- Motivation and challenges
- Architecture
- Production performance
- Assessment and outlook

WLCG Data Processing on Piz Daint at CSCS

SWISS NATIONAL SUPERCOMPUTING CENTRE

Gianfranco Sciacca

AEC - Laboratory for High Energy Physics, University of Bern, Switzerland

Nordugrid Conference 2018 - 4-7 June 2018, Garching bei München, Germany

UNIVERSITÄT BERN
AEC ALBERT EINSTEIN CENTER FOR FUNDAMENTAL PHYSICS

LABORATORIUM FÜR HOCHENERGIEPHYSIK LHEP UNIVERSITÄT BERN

MOTIVATION AND CHALLENGES

WLCG computing on HPC systems

- **The HENP computing community will face several challenges with respect to the computing requirements for the next decade and beyond**
- **For the LHC community, requirements for the High-Luminosity LHC runs (2025-2034) are expected to be a factor of ~50 higher than today**
- Available **budgets** are expected to be **flat at best**
- Novel computing models making a more dynamic use of heterogeneous resources (supercomputers and clouds) need to be evaluated in order to address such challenges
- HPC machines are increasingly powerful, could play a crucial role in delivering **more computing for the same price**
- **One of the biggest challenges is the transparent integration with the complex experiment data processing frameworks**

MOTIVATION AND CHALLENGES

WLCG computing on HPC systems

- **HPC is awesome**
  - **Piz Daint Cray XC50 / XC40 @ CSCS**

    - ‘cpu/gpu hybrid’ (5320 nodes) and ‘multicore’ (1431 nodes), 361,760 cores

    - NVIDIA Tesla P100, Xeon E5-2690v3 2.6 GHz, 521 TB of RAM, 25k Tflops

    - Cray Aries high-speed “dragonfly” topology interconnect

    - Lustre Sonnexion 3000, 6.2 PB, Sonnexion 1600, 2.5 PB

    - GPFS, 700 TB + 90 TB SSD transparent cache

    - DVS, Burst Buffer, Data Warp (POSIX filesystem on demand on SSDs)

MOTIVATION AND CHALLENGES

WLCG computing on HPC systems

- **HPC is Awkward**
  - **No local disk**
    - Breaks a lot of standard Linux workflows
  - **Minimal OS**
    - Designed to accelerate parallel software
    - Many "expected" Linux tools are absent
    - Runs SUSE, and doesn't upgrade often
  - **Limited RAM**
    - Most of the CPU only nodes with 1 GB / core, some with 2 GB / core
    - No swap
  - **Network connectivity not guaranteed**
    - Must be negotiated
    - Needs gateways
    - Interfacing external services is not straightforward (e.g. mount a directory)

MOTIVATION AND CHALLENGES

WLCG computing on HPC systems

- **The Swiss HEP computing community and CSCS have started working on the HPC integration with the LHC experiment Tier-2 facilities in 2014**

  - **ATLAS Geant 4 simulation**
    - Ran for 6 months on a Cray XK7
    - Integrated by means of a modified ARC CE, submitting remotely to CSCS

