The DOMA project: requirements

LHCONE/LHCOPN workshop – CERN – 2020

S. Campana (CERN) on behalf of WLCG DOMA

DOMA in a nutshell

DOMA project
(Data Organization, Management, Access)

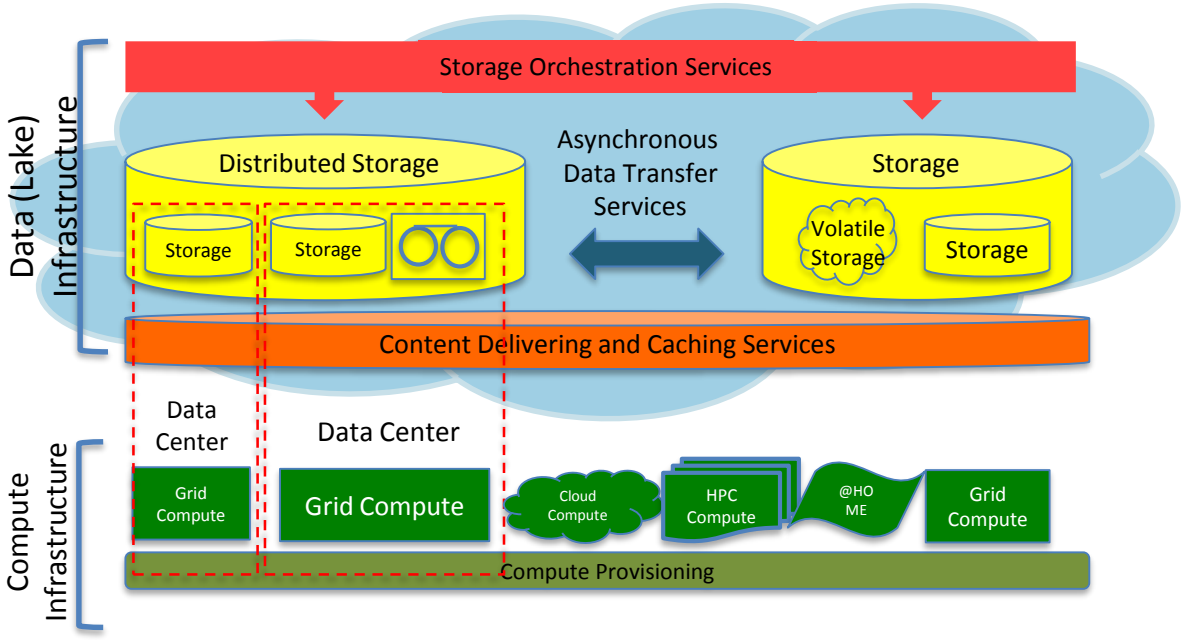
https://twiki.cern.ch/twiki/bin/view/LCG/DomaActivities

A set of R&D activities evaluating components and techniques to build a common HEP data cloud

Three Working Groups

- ACCESS for Content Delivery and Caching
- TPC for Third Party Copy
- QoS for storage Quality of Service

And many activities, reporting regularly



Network needs in HL-LHC (1)

Connectivity in HL-LHC from the DOMA experience and the experiment inputs. This is really handwaving, so we talk about orders of magnitude here:

The LHC experiments produced ~20PB of RAW data / year in Run-2 . They will produce ~ 1EB /year in Run-4 (so 50 times).

Today the OPN is at multiple 10Gb/s per link. We will need AT LEAST **O(1Tb/s)** per link in HL-LHC.

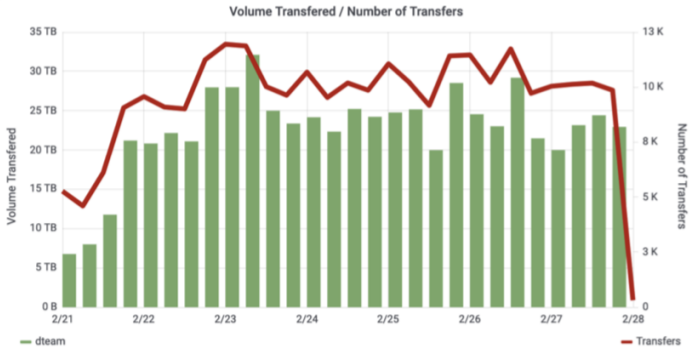
BTW, special attention is needed for the US, which receive > 30% of the RAW data in quasi real time from CERN, so 0.5Tb/s **JUST FOR THE RAW DATA**

