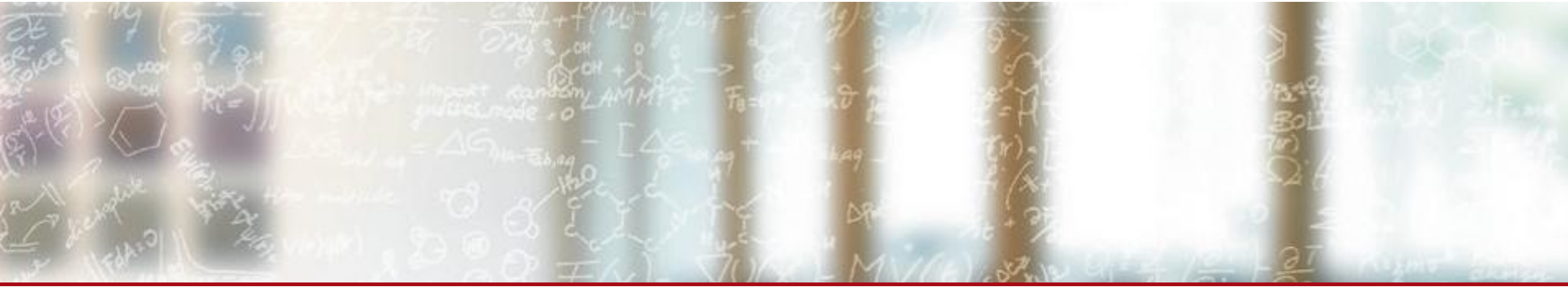
CSCS
Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre