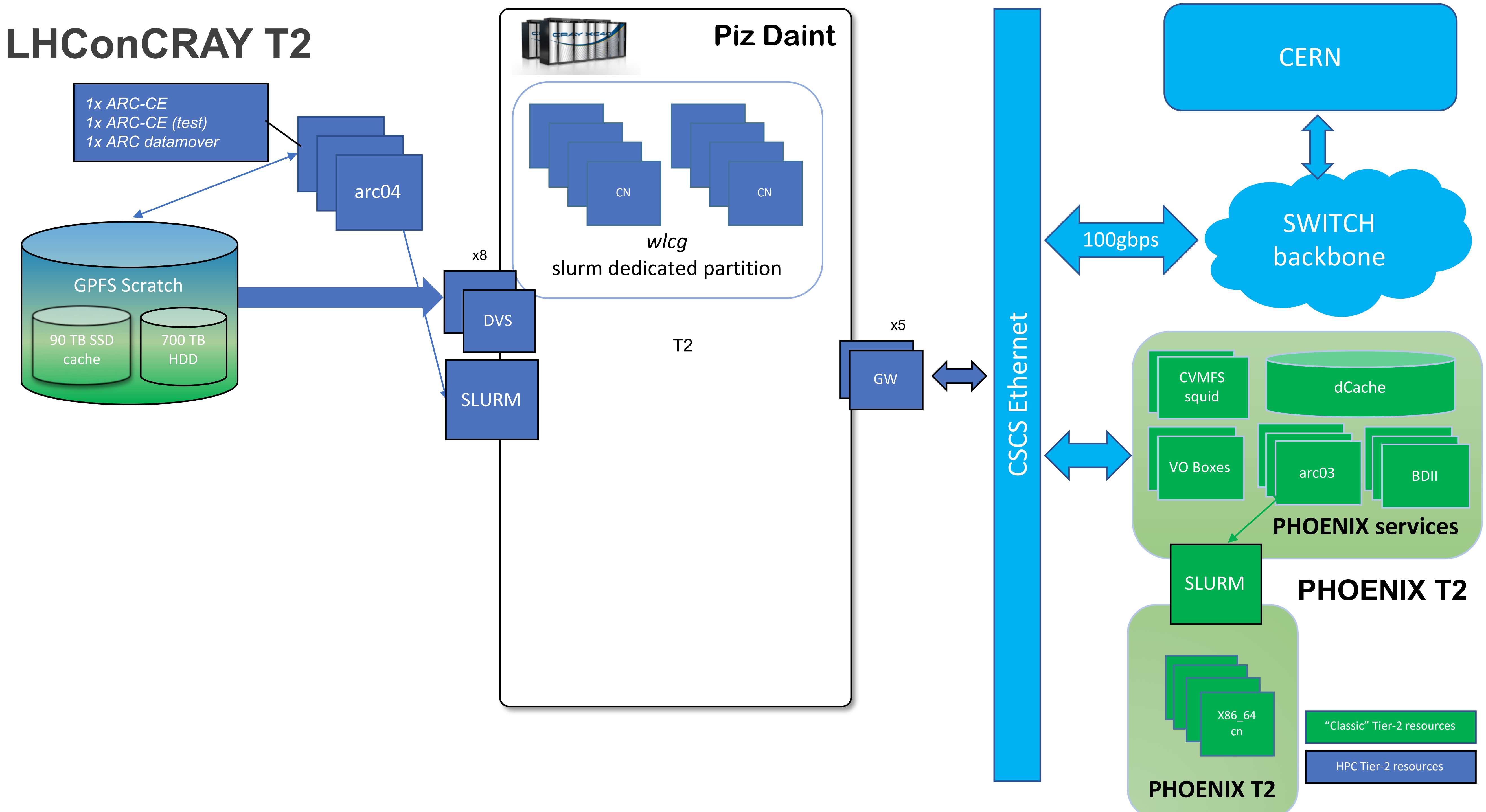
  - **LHConCray project (ATLAS, CMS, LHCb)**
    - Ran for about 2 years in 2016-17
    - Aimed at integrating Piz Daint with the LHC experiment frameworks
    - Targeted all experiment workflows (including user analysis)
    - Went in production with 1.6k cores in 2017

  - **WLCG Tier-2 facilities migrated to Piz Daint**
    - Decision taken at the end of 2017
    - ~4k cores by April 2018, >10k by April 2019

