Prague TIER2

Site Report

Jiří Chudoba

Lukáš Fiala, Tomáš Kouba, Jan Kundrát, Miloš Lokajíček, Jan Švec

HEPIX 2009, NERSC, Berkeley
31 Oct 2009

Outline

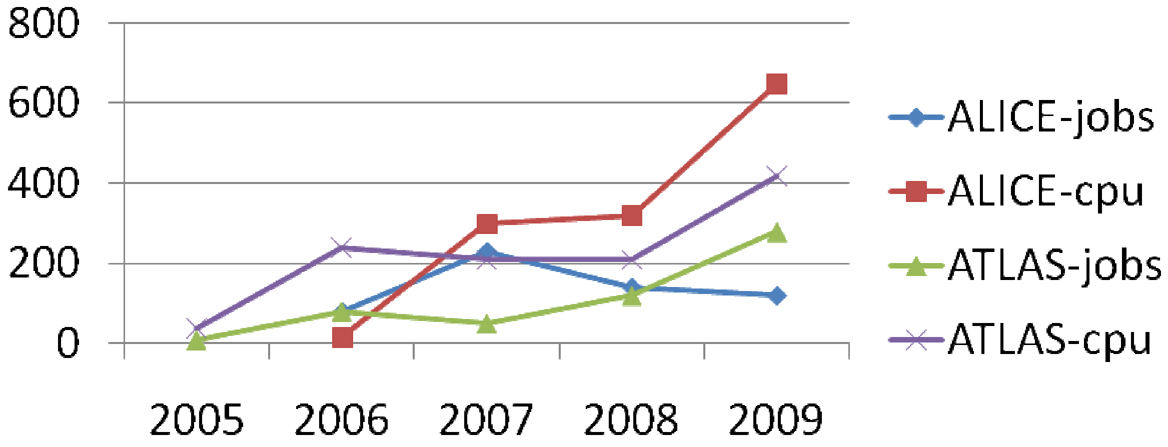
- Who we are, What are we doing
- Computing Centre Evolution
- General infrastructure (electricity, cooling, network)
- HW, SW and tasks, management and status
- Services failover
- Conclusion

Who we are, What are we doing, Our users

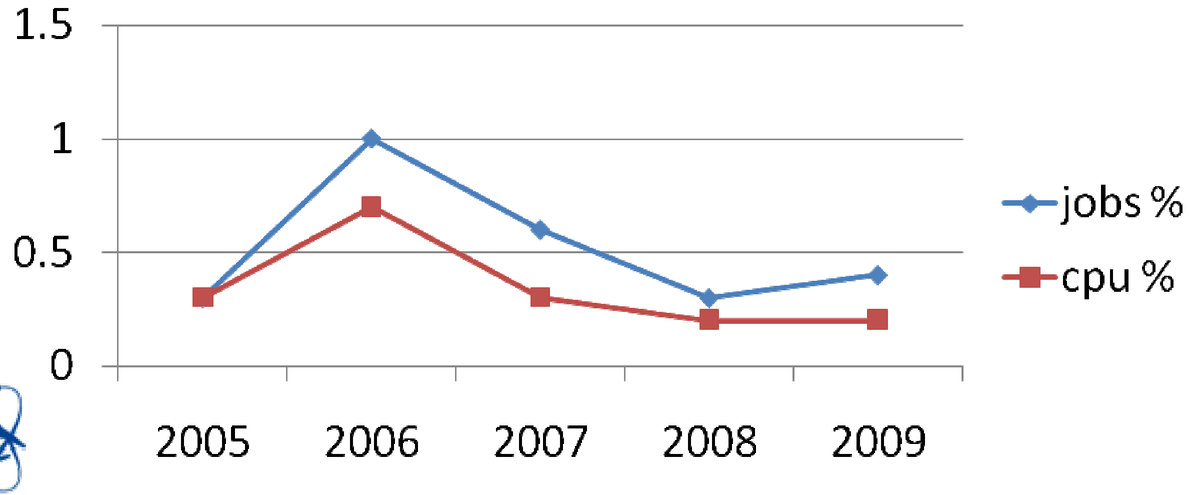
- Who we are?
  - Regional Computing Centre for Particle Physics
    Institute of Physics of the Academy of Sciences of the Czech Republic, Prague
    - Basic research in particle physics, solid state physics and optics
- What are we doing?
  - Computing support for big international Particle Physics, Nuclear Physics and Astro-particle Physics experiments using grid environment
    - D0, ATLAS, ALICE, STAR, AUGER, Belle (CESNET)
    - WLCG TIER2 centre
  - Solid State Physics computing
  - From the computing point of view: High Throughput Computing (HPC), large data samples processing, chaotic data analysis (by physicists), parallel computing
- Our users?
  - Collaborating scientists from institutes of the Academy of Sciences of the Czech Republic, Charles University and Czech Technical University
  - Big experiments (grid environment), individual members of the international experiments, local physicists