ETH zürich



Swiss HPC Tier-2 @ CSCS

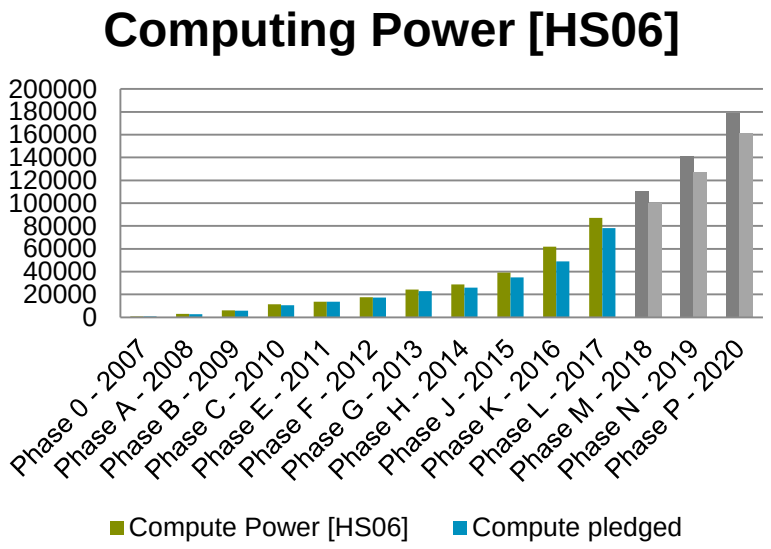
Pablo Fernandez, Miguel Gila, CSCS

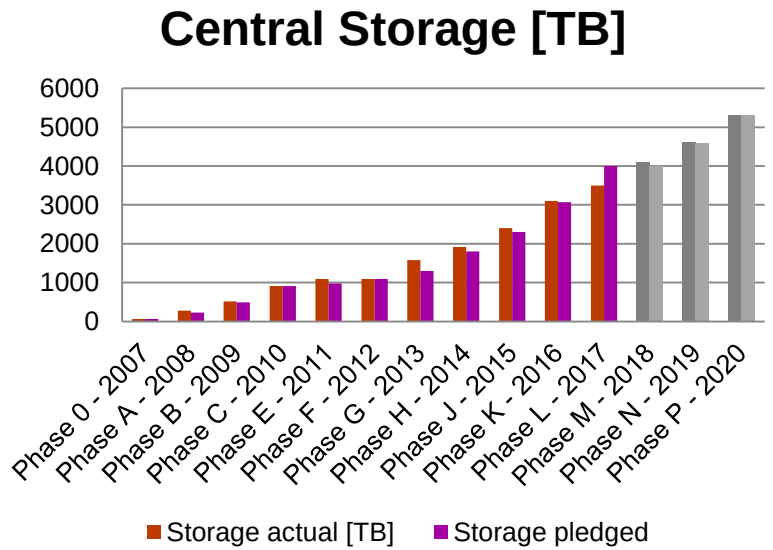
February 14th, 2018

Swiss Institute of Particle Physics

The Swiss Tier-2 @ CSCS

- CSCS provides ATLAS, CMS and LHCb resources since 2007 (~100 kHS06 & 4 PB)
  - https://wiki.chipp.ch/twiki/bin/view/LCGTier2/
- Funded by SNF (mostly) and ETH Zurich



