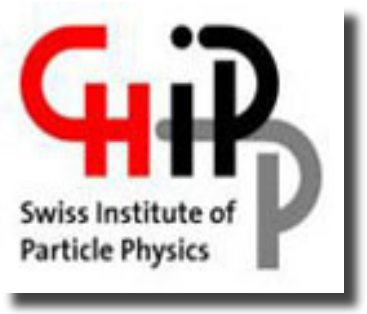
# HEP Computing in Switzerland




Christoph Grab (ETH)

Head of CHIPP Computing Group

CHIPP, August 23, 2016

# Status of WLCG Tier-2 and Tier-3 computing resources in Switzerland

Only few updates with respect to ECFA status report on 1.4.2016

# Overview Swiss LHC Computing Resources

- **Switzerland operates Tier-2 Regional Centre at CSCS and AEC**
  - Maintain our own dedicated compute-cluster integrated into “WLCG” .
  - Switzerland is committed as full member to contribute resources; signed MoU
- **Tier-2 operated by CSCS, serves all 3 experiments: ATLAS,CMS, LHCb**
- **Tier-2 operated at AEC-UNIBE serves ATLAS only**

  - Collaboration agreement for operation of T2 between CHIPP and CSCS/ETHZ (2007-2018 with additional ETHZ funding secured)
  - Available resources provided to WLCG and exploited centrally by experiments