- CESNET (Czech Research Network Provider) contributes with small part of resources in the framework of collaboration on EGEE project (partner)

ALICE and ATLAS - jobs and CPU produced in Prague



- Thousands of jobs
- Thousands of CPU normalized hours



- Prague share on total LCG computing (all experiments)
- Delayed financing of computing last 3 years

Prague LCG Farm Evolution

Year	# cores	Storage TB	network	experiment
1998	Unix, Win PCs	0,3	TEN34>TEN155	D0 simulation
1999			TEN155	D0
2000	8 FZU	1	GEANT, Gb	D0, EDG prep.
2001	8 , 32 CESNET	1		EDG
2002	64 ,32+32	1	2.5 Gb backbone in CZ	LCG
2003		10		
2004	160 , 64	40	10 Gb bb/ 2.5 Gb to Prg Tier3	EGEE, CESNET as partner
2005	200 , 32			
2006	250 , 32		Gb links FNAL, ASGC, FZK	
2007	460 , 32	60	Gb link to BNL	
2008	1300, 32	200 + 24 Tier3		WLCG signed

Current status

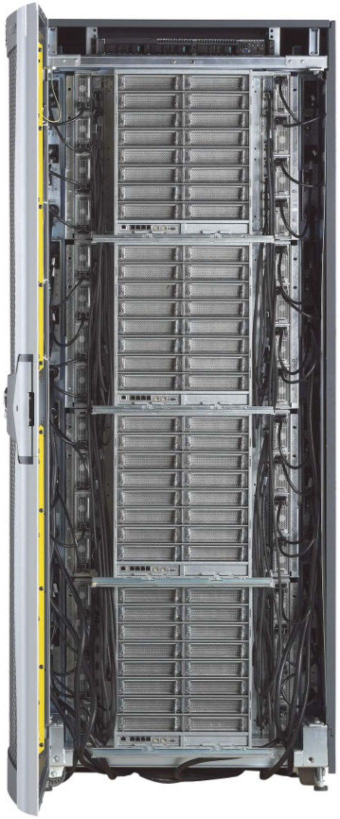
- TIER2 center with 5 off site user groups
  - One TIER3 has computing resources and storage (1 Gb optical link) – distributed TIER2
  - **1500 cores, 10 000 HEP_SPEC, 200 TB**
    - Plus 800 cores for solid state physics
  - **36 TB (xrootd) at TIER3** in NPI, 1 Gbps link

- Plan to add at end 2009
  - App. **4 800 HEP_SPEC , 200 TB**

FZÚ

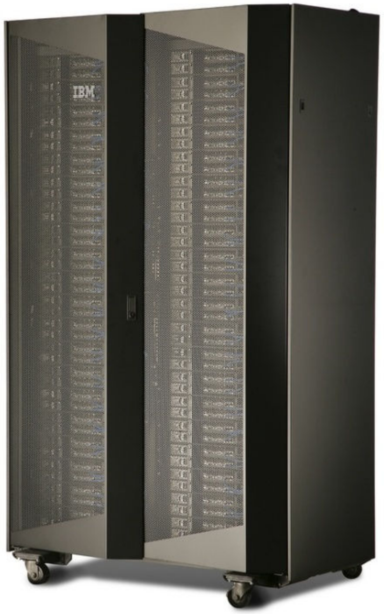
Computing capacity

ICE 8200



- **HP**
  - Blades: BL35p (36x), BL20p (6x), BL460c (9x4 cores), BL465c (12x), HP BL 460C (10x8cores)
  - U1: DL140 (67x), DL145 (3x), DL360 (2x), LP1000r (34x)
  - **Together 800 kSI2K**
- **SGI** Altix ICE 8200, infiniband
  - 64 x 8 cores, E5420 2.5GHz, 512 GB RAM, for solid state physics
- **SGI** Altix XE 310
  - 40x8 cores, E5420 2.5GHz, 640 GB RAM
- **IBM** iDataPlex dx340
  - 84 x 8 core, E5440 2.83GHz, 1344 GB RAM
- **Together 13 000 kSI2k, 2 100 cores (LCG 1 300 cores)**

