Production pilot of Swiss ATLAS federated storage with DPM and ARC caches

Gianfranco Sciacca

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

HSF WLCG Virtual Workshop - 20 November 2020

SWISS WLCG RESOURCES

Geneva DPNC/University
Tier-3 / ATLAS

Institutional HCP
under commissioning

ARC CE push mode

BeeGFS cluster FS
IB interconnect

DPM pools
0.4PB RAID6

Bern LHEP
Tier-2 / ATLAS-others

Commodity cluster
up to ~6k cores

ARC CEs push mode

Lustre cluster FS
IB interconnect

DPM head/pools
1.7PB RAID6 (1PB ATLAS)

Bern University
Tier-2 / ATLAS

Institutional HPC
up to 2k cores

ARC CE push mode

GPFS cluster FS
IB interconnect

CSCS
Tier-2 / ATLAS-CMS-LHCb

Large scale HPC
15k cores (6k ATLAS)

ARC CEs push mode

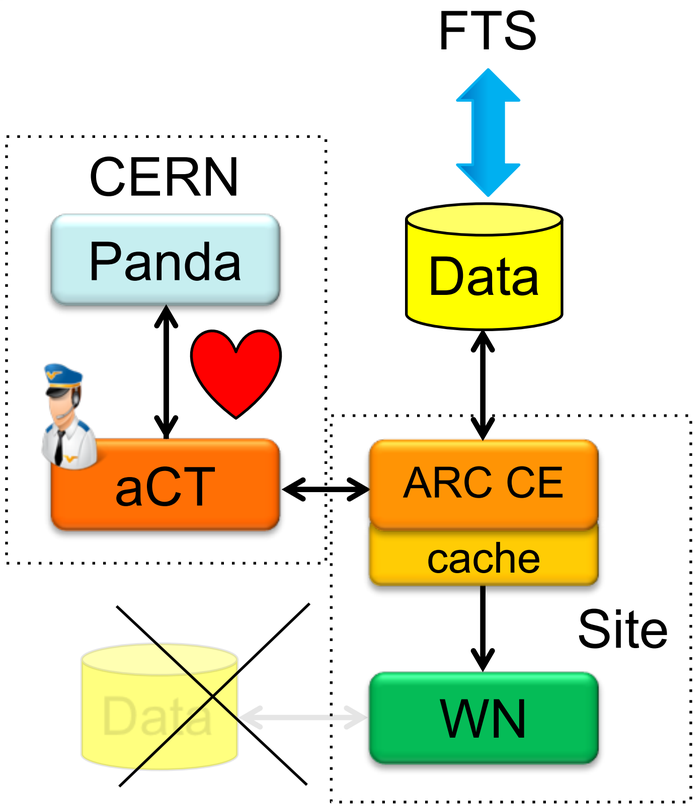
GPFS cluster FS
IB/FC interconnect

dCache head/pools
5.3PB GPFS (2.1PB ATLAS)