CSCS
Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre

Swiss Tier-2 Phoenix cluster at Lugano

- **Complemented by local Tier-3 clusters** at PSI, UBe+UGe, UZH+EFL

# Overview Swiss LHC Computing Resources

**EGI** (European Grid Infrastructure); NGI **WLCG** (Wordwide LHC compute Grid)

**CSCS T2 -ATLAS, CMS,LHCb**
- 6240 cores; 69 kHS06
- ~3500 TB disk

CPU / Disk share~ 35:35:29 / 39:39:22

- 40 Gb/s to 100 Gb/s backbone

**AEC-UNIBE T2 (ATLAS)**
- 2300 cores; 20 kHS06 (2 clusters);
- 500 TB disk ; 10 Gb/s

**PSI-ETHZ-UZH T3 (CMS)**
1040 cores; 12.5 kHS06;755 TB

**DPNC-UNIGE T3 (ATLAS)**
784 cores; 6.0 kHS06; 828 TB
Direct 10 Gb/s to CERN IT

**UZH T3 (LHCb)**
480 cores; 5.2 kHS06; 250 TB

**EPFL T3 (LHCb)**
410 cores; 6.8 kHS06; 80 TB

**Operation**
- Monthly meetings; (CH; EGI/GDB).

**Networking by SWITCH)**

**Coordination**
CHIPP Computing Board
Regular F2F meetings of CCB
(also SwiNG as NGI interface to EGI)

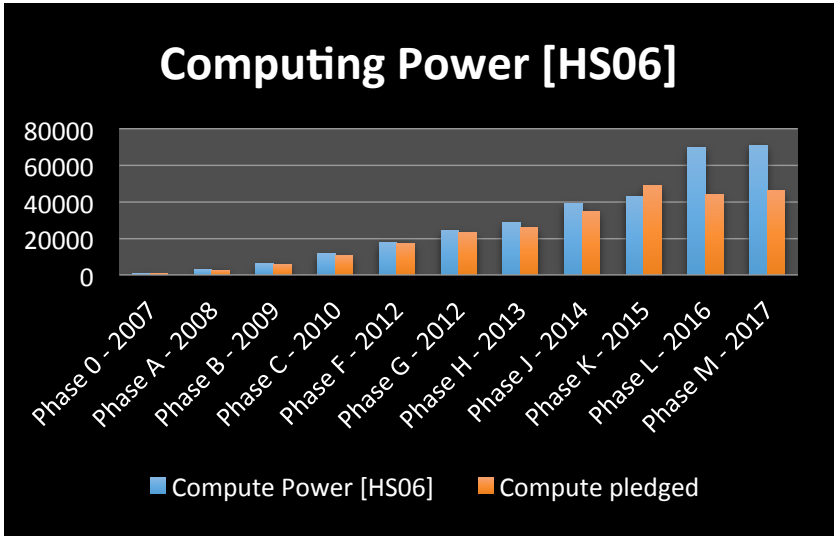
CG 8/16

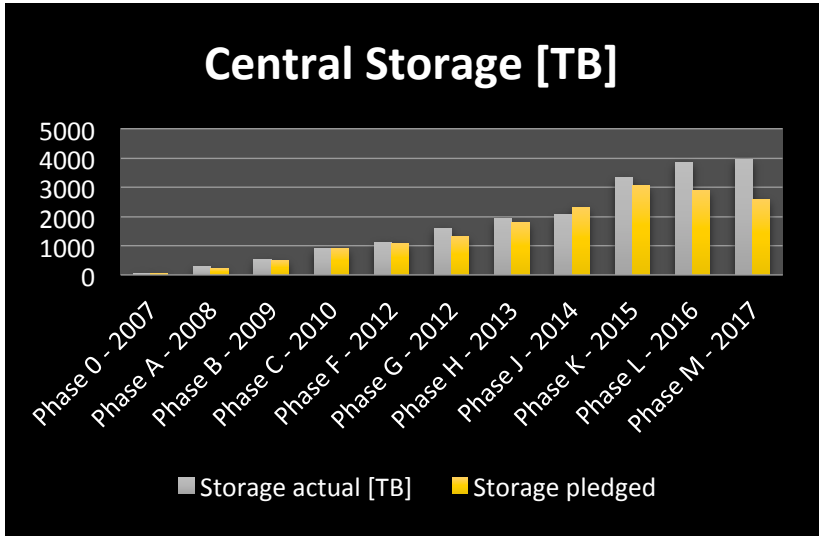
Note: sum of Tier-3 resources [~30 kHS06; 1.5 PB]
equals ~ 2/3 of Tier-2 resources (except ATLAS)

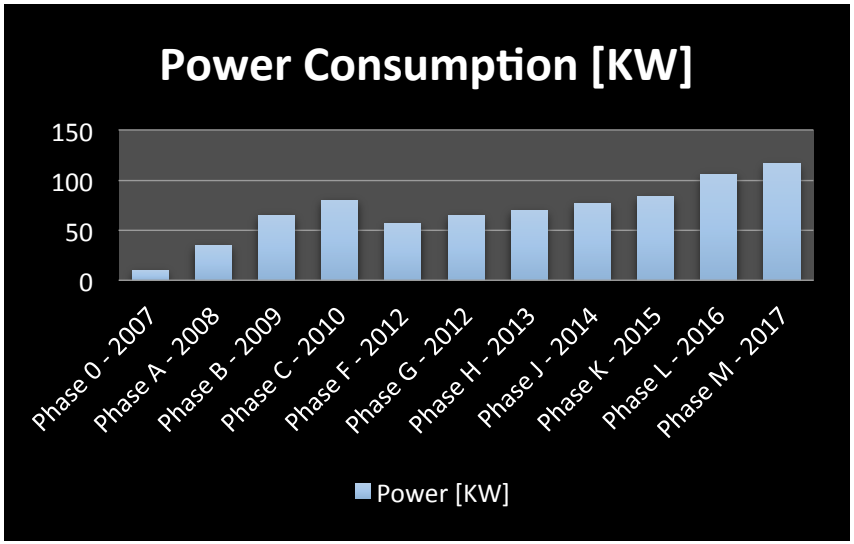
# Swiss Tier-2 Resources Evolution

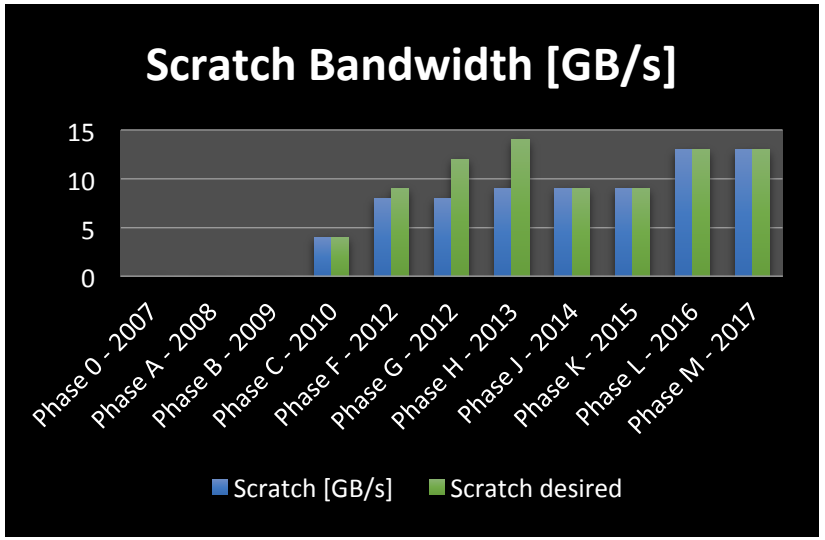
## Evolution for 2007 – 2017

(phase K; met pledges 1.4.2016; now in phase L)