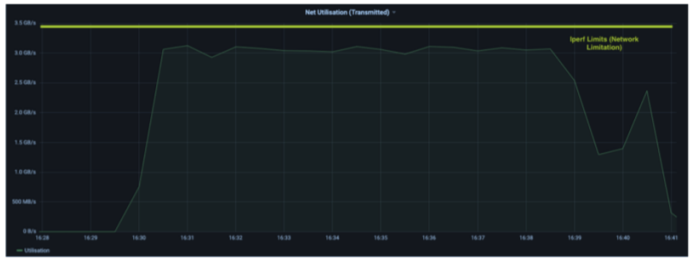
Third Party Copy in DOMA

Goal: commission non-gridFTP protocols for asynchronous data transfer (Third Party Copy)

- Phase-2: all sites providing > 3PB of storage to WLCG should provide a non gridFTP endpoint in production

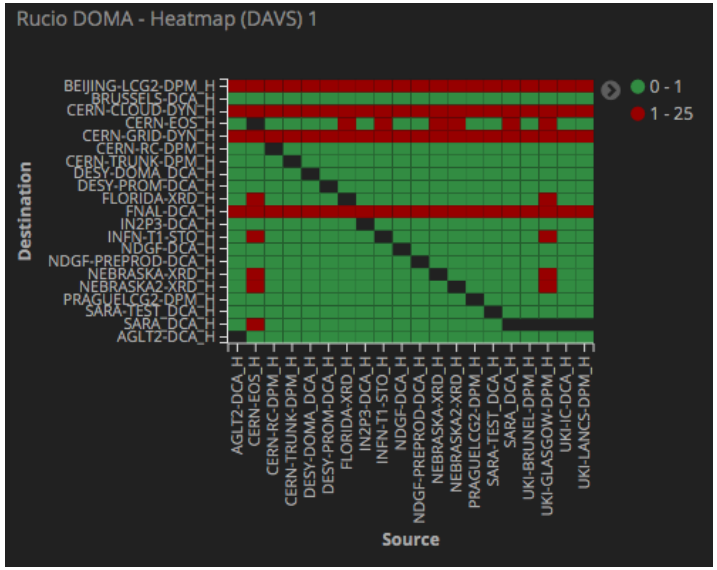
Functional and Stress testing



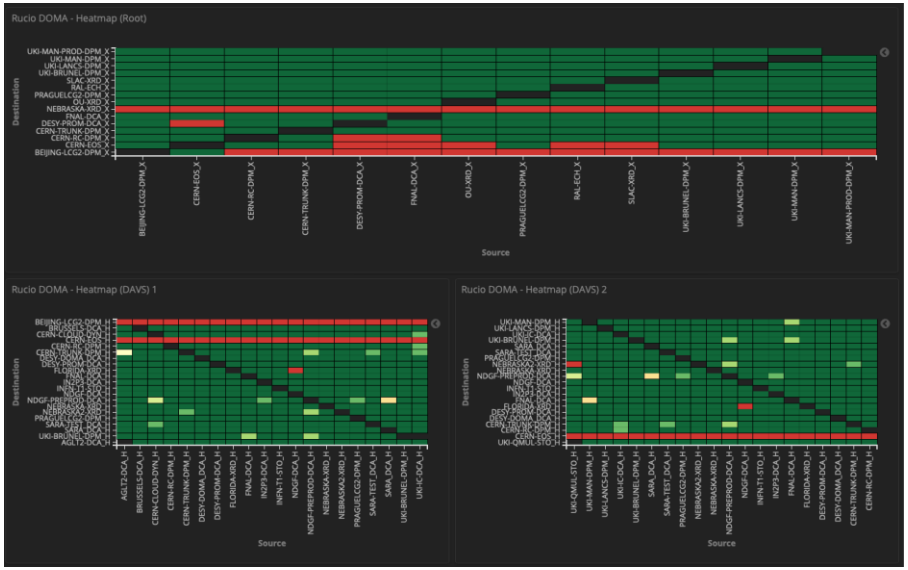


Capable to fill available bandwidth

Functional Tests in June



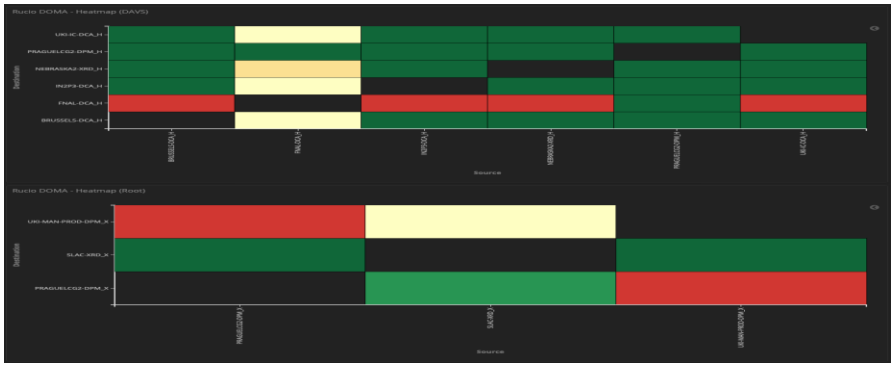
Functional tests in December



We need to demonstrate we can fill this bandwidth with the services and protocols we are using (e.g. moving to HTTP/xrootd for asynch transfers).

Commissioning and parameter tuning in the next years, surely before HL-LHC will require help of network experts

