Swiss Institute of Particle Physics

ETH Institute for Particle Physics

Computing Board Report

Christoph Grab and Derek Feichtinger

CHIPP Plenary 15./16.10.2007

Contents

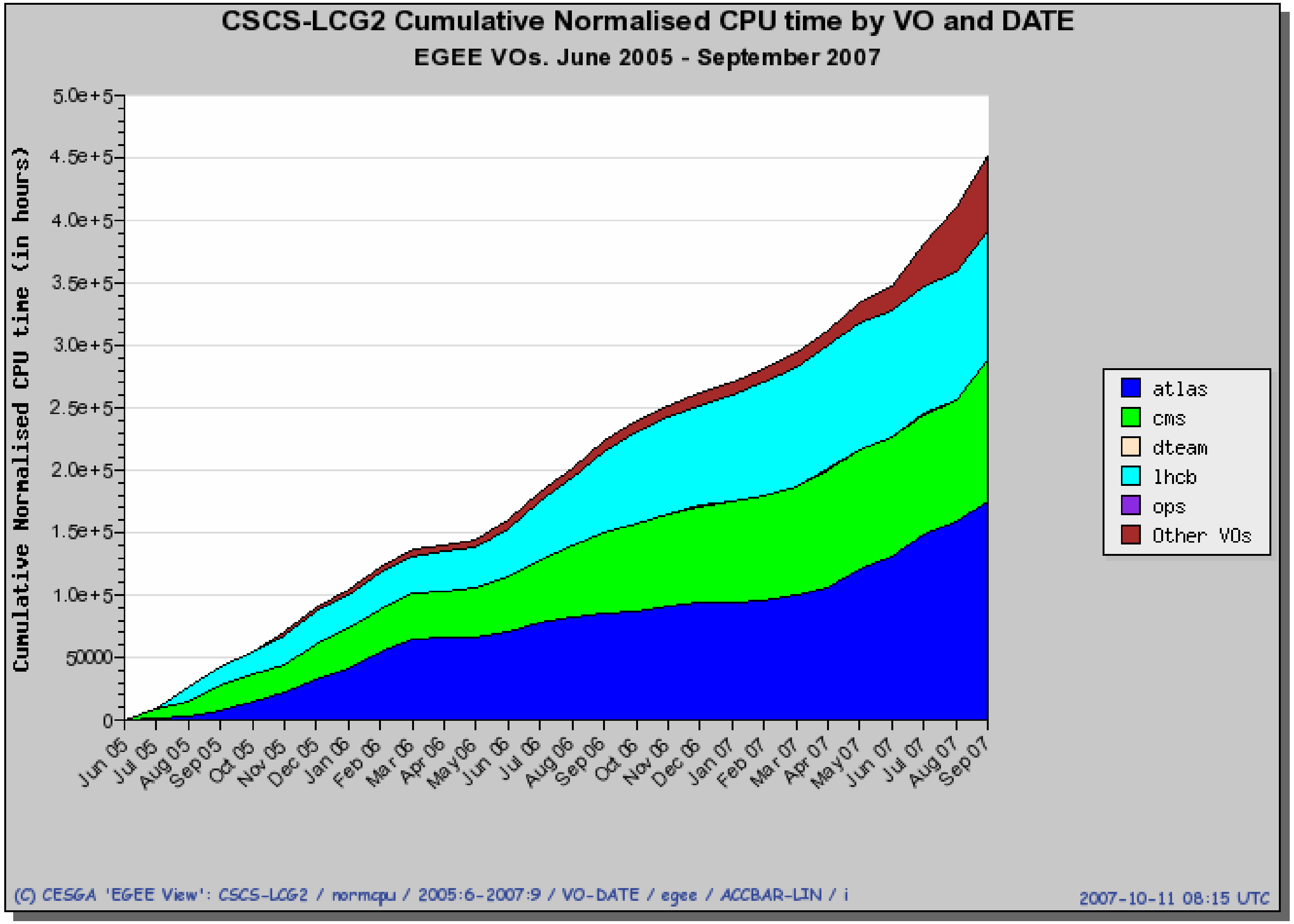
- Status and Plans : Swiss Tier-2 (CG)
  - Overview cluster hardware setup and upgrade path
  - Issues with Tier-3s
  - Personnel issues

- Status technical operations (D.Feichtinger)
  - Status operation and lessons learned
  - Participation in data and analysis challenges of experiments (e.g. CSA07 ..)

Status of the
Swiss Tier-2
PHOENIX cluster

PHOENIX LCG usage (6/05-7/07)

shown is **LCG-usage over last 2 years**



(C) CESGA 'EGEE View': CSCS-LCG2 / normcpu / 2005:6-2007:9 / VO-DATE / egee / ACCBAR-LIN / i