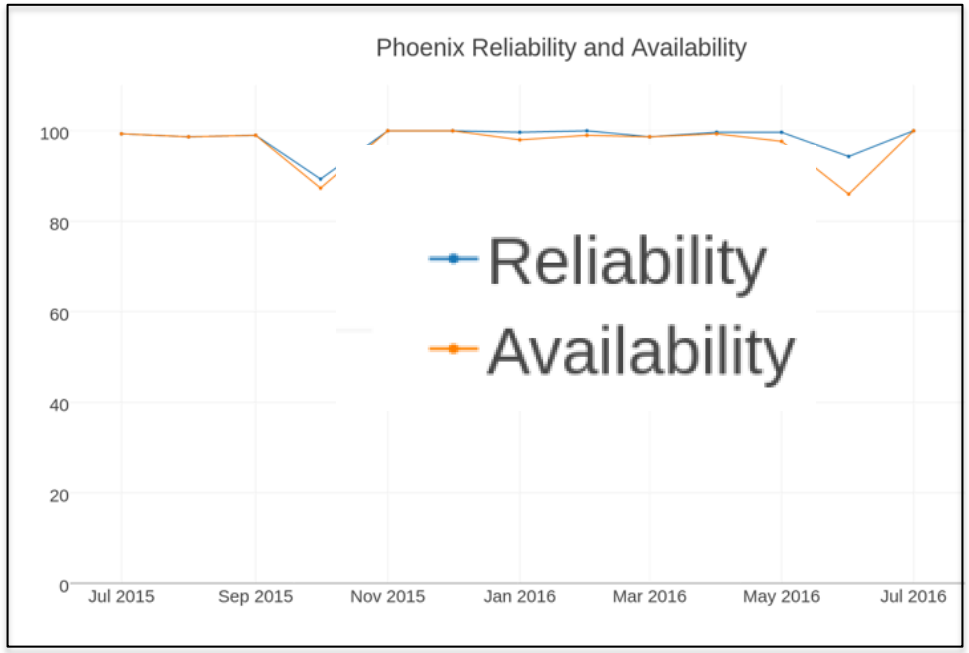



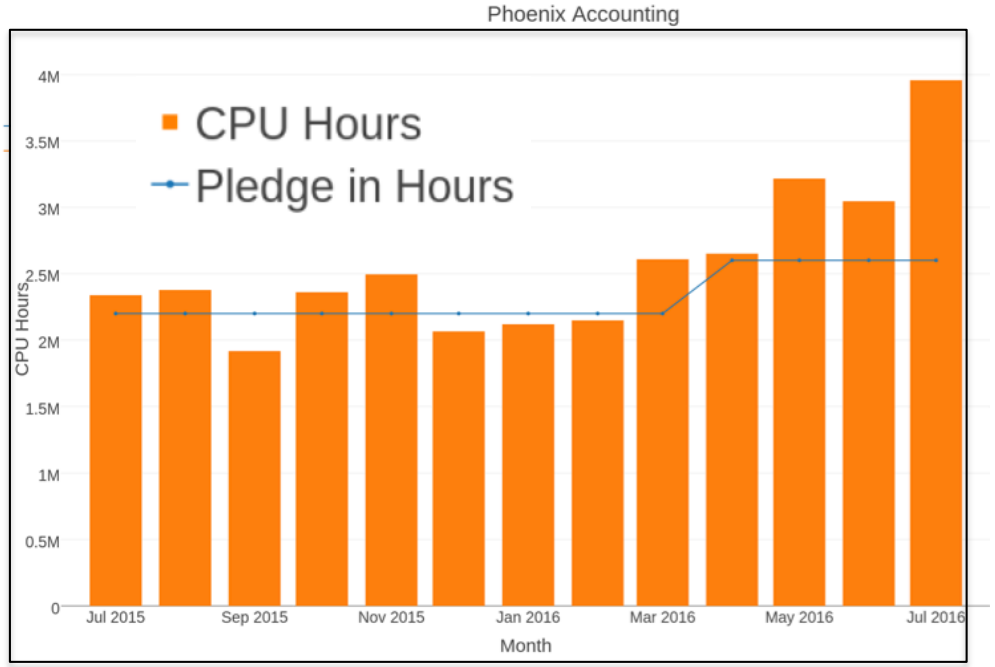

**actual CPU+Disk > pledges**

8.2016

# Cluster CPU statistics (2015-2016)

## For period 2015-2016



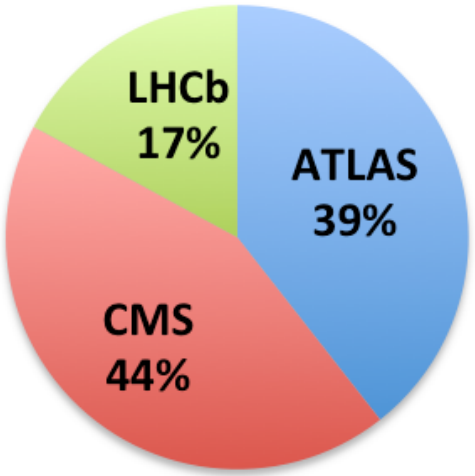

**Availability and reliability of the cluster to experiments (WLCG)**

Overall high (>95%).