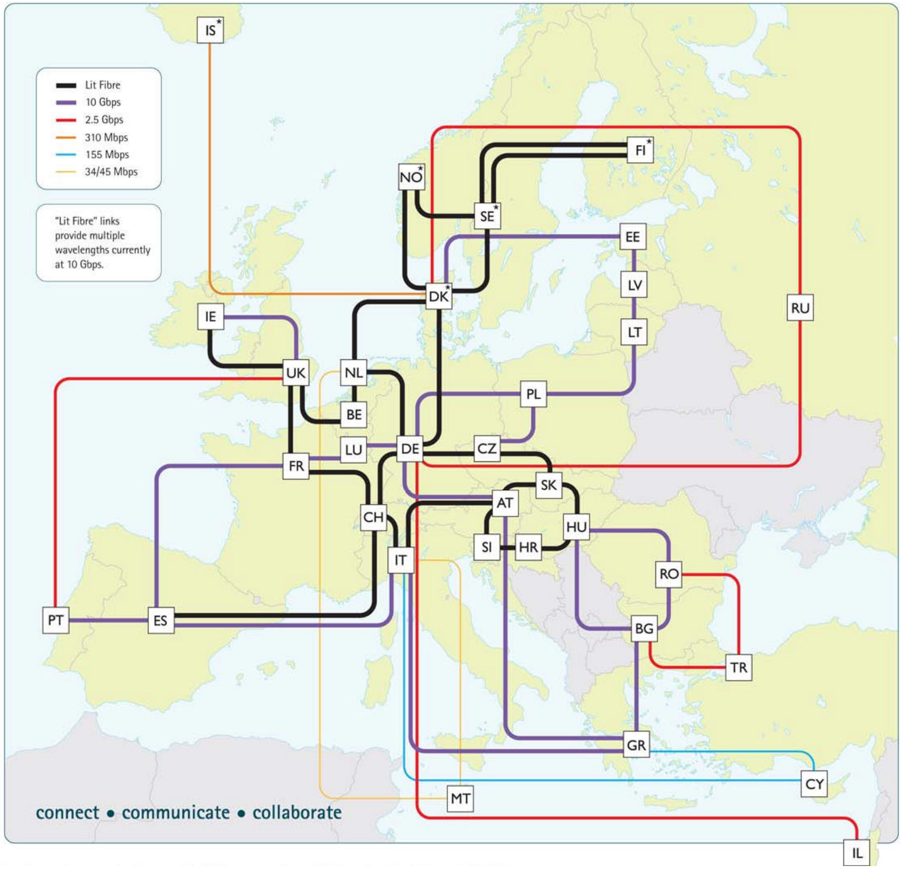
iDataPlex

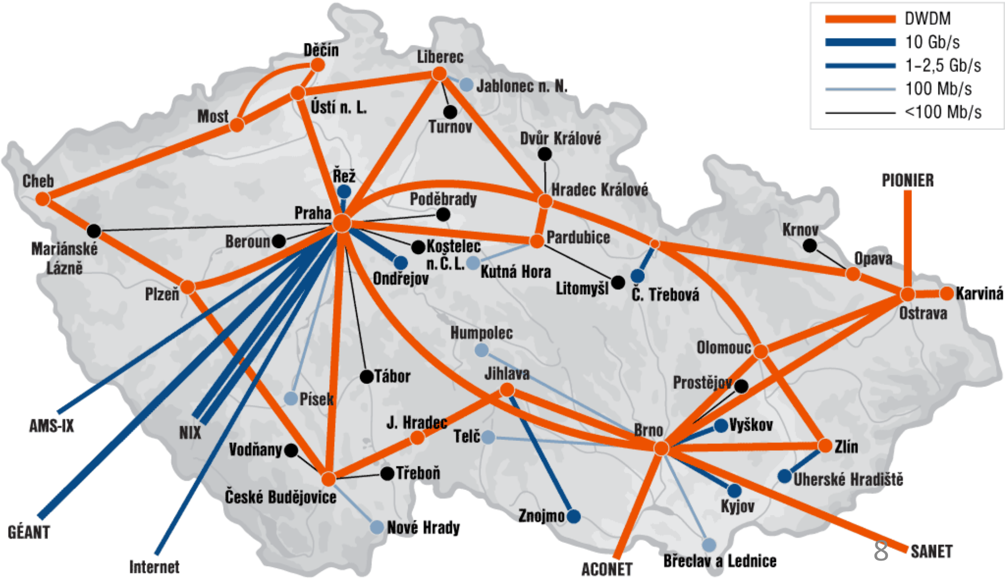


FZÚ

Networking

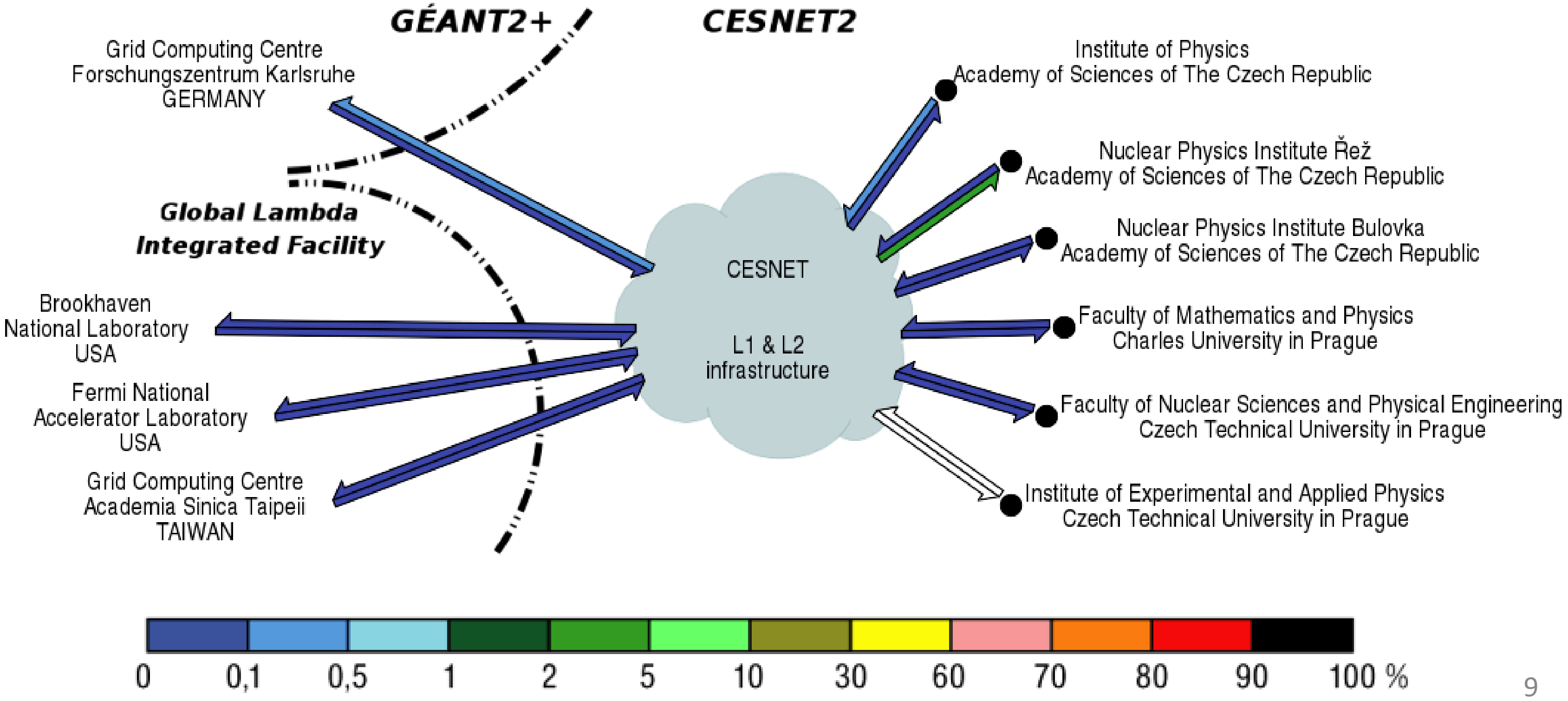
- Czech Republic well integrated into GEANT infrastrusture
- CESNET internal infrastructe – multi 10 Gbps lines