Stress tests in December



NOTED: orchestrating networks

See the dedicated talk from Edoardo

Transfer broker:

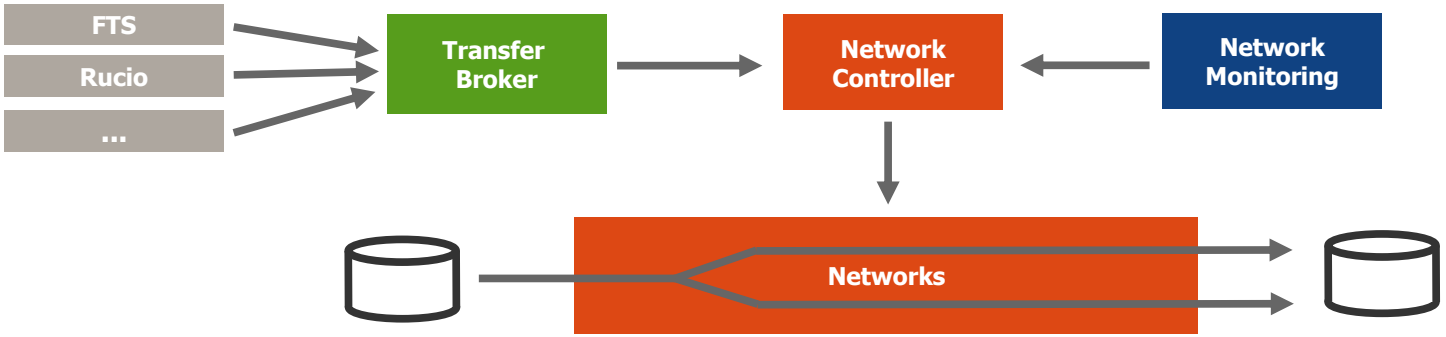
- Identify upcoming and on-going substantial data transfers
- get information from transfer services (FTS, Rucio …)
- map transfers to network endpoints
- make transfers info available to network providers

Network Controller:

- takes input from Transfer Broker
- modify network behavior to increase transfer efficiency
- take into account real-time network status information

LHCOPN/LHCONE interplay: CERN-RAL example, moving from multiple 10Gb/s links to larger (100Gb/s) single link with LHCONE as backup. Does it have to be backup only?

