Distributed Computing and Data Management for HEP in the next decade

Xavier Espinal (CERN)

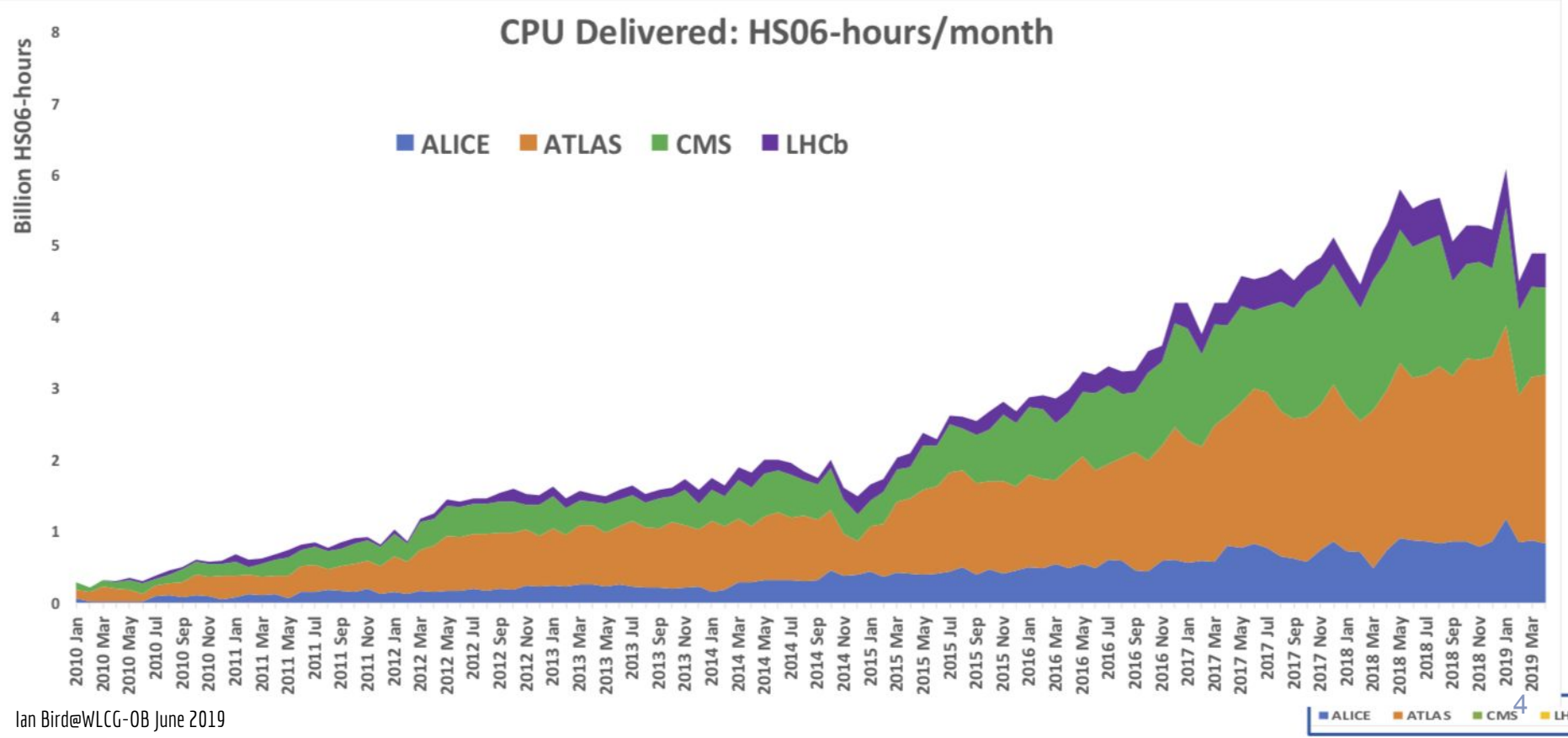
2000: Scientific computing goes distributed

- LHC brought unprecedented computing needs, too much for a single datacenter
  - Change of paradigm: **scientific computing goes distributed**
- Big challenge. Years of preparation to design and set-up the Worldwide LHC Computing Grid (*the grid*)
  - Common auth/authz, data transfers across sites with different technologies, workload management systems (grid vs. batch systems), networking, operations model, etc.
- Big success: in full production since the start of LHC enabling science worldwide