CESNET Link Monitoring System

- Detailed monitoring of all links in both directions delivered by CESNET http://www.ces.net/netreport/hep-cesnet-experimental-facility/
- Lines used up to nominal capacity 1 Gbps
  Extremely important for continuous work of 1 300 cores


GÉANT2+
CESNET2
Grid Computing Centre
Forschungszentrum Karlsruhe
GERMANY
Global Lambda
Integrated Facility
Brookhaven
National Laboratory
USA
Fermi National
Accelerator Laboratory
USA
Grid Computing Centre
Academia Sinica Taipeii
TAIWAN
CESNET
L1 & L2
infrastructure
Institute of Physics
Academy of Sciences of The Czech Republic
Nuclear Physics Institute Řež
Academy of Sciences of The Czech Republic
Nuclear Physics Institute Bulovka
Academy of Sciences of The Czech Republic
Faculty of Mathematics and Physics
Charles University in Prague
Faculty of Nuclear Sciences and Physical Engineering
Czech Technical University in Prague
Institute of Experimental and Applied Physics
Czech Technical University in Prague
0
0,1
0,5
1
2
5
10
30
60
70
80
90
100 %

Equipment – network, storage

- gigabit infrastructure in transition to 10 Gbps (Cisco C6506 routing, planned Force10 S2410p switching)
- each “high density” rack (iDataplex, Altix XE310, storage) => 1Gb switch with 10Gb uplink
- Storage
  - Tape Library Overland Neo 8000, LTO4, 100 TB (max non compression capacity 400 TB) with disk cache Overland , Ultamus 1200
  - Disk array HP EVA 6100, 80 TB
  - Disk array *EasySTOR*, 40 TB
  - Disk array VTrak M610p (CESNET), 16 TB
  - Disk Array Overland Ultamus 4800, 144 TB

    Together 100 TB tape space, 280 TB raw disk space

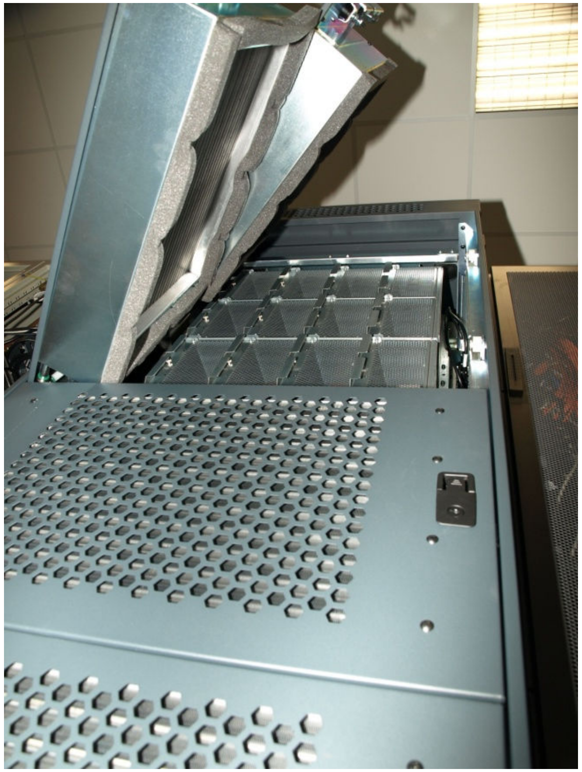
Electric power and cooling

- Computing room
  - 18 racks
  - room size 65 m^2
- UPS Newave Maxi **200kVA**
- diesel F.G.Wilson 380 kVA
- Air conditioning Liebert-Hiross 2x56 kW
- **New computing HW from 2008 could not be completely switched on**
- 2009 - added two units Water chillers STULZ CLO 781A 2x 88 kW
- Today complete cooling power **290 kW (N+1)**
- **Further computing HW must be delivered with water cooled racks**