CONTENT DELIVERY AND CACHING

NORDUGRID
Grid Solution for Wide Area Computing and Data Handling

asyncronous data staging

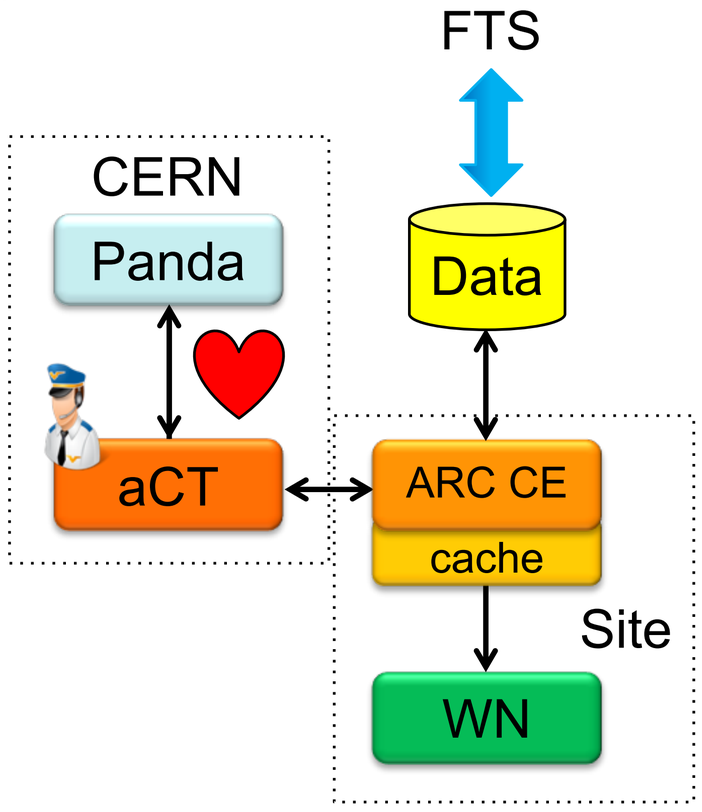


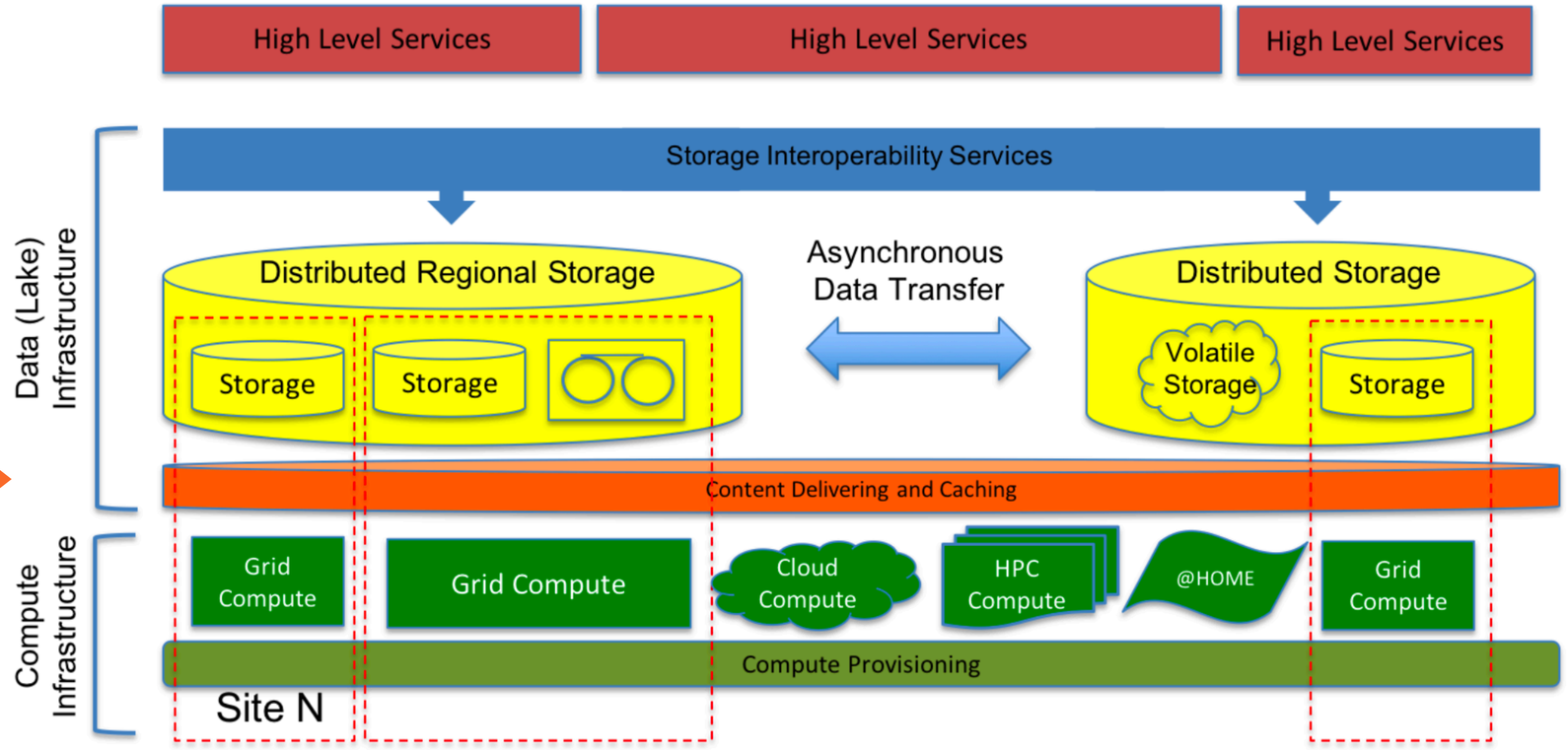
Decouples data delivery from job execution layer