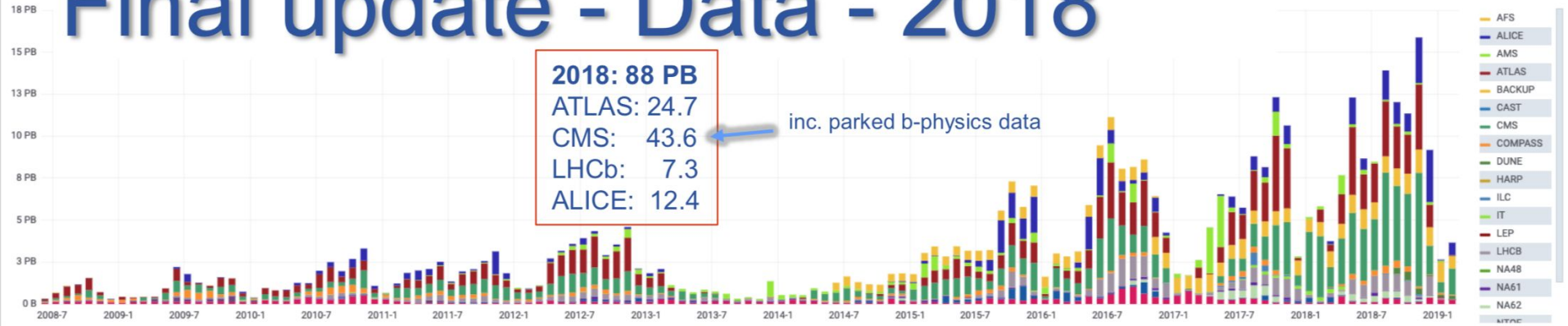
2019: The Worldwide LHC Computing Grid

- 170+ centers in 42 countries
  - Interoperability with architectural ”freedom of choice” for the sites and common “middleware”
- 4 LHC experiments with a large community of scientists
- 900,000 cores and 1EB of storage. EB-scale annual data flow (avg 60GB/s)
  - CERN provides about 20% of the WLCG resources
- Several solutions arose to address same problems (Data Management, Workload Management). Converging with time.
- Constant quest for optimization in computing resources and operations (manpower)
  - From sites to infrastructure to experiments
- Some current tools as building blocks for the future in scientific computing:
  - FTS, RUCIO, DIRAC, token-based AAI, etc.

WLCG: data processing



Final update - Data - 2018



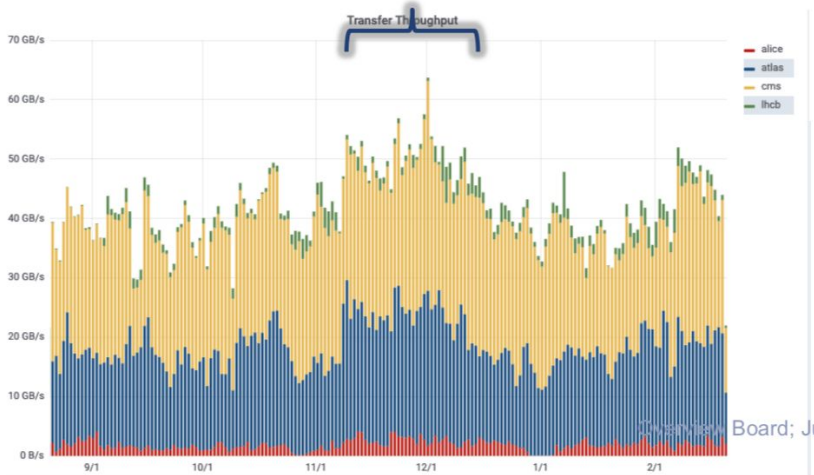
2018: 88 PB
ATLAS: 24.7
CMS: 43.6
LHCb: 7.3
ALICE: 12.4