Powered by ARC

SHARED ARCHITECTURE

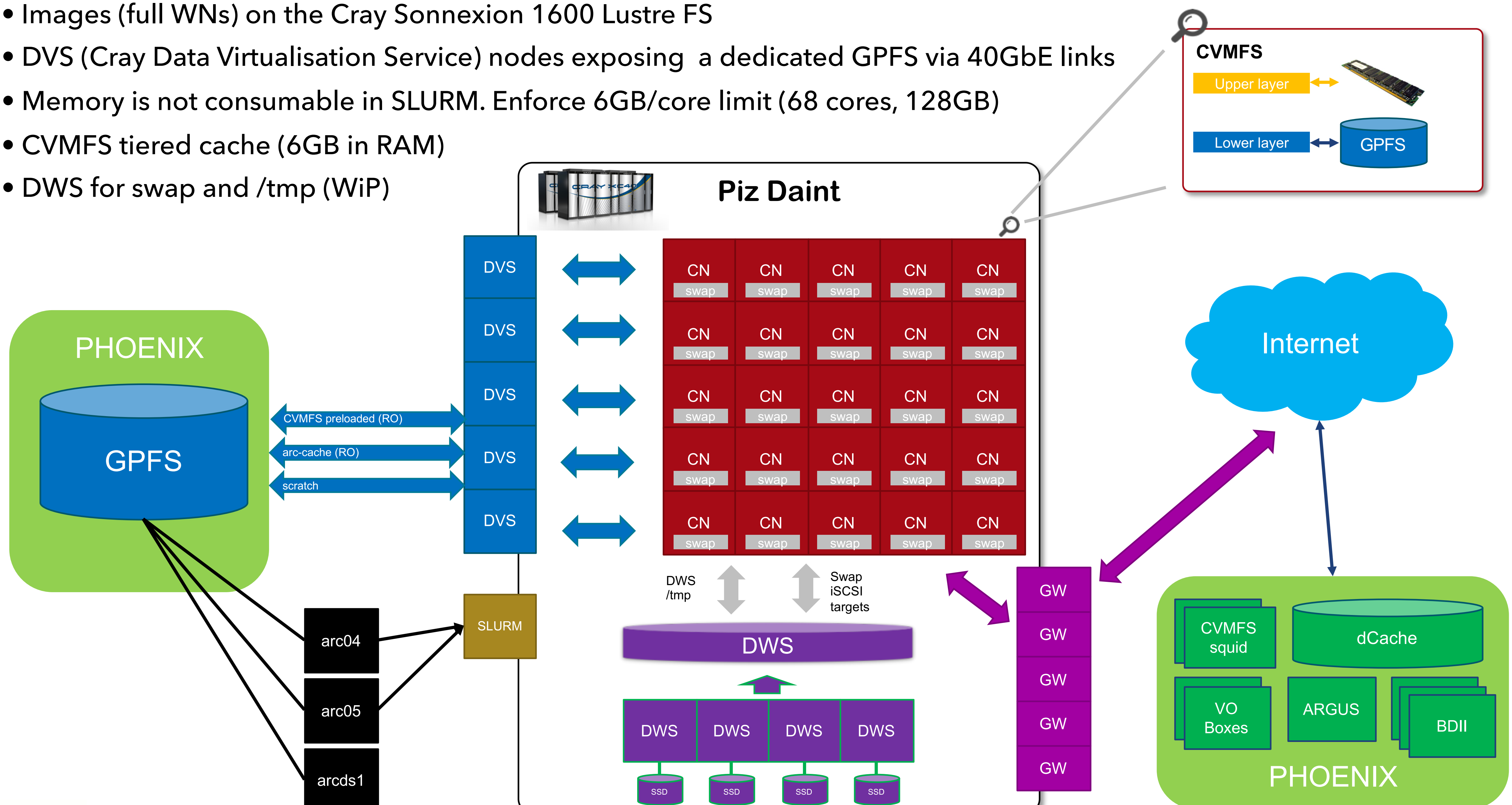
Classic x86_64 and HPC systems



SHARED ARCHITECTURE

HPC system highlights

- Jobs run in Docker containers using Shifter
- Images (full WNs) on the Cray Sonnexion 1600 Lustre FS
- DVS (Cray Data Virtualisation Service) nodes exposing a dedicated GPFS via 40GbE links
- Memory is not consumable in SLURM. Enforce 6GB/core limit (68 cores, 128GB)
- CVMFS tiered cache (6GB in RAM)
- DWS for swap and /tmp (WiP)