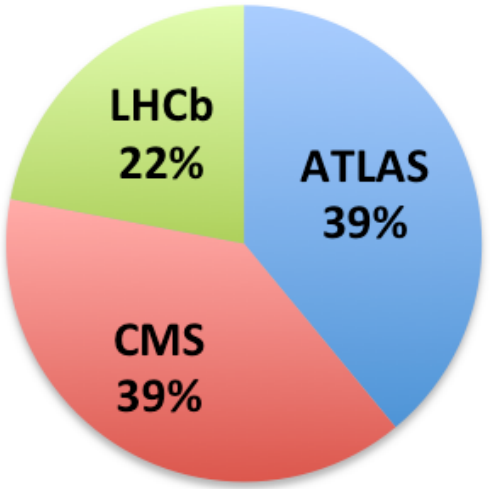
Pledges (in CPU hours ) are met.

# Resource distribution @CSCS (8/2016)

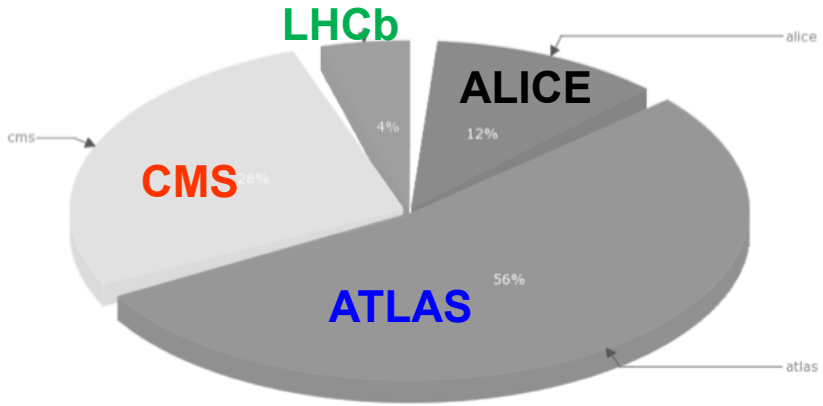
## Walltime CPU usage 1.2015-1. 2016




## Storage usage on 8.2016




## Compare to worldwide T2 usage




**Resource ratios at CSCS:** (8.2016)
ATLAS:CMS:LHCb

- CSCS fairshare ratio 40:40:20
- effective CPU usage: 35:35:29
- CSCS disk ratio: 39:39:22

# Comments on Resources

- **HW investments at CSCS (**replacements and additions) are based on C-RRB recommendations of a **"flat budget"**. Funded by FLARE/SNF
  Provides typically 15-20% increase of resource "power" per year.
- HW investment at AEC: HW + FTE from Institute/SNSF;
  power/infrastructure covered by the UniBe

**~ 3 FTE for T2**

- **Personnel for operation** :
  - 1.5 FTE to support Tier-2 operation at CSCS, covered by SNF/FLARE
  - 1 FTE covered by ETH internal funds for T2 at CSCS
  - + 0.5 for AEC-ATLAS only Tier-2 operation
  - Additional ~0.3 FTE per experiment as user- and experiment-specific software support, covered by institutes
  - Overall management and coordination tasks covered by ETH

**~1 FTE for expt. contact**

- **Other resource items T2 and T3**
  - Recurring power/infrastructure costs at CSCC are carried by ETH; at AEC-UniBe carried by UniBe.
  - Tier-3 hardware costs covered by institutes
  - specific Tier-3 manpower covered by institutes, partly by SNF

# Swiss Tier-2 Resource Planning

**COMPUTE Resources, in kHS06 units**

| Experiment | 2013 (delivered) Phase G | 2014 (delivered) Phase H | 2015 (delivered) Phase J | 2016 (in progress) Phase K | 2017 (planned) Phase L | 2018 (future) Phase M |
|---|---|---|---|---|---|---|
| ATLAS | 9.2 | 10.4 | 14 | 18 | 22 | 25 |
| CMS | 9.2 | 10.4 | 14 | 18 | 22 | 25 |
| LHCb | 4.6 | 5.2 | 7 | 13 | 18 | 22 |
| **TOTAL (kHS06)** | **23** | **26** | **35** | **49** | **62** | **72** |

Table 1: CPU resources, quoted in kHS06 units, as delivered or pledged to be available on April 1 of the year by the Swiss Tier-2 at CSCS as our national contribution (CH_Tier2) to WLCG[4].

**STORAGE Resources, in Terabyte units**

| Experiment | 2013 (delivered) Phase G | 2014 (delivered) Phase H | 2015 (delivered) Phase J | 2016 (in progress) Phase K | 2017 (planned) Phase L | 2018 (future) Phase M |
|---|---|---|---|---|---|---|
| ATLAS | 649 | 792/350 | 875/350 | 1200 | 1375 | 1500 |
| CMS | 649 | 792 | 875 | 1200 | 1375 | 1500 |
| LHCb | 2 | 216 | 550 | 670 | 750 | 1000 |
| **TOTAL (TB)** | **1300** | **1800** | **2300** | **3070** | **3500** | **4000** |

Table 2: Effective Grid Storage resources in Terabytes, as delivered or pledged to be available on April 1st of the year by the Swiss Tier-2 at CSCS as our national contribution (CH_Tier2) to WLCG [4].