inc. parked b-physics data

Ian Bird@WLCG-OB June 2019

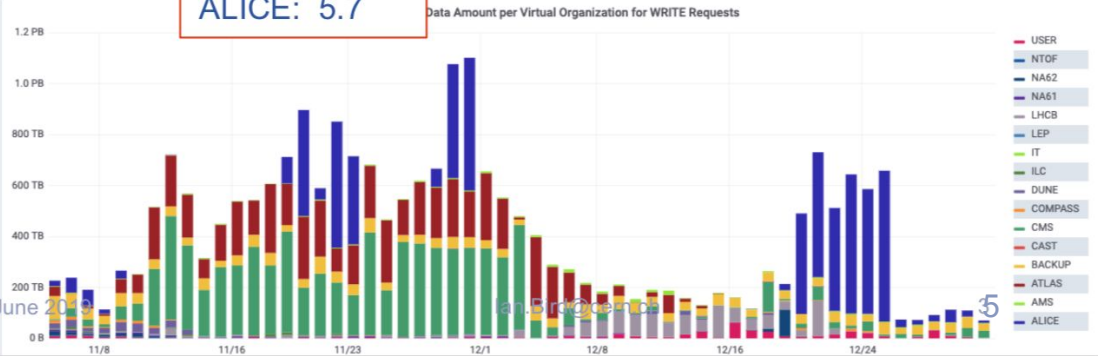
Data transfers

HI Run

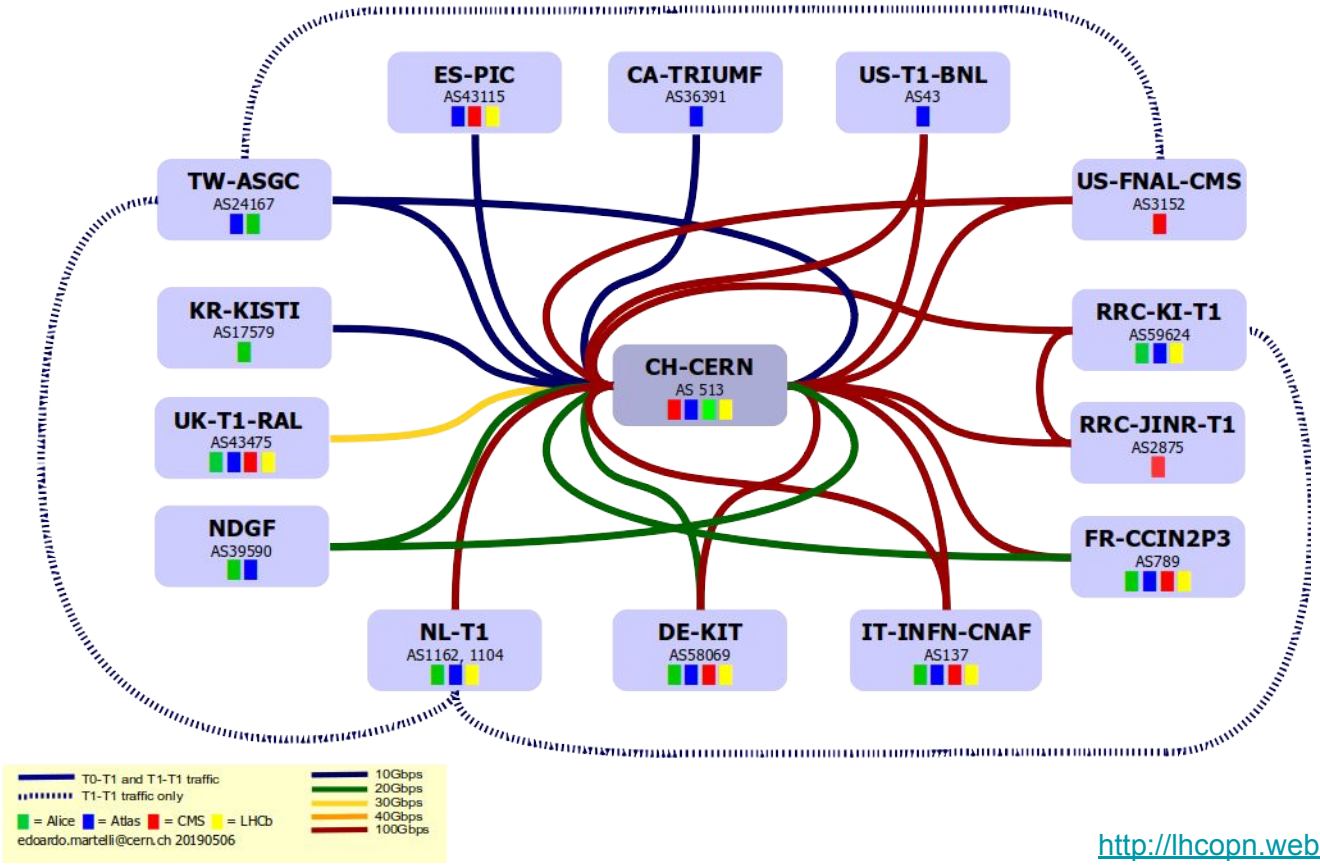


2018: 19.8 PB
ATLAS: 5.2
CMS: 7.7
LHCb: 1.2
ALICE: 5.7

HI Run



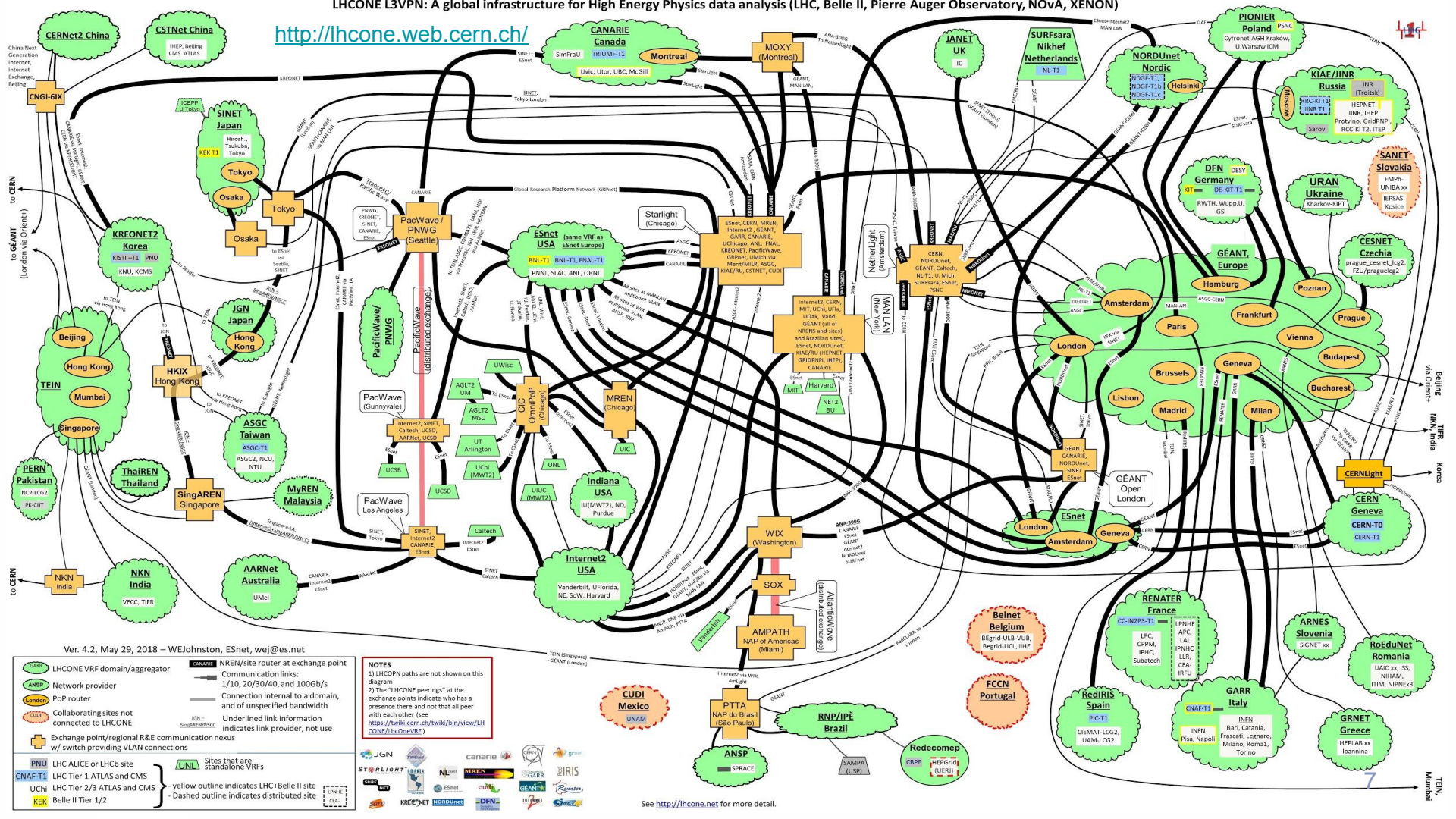
LHCOPN



http://lhcopn.web.cern.ch/lhcopn/

LHCONE L3VPN: A global infrastructure for High Energy Physics data analysis (LHC, Belle II, Pierre Auger Observatory, NOvA, XENON)