CONTENT DELIVERY AND CACHING

NORDUGRID
Grid Solution for Wide Area Computing and Data Handling

asyncronous data staging





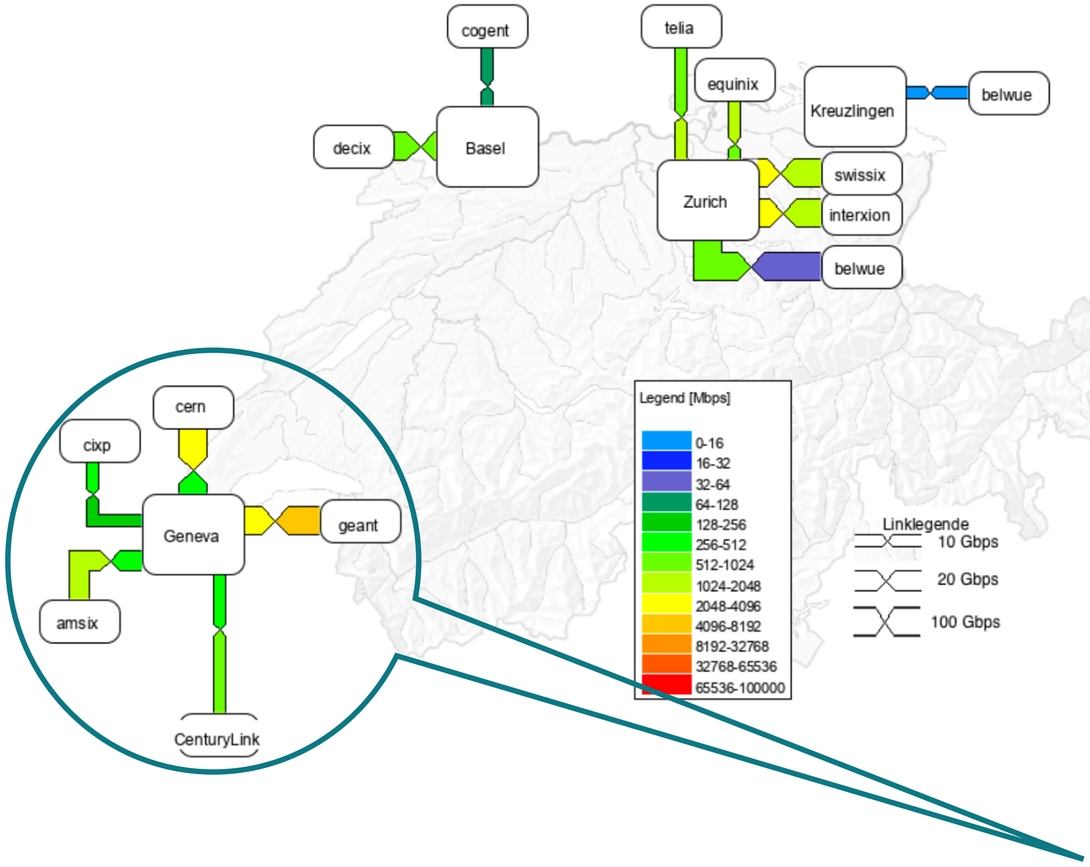
S. Campana

Decouples data delivery from job execution layer

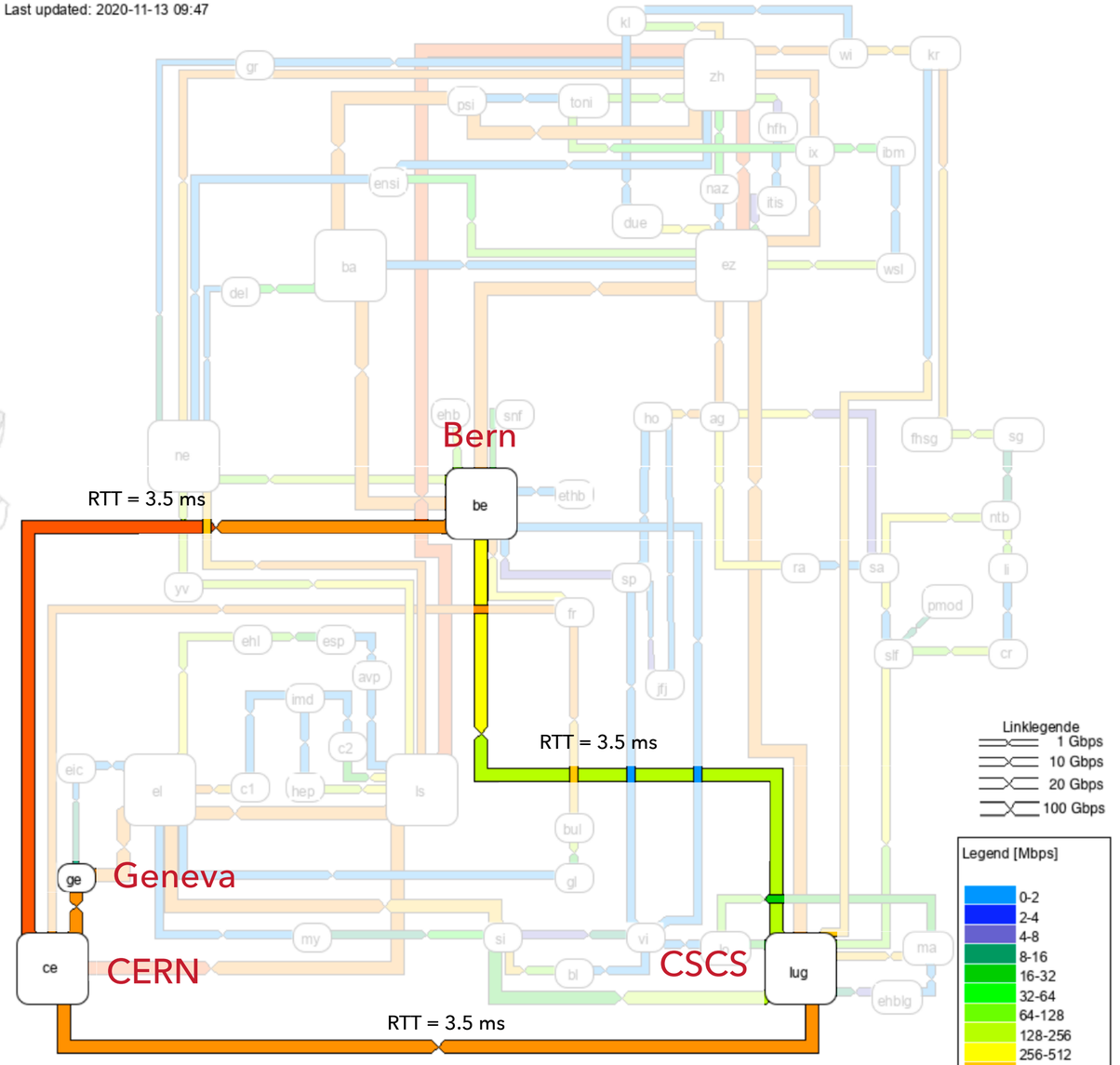
SWISS NETWORK INFRASTRUCTURE

SWITCH

Last updated: 2020-11-15 21:47



Last updated: 2020-11-13 09:47



SWISS NETWORK INFRASTRUCTURE

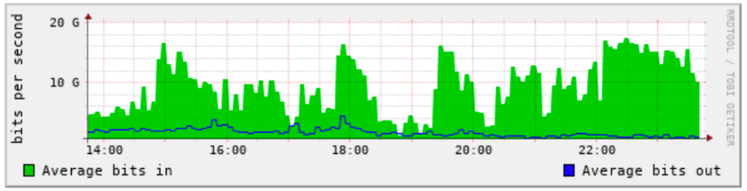
Graphs for be-ce **BERN-CERN**

Summary