**SUMMARY of Resources**

| Resource | 2013 (delivered) Phase G | 2014 (delivered) Phase H | 2015 (delivered) Phase J | 2016 (in progress) Phase K | 2017 (planned) Phase L | 2018 (future) Phase M |
|---|---|---|---|---|---|---|
| CPU (kHS06) | 23 | 26 | 35 | 49 | 62 | 72 |
| Effective disk (TB) | 1300 | 1800 | 2300 | 3070 | 3500 | 4000 |

Table 3: Summary of the resources as delivered or pledged for April 1 of the year by our Swiss Tier-2 at CSCS to the international LHC community WLCG.

phase K:= completed;
met pledges 1.4.2016

**Phase L: =**
Install in 2016; to meet pledges in 1.4.2017.

Note: surpassed our expectations on CPU due to better deal, and faster HW (Haswell CPU, Xeon E5-2680 v3 2.5GHz; vs IvyBridge v2).

**Planning + Pledges:**
To be discussed at the next CHIPP computing board on 1.9.

# Swiss CHIPP LHC Computing Pledges

CHIPP Tier-2 pledged vs delivered resources on 8.2016 ( source REBUS)

**Installed Capacities**

Year: 2016 Month: 8

| Infrastructure | Site Name | Physical CPU | Logical CPU | HEPSPEC06 | Disk (GB) | Tape (GB) |
|---|---|---|---|---|---|---|
| EGI | CSCS-LCG2 | 304 | 6,208 | 69,513 | 3,680,812 | 0 |
| EGI | UNIBE-LHEP | 179 | 2,281 | 0 | 0 | 0 |
| Total | | 483 | 8,489 | 69,513 | 3,680,812 | 0 |

Showing 1 to 2 of 2 entries

Note:
We surpassed our expectations on CPU due to better deal (bulk order CSCS) and faster CPU-HW (Haswell CPU vs IvyBridge).

**Status 8/2016**

**Federation Pledges**

Year: 2016

| Pledge Type | ALICE | % of Req. | ATLAS | % of Req. | CMS | % of Req. | LHCb | % of Req. | SUM | % of Req. |
|---|---|---|---|---|---|---|---|---|---|---|
| CPU (HEP-SPEC06) | | | 24,500 | 4% | 14,500 | 2% | 10,000 | 12% | 49,000 | 4% |
| Disk (Tbytes) | | | 1,380 | 2% | 1,030 | 3% | 540 | 19% | 2,950 | 3% |

Showing 1 to 2 of 2 entries

https://wlcg-rebus.cern.ch/apps/topology/federation/259/

➔ Pledges are met; sources vary over time (due to replacement steps)