http://lhcone.web.cern.ch/



WLCG strategy towards HL-LHC

- Set out the path towards **computing for HL-LHC** in 2026/7
- Estimates of the data volumes and computing show a **major step** up from the current needs
- Program of work from the WLCG point of view, with a focus on HL-LHC, building on all of the background work provided in the CWP
  - **WLCG DOMA project**
    - DOMA TPC
    - DOMA QoS
    - DOMA ACCESS

- **Reducing data volumes:** A key cost today is the amount of storage required. Investigating mechanisms for reducing that volume will have a direct effect on cost: removing or reducing the need for intermediate data products that must be stored, managing the sizes of derived data formats, for example with "nanoAOD"-style even for some fraction of the analyses will have an important effect. There is a big potential here, but needs work from the experiments.
- **Managing operations costs:** Here there are a number of strategies. Investigating the opportunities with storage consolidation is a high priority. The idea of a "data-lake" where few large centres manage the long-term data, while needs for processing are managed through streaming, caching, and related tools, allows the cost of managing and operating large complex storage systems to be minimised. It also reduces complexity for the experiment. Importantly, such a structure gives the opportunity to move common data management tools out of the experiments and into a common layer. This allows better optimisation of performance and data volumes, easier operations, and common solutions. It also makes it easier to introduce common workflow solutions. Storage consolidation can save cost on expensive managed storage, but requires that we are able to hide the latency via streaming and caching solutions. This is feasible as many of our workloads are not I/O bound, and data can be streamed to a remote processor effectively with the right tools.
- **Optimising hardware costs:** There is an opportunity to reduce storage cost also by more actively using tape (or cold storage). With a highly organised access to tape it could replace the need to keep a lot of data that is today kept on disk. The judicious use of virtual data (re-create samples rather than store) is another opportunity. This could save significant cost, but requires the experiment workflows to be highly organised and planned. Moving away as far as possible from random access to data before the final highly refined analysis formats. Other considerations include the optimisation of the amount of storage vs compute, and optimising the granularity of data that is moved - between dataset level and event level.

Future in Scientific Computing (1/2)

- HL-LHC requirements (2025+) pose (again) a challenge in terms of data volumes and processing capacity
  - technology evolution and funding not helping
- Other experiments on a similar scale
  - Astrophysics, Cosmology, Gravitational Waves and Neutrino physics
- Global interest for collaboration: common tools and infrastructures
  - Storage consolidations (Datalakes), Auth/Authz (AAI), Data Management, Workload Management, Data Transfers, protocols, etc.

Future in Scientific Computing (2/2)

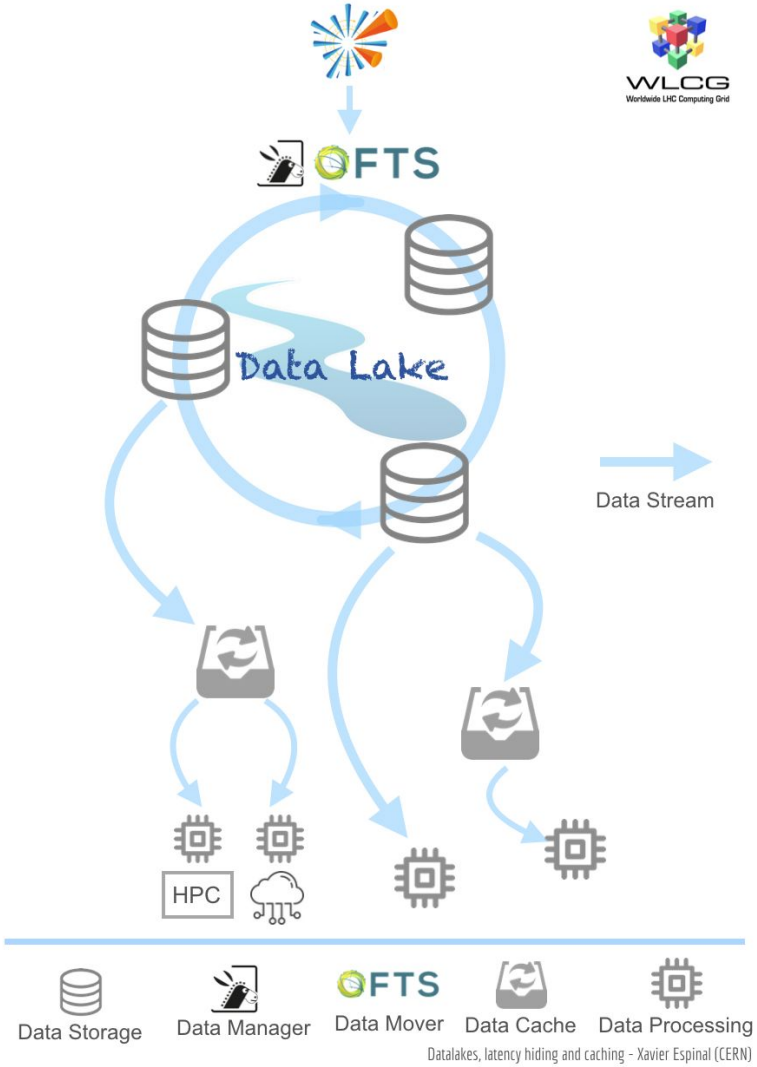
- New scenario: towards shared infrastructures for multi-science
  - Enhancing scientific diversity while maximizing commonality of tools
  - Computing infrastructures flexible enough to embrace different types of: **workloads**, **sites** and **resources** (on-premise, HPC, clouds)
- New possibilities: compact analysis datasets and caching “infrastructures”
  - Opening new possibilities in data processing models and infrastructure
  - Event streaming services, data frames as alternative to standard event based formats for analysis
- New paradigms: can we make storage smarter/optimal/cheaper?
  - Towards storage consolidation and datalakes
- Review/rethink data placement and sites’ roles: network impact, interactive analysis evolution