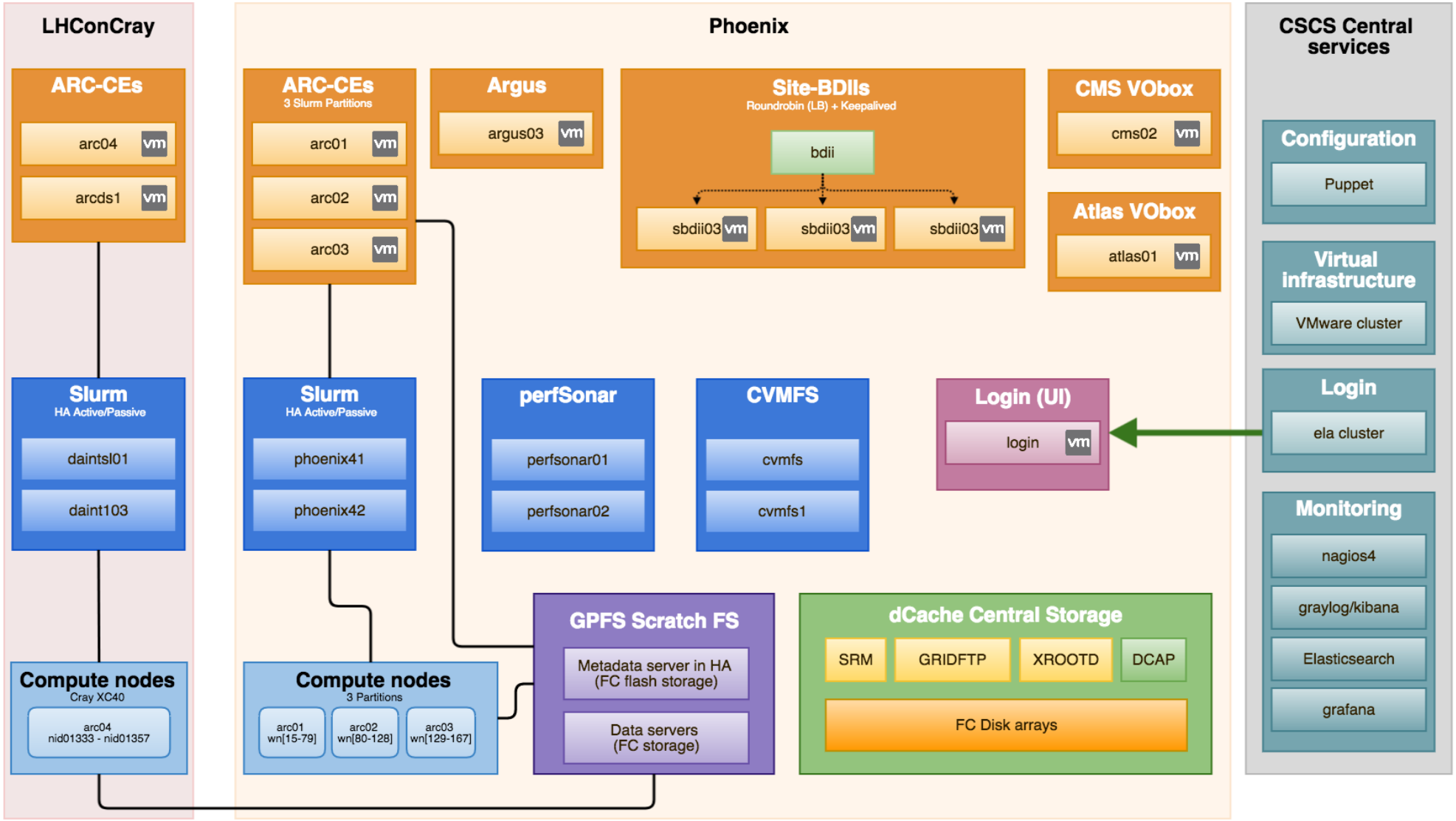


Services: logical view as of today

Middleware

Auxiliary services

Resources

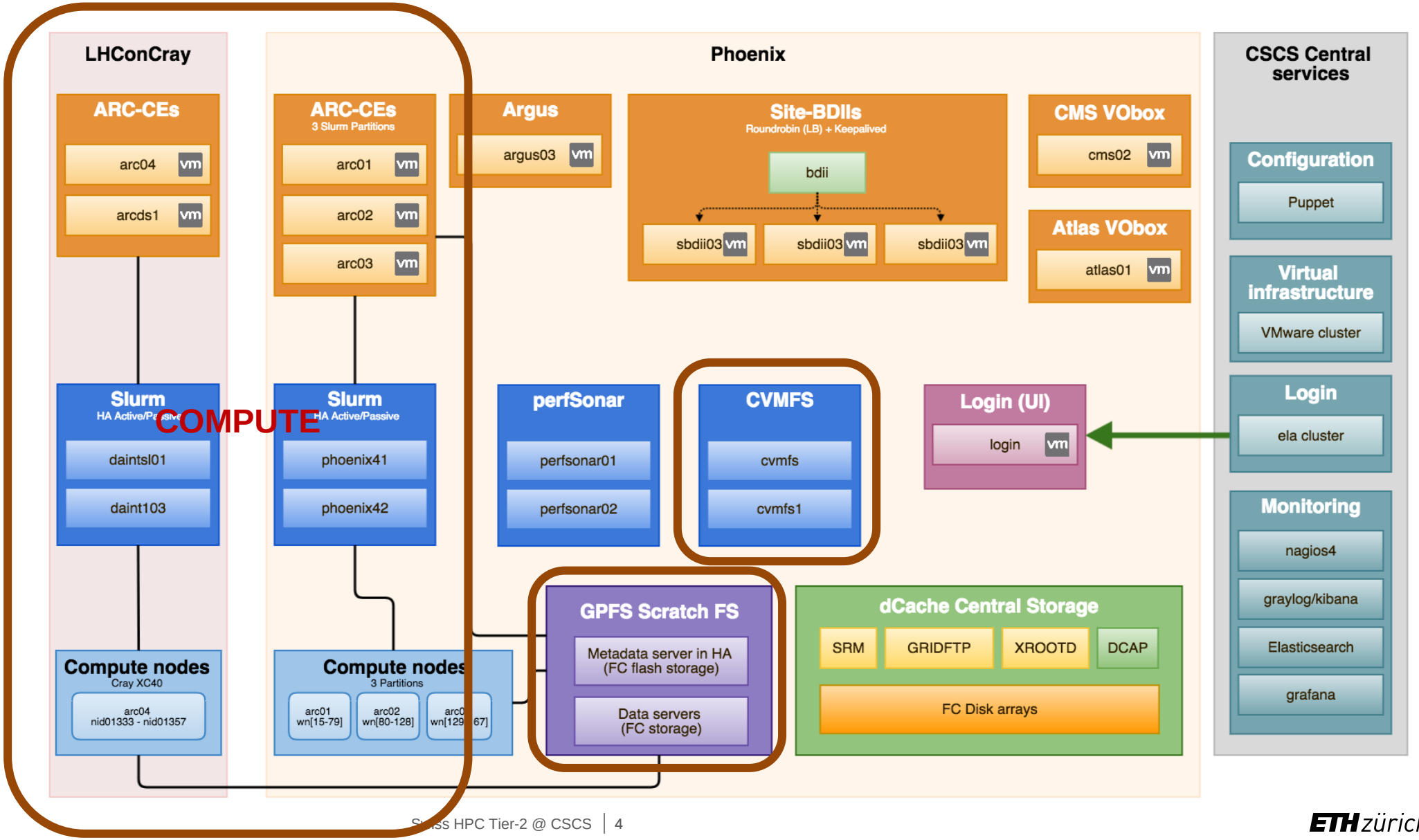


But let‘s just talk about the Computing side...