LHConCRAY - Acceptance Tests 2017 – Configuration evolution 1

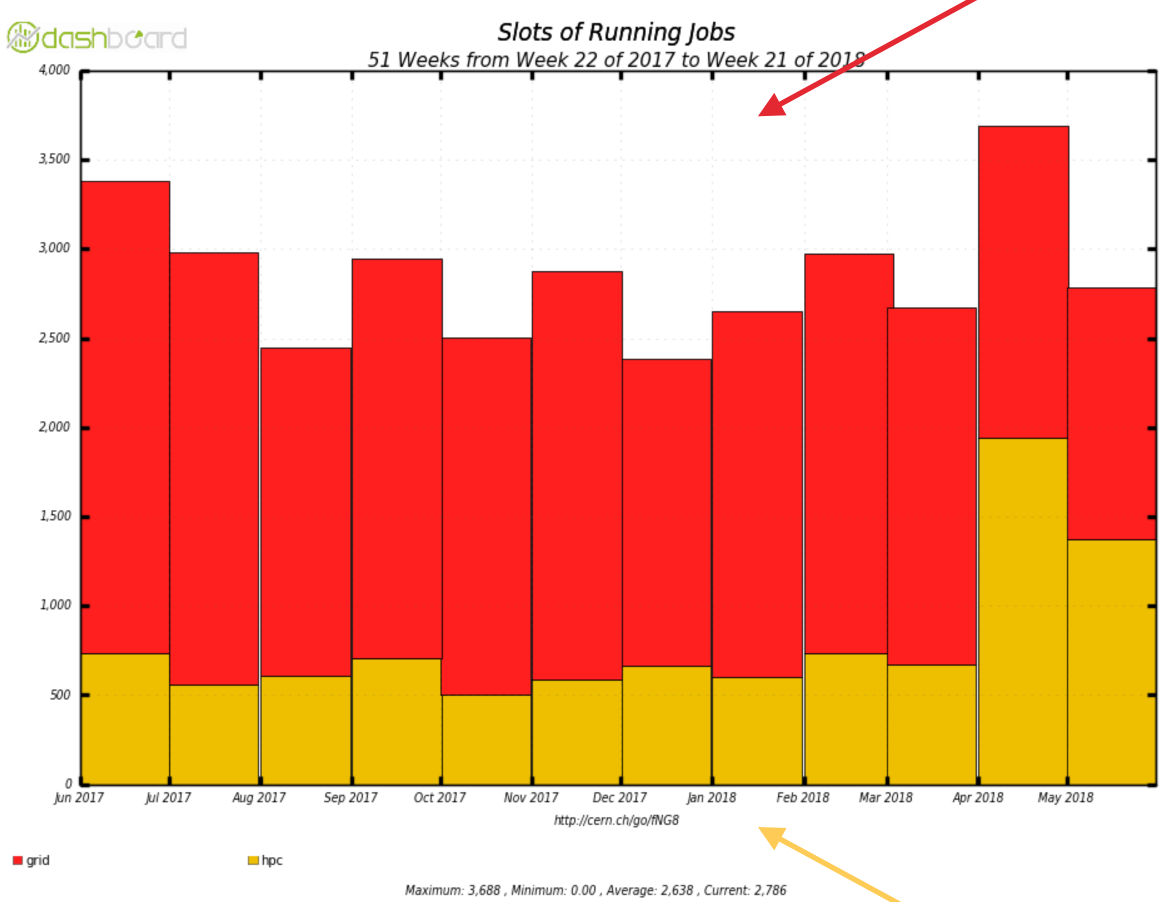
PRODUCTION PERFORMANCE

WLCG computing on HPC systems

Used slots - 1 year

Processing share - 1 year