WLCG DOMA R&D Project

- **R&D activities** evaluating components and techniques to build a common HEP computing infrastructure:
  - https://twiki.cern.ch/twiki/bin/view/LCG/DomaActivities
- Three main areas covered by three **Working Groups**
  - **ACCESS**
    - Datalakes, Content Delivery and Caching
  - **TPC**
    - Third Party Copy and Protocols
  - **QoS**
    - Storage Quality of Service
- There are also other activities regularly reporting to the DOMA project:
  - **Networking**
  - **AAI**: Authorization, Authentication and Identity Management
  - ...

DOMA ACCESS WG

- Exploring models to bridge data with cpu
  - Datalakes, content delivery and caching
- Studies on:
  - “datalake” models
  - XCache deployment and performance
  - Data access/patterns, file usage
- Strong involvement from the community:
  - Experiments
  - Sites
  - Software
- Synergies with all DOMA activities

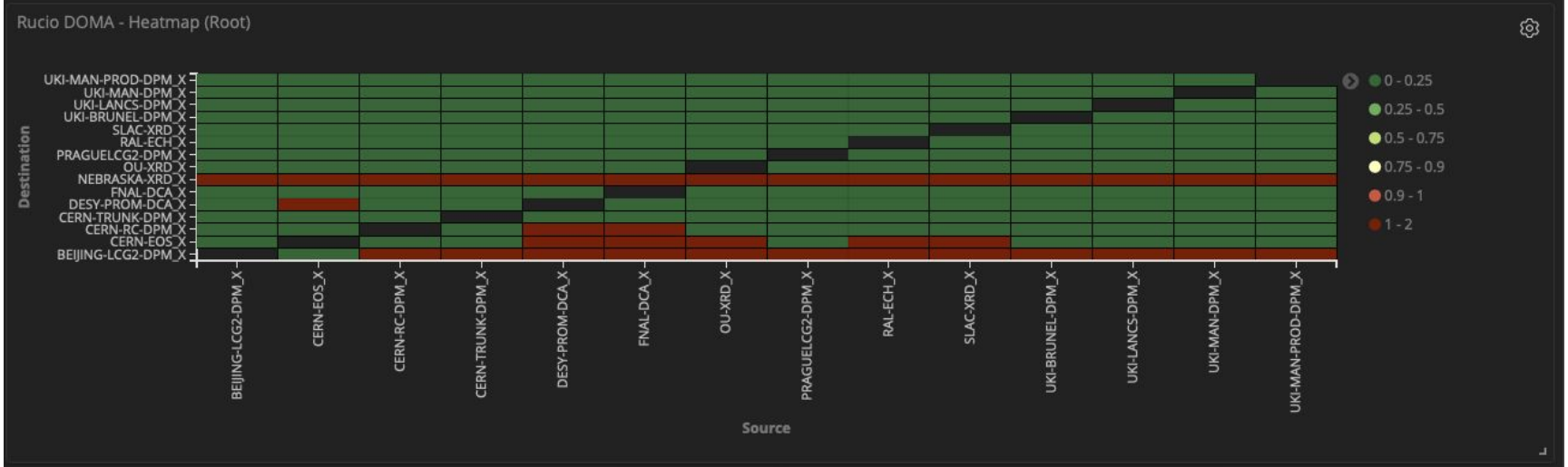


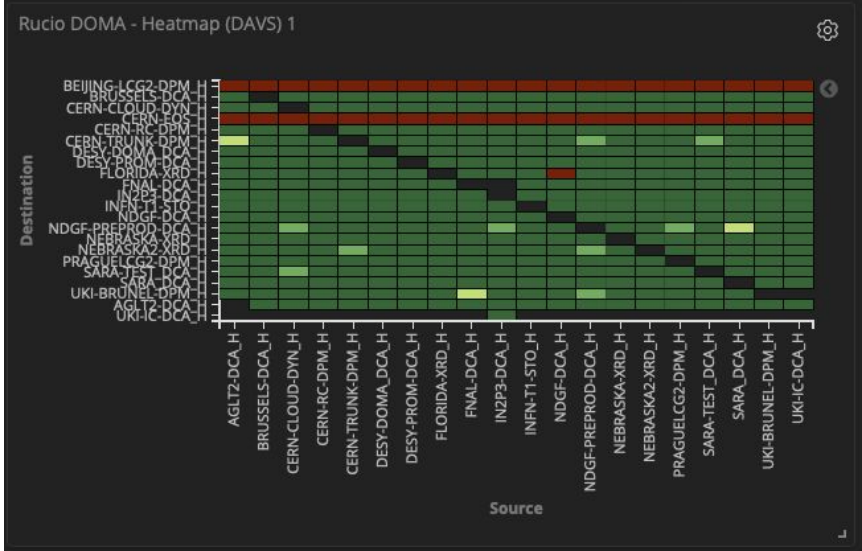
DOMA ACCESS: One year of activity

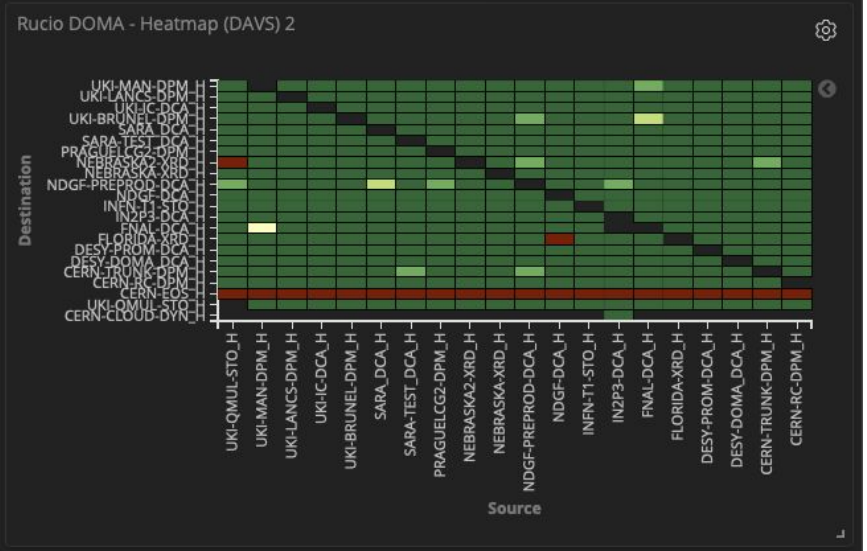
- Focussed on **collecting information**
  - R&D activities on “caching”: mainly based on XCache
  - Future computing models and file formats (nanoAOD, phys_lite)
- **Understanding** of data access and file usage
  - Simulation (and emulation) of data access patterns
  - Emulations of file usage
  - Network usage implications
- Strawman model: data access on a datalake
  - Early straw man model addressing (too) many topics
  - To be re-visited with results from R&D activities, new data formats and new ideas

DOMA **TPC** WG - **T**hird **P**arty **C**opy

- Third party copy address a core grid activity: moving data from site A to site B
- Need to replace the functionality found in the Globus Toolkit (GridFTP & GSI)