2007-10-11 08:15 UTC

Rest (not shown) are

- NorduGrid + user jobs

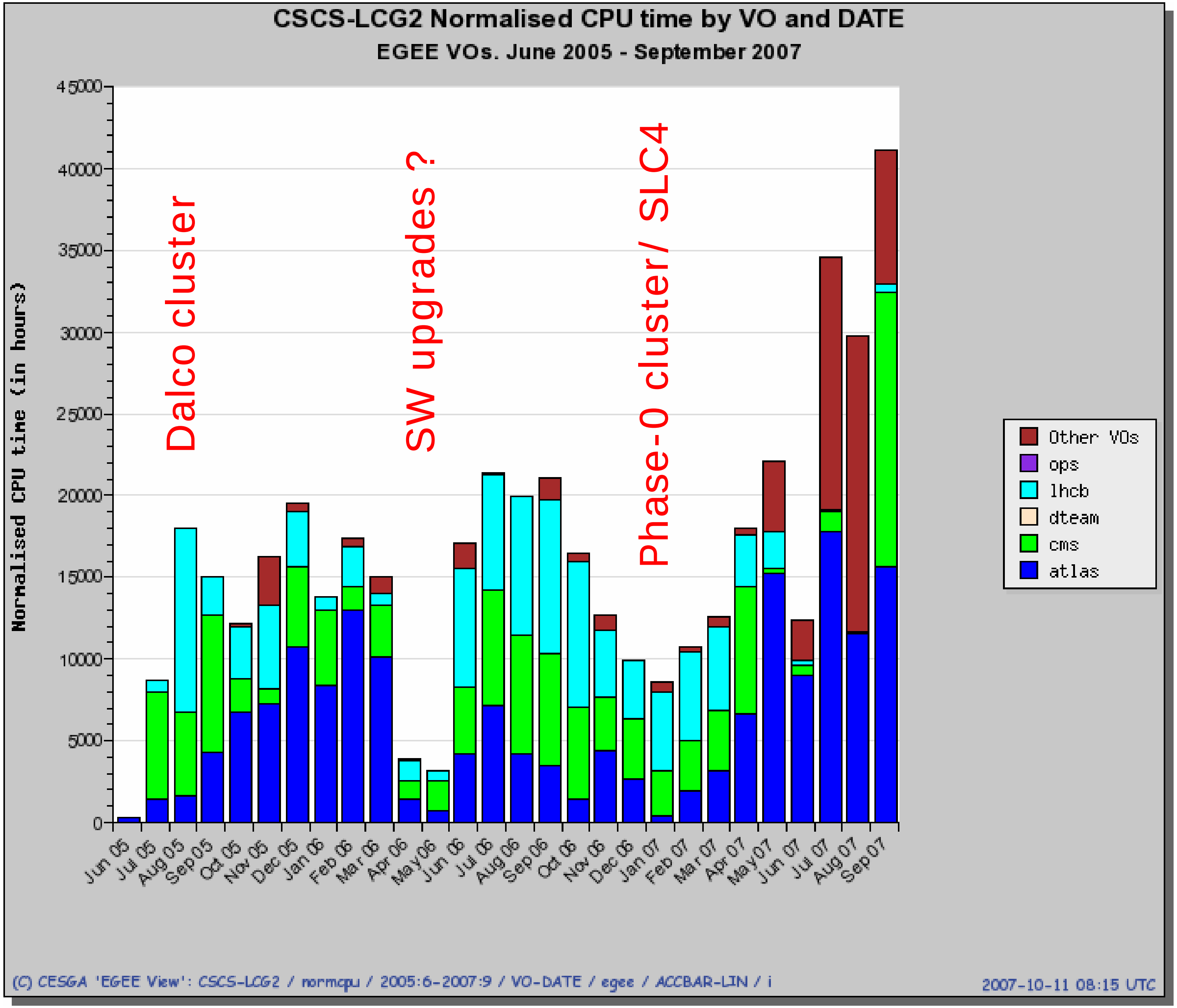
- High efficiency
  user analysis jobs now
  ~100% efficiency

- You can get any plots via Phoenix Wiki

➔ **Cluster operates stably (with high efficiency)**

➔ **chosen architecture functions well**

PHOENIX usage (9/06-7/07) [ref]



(C) CESGA 'EGEE View': CSCS-LCG2 / normcpu / 2005:6-2007:9 / VO-DATE / egee / ACCBAR-LIN / i