NOTED project interesting for T1s/T0 at the least



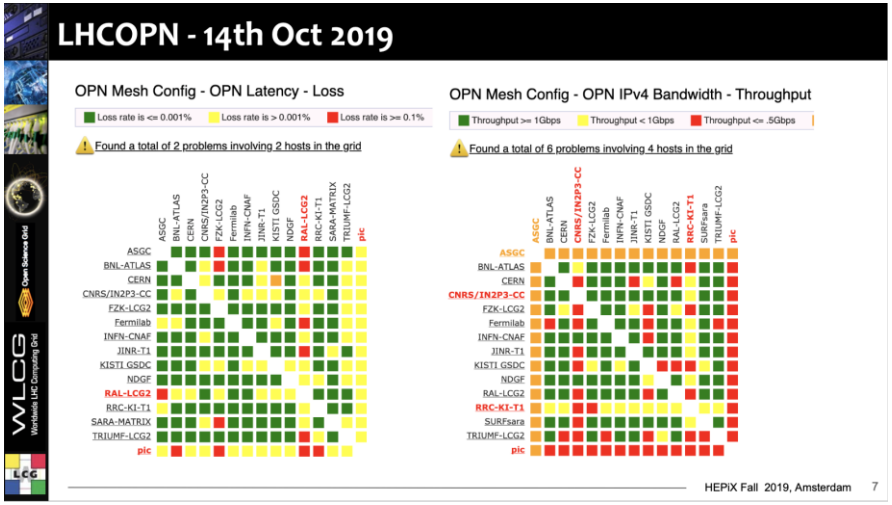
Network Monitoring

perfSONAR infrastructure deployed for WLCG: a very useful diagnostic tool

Steps forward to use it as “Monitoring”? (e.g. “red” in packet loss always an indicator of a problem?)

Monitoring of used and available bandwidth (e.g. for transfer scheduling) not available today

Is this something the LHCOPN/LHCONE community could provide?
(e.g. collect from routers and aggregate?)



Data ACCESS strawman model

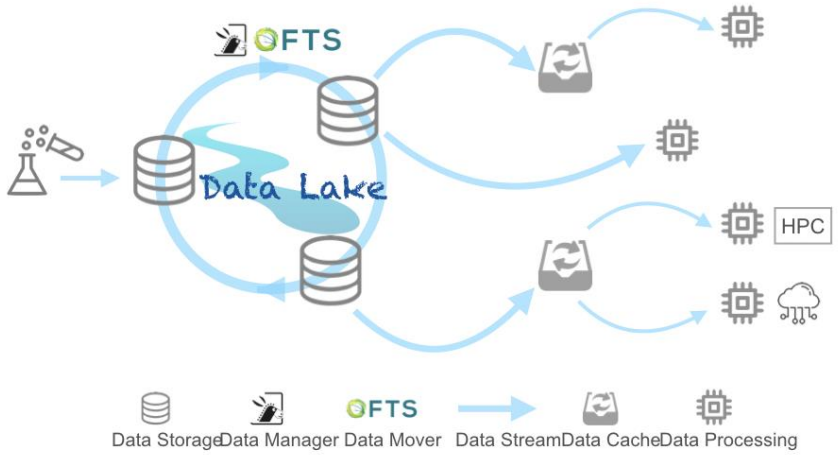
Fewer number of facilities operating storage services, more data traffic for those

Some storage services will be distributed: different data nodes at different sites

CPUs and storage not necessarily co-located: need to deliver the content over the WAN and/or cache it

What today is LAN activity becomes WAN activity in the datalake – at least within the NREN

How do we see this network shaped as part of the global R&E network



Network needs in HL-LHC (2)

From the likely analysis model nanoAODs will be primary analysis format. They are O(1000) times smaller than RAW: 1PB/year single version single copy

Unlikely network bandwidth challenges will come from analysis of nanoAODs

Network challenges for Analysis:

- latency for remote reading: need latency hiding mechanism in the content delivery layer
- chaotic nature: need caching in our content delivery layer
- **reliability and redundancy of the network**

Courtesy David Lange
Present Model of CMS
HL-LHC resource planning

Data Tier	Data
RAW [MB]	7.4
AOD [MB]	2.0
MiniAOD [kB]	200
NanoAOD [kB]	4

ACCESS: xCache performance

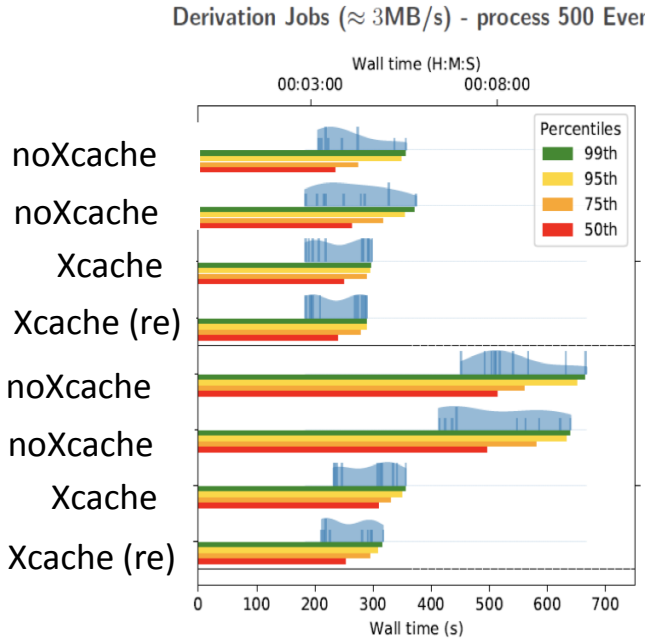
ATLAS Derivation jobs. Metric: WallTime / 500 Evts