  - The goal is to commission non-gridFTP protocols for asynchronous data transfer
- Roadmap
  - **Phase-1** completed
    - Common storage implementations have (at least) one production site enabling non-gridFTP TPC
    - Compatibility and performance tests are performed
  - **Phase-2** (deadline ~now):
    - All sites providing > 3PB of storage to WLCG should provide a non-gridFTP endpoint in production
  - **Phase-3** (deadline ~Q1 2020): all sites to have a non-gridFTP endpoint
    - Some features needed for TPC only available only in recent versions of storage
    - Upgrade campaigns now ongoing
- Spin-off: progressing toward a SRM-less world


Rucio DOMA - Heatmap (Root)
Destination
UKI-MAN-PROD-DPM_X
UKI-MAN-DPM_X
UKI-LANCS-DPM_X
UKI-BRUNEL-DPM_X
SLAC-XRD_X
RAL-ECH_X
PRAGUELCG2-DPM_X
OU-XRD_X
NEBRASKA-XRD_X
FNAL-DCA_X
DESY-PROM-DCA_X
CERN-TRUNK-DPM_X
CERN-RC-DPM_X
CERN-EOS_X
BEIJING-LCG2-DPM_X
Source
0 - 0.25
0.25 - 0.5
0.5 - 0.75
0.75 - 0.9
0.9 - 1
1 - 2


Rucio DOMA - Heatmap (DAVS) 1
Destination
BEIJING-LCG2-DPM_H
BRUSSELS-DCA_H
CERN-CLOUD-DYN_H
CERN-EOS_H
CERN-RC-DPM_H
CERN-TRUNK-DPM_H
DESY-DOMA_DCA_H
DESY-PROM-DCA_H
FLORIDA-XRD_H
FNAL-DCA_H
IN2P3-DCA_H
INFN-T1-STO_H
NDGF-DCA_H
NDGF-PREPROD-DCA_H
NEBRASKA-XRD_H
NEBRASKA2-XRD_H
PRAGUELCG2-DPM_H
SARA-TEST_DCA_H
SARA_DCA_H
UKI-BRUNEL-DPM_H
AGLT2-DCA_H
UKI-IC-DCA_H
Source


Rucio DOMA - Heatmap (DAVS) 2
Destination
UKI-MAN-DPM_H
UKI-LANCS-DPM_H
UKI-IC-DCA_H
UKI-BRUNEL-DPM_H
SARA_DCA_H
SARA-TEST_DCA_H
PRAGUELCG2-DPM_H
NEBRASKA2-XRD_H
NEBRASKA-XRD_H
NDGF-PREPROD-DCA_H
NDGF-DCA_H
INFN-T1-STO_H
IN2P3-DCA_H
FNAL-DCA_H
FLORIDA-XRD_H
DESY-PROM-DCA_H
DESY-DOMA_DCA_H
CERN-TRUNK-DPM_H
CERN-RC-DPM_H
CERN-EOS_H
UKI-QMUL-STO_H
CERN-CLOUD-DYN_H
Source