2007-10-11 08:15 UTC

LCG-usage per VO

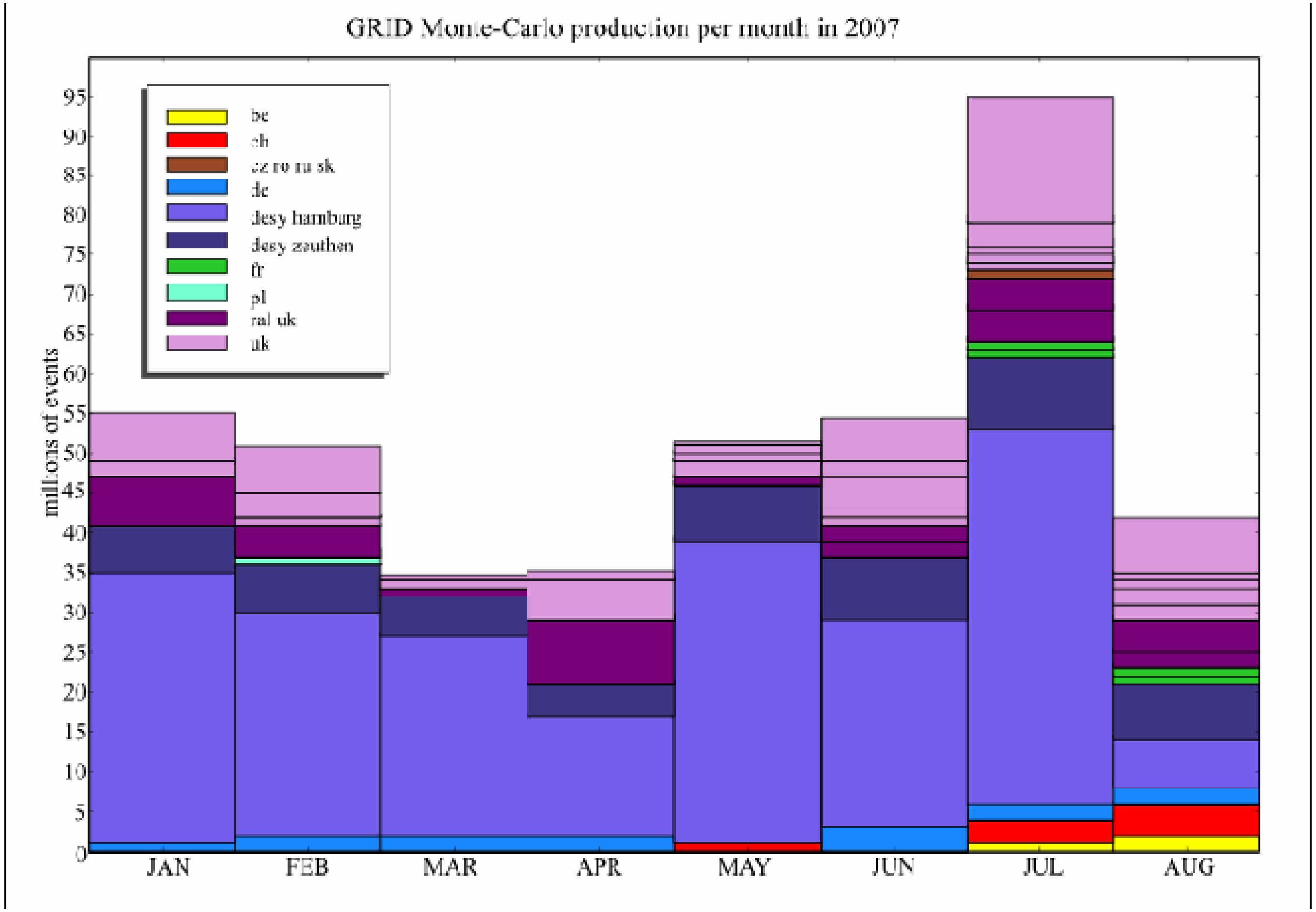
Nordugrid and user jobs not shown

Idle cycles given to others

Comment on PHOENIX usage

Idle cycles are not wasted…
but made available to registered VOs

Illustration: H1 MC production



CH / CSCS

HW Status of PHOENIX @CSCS ('07)

- Service nodes running services:
  - 4 management nodes for the Storage Elements SE/dCache: ➔ **effective 36 TB RAID5**
  - User interface, LFC, MON, GANGLIA,…
- 24 Worker Nodes (126 active cores)
- Total sum is **214 kSI2k**
- 3 VO-boxes: **CMS, Atlas, ARC(Nordugrid) and** some test nodes
- see D.Feichtinger on operations details

PHOENIX Phase-0

Evolution of PHOENIX cluster

Ramp-up Plan CH-Tier-2: status Q3/07

- Actual ramp-up schedule ; calculations based on Q2/07 pricing, to be installed in collaboration with SUN (phase-A)

Phases	Latest installation time	Minimum *aggregate* compute capacity [kSi2000]	Minimum *available* disc space [TB]	Cost estimates in kCHF
Existing cluster +CSCS GRID	operational	214	52	250
Phase A	End 2007	820	280	1120
Phase B	End 2008	1500	420	~ 1300
Phase C	End 2009	2600	800	~ 1300

* 1 XEON 3 GHz ≈ 1.5 kSI2000

- WAN: need > 3 Gbps in 2008

have already now 1 Gbps, and can get 2x10 Gbps anytime ☺

Projected Components PHOENIX Phases

Assumptions:

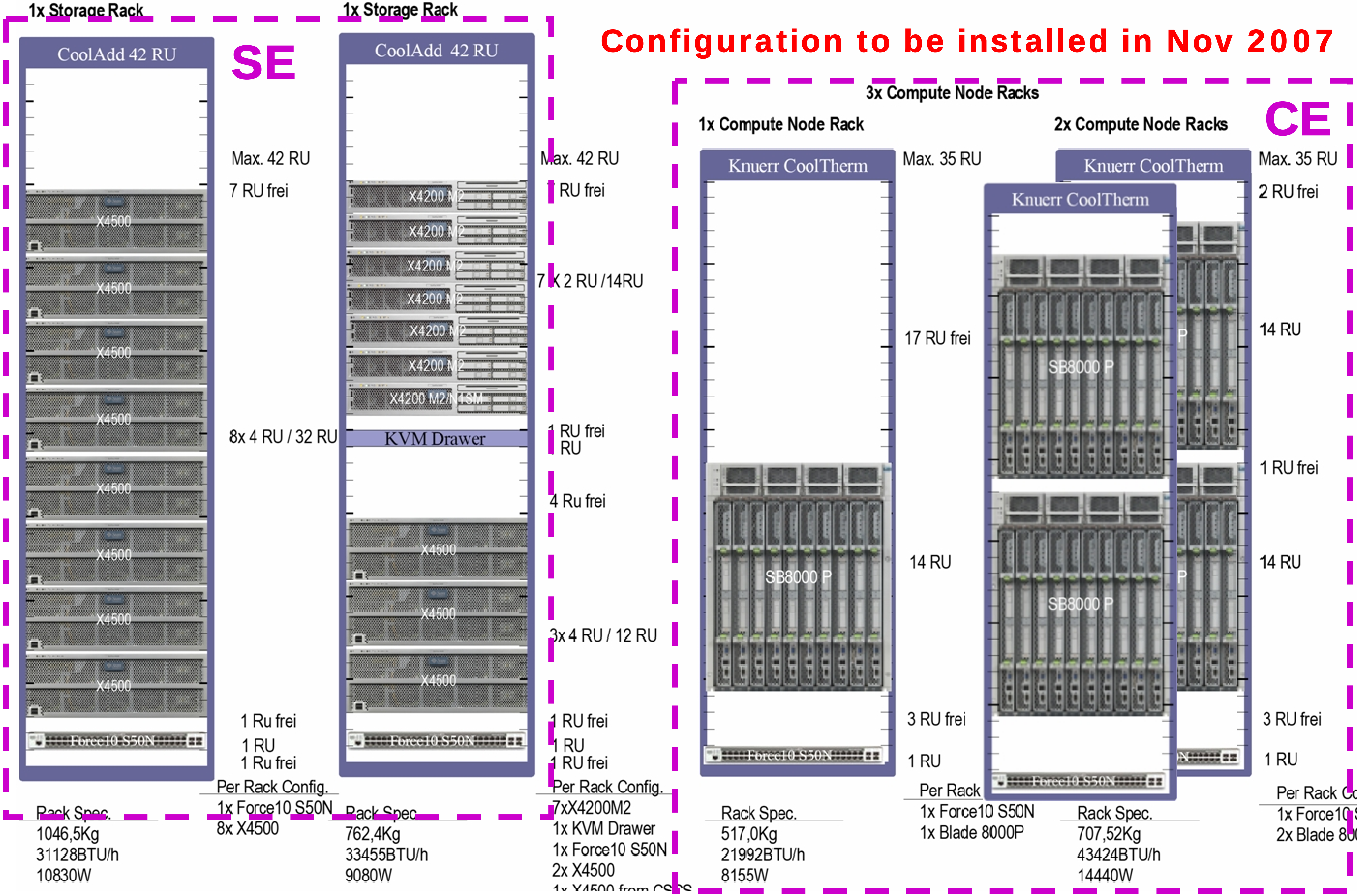
- **CPUs:** Opteron 2.6 GHz CPU with 1.5 kSi2k per core.
  - Phase A Dual Core CPUs
  - Phase B Quad Core with Phase A duals upgraded to Quads
  - Phase C Quad Core
- **Disk: X4500 :** net capacity of 17.85 TB (incl. raid, mirror, spares, overhead)