# Swiss Tier-3 resources (8/16)

**Swiss Tier-3 resources are undispensible tools**
and exist in quite different “flavours” for :

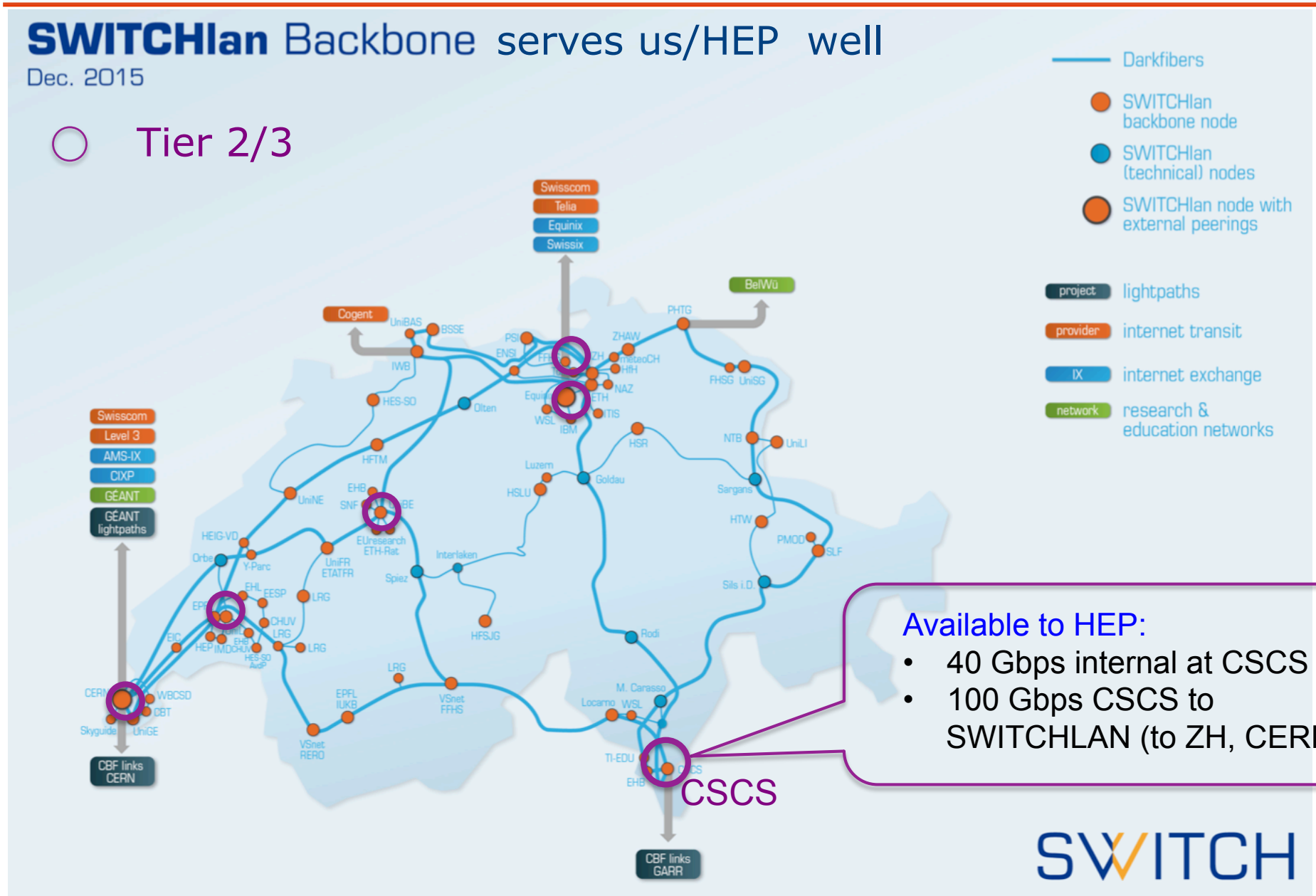
- ATLAS: each at UBern and at UGe
- CMS: common T3 for ETHZ, UZH, PSI at PSI
- LHCb: each at UZH and EPFL.

- Their capacity sum up to ~50% and 70% of CPU and storage of Tier-2 (at CSCS w/out AEC) .

**Updated details (8/16) of each Tier-3 in backup slides ....**

# Network in Switzerland

# Swiss National Network



# Efforts towards a Future Model of improved resource sharing

Continuously growing demand for computing resources requires rethinking of traditional HEP-only High Throughput Computing (HTC) clusters

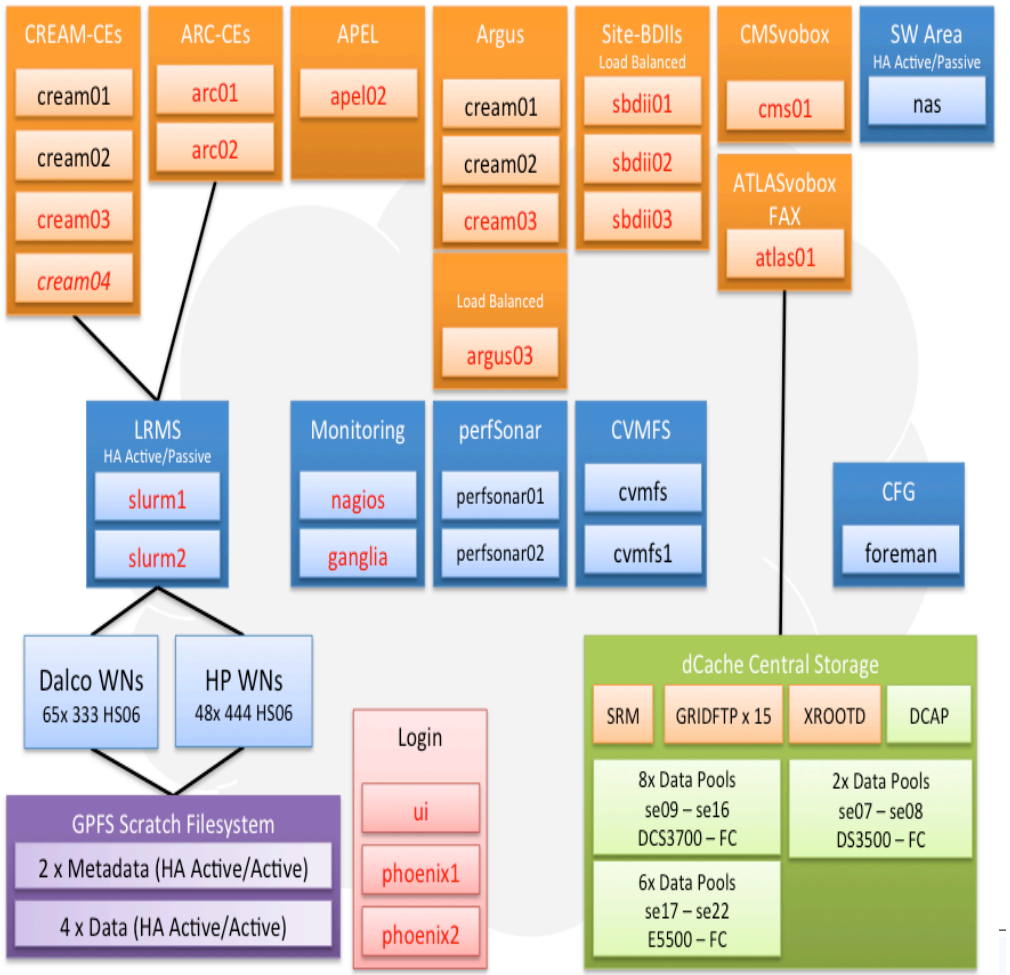
Approach: Make High Performance Computing (HPC) resources available to the HEP community.