DOMA QoS: storage Quality of Service

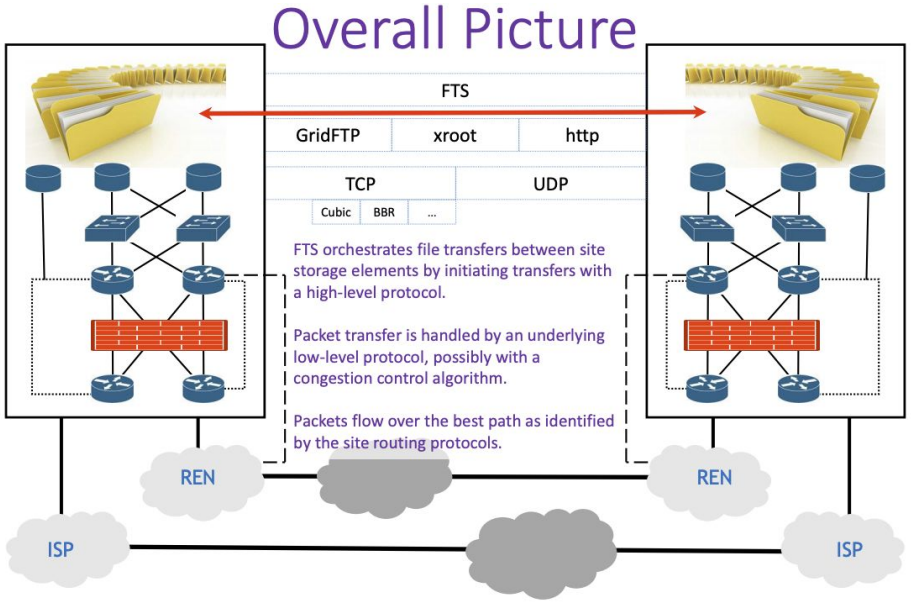
Eierlegende wollmilchsau

DOMA QoS: storage Quality of Service

- Storage site survey
  - Around 80 sites responded, analysis underway:
    - https://twiki.cern.ch/twiki/bin/view/LCG/QoSSurveyAnswers
- Objective:
  - Identify a small set of topics on which the WG will concentrate future effort
    - **Common directions and interests across sites**
  - Possible list (see twiki for more detail)
    - Procurement, densification and media
    - Software defined storage
    - Client-driven QoS
    - WLCG QoS classes

Network R&D activities

- Focusing on DTNs, low level transfer protocols, bandwidth on demand, P2P channels and SDNs
  - Collaboration with the SKA AENEAS project and HEPIX
- Projects:
  - NOTED
  - multiONE



Cass@DOMA
https://indico.cern.ch/event/748636/

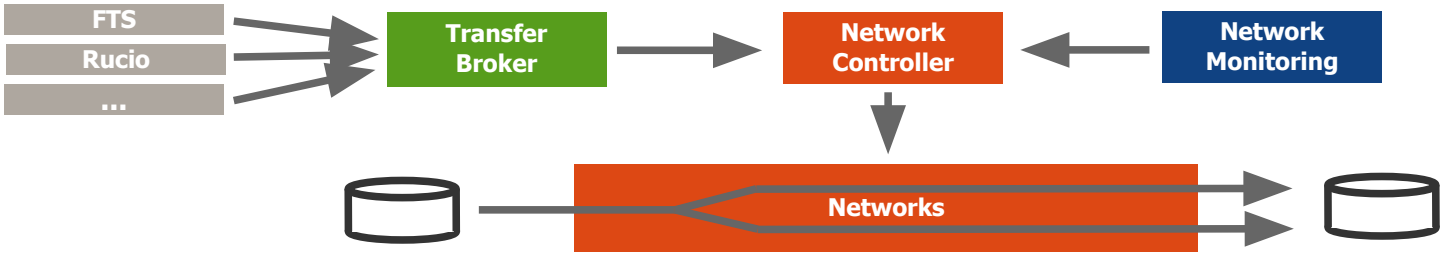
NOTED: shaping networks

Description

- Implement a Transfer broker:
  - **Identify** upcoming and ongoing substantial data transfers
  - **Publish** transfers information to network providers
- Demonstrate a Network Controller:
  - Takes input from Transfer Broker
  - **Modify** network behavior to increase transfer efficiency
  - Take into account real-time network status information

Status

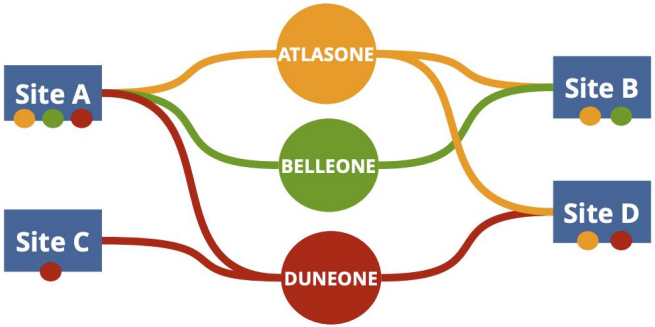
- Transfer Broker:
  - Interpreting information provided by Rucio to estimate volume of upcoming data transfers and to identify source-destination storage elements
- Network Controller, currently evaluating:
  - Stackstorm for network controller
  - Segment Routing and SCION for traffic engineering



multiONE: from LHCONE success

From the recent Discussion at FNAL's [Grid Deployment Board]