x86_64



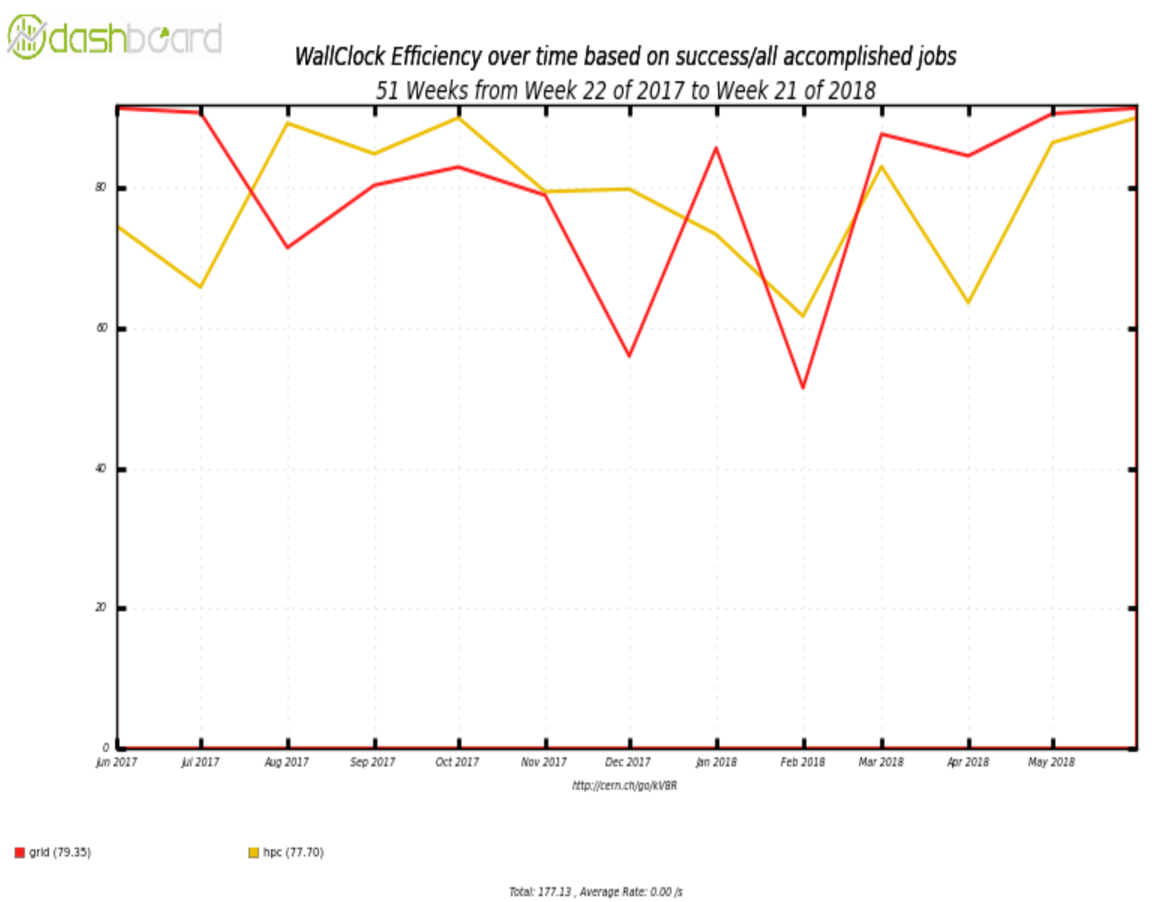
HPC

ATLAS only (~40% of the total)

PRODUCTION PERFORMANCE

WLCG computing on HPC systems

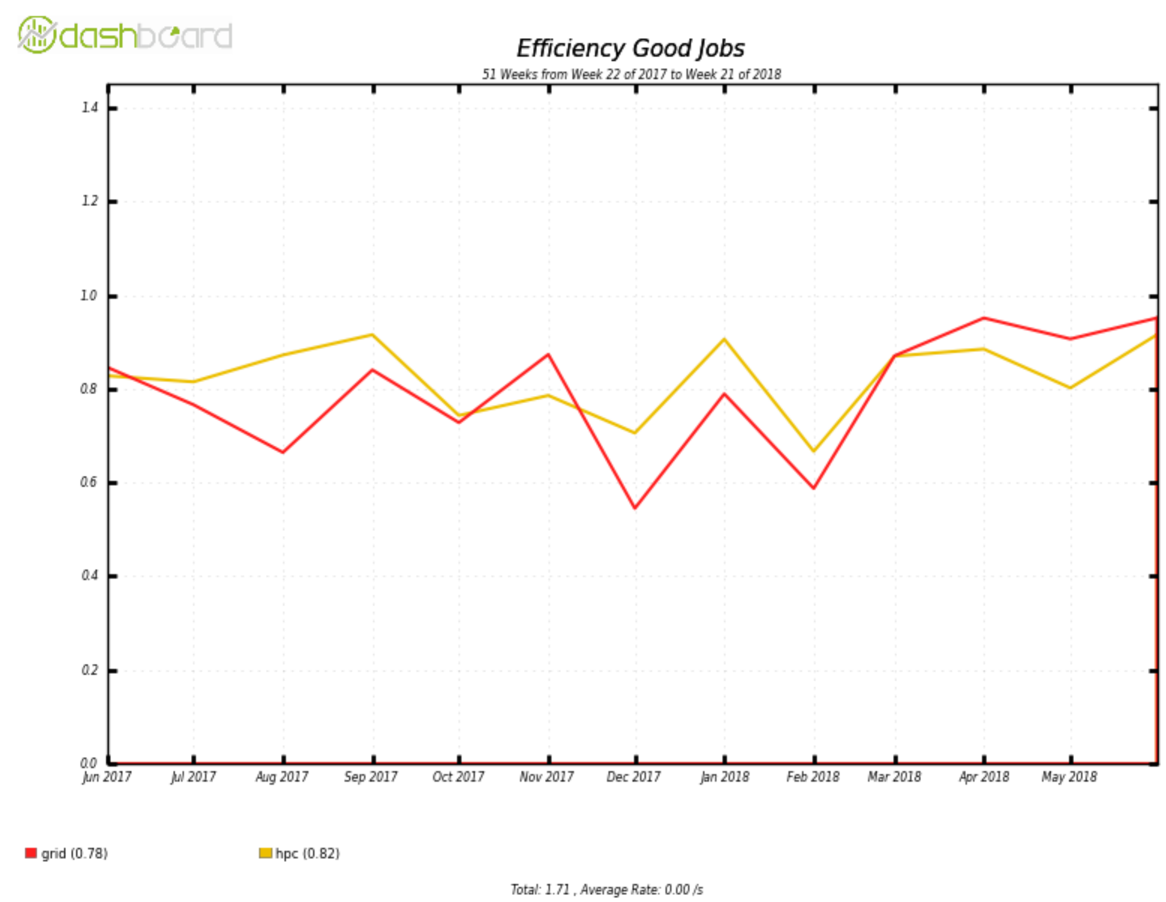
WC efficiency (success/all)