	Aggr. CPU Performance in Ksi	Additional Cores needed	Cores delivered	Aggr. Net Storage Capacity in TB	Storage needed	Storage delivered
Phase A	680 * *	453	400 (5 Blade Centre)	225	225TB (12.6 X4500)	178,5TB (10 X4500) +1 CSCS
Phase B	1440	+453 + 54 (507)	+(400 + 160) (+ a 6th Blade Centre)	490	+265TB (+14.8 X4500)	+303.4TB (+17 X4500)
Phase C	2640	800	+800 (5 Blade Chassis)	910	+420TB (+23.5 X4500)	+410TB (+23 X4500)
Total	2640	1760	1760 (11 Blade Centre)	910	910TB (50.9 X4500)	910.3TB (51 X4500)

* * new processors (in 2007) will boost this 680 to about 800 kSI2k

PHOENIX Phase-A Hardware

Configuration to be installed in Nov 2007



SE

1x Storage Rack

Max. 42 RU

7 RU frei

8x 4 RU / 32 RU

1 Ru frei

1 RU

1 Ru frei

Per Rack Config.
1x Force10 S50N
8x X4500

Rack Spec.
1046,5Kg
31128BTU/h
10830W

1x Storage Rack

Max. 42 RU

1 RU frei

7 X 2 RU /14RU

1 RU frei
1 RU

4 Ru frei

3x 4 RU / 12 RU

1 RU frei

1 RU

1 RU frei

Per Rack Config.
7xX4200M2
1x KVM Drawer
1x Force10 S50N
2x X4500
1x X4500 from CSCS

Rack Spec.
762,4Kg
33455BTU/h
9080W

CE

3x Compute Node Racks

1x Compute Node Rack

Max. 35 RU

17 RU frei

14 RU

3 RU frei

1 RU

Per Rack
1x Force10 S50N
1x Blade 8000P

Rack Spec.
517,0Kg
21992BTU/h
8155W

2x Compute Node Racks

Max. 35 RU

2 RU frei

14 RU

1 RU frei

14 RU

3 RU frei

1 RU

Per Rack Config.
1x Force10 S50N
2x Blade 8000 P

Rack Spec.
707,52Kg
43424BTU/h
14440W

Components Present Cluster [ref]

For tech-experts ...

PHOENIX

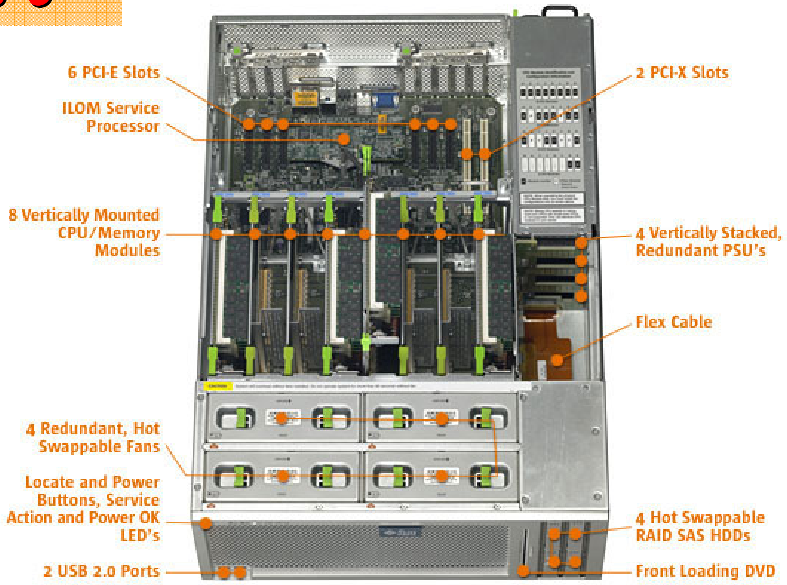
Phase-0

CPU: SUNfire X2200 M2 x64 servers:

total of 26 server modules;

2x AMD Opteron dual-core 2.7 GHz with 2 GB/core

(~ 1.7 kSI2k per core)


6 PCI-E Slots
ILOM Service Processor
8 Vertically Mounted CPU/Memory Modules
4 Redundant, Hot Swappable Fans
Locate and Power Buttons, Service Action and Power OK LED's
2 USB 2.0 Ports
2 PCI-X Slots
4 Vertically Stacked, Redundant PSU's
Flex Cable
4 Hot Swappable RAID SAS HDDs
Front Loading DVD

→ total of ~180 kSI2k

Disk : SUN Fire 4500 x64 Servers:

2x AMD Opteron dual-core 2.7 GHz with 4 GB/core

total 2 units with each 48 disks,

→ ~ effective 35 TB

Phase-A components ~ same

(or newer version, if exist)

Status Financing

Financing Tier-2 - Timeline

- Prototype financing by institutes: in 1.2004
  for CSCS-LCG cluster [10 nodes]; granted 50 kFr
- 1st FORCE grant : requested 128 kFr. in 3.2004;
  →for "Phoenix cluster" [Prototype 15 nodes] granted 128 kCHF
- 2nd FORCE grant : requested 670 kFr in 9.2005;
  → "Phoenix cluster" [Phase A/1] in 3.2006 granted 300 kCHF
- 2nd +: addendum in 9.2006
  → "Phoenix cluster" [Phase A/1+] granted 190 kCHF
- 3rd FORCE grant : requested 670 kFr in 9.2006;
  → "Phoenix cluster" [Phase A/2] in 3.2007 granted 500 kCHF
- 4th FORCE grant : requested 1300 kFr in 9.2007;
  → "Phoenix cluster" [Phase B]
- Planned 5th FORCE grant : ask for ~1300 kFr in 9.2008;
  → "Phoenix cluster" [Phase C]
- from then on... rolling replacement; order 500 kFr / year

Phase A is financed

asked for phase-B

Other Financial Contributions

Additional contributions so far:

- Unis+ETH: granted in 1.2004 50 kFr.
  invested in prototypes CSCS-LCG cluster [10 nodes]
- ETH-IPP : granted in 12.2006 50 kCHF
  ➔ invested in thumper addition for PHOENIX [24 TB]
- ETH-PSI : granted in 2007: 50 kCHF
  ➔ invested in PHOENIX phase A
- UNIZ : granted in 2007: 30 kCHF
  ➔ invested in PHOENIX phase A

Note: also in the future:
contributions by Unis and ETH are strongly suggested !