**Swiss R&D project : "LHConCray at CSCS"**

# LHConCray at CSCS – present situation

- CHiPP currently operates its own dedicated Tier-2 hardware cluster at CSCS. This requires:
  - Maintain multiple middleware interfaces (compute, storage, info)
  - All tailored specifically for CHiPP
  - System/Interfaces at CSCS, VO representatives outside

Overall efficiency can potentially be much improved by sharing resources with many other communities

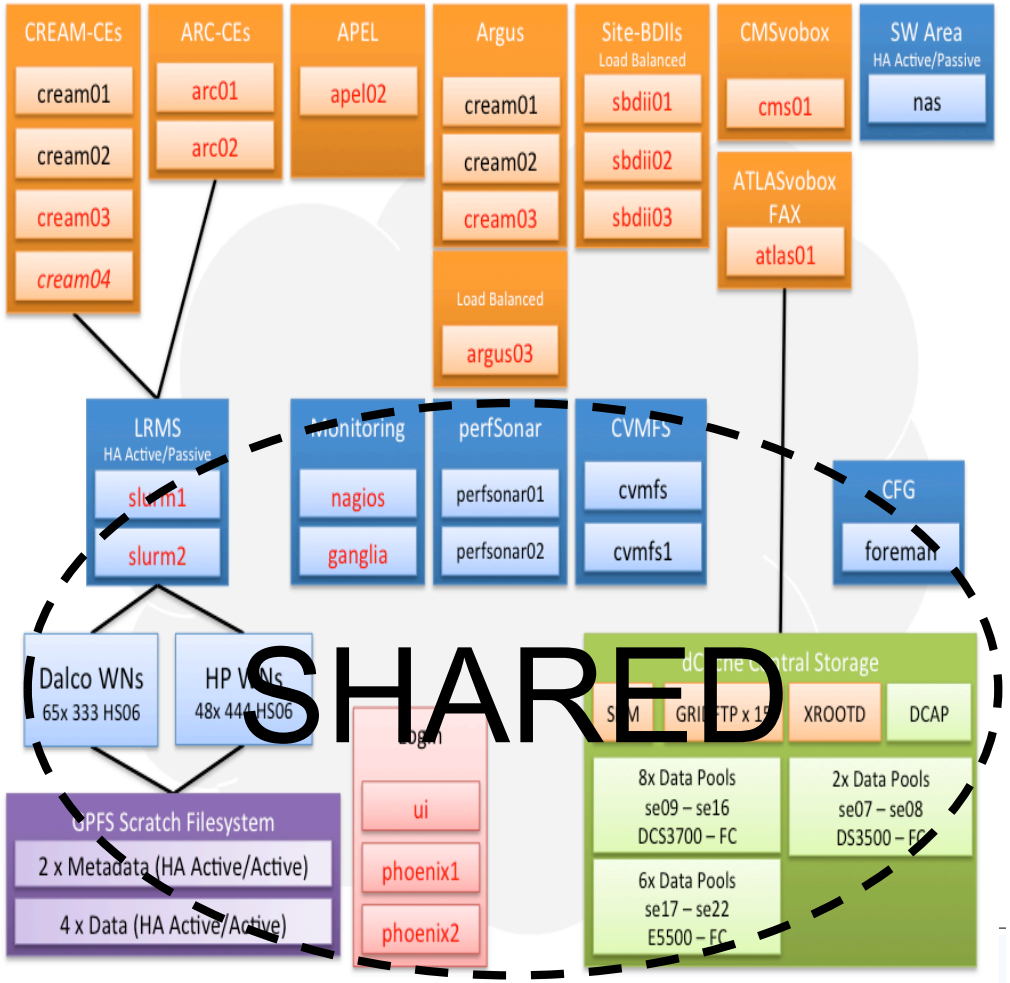


# LHConCray at CSCS – goal of R&D

- Goal is to share resources and efforts with other communities.
- In our specific case this would mean profit from the shared HPC Systems at CSCS (with >6500 nodes and >10 PB of storage) while keeping the interfaces to the Grid World WLCG