be-ce: be3-hundredgige0_1_0_1
Values at last update:

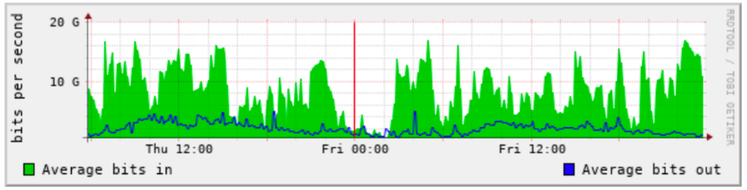
Average bits in: 10.19 Gbits/sec Average bits out: 905.06 Mbits/sec

Hourly graph



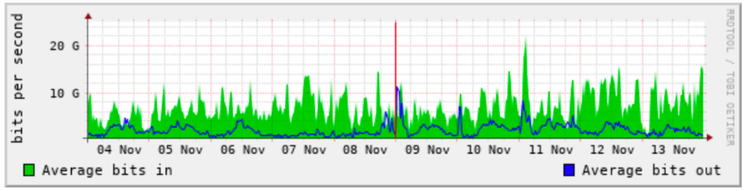
←20 G

Daily graph



←20 G

Weekly graph



←20 G

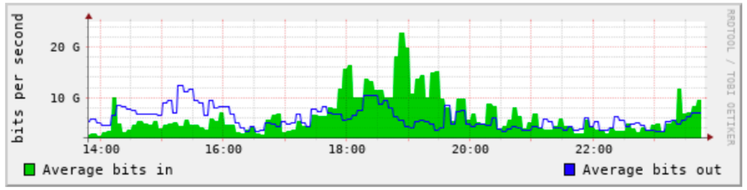
Graphs for ce-lug **CSCS-CERN**

Summary

ce-lug: ce4-hundredgige0_0_0_9
Values at last update:

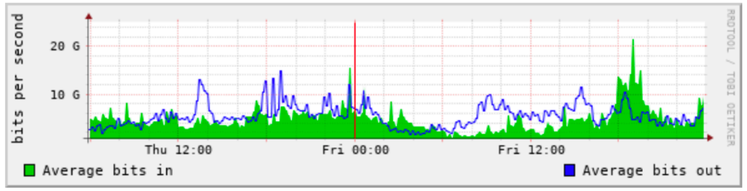
Average bits in: 9.52 Gbits/sec Average bits out: 7.01 Gbits/sec