Compares direct read from storage (directIO) with read through xCache in Munich

Conclusion:

- xCache hides latency for high RTT. Data access seen as “quasi-local”
- Further benefit in case of re-use (caching)

Processing Nodes in Munich

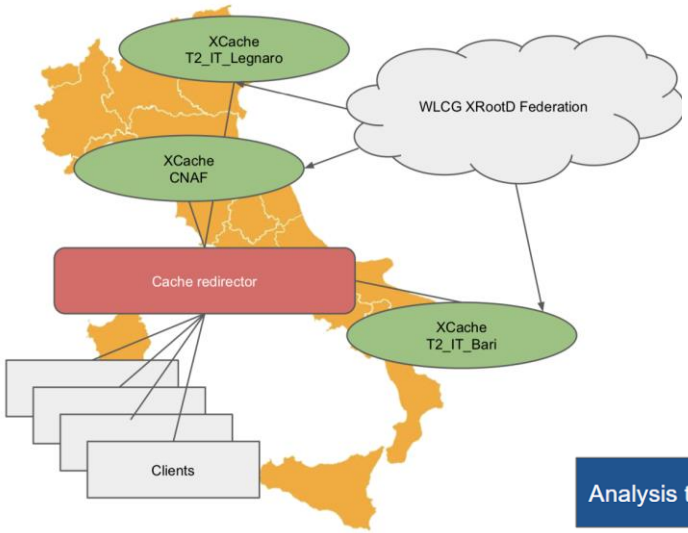


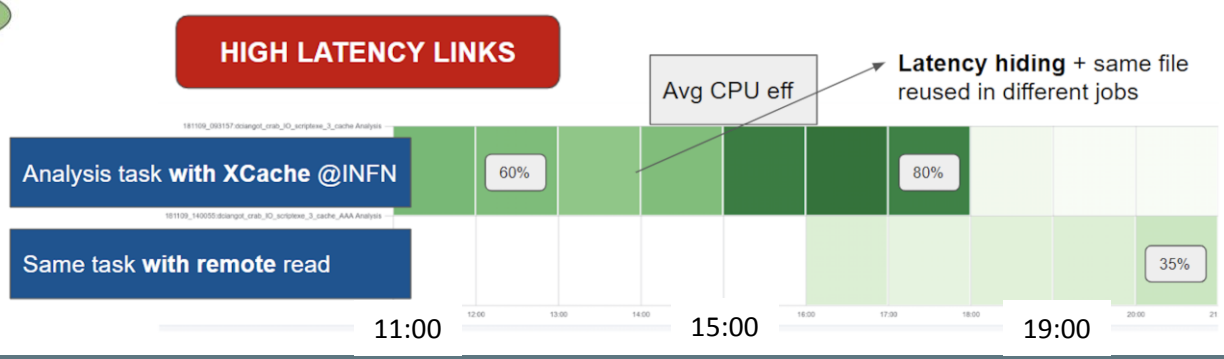
data from DESY

data from Beijing

ACCESS: caching layer prototype

A distributed caching system in INFN



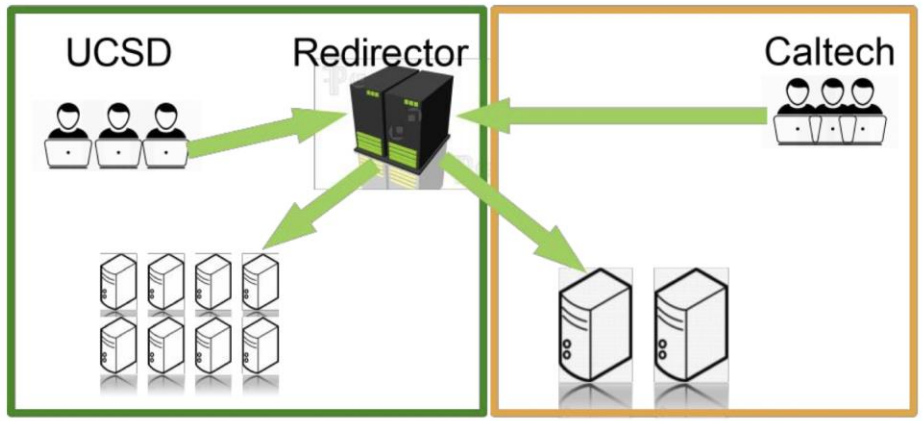


ACCESS: SoCal

Courtesy David Lange
Present Model of CMS
HL-LHC resource planning

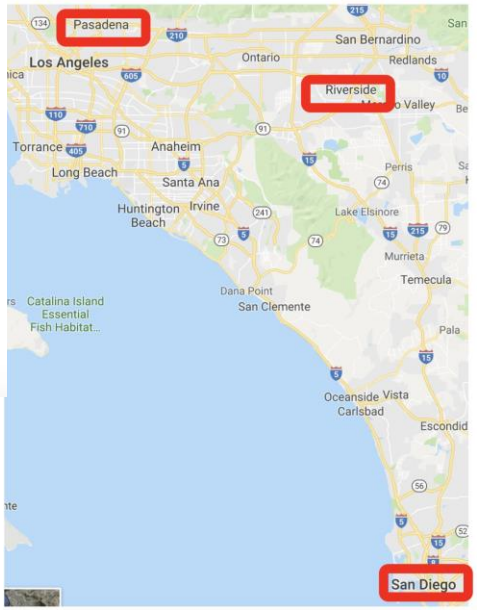
Data Tier	Data
RAW [MB]	7.4
AOD [MB]	2.0
MiniAOD [kB]	200
NanoAOD [kB]	4

SoCal XRootD Cache



	UCSD	Caltech
Nodes	11 (10 more coming)	2
Disk Capacity per node	12x2TB = 24TB	30x6TB (HGST Ultrastar 7K600)
Network Card per node	10 Gbps	40 Gbps
Total Disk Capacity	264 TB	360TB

Last Year



SoCal serves data to CPUs in all South California CMS sites
We now better understand trading of storage and network

Network needs in HL-LHC (3)

Producing AODs and derived formats will be the dominating network activity

In the datalake model, the network traffic will increase in the “region” (we already have reports about that e.g. in the UK)

Network within a region today on the order of tens of Gb/s (2x100Gb/s for RAL to JANET), should scale up also **to the Tb/s level**

Courtesy David Lange
Present Model of CMS
HL-LHC resource planning

Data Tier	Data
RAW [MB]	7.4
AOD [MB]	2.0
MiniAOD [kB]	200
NanoAOD [kB]	4