FZÚ

Water cooling accessories for IBM and SGI racks

IBM – one big radiator 200x120 cm



SGI – independent smaller radiators for each crate

Cooling SGI and IBM

- SGI: SGI Altix ICE 8200
  - 64 servers: 2x Intel QC 2.5GHz E5420, 16GB RAM per server diskless servers , infiniband,
  - RAID 18x400GB SAS 15k RPM (infiniband), administration via SGI Tempo sw, needs 2 servers (admin node, rack leader)
  - Peak consumption given by producer **23.2kW**,
    max measured consumption by us **17.5kW.**
    Measured SPEC 06 - 67.51 per node

- IBM: iDataPlex
  - 84x 2 x Xeon E5440 2.83GHz (dx340 nodes), 16GB RAM, 1x 300GB SAS disk 15k RPM, integrated in special rack with switches.
  - IBM gives max consumption **25.4kW**,
    measured max consumption by us **22.9kW**.
    Measured HEP-SPEC 06: 69.76 per node

SW and tasks, management and status

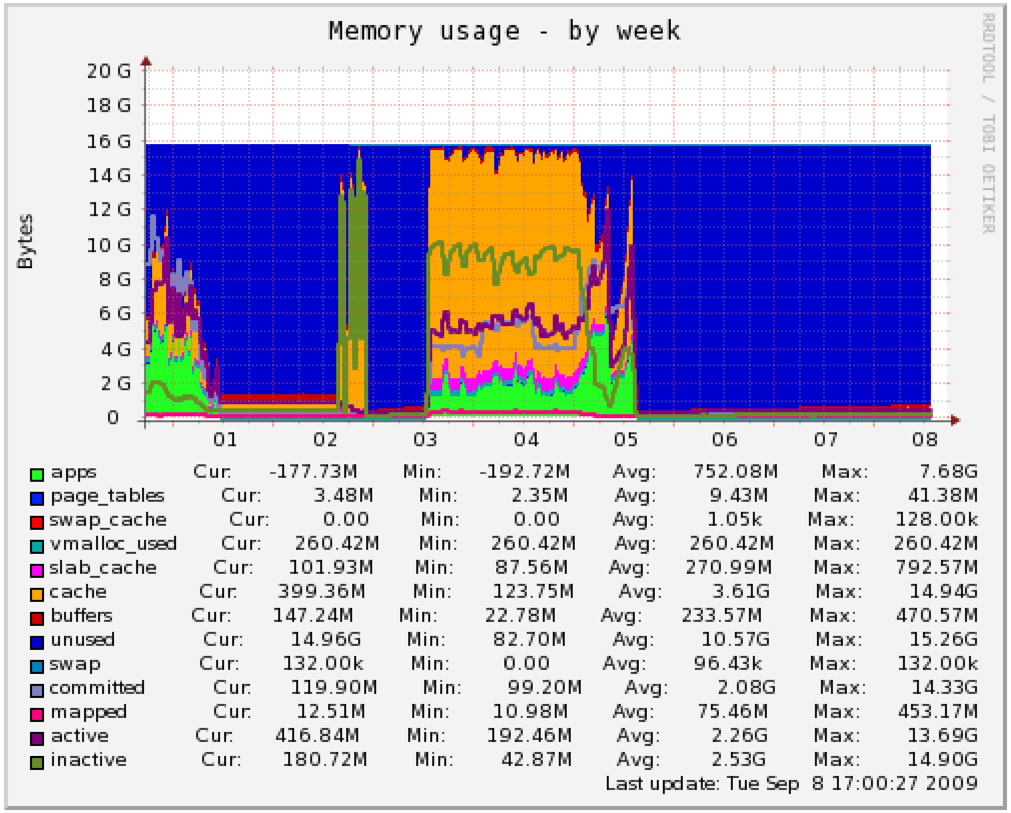
- SL4.8, SL5.3 (RedHat); Altix ICE 8200: SLES10
- PBSPro server 9.2, 480 licenses
- Torque, Maui – free for rest
- EGEE grid – gLite 3.1
  - Computing Elements, Storage Element, sBDII, vobox, UI
  - Virtual Organizations: ATLAS, ALICE, AUGER, D0, CALICE, HONE
- ALL users (local and from outside) share same resources
- Altix ICE 8200 – reserved for parallel tasks
- Management
  - Network installation,
  - Cfengine – for automatic installation and configuration of different nodes
  - ILO (HP), IPMI (IBM, SGI)
  - Management locally and remotely over internet