Middleware

Auxiliary services

Resources



Why are we moving computing towards HPC?

- Challenge for LHC computing for the HL-LHC
  - HPC is considered one of the main workforces for the future
  - Increased needs: 50x in 8-10 years
  - Technology improvement will only provide ~5x

- HPC could help solving the problem
  - CSCS is running a much bigger (~100 times) shared HPC system
  - Can grow/scale much better
  - Plenty of on-site expertise (HPC & Grid)
  - Benefit from economies of scales
  - Access to innovative technologies (accelerators, in-node flash drives, high-speed networks…)

… but the highest risk we see is to be left “aside” from the WLCG community

Piz Daint

Our flagship supercomputer!

Model	Cray XC40/XC50
>5000 XC50 Compute Nodes >1400 XC40 Compute Nodes	Intel® Xeon® @ 2.60GHz (12 cores, 64GB RAM) and NVIDIA® Tesla® P100 16GB 2x Intel® Xeon® @ 2.10GHz (2 x 18 HT cores, 64/128 GB RAM)
Interconnect Configuration	Aries and Dragonfly (~5 GB/s per CPU socket)
Scratch capacity/performance	/scratch/snx3000 6.2 PB / 112GB/s /scratch/snx1600 2.5 PB / 138GB/s /scratch/lhc 672TB / 10GB/s (shared with other system)
Peak performance	25.325 PF (XC50) / 1.731 PF (XC40)