Some extreme scenarios should also be considered, such as large allocations of compute capacity (say at an HPC) for a short period of time. That could allow to process multi-year data in few months, and again the connectivity required would scale at the Tb/s level **to that center**

LHCONE, DUNEONE, SKAONE, ALLINONE

Threats and Opportunities from many (scientific) communities competing for networks (say SKA, LSST, DUNE, Belle-2, 3rd Generation G-Waves, CTA, …)

Threats: competing for bandwidth (really a funding issue) and increasing complexity (security, “QualityOfService”, ..). We have very good experience with LHCONE, but how does it extend to those other communities? Invitation to the network community to define the problem scope and look for solutions. Notice this is not a point-2-point problem but a global problem and the solution needs to be simple enough

Opportunities: increase the global worldwide connectivity particularly regions with a complicated connectivity (Asia is an example – synergy with ATCF). Expand the LHCOPN/LHCONE experience and ecosystem to new scientific communities (operations, policies, …). Global scientific network operations ?

Carefully craft a solution simple for everyone: experiments, NRENs, sites

Take Home

In preparation for HL-LHC

- Expect an increase to 1Tb/s for LHCOPN
- Expect an increase to 1TB/s for regional traffic

- Improve further the resilience of network connectivity
- Improve further the network monitoring

- Leverage collaboration between scientific computing experts and network experts to use network at its best e.g. data challenges

- Consider HL-LHC and Scientific Computing in the 2020s as an opportunity to invest in network R&D: scope the problem and do not increase complexity