Monitoring

- Key for effective management
- Nearly all functions with different tools
  - Nagios, Ganglia, Munin, MRTG, RRD (graphs)
  - Grid functions : SAM tests, …

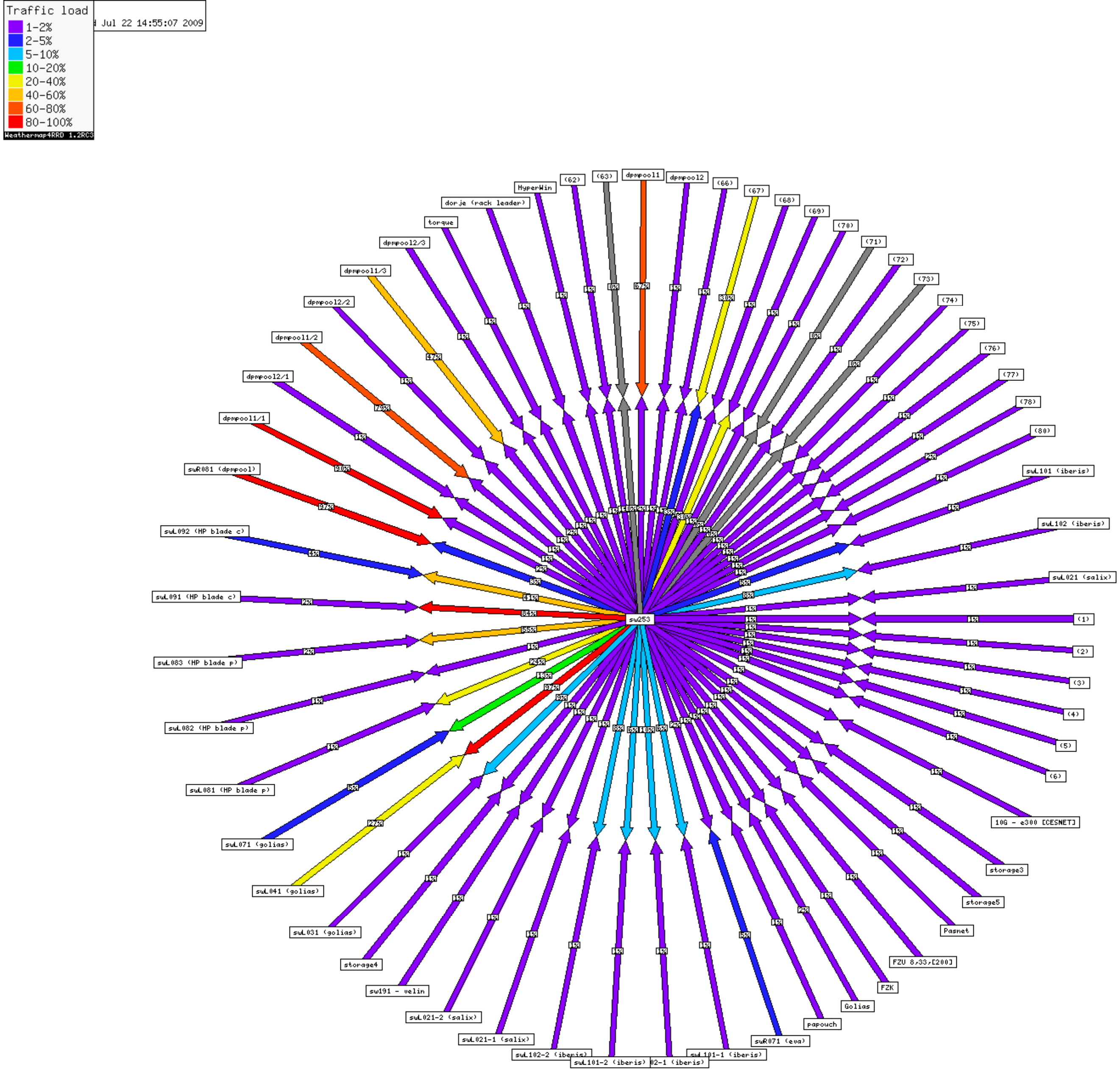
  - Examples
  - Disk
    - Filesystem usage (in %)
    - IOstat
  - Network
    - eth0 traffic
    - eth1 traffic
    - Netstat
  - Processes
    - Number of Processes
    - VMstat
  - System
    - CPU usage
    - Load average
    - Memory usage
  - Sensors
    - HDD temperature