Internet @ 100 Gbps

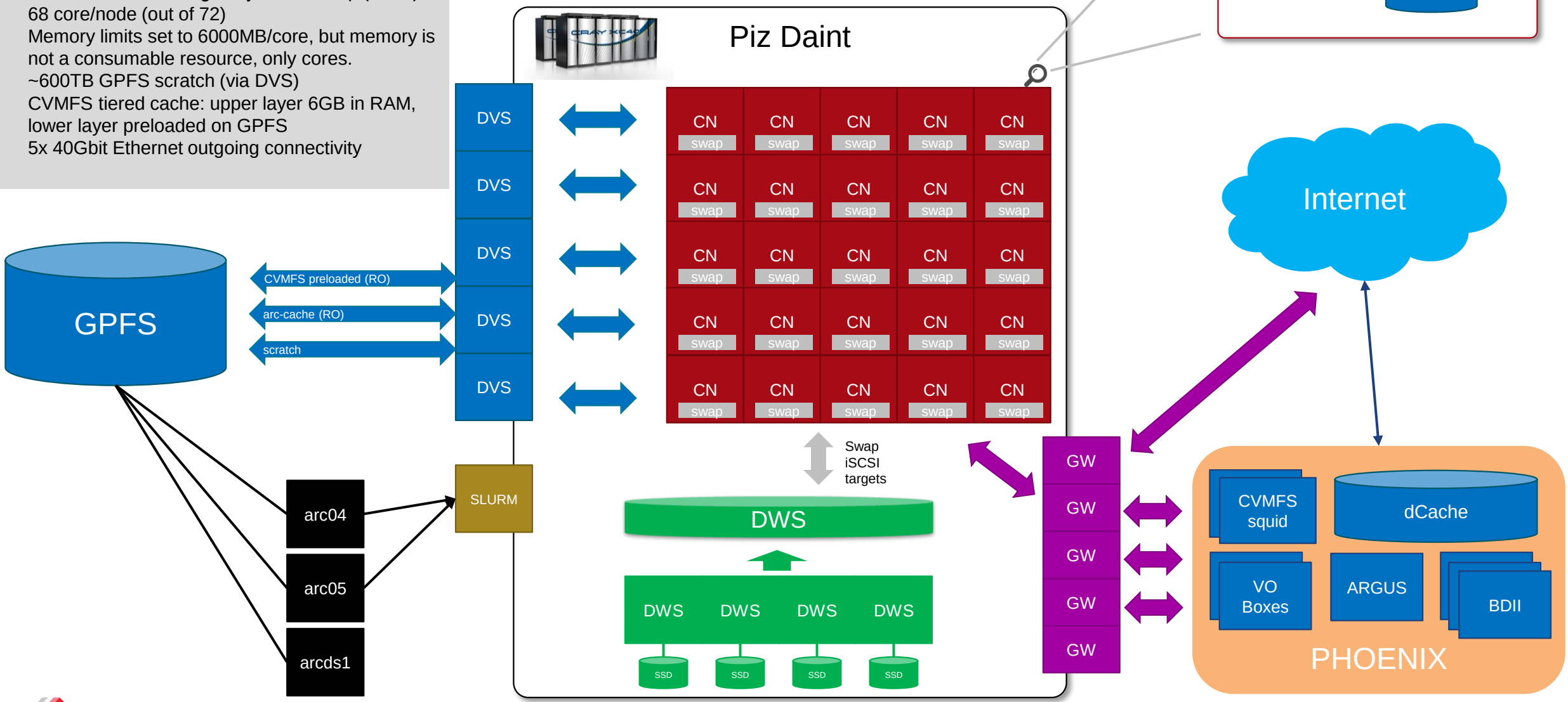
1 GFLOP ~= 1 HS06

Solution Highlights

- Uses standard WLCG middleware (fully compatible with VO factories)
  - Regular ARC CE in front of Piz Daint
- OS/Software on the compute nodes is **containerized**
- Most middleware components are shared with the previous cluster (Phoenix)
- Small/standard customizations are needed
  - Mainly on ARC CEs
- No local disks on the nodes

Overall LHConCRAY configuration

- 16 GB of SWAP using Cray's DataWarp (DWS)
- 68 core/node (out of 72)
- Memory limits set to 6000MB/core, but memory is not a consumable resource, only cores.
- ~600TB GPFS scratch (via DVS)
- CVMFS tiered cache: upper layer 6GB in RAM, lower layer preloaded on GPFS
- 5x 40Gbit Ethernet outgoing connectivity



Customizations on Piz Daint

- CVMFS is installed on LHC compute nodes. It uses a preloaded cache on disk + RAM cache

- Dedicated GPFS scratch filesystem for increased metadata performance
  - Uses the Cray's DVS nodes
  - Used for Tmpdir, Sessiondir, arc_cache and cvmf_preloadedcache

- SLURM dedicated partition
  - Cores are the only consumable resource
  - This permits jobs to use more than the base 2GB/core
  - Overall memory consumption is strangely balanced
  - Specific QoS and PriorityTier to make the backfill scheduler pass 'fast' over LHC jobs

- SWAP using DataWarp. Each node has 16GB of swap space to prevent OOMs

Customizations ARC-CE

- GLUE 1.2 publisher needed a bit of ‘tuning’ to publish the right values.
  - Glue2 was fine interestingly enough.

- SLURM batch generator modified to create batch jobs that
  - Have the proper flags (QoS, partition, nice, etc.)
  - Run the payload within a Shifter container
  - Needed a custom Slurm Client RPM compatible with Cray