Local CSCS Infrastructure Details ...

Infrastructure Upgrades at CSCS

Upgrades of infrastructure in machine room at CSCS in preparation for PHOENIX (and other clusters) installation :

New cooling aggregate	done
new dynamical UPS	done
New hydraulics	in progress (due 28.10.07)
Electrical installation	due 4.Nov.07

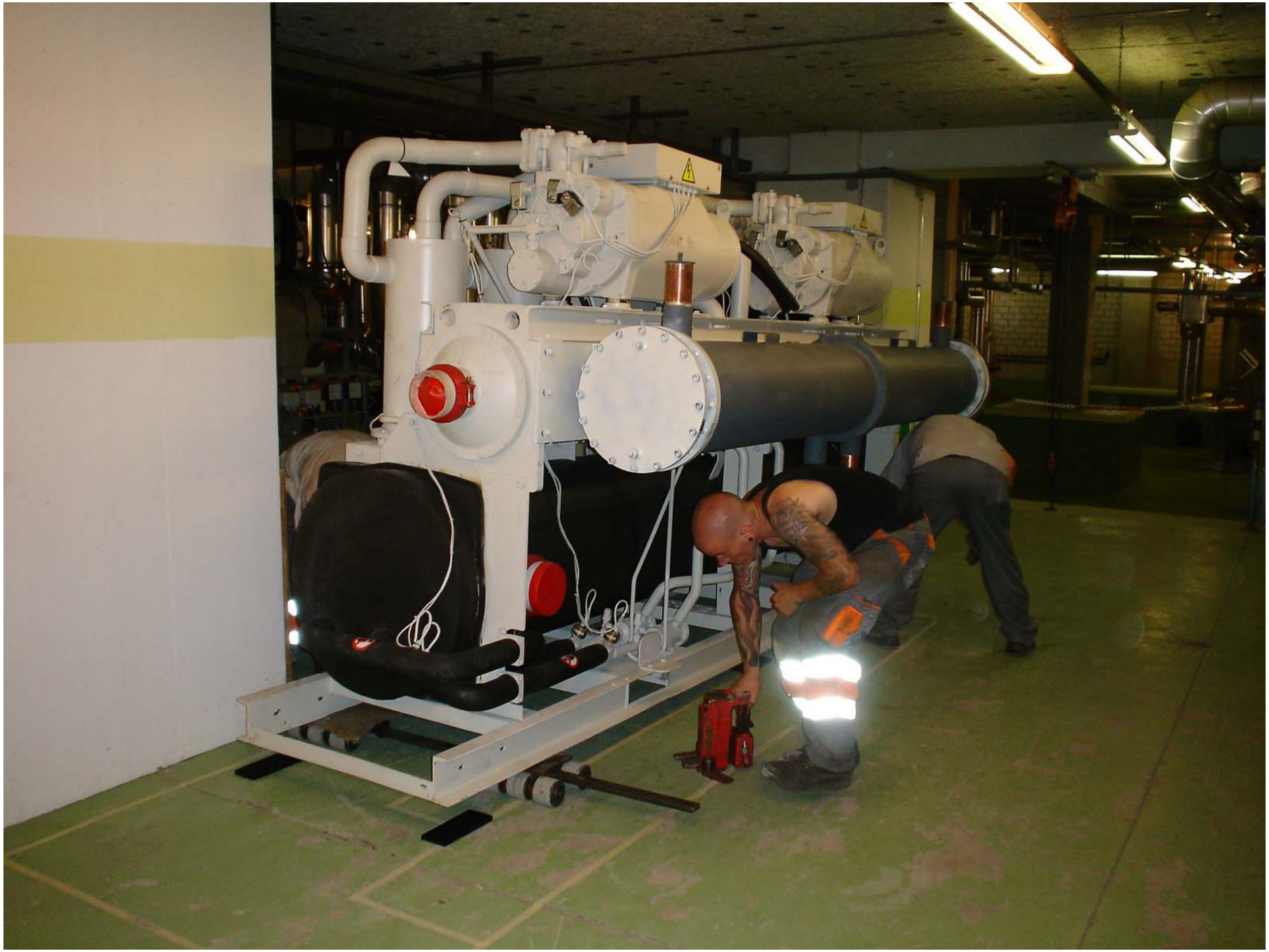
Arrival PHOENIX HW from SUN to complete the phase-A installation	12.Nov.07
Installation and commissioning phase-A	Nov. 07
Goal: full operation of phase-A in	Dec. 07

thanks to CSCS (Kunszt,Guptill)