Hourly graph



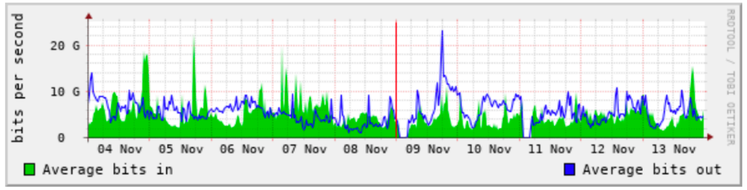
←20 G

Daily graph



←20 G

Weekly graph



←20 G

MOTIVATIONS

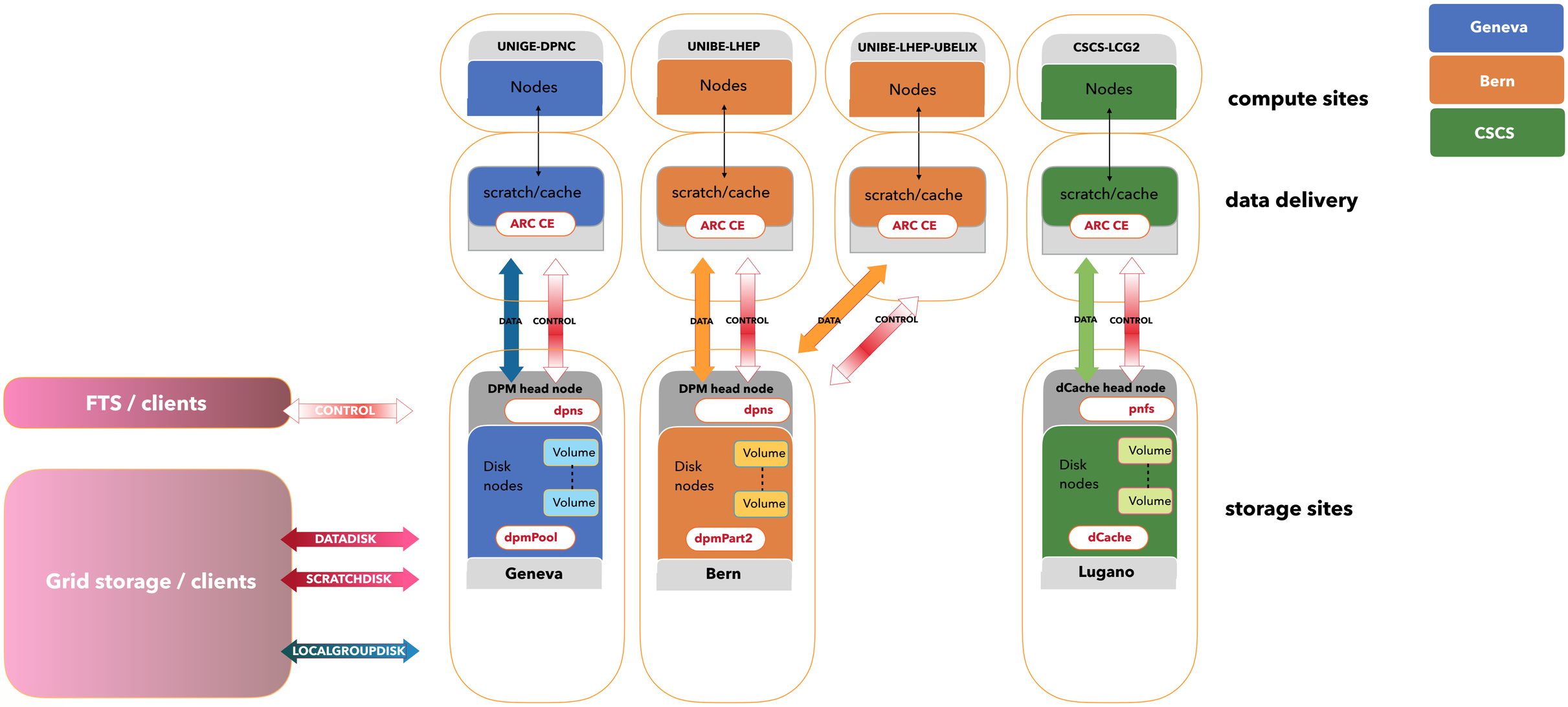
- **ATLAS long standing wish: reduce nr. of storage endpoints**
  - ✳ Reduce operational load on DDM ops
  - ✳ One larger storage more useful than the sum of its components
- **Consolidation of resources at the national level**
  - ✳ Discontinuing small storage => loss of funding
- **Optimise Human Resources**
  - ✳ Centralise MW administration, where the experiment specific expertise is
- **Aggregate diversified resources and technologies**
  - ✳ Provides flexibility in adapting to ongoing evolution
- **Step towards the WLCG data lake direction leveraging current technologies**
- **Exercise the federated model**
  - ✳ More than aggregating computers and services: policies, inter-site mechanics, trust, etc.