**Project requires:**

- Porting different workflows (VO Job factories and such) into the shared systems
- Render Grid Middleware CRAY-enabled
- **TEST, TEST, TEST ..**



# LHConCray at CSCS – status

- *Objective: Run LHC workflows in production on the CSCS Cray in shared mode*

- **First tests on prototype HW were successful (ATLAS MC production).**
- **Extensive tests** on actual CRAY systems in "real production environment" are now **in progress** for ATLAS, CMS and LHCb.
- Currently, however CRAY-HW at CSCS is being upgraded ...
- Need to do:
  - Demonstrate technical feasability of running for all 3 experiments
  - Testing, testing ... of ~1 month for stability
  - Determine operation efficiency, costs, study system behaviour...
  - Do cost comparisons of present vs future LHConCRAY, based on data.
  - Then decide on switching systems, stop, or have independent R&D.

**Profits**:

✓ Broaden availability of resources worldwide far beyond present technologies
✓ Leverage from economy of scales procuring and operating hardware
✓ Cooperate and involve other communities (HPC)...
✓ Other, similar efforts ongoing in ATLAS, CMS, (ALICE)

**Questions: Long-term strategy >2020? Switch completely ? Funding?**

# Other efforts: HPC and Clouds

- *From* Liz Sexton-Kennedy of USCMS (june 16):

We are not the only one investigating HPCs

U.S. CMS Operations Program

## NSF/DOE "Open" HPC sites

| Name | Institution | Architecture | Start Date |
|---|---|---|---|
| Mira * | ANL | 786k core IBM PowerPC | 2012 |
| Titan | Oak Ridge | 299k core AMD Opteron | 2012 |
| Stampede | TACC | 100k core Intel Sandy Bridge | 2013 |
| Comet | SDSC | 47k core Intel Haswell | 2015 |
| Edison * ** | NERSC | 133k core Intel Ivy Bridge | 2015 |
| Cori 1 * ** | NERSC | 52k core Intel Haswell | 2015 |
| Cori 2 | NERSC | 500k core Intel Knights Landing | 2016 |
| Theta | ANL | 150k core Intel Knights Landing | 2016 |
| Summit | Oak Ridge | ~3400 nodes IBM Power 9 | 2017 |
| Aurora | ANL | ~50k nodes Intel XEON Phi Gen 3 | 2018 |
| Stampede 2 | TACC | TBD ("twice the performance of Stampede") | TBA (announced June 2016) |

NSF DOE * = CMS has an allocation,
** = have run standard workflows

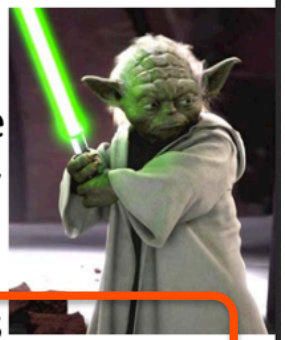
**CMS**

# Other efforts: HPC and Clouds

- *From* Torre Wenaus (BNL) of ATLAS (May 16):

## Yoda: Event Service on Supercomputers

- While PanDA was originally developed for the Grid, BigPanDA and ATLAS have extended it to operate also as an HPC internal system
  - Designed for efficient and flexible resource allocation and management of **MPI-based parallel workloads within HPC**
- **Yoda** - HPC-internal version of PanDA - leverages the experience acquired in massively scaled data Intensive processing for efficient utilization of a single massively scaled HPC machine



- The PanDA team is working with computing specialists at **NERSC**, **OLCF** and **ALCF** on implementing several approaches towards fine-grained, adaptive, flexible workflows to achieve the highest possible system utilizations
  - Both **backfill** and **scheduled** allocation modes

BROOKHAVEN NATIONAL LABORATORY

BERKELEY LAB

**ATLAS**

Others:
- RZG, Hydra

- ALCF =Argonne leadership computing facilty

- OLCF: Oak Ridge leadership ...

# CHIPP Computing Board

**Coordinates the tier-2 and tier-3 activities**

includes representatives of all institutions and experiments, CSCS, and tier-3 experts


T.Golling, Luis M.Ruiz (UNI Ge)

S.Haug, G.Sciacca (UNI Bern)


**C.Grab (ETHZ) chair CCB**

D.Feichtinger (PSI) vice-chair CCB

J.Pata (ETHZ), F.Martinelli (PSI)


R.Bernet (UNIZH)

A.Bay, M.Tobin (EPFL)


P.Fernandez, M.Gila, M.Ricciardi,
M. De Lorenzi (CSCS)

**Thank you …**