Infrastructure Upgrades at CSCS

- Arrival new cooling components

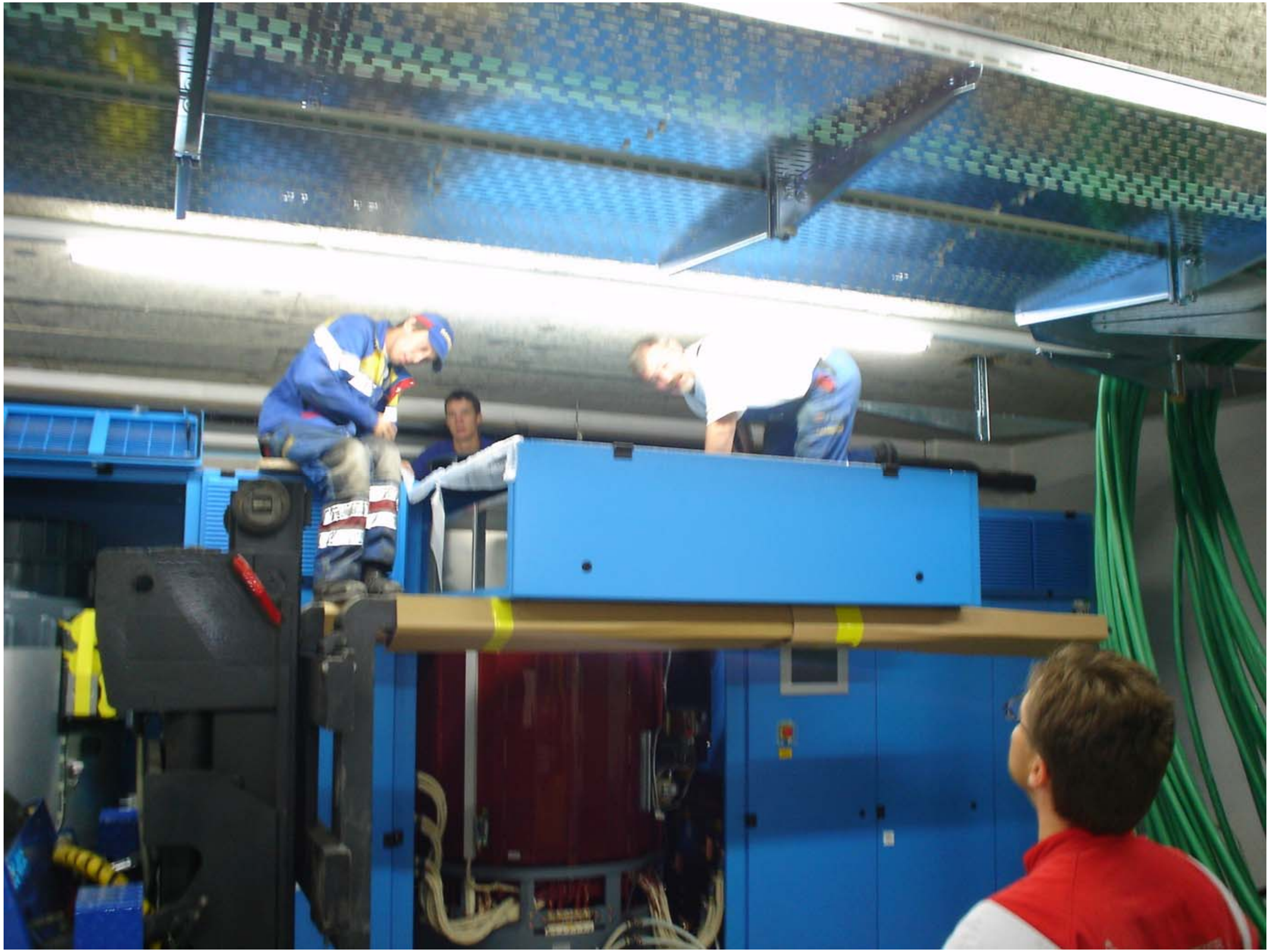
20.6.07



Infrastructure Upgrades at CSCS

- Installation of new dynamical UPS

22.8.07



Tier-3 Issues ...

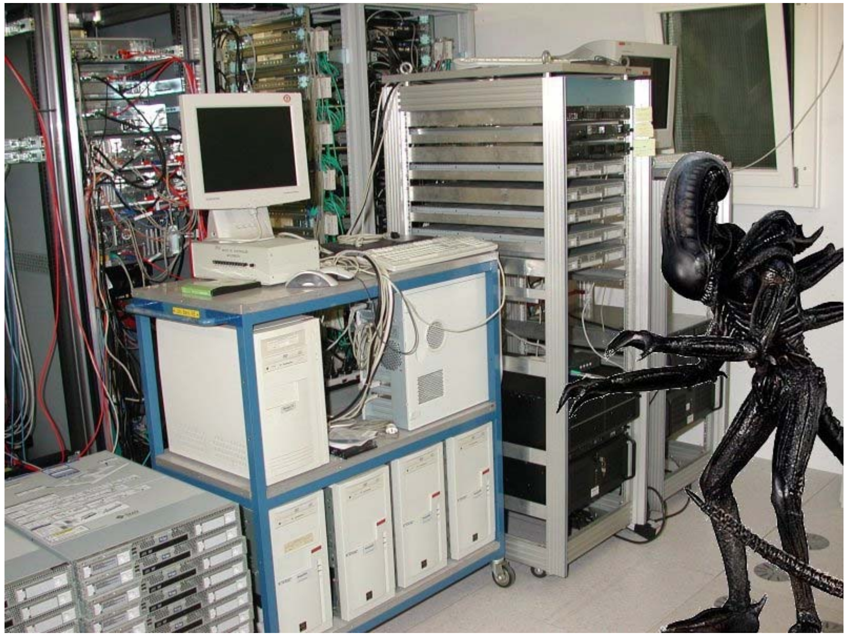
ATLAS Tier 3 in Geneva

- a system in production since 2005, mainly as a grid batch job facility
  - over 100’000 CPU hours in 2007 for ATLAS
- recently more interactive use by the Geneva group
  - development and testing of code, analysis of n-tuples, short batch
- size
  - now 84 CPU, 26 TB
  - for 1st data 188 CPU, 75 TB

thanks to S.Gadomski

The Bern ATLAS Tier 3 in 2007

Two clusters with NorduGrid front ends in production since 2005. For local physics analysis and simulation. Fills up with ATLAS central production jobs when not used by locals.



Size

~130 cores for ATLAS.
~ 33 TB disk (end of 2008 44 TB).

Usage

~ 120 000 Wall Time Hours in 2006.
~ 130 000 Wall Time Hours in 2007 so far.

thanks to S.Haug

Status CMS - Tier 3

- **Presently operating**: CMS-nodes at ETH (Trueb, Dambach):
  5 Servers (each: 2 Dual-Xeons 3.2 GHz + eff. 4.5 TB) :
- **Planned: common CMS Tier-3** for ETHZ, PSI + UNIZ;
  located and operated at PSI; choose similar architecture as Tier-2
  (planning D.F. + C.G.)