INTER-SITE AGREEMENTS

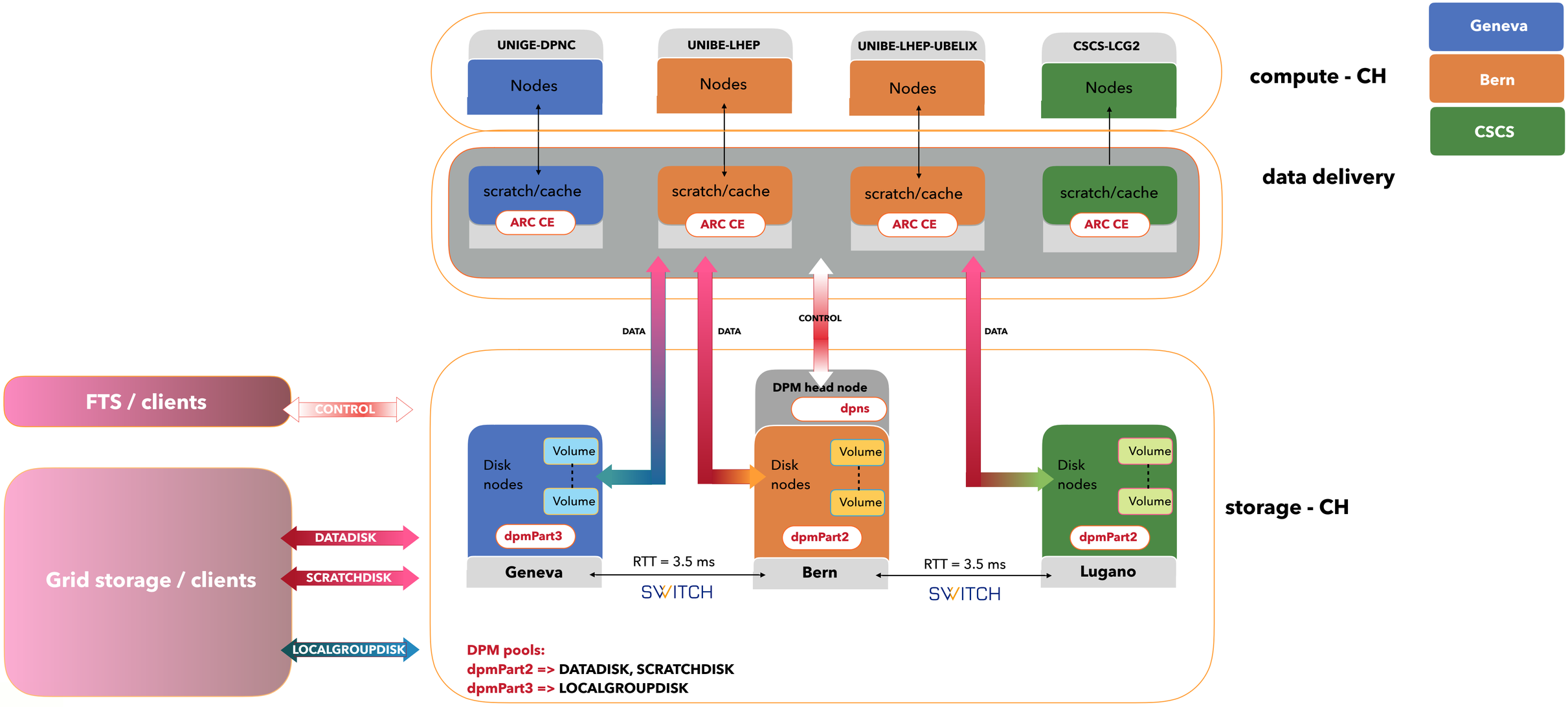
- **Each site's responsibilities**
  - Storage server provisioning and maintenance (base OS, bad disks, etc.)
  - Network, firewall …
  - Host certificates

- **Head site responsibilities (Bern)**
  - Head node
  - Middleware stack on all storage servers inc. remote
  - OS updates, security, monitoring …
  - Admin access by ssh key, granted from one Bern IP address or subnet