x86_64: 79%

HPC: 78%

CPU/WC efficiency (good jobs)



x86_64: 78%

HPC: 82%

ATLAS only (~40% of the total)

UNIVERSITÄT BERN
AEC
ALBERT EINSTEIN CENTER FOR FUNDAMENTAL PHYSICS

LABORATORIUM FÜR HOCHENERGIEPHYSIK
LHEP
UNIVERSITÄT BERN

SHORT TERM ASSESSMENT

WLCG computing on HPC systems

- **Cost of resources marginally lower**
- **Comparable performance**
- **High integration costs and ongoing challenges**
- **Where is the big deal then?**

SHORT TERM ASSESSMENT

WLCG computing on HPC systems

- **Cost of resources marginally lower**
- **Comparable performance**
- **High integration costs and ongoing challenges**
- **Where is the big deal then?**

- **Short term answer:**

  - There is NO big deal under such conditions and business model

  - No opportunistic usage means we get what we pay for (not elastic either)

  - It draws lots of attention, but the implementation effort and the operational pressure are considerably higher

- **HPC is sexy**

  - Could be made sexier

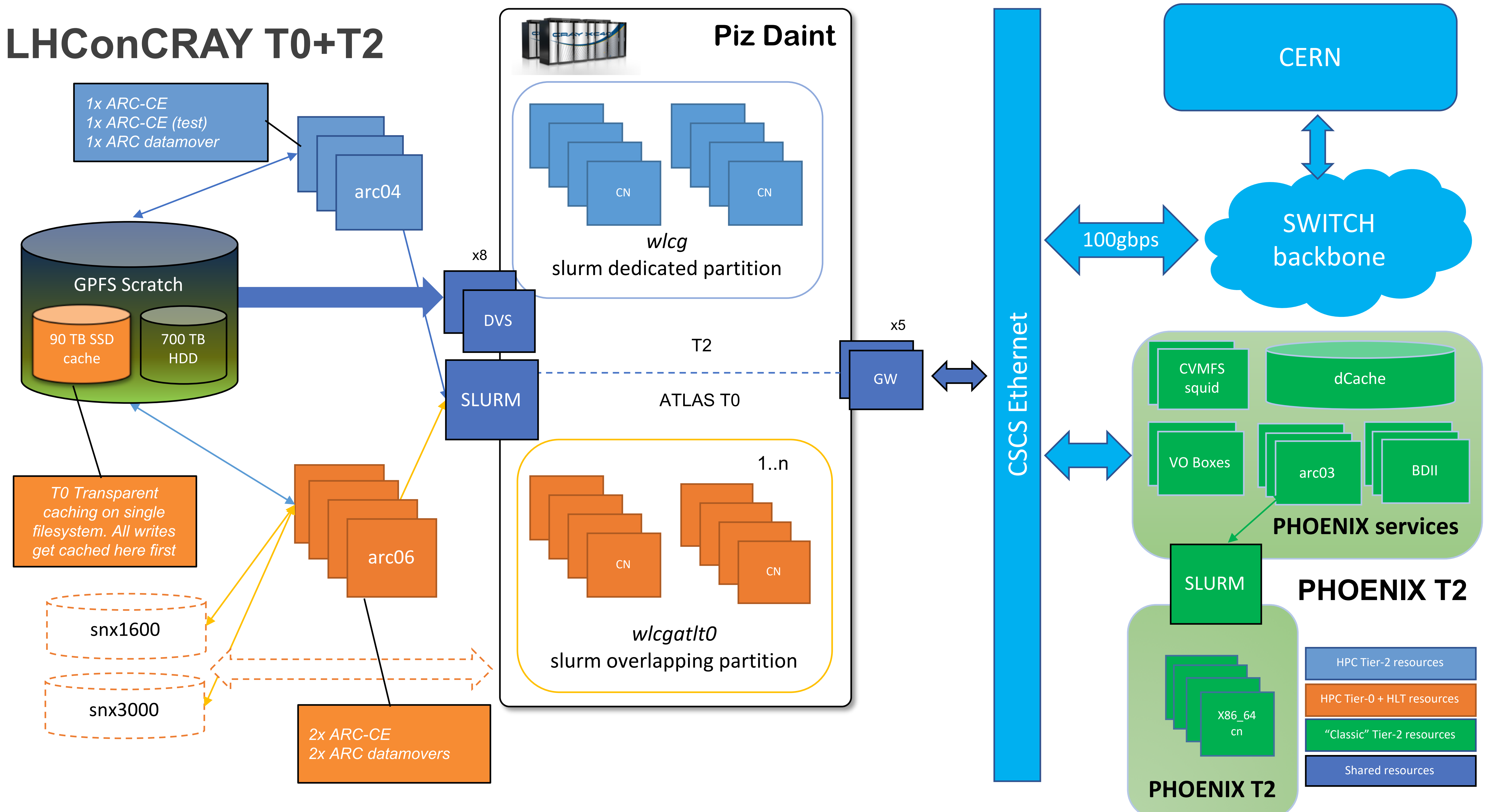
MID AND LONG TERM ASSESSMENT

WLCG computing on HPC systems

- **The architecture solutions may pave the way to future computing models**
  - E.G. shifter, CVMFS solutions
  - Potential in DWS
  - …
- **And the other way around**
  - Community feedback to the next generation (exa-scale?) machine
- **The provisioning model is more flexible**
  - Still largely left to the good-will of the resource provider
  - Some freed up effort can be re-directed toward expert experiment support
- **The integration effort could pay back in the long term**