Year	**2008**	**2009**
CPU / kSI2k	**180**	**500**
Disk / TB	**75**	**250**
No of Worker Nodes	10	28
No of CPU Cores	80	224
No of Storage Nodes	4	14
No of Racks	1	3

→ emphasis on storage

→ **to be operational in Q1/08**

Network: aim for 6 MB/s read access per job and kSI2k;
Connection of 1 Gb between PSI and CSCS.

D.Feichtinger+C.Grab

Other issues …

Comments on Personnel

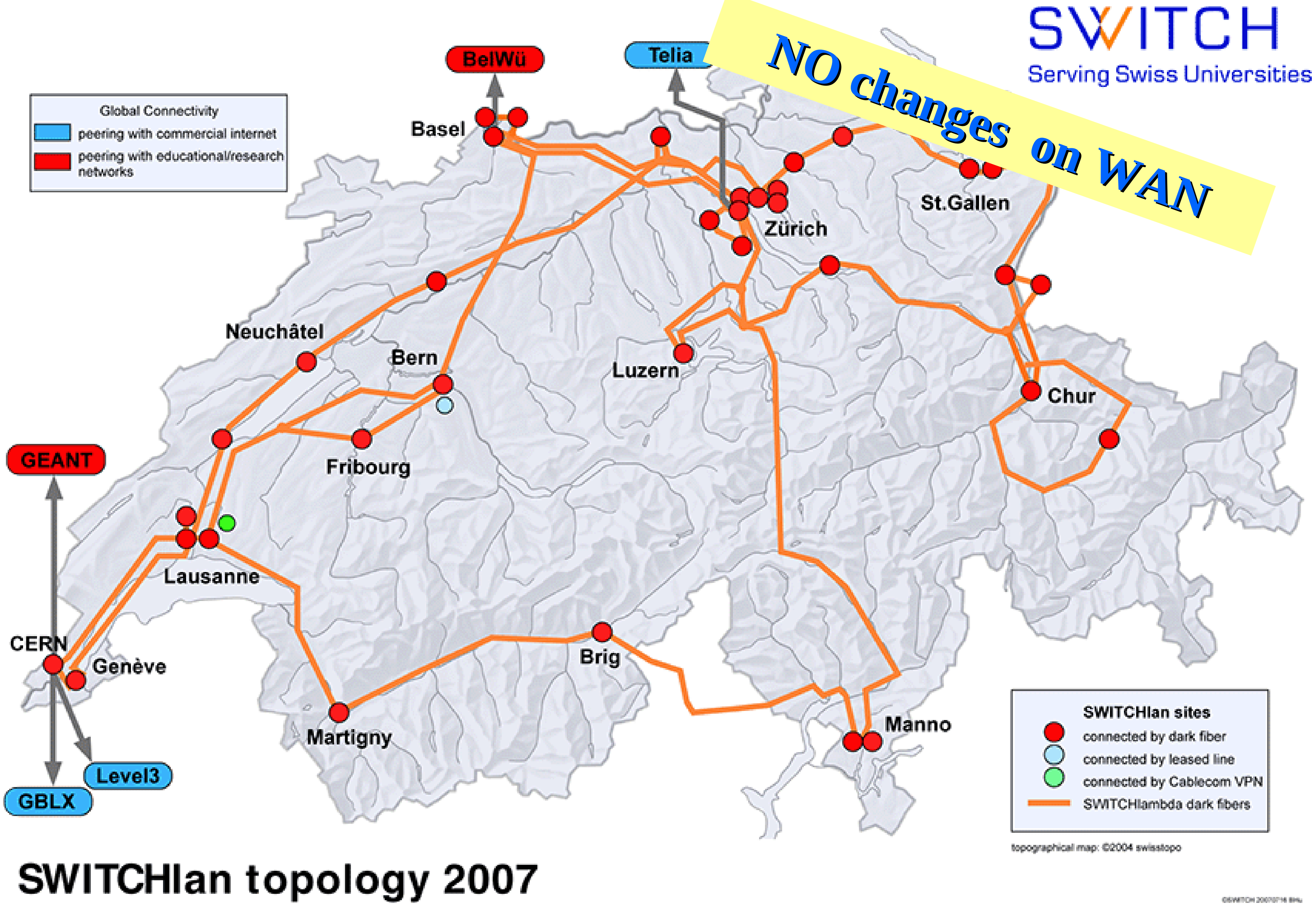
- **Personnel at CSCS for GRID computing :**
  - Peter Kunszt; S.Maffioletti, Tom Guptill
  - new dCache expert needed Q4/07... position needs to be filled !!!
  - total: 2 GRID-FTE + about 2 FTE for operational/technical support

- **CHIPP personell for experiment's related operations**
  - **CMS:** D.Feichtinger (100%, PSI);
    → crucial contributions to the overall T2- PHOENIX operation – thanks !
    Tier-3: DF for operation of planned T3-cluster at PSI; Dambach/Trueb now;
  - **ATLAS**: Szymon Gadomski (Geneva) (100%), S.Haug (Bern)
  - **LHCb**: Roland Bernet (UniZ), parttime (no need for fulltime now)
  - **CHIPP**: C.Grab parttime for all expts./CHIPP: coordination Tier-2 (+ CMS-T3)

- Note: we will need a second person to support middleware for each experiment at the Tier-3 !