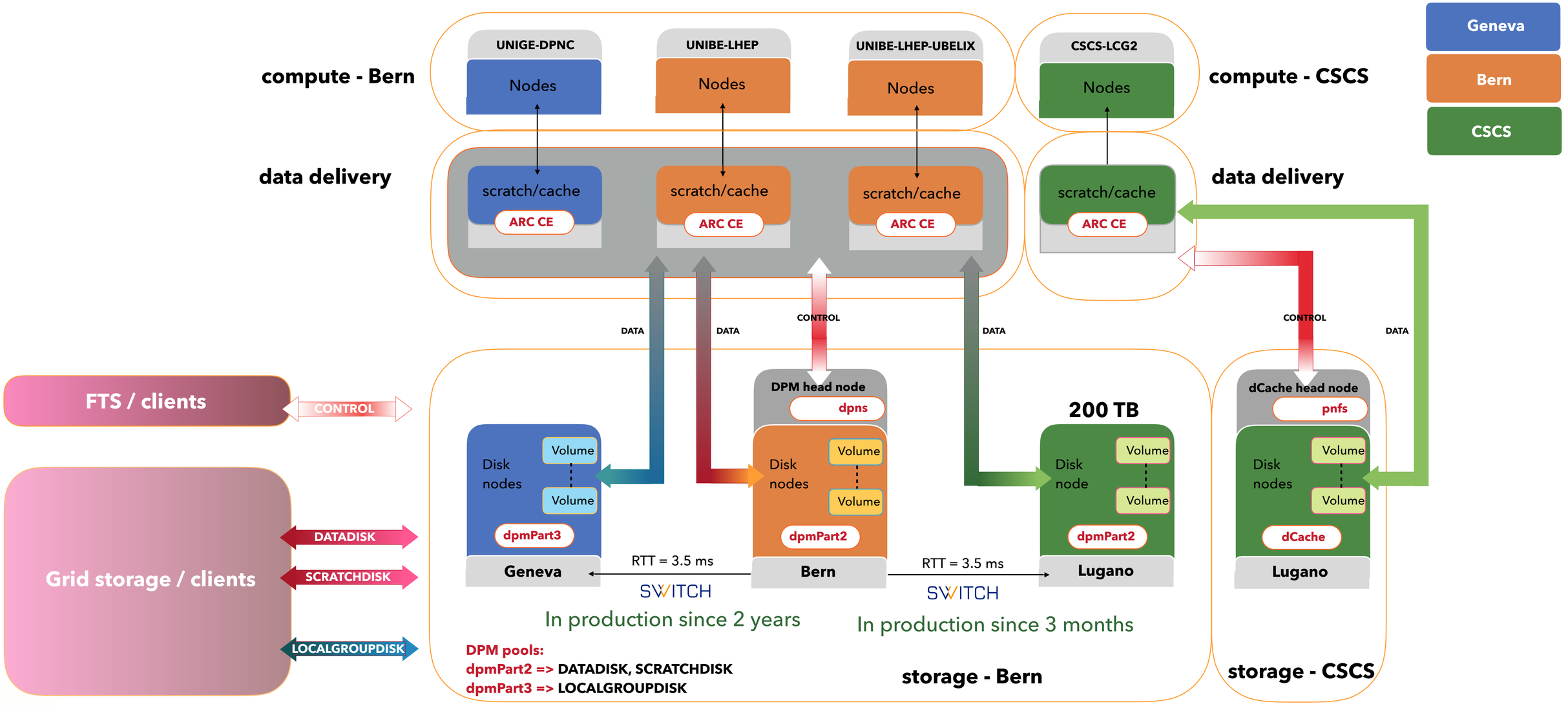
SWISS ATLAS RESOURCES



SWISS ATLAS FEDERATION LAYOUT

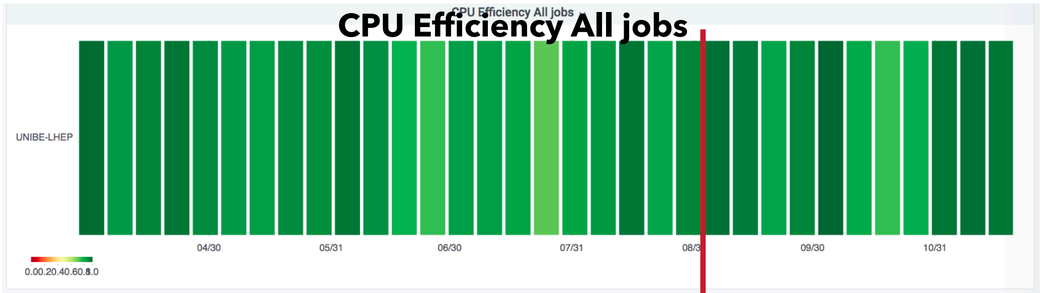


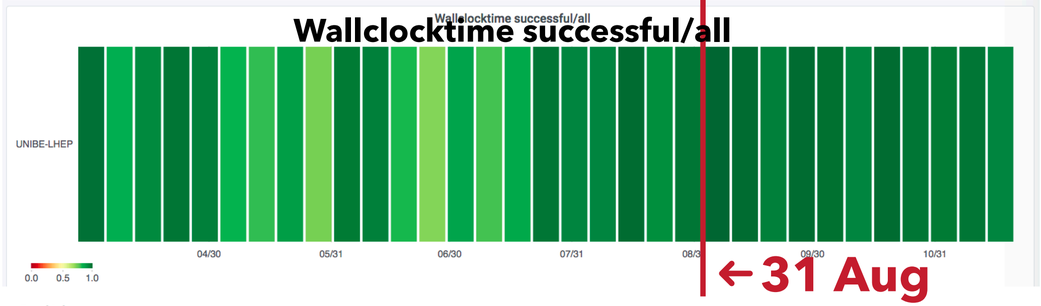
SWISS ATLAS FEDERATION CURRENT STATUS

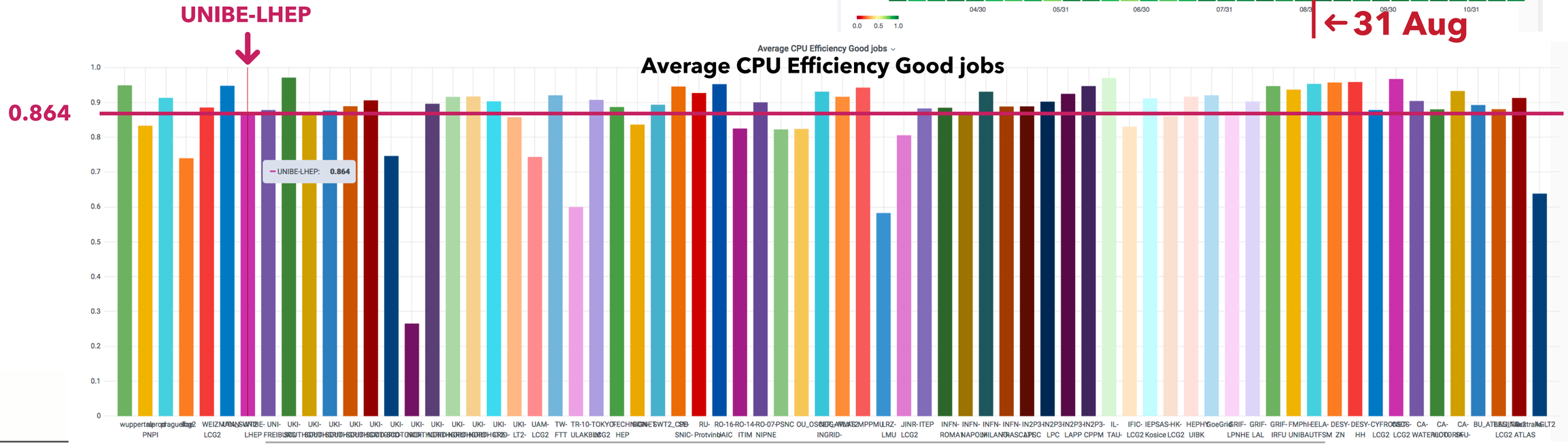


SWISS ATLAS FEDERATION: PERFORMANCE CONSIDERATIONS

UNIBE-LHEP	Before 31/8	After 31/8	T2 average
WCtime successful/all	0.76 →	0.89	0.89
CPU eff good jobs	0.89 →	0.86	0.87
CPU eff all jobs	0.77 →	0.79	0.84







ARC CACHE CONSIDERATIONS

- **ARC cache is *not* an additional service at the site**
  - ✳ Data and cache management are built-in in ARC
  - ✳ Switched on and tuned with 4 lines of configuration

```
[arex/cache]
cachedir=/grid/lustre/cache
[arex/cache/cleaner]
cachesize=75 65
```

← **set location**

← **set size**

  - ✳ Can be anything: cluster FS, NFS server(s), dir in existing vol., dedicated vol., etc.

- **Example at UNIBE-LHEP**
  - ✳ *Shared FS:* Lustre on commodity JBODs, IB, 200 spindles, 200 TB, for 6k cores
  - ✳ Cache tuned to 40% of total capacity (80 TB): >50% hit rate (~6 months, 7M files)
  - ✳ Increasing limit to 75% (150 TB) doesn't yield a higher rate (1 week statistics)

OUTLOOK AND CONCERNS

- **Federation provides more than the sum of its components**
  - ✳ Equally true for the underlying challenges (funding schemes, agreements, etc)
- **Technical challenges are second order**
- **But: additional boundary conditions, e.g. middleware sustainability**
  - ✳ DPM future at risk
  - ✳ ARC has a roadmap of development aiming at meeting requirements for the HL-LHC
- **Human Resources challenge**
  - ✳ Switching established technologies is not trivial and is costly
  - ✳ Plenty of long term expertise is lost, not guaranteed to be easily replaceable
  - ✳ Any big scale changes should be established well in advance
- **Also impact on infrastructure design and developments ($$$ long term investments)**

THANK YOU FOR YOUR ATTENTION

BACKUP