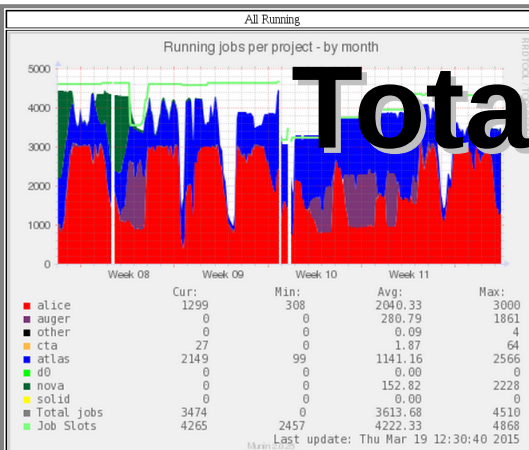
18,635



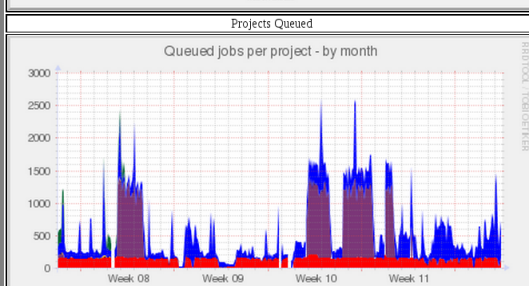
Monitoring

- Munin v2
 - rrd databases with 10y history
 - Plugins for jobs, ups, temp, power...
 - Total overview webpage
- Ganglia
 - Nodes -> aggregators on ipv6 (UDP)
 - Aggregators -> gmetad: ipv6 not implemented (TCP)

Total overview + node-page

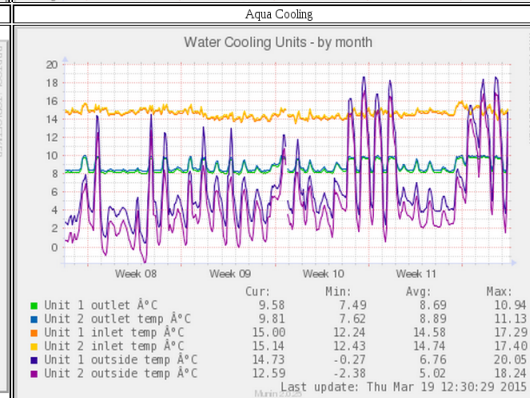
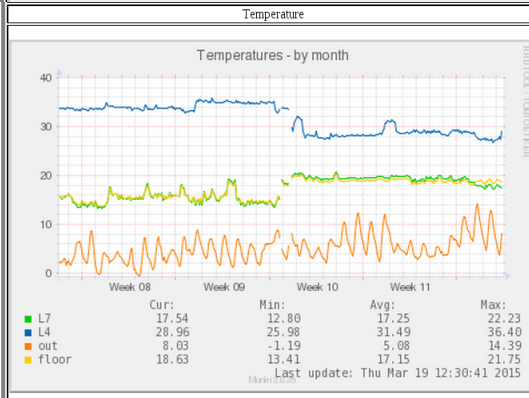


aplex01	aplex02	aplex03	aplex04	aplex05	aplex06	aplex07	aplex08	aplex09	aplex10
aplex11	aplex12	aplex13	aplex14	aplex15	aplex16	aplex17	aplex18	aplex19	aplex20
aplex21	aplex22	aplex23	aplex24	aplex25	aplex26	aplex27	aplex28	aplex29	aplex30
aplex31	aplex32								



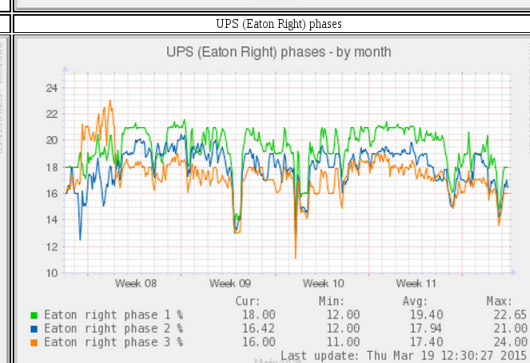
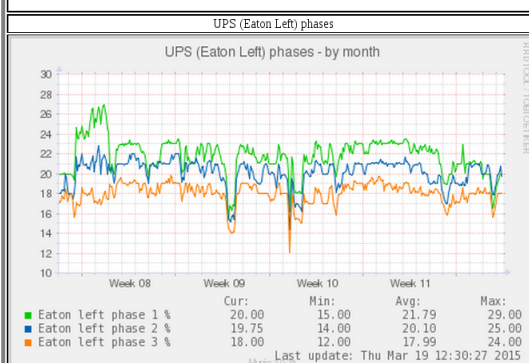
ib01	ib02	ib03	ib04	ib05	ib06	ib07	ib08	ib09	ib10
ib11	ib12	ib13	ib14	ib15	ib16	ib17	ib18	ib19	ib20
ib21	ib22	ib23	ib24	ib25	ib26				