Network: Switch backbone [ref]

NO changes on WAN


SWITCH
Serving Swiss Universities
Global Connectivity
🟦 peering with commercial internet
🟥 peering with educational/research networks
BelWü
Telia
GEANT
Level3
GBLX
Basel
Zürich
St.Gallen
Neuchâtel
Bern
Luzern
Chur
Fribourg
Lausanne
CERN
Genève
Brig
Martigny
Manno
SWITCHlan sites
● connected by dark fiber
○ connected by leased line
○ connected by Cablecom VPN
— SWITCHlambda dark fibers
topographical map: ©2004 swisstopo

SWITCHlan topology 2007

Status: Networking

- CSCS to the world through SWITCH
  - Currently 1 Gbps is operational : CSCS → ZH → CERN
    10 Gbs within CSCS done; 10 Gbps access needed at ZH-Hub.
  - Connection through Domodossola/Brig ready to be equipped as soon as needed (Switchlambda dark fibers are available)
  - All GEANT (Europe) connections are through the CERN POP/CIXP
  - GEANT can also provide dedicated links to FZK, IN2P3, CNAF, RAL (if really needed)
  - CSA07 exercise for CMS ongoing NOW …
    provides valuable input to estimate future needs …

Where to get information? [ref]

- **Start from the Wiki pages at CH-T2 CSCS …**
  - **https://twiki.cscs.ch//bin/view/LCGTier2/WebHome**
  - **contains lots of references: expt's, Hypernews, trouble tickets..**

- **Monitoring sites: which sites are up and running, free slots...**
  - **http://goc.grid-support.ac.uk/gridsite/monitoring/**

- **Whom do I contact, if I have problems with PHOENIX at CSCS?**
  - **Experiment specific site contact persons (DF, SG, RB)**
  - **GGUS (Global Grid User Support): https://gus.fzk.de/pages/home.php**

Conclusions

- **Message 1**:
  - the **Swiss Tier-2 “PHOENIX” cluster has been up and running routinely for over 2 years, servicing the experiments !!**
  - **Stability has increased over time - now high efficiency reached.**
    Thanks to D.Feichtinger and P.Kunszt etal @ CSCS
  - **Present hardware architecture proved a good choice, continue along these lines ...**

- **Message 2:**
  - **Many lessons learned from daily operation**
    and from participations in CSA ...
    ... covered by D. Feichtinger

We’ll be ready for LHC data ☺

CHIPP Computing Board

CSCS
Swiss National Supercomputing Centre

A.Clark, S.Gadomski (UNI Ge)
H.P.Beck, S.Haug (UNI Bern)

C.Grab (chair CCB, ETHZ)
U. Langenegger (ETHZ)
D.Feichtinger (PSI) (vice-chair CCB)

R.Bernet (UNIZH)
J. Van Hunen (EPFL)

M-C. Sawley (CSCS general manager)
P. Kunszt (CSCS Swiss Grid Initiative)

Technical issues and Lessons learned ...

Derek Feichtinger

Swiss Tier-2 at CSCS

In case of questions ...

Points for discussion (opt)

- Amount of user resources for official production at Tier-2:
  - CMS: 1/3 production to 2/3 users for CPU
  - ATLAS: no CPU for users; only about 1/3 user disk at tier-2 (CPU only at tier-3)
  - LHCb: disks at tier-2 only for user analysis (no production storage)

- Our total estimates are based on the TDR which foresaw start-up in 2008.
However the TDR-numbers and time-scales have changed.
Q: do we want to increase/revise the total requirements now?
→ this is a moving target (requests always increase…)

- Need to setup a resource allocation structure
  - provide tools in SW
  - define rules how to deal with this

Timeline Swiss-T2 hardware [ref]

- April 2004: Prototypes installed at CSCS (20 nodes AMD MP)
  => 20 WN (12 kSI2k); 3.2 TB SE; → very unstable operation
- Jan 2005 : decide to acquire PHOENIX-cluster from Dalco
- June 2005: operational => 15 WN (45 kSI2k), 8 TB SE;
- Nov 2006: CSCS buys CSCS GRID-cluster (phase-0; 250 kCHF)
  CHIPP receives full access to this cluster for LHC-purposes
- Dec 2006: IPP/ETH buys SUN thumper 24 TB as addon
- Mar/April 2007: phase-0 GRID-cluster operational
  => 52 x 2-CPU-dualcore AMD WNs : 180 kSI2k + 1 SE a 24 TB
- Nov 12, 2007: full phase-A PHOENIX cluster delivered
- Q3/2008: phase-B PHOENIX cluster delivered ?
- Q3/2009: phase-C PHOENIX cluster delivered ?

Original: Ramp-up Plan CH-Tier-2 (opt)

- Gradual increase according to percentage of CPU to reach size of the full Tier-2 RC by (end) 2008.

Original Plans in '04

	2005	2006	2007	**2008**
CPU (kSI2000)	45	231	692	**2307**
Disk (TB)	9	79	236	**787**
Tape (TB)	0	0	0	**0**
CPU fraction w.r.t. 2008 (%)	2	10	30	**100**

- **The final goal has not changed.. but the timeline and steps ...**

*** 1 XEON 3 GHz ≈ 1.5 kSI2000**

Status of PHOENIX @CSCS (06) [ref]

PHOENIX

- 5 Service nodes running services:
  - Compute Element **(CE)**
  - Storage Element **(SE/DPM: 8TB RAID5**)
  - User interface
  - LFC, MON, GANGLIA
- 15 Worker Nodes (30 CPUs)
- 3 VO-boxes: **CMS, Atlas, ARC(Nordugrid)**
- 1 management node + test nodes

Node Architecture:
Intel dual-Xeon 3.0 GHz, 4 GB RAM
Intel Pentium4 3.0 GHz, 1 GB RAM

45 kSI2000
(1 Xeon 3.0 GHz ≈ 1500 kSI2000)
2 x 1 Gbps network access

ATLAS Tier-3 Federation [ref]

- Swiss ATLAS Cluster Resources (see also A.Straessner)
  - compute elements (CE): typical number of worker cores
  - Storage elements (SE) disk space in TB
  - Shared resources: ATLAS shares, with full resource sizes in parentheses.
  - The Bern Tier-3 Shared is the central university facility.

	CSCS T2 shared	Ge T3	Be T3	Be T3 Shared	SUM
CE Cores 2006	10 (of 30)	24	36	100 (of 288)	170
CE Cores 2007	50 (of 140)	108	32	100 (of 506)	290
CE Cores 2008	170 (of 500)	188	59	100 (OF 506)	508
SE [TB] 2006	2 (of 10)	7	10	0	19
SE [TB] 2007	14 (OF 40)	25	10	0	49
SE [TB] 2008	90 (OF 280)	75	25	0	190