  - Opens doors to "*beyond pledge* activities"
  - E.G. elastic provisioning of resources for specific activities
  - One example: run Tier-0 processing tasks (T0 spill-over or extension)
  - Several r&d projects are being proposed by CERN & the experiments
  - Novel architectures, data lakes, code optimisation, porting to GPUs, etc.
  - Aiming at developing novel computing models

T0 / T2 SHARED ARCHITECTURE

From T2 shared architecture to T2 / T0 shared



TIER-0 RESOURCE COMMISSIONING

Commissioning status

- **Dealing with grinding teeth problems**
  - Shortage of node memory for the type of workload
  - High and sustained I/O pressure
  - I/O patterns not usual for HPC environments
  - Exposing bugs in DVS, DWS Cray technologies
  - …

- **The Tier-2 HPC performance is good**
  - Comparable to the one of the "classic" PHOENIX cluster
  - The mix of workloads smooths out the edge requirements of a single one
  - Both in terms of memory and I/O

CONCLUSIONS

WLCG computing on HPC systems

- **LHC future computing needs should benefit from HPCs**
- **Integration with the experiment data processing frameworks is not trivial**
- **Efforts ongoing between the Swiss HEP community and CSCS since a few years, successfully ingrated Piz Daint with the WLCG frameworks**
- **CERN and CSCS have mutual interest in developing new models**
- **R&D projects have been launched**
  – Novel architectures and software optimisation
- **Next generation (exa-scale) machines should accommodate the requirements of the HEP community applications**

Thank you for your attention!