iberis01	iberis02	iberis03	iberis04	iberis05	iberis06	iberis07	iberis08	iberis09	iberis10
iberis11	iberis13	iberis14	iberis15	iberis16	iberis17	iberis18	iberis19	iberis20	iberis21
iberis22	iberis23	iberis24	iberis25	iberis26	iberis27	iberis28	iberis29	iberis30	iberis31
iberis32	iberis33	iberis34	iberis35	iberis36	iberis37	iberis38	iberis39	iberis40	iberis41
iberis42	iberis73	iberis75	iberis76						



ibis01	ibis02	ibis03	ibis04	ibis05	ibis06	ibis07	ibis08	ibis09	ibis10
ibis11	ibis12	ibis13	ibis14	ibis15	ibis16	ibis17	ibis18	ibis19	ibis20
ibis21	ibis22	ibis23	ibis24	ibis25	ibis26	ibis27	ibis28	ibis29	ibis30
ibis31	ibis32	ibis33	ibis34	ibis35	ibis36	ibis37	ibis38	ibis39	ibis40
ibis41	ibis42	ibis43	ibis44	ibis45	ibis46	ibis47	ibis48	ibis49	ibis50
ibis51	ibis52	ibis53	ibis54	ibis55	ibis56	ibis57	ibis58	ibis59	ibis60
ibis61	ibis62	ibis63	ibis64	ibis65					

malva01	malva02	malva03	malva04	malva05	malva06	malva07	malva08	malva09	malva10
malva11	malva12								



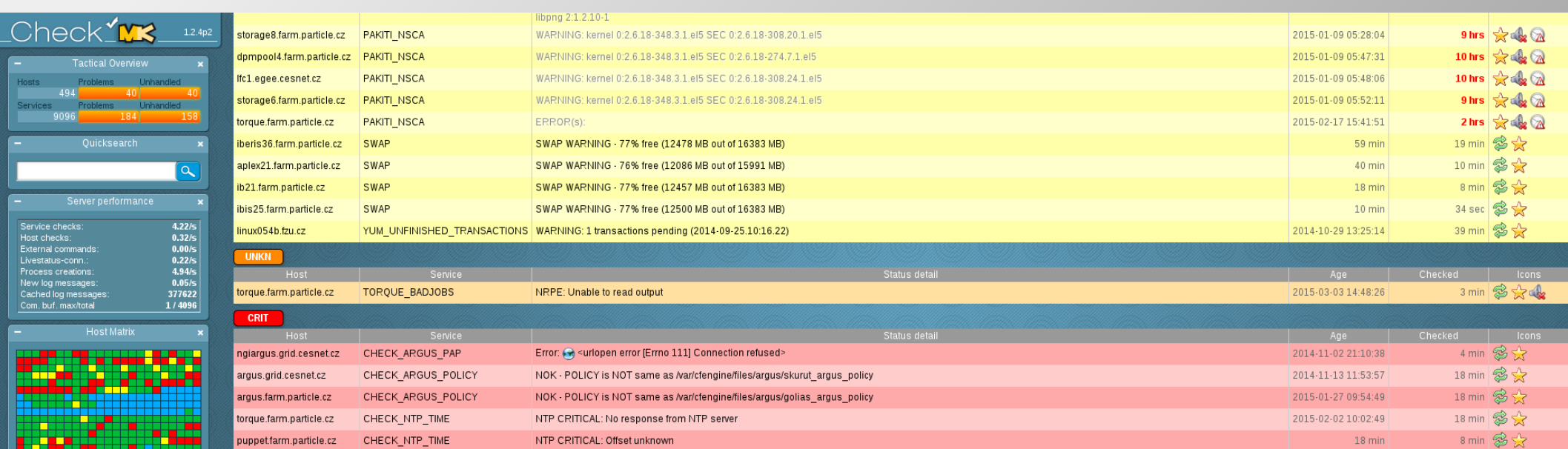
rubuk01	rubuk02								
rubul01	rubul02								
rubus01	rubus02	rubus03	rubus04	rubus05	rubus06	rubus07	rubus08	rubus09	rubus10
rubus11	rubus12	rubus13	rubus14	rubus15	rubus16	rubus17	rubus18	rubus19	rubus20
rubus21	rubus22	rubus23							

phase 1 %	48.00	15.16	49.92	60.81
phase 2 %	46.58	16.00	47.81	59.00
phase 3 %	39.50	18.00	45.20	56.81

Last update: Thu Mar 19 12:30:27 2015

Monitoring

- Nagios + Multisite interface
 - Warning
- MRTG with weathermap interface
 - Network load
- Netflow, Smokeping, Zabbix...



Thank you for your attention!

Questions?