- Configured using a data staging server for increased performance for ATLAS

All in all, not really different than a normal site

Limitations and risks

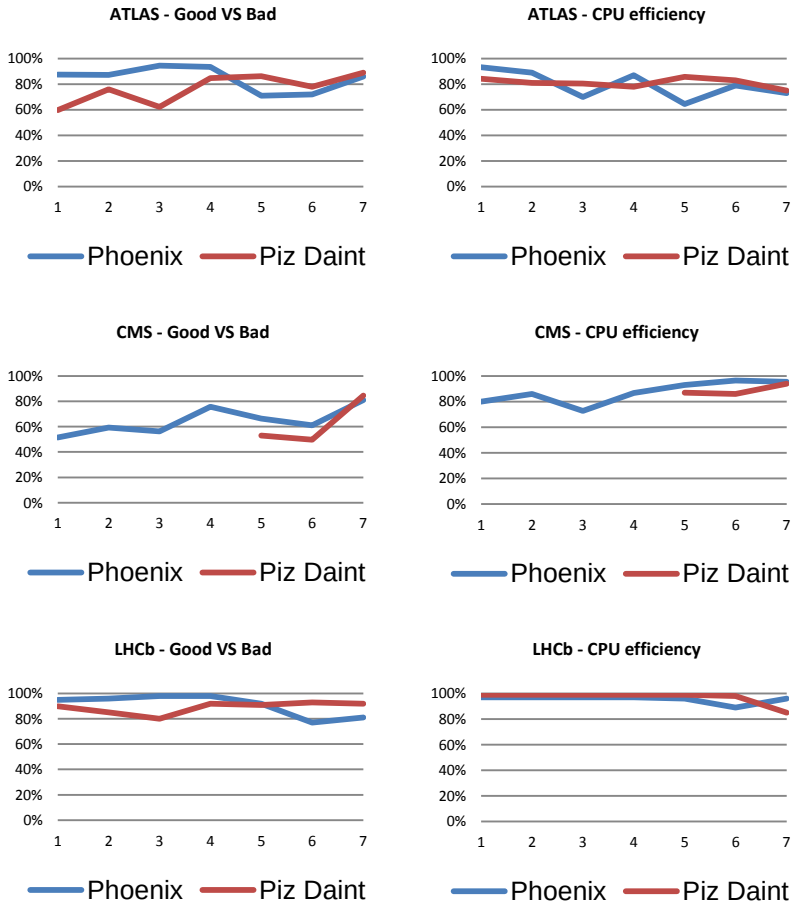
- Getting Fair-share right on a complex machine is… complex!
- GPFS is nice because it has multiple metadata servers... but WLCG jobs can *really* generate a lot of files.
- WLCG is now moving towards Singularity... But we run Shifter.
  - Not nice that CMS announces that, within 4 months, either you install Singularity or you go down (what happened with avoiding vendor-locks? Other solutions are possible)
- FUSE is ok for us, but some sites don't allow this to run on CNs; which is a limitation for instance on DOE machines
- 5-day jobs on an HPC system is a bit too long for what we are used to.

We need to be creative in finding workarounds

and need to work together with Sites and VOs

Performance comparison

- Both systems running since April'17 (Piz Daint resources are 4 times smaller)
- Compared all jobs (regular workloads) during 3 months, after finalizing last changes



Piz Daint	5	6	7	AVG
Good VS Bad %	74%	70%	88%	77%
CPU Efficiency %	89%	87%	85%	87%
Effective Efficiency	66%	61%	75%	**67%**

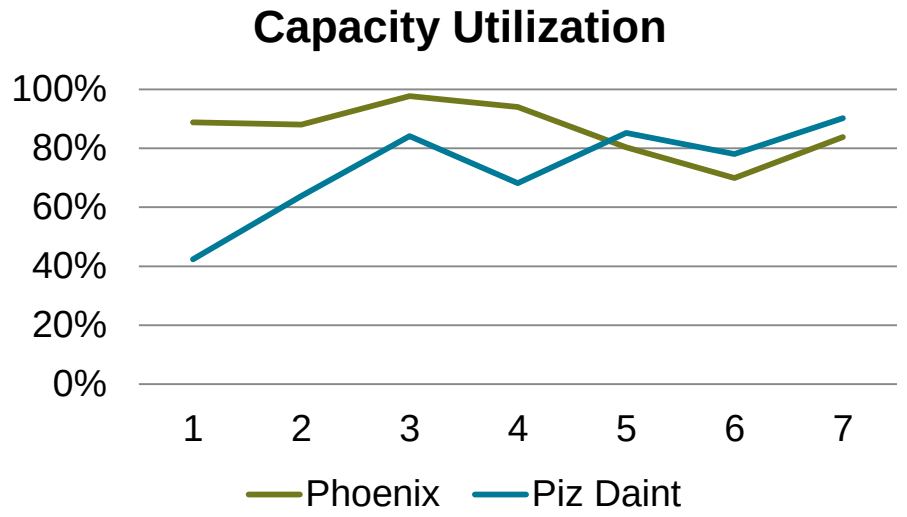
PHOENIX	5	6	7	AVG
Good VS Bad %	73%	69%	83%	75%
CPU Efficiency %	82%	88%	87%	86%
Effective Efficiency	60%	61%	72%	**65%**