- UPS, Diesel, Cooling, Temperatures
- Warnings sent-out (SNMP traps, SMS alerts in near future)
- Graphs for everything available



🟩 apps	Cur.	-177.73M	Min:	-192.72M	Avg:	752.08M	Max:	7.68G
🟦 page_tables	Cur:	3.48M	Min:	2.35M	Avg:	9.43M	Max:	41.38M
🟥 swap_cache	Cur:	0.00	Min:	0.00	Avg:	1.05k	Max:	128.00k
🟦 vmalloc_used	Cur:	260.42M	Min:	260.42M	Avg:	260.42M	Max:	260.42M
🟪 slab_cache	Cur:	101.93M	Min:	87.56M	Avg:	270.99M	Max:	792.57M
🟧 cache	Cur.	399.36M	Min:	123.75M	Avg:	3.61G	Max:	14.94G
🟥 buffers	Cur:	147.24M	Min:	22.78M	Avg:	233.57M	Max:	470.57M
🟦 unused	Cur:	14.96G	Min:	82.70M	Avg:	10.57G	Max:	15.26G
🟦 swap	Cur:	132.00k	Min:	0.00	Avg:	96.43k	Max:	132.00k
🟪 committed	Cur.	119.90M	Min:	99.20M	Avg:	2.08G	Max:	14.33G
🟥 mapped	Cur.	12.51M	Min:	10.98M	Avg:	75.46M	Max:	453.17M
🟪 active	Cur.	416.84M	Min:	192.46M	Avg:	2.26G	Max:	13.69G
🟩 inactive	Cur:	180.72M	Min:	42.87M	Avg:	2.53G	Max:	14.90G

Last update: Tue Sep 8 17:00:27 2009


Traffic load
🟪 1-2%
🟦 2-5%
🟦 5-10%
🟩 10-20%
🟨 20-40%
🟧 40-60%
🟧 60-80%
🟥 80-100%
Weathermap4RRD 1.2RC3
d Jul 22 14:55:07 2009
su253
HyperWin
(62)
(63)
dpmpool1
dpmpool2
(66)
(67)
(68)
(69)
(70)
(71)
(72)
(73)
(74)
(75)
(76)
(77)
(78)
(80)
suL101 (iberis)
suL102 (iberis)
suL021 (salix)
(1)
(2)
(3)
(4)
(5)
(6)
10G - e300 [CESNET]
storage3
storage5
Pasnet
FZU 8,33,[200]
FZK
Golias
papouch
suR071 (eva)
suL101-1 (iberis)
02-1 (iberis)
suL101-2 (iberis)
suL102-2 (iberis)
suL021-1 (salix)
suL021-2 (salix)
sw191 - velin
storage4
suL031 (golias)
suL041 (golias)
suL071 (golias)
suL081 (HP blade p)
suL082 (HP blade p)
suL083 (HP blade p)
suL091 (HP blade c)
suL092 (HP blade c)
suR081 (dpmpool)
dpmpool1/1
dpmpool2/1
dpmpool1/2
dpmpool2/2
dpmpool1/3
dpmpool2/3
torque
dorje (rack leader)

Services failover

- Important feature, not systematically implemented
  - Trying to run two copies of important (e.g. Grid) services
    - DNS, LDAP, computing element (CE), User Interface (UI)
    - DHCP server – two instances
    - Storage dual connections over fibre channel
  - Or a copy of the service in the virtual machine (Xen)
    - Quick deployment of other copy in case of problems
    - …
  - Hardware copy of important server with special interfaces

Conclusion

- Relatively slow centre development in previous years
- Accelerated when receiving financial resources for LHC computing last year, immediate serious problems
  - Cooling insufficient, substantial upgrade
  - Electric UPS power coming to limits
- 10 Gbps network backbone planned this year
- Resources used both with international communities and local users
- New solutions like iDataPlex and Altix ICE are effective, powerful, simple to install, space economic, electricity effective