N.B. looking only at used ratios (and not 40:40:20) the difference is of 9% in favor of Piz Daint

Economic comparison

The average capacity utilization during the last three Runs is of 78% on Phoenix and 84.5% on Piz Daint

- Not very significant because depends on many external factors



Cost comparison between proposed plan and going forward just with Phoenix:

- Average growth ratios:
  - Phoenix (compute +17%, storage +11%)
  - Piz Daint (compute +27%, storage +15%)
- Over the time this has an impact of:
  - 2018: +2.7% compute, 0% same storage
  - 2019: +14.4% compute, +7% storage
  - **2020: +26.6% compute, +10% storage**

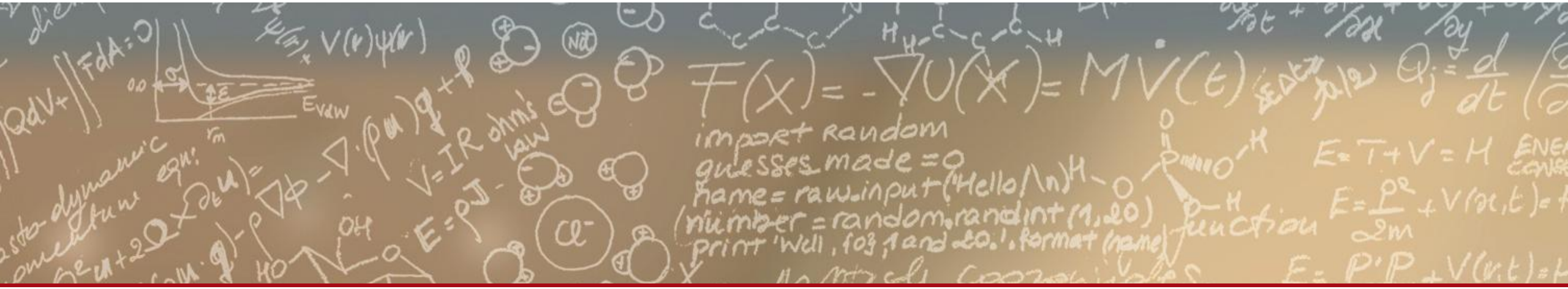
		2017	2018	2019	2020
SHARED	Compute Capacity	78’687	99’932	126’914	161’181
	Storage Capacity	3’500	4’015	4’617	5’310
PHOENIX	Compute Capacity	79’091	97’273	110’909	127’273
	Storage Capacity	3’500	4’000	4’300	4’800

Take home…

- HPC systems are very important to WLCG and can be used for more than just “scavenging” resources
- CSCS and the three VO Reps worked hard to make a standard Tier-2 work on top of a ‘classic’ HPC system
- Similar performance compared to a normal cluster, a better economy of scales, and opens the door to HPC technologies (e.g. GPUs…)
- Main risk is to drift away from the other sites. This is why we’re here!
  - E.g. we already use containers
  - E.g. we don’t have local disks on the nodes
  - We need the GDB to take into account that such sites exist

CSCS
Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre

ETH zürich



Thank you for your attention.

Follow-up questions welcome!

pablo.fernandez@cscs.ch (CSCS)
miguel.gila@cscs.ch (CSCS)
gianfranco.sciacca@lhep.unibe.ch (ATLAS)
grab@phys.ethz.ch (Chair Mgmt. board)
derek.feichtinger@psi.ch (CMS)
thomas.klijnsma@cern.ch (CMS)
bernet@physik.uzh.ch (LHCb)