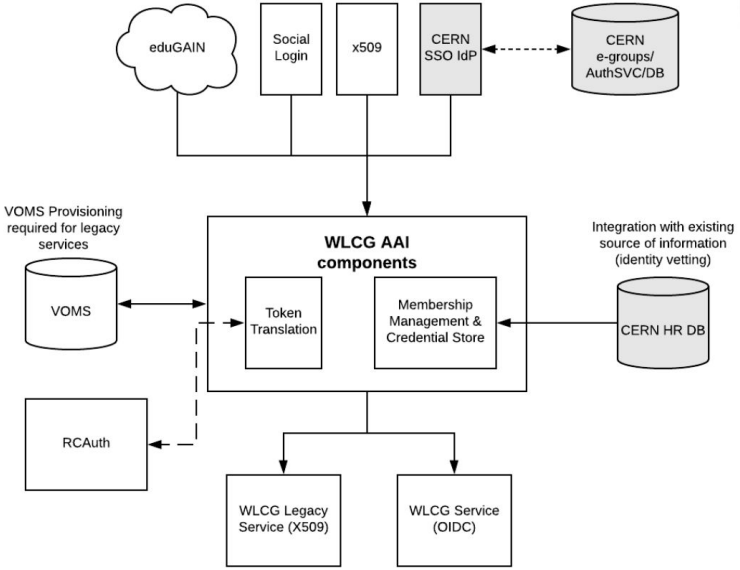
- LHCONE is a worldwide trusted VPN implemented by NRENs connecting Tier1 and Tier2 centres
- Other physics collaborations have benefited of LHCONE, although undermining its primary benefit (connecting few trusted sites)
- Should future major Collaborations should get their own "ONE" ?
- Working on prototype for the DUNE experiment between CERN and Fermilab



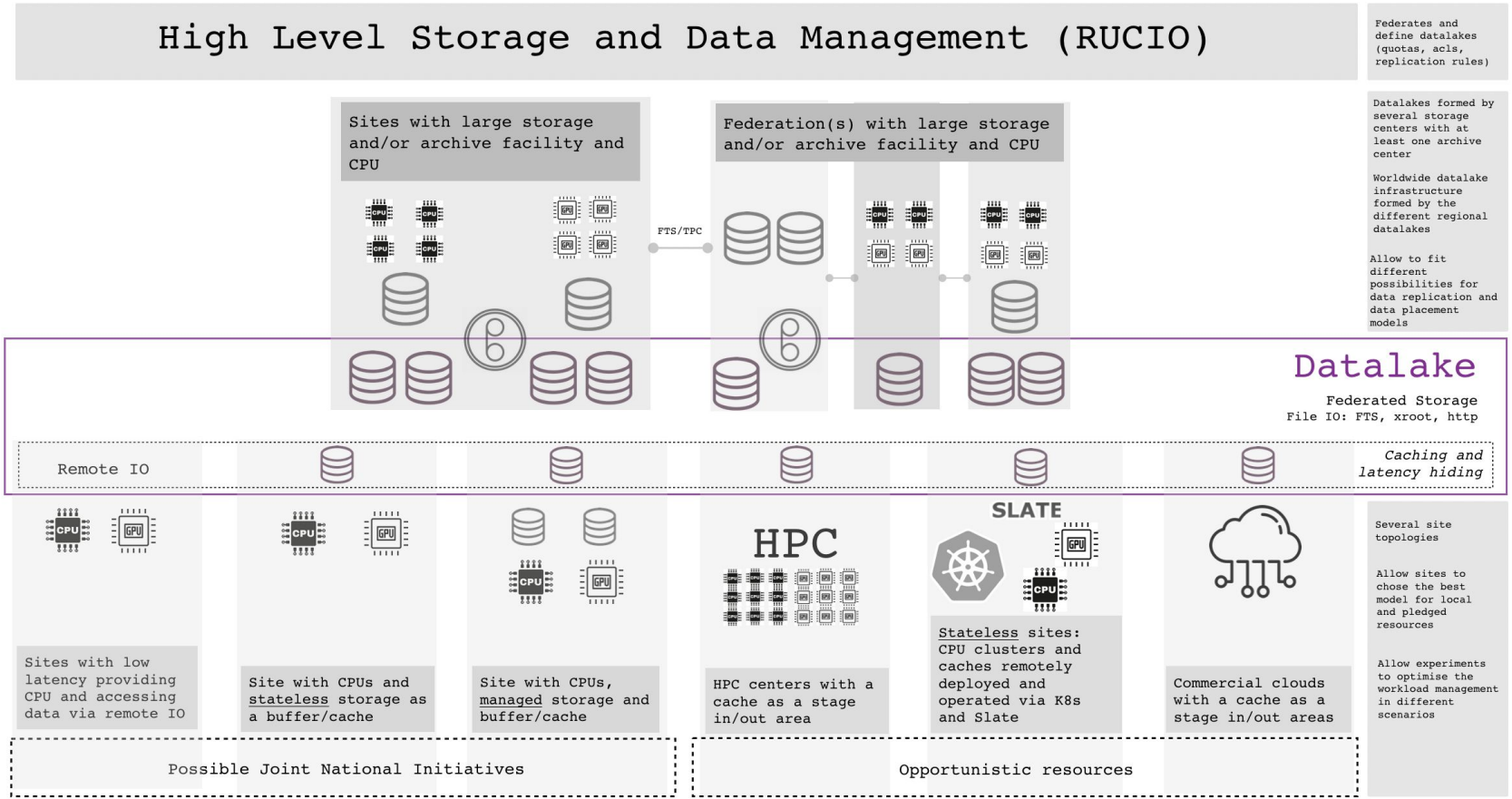
Martelli, Cass@GDB/FNAL
https://indico.cern.ch/event/739882/contributions/3520004/attachments/1906199/3148167/EM-multiONE-GDB.pdf

AAI

- Moving away from X509, embracing a token based architecture
- WLCG MB approved the document describing [WLCG Common JWT Profiles](#) provided by the AAI WG
  - The WG defined the details of a token's format (JWT), content and trust establishment
- The WG has not yet defined:
  - How the tokens should be mapped onto use cases
  - How they should co-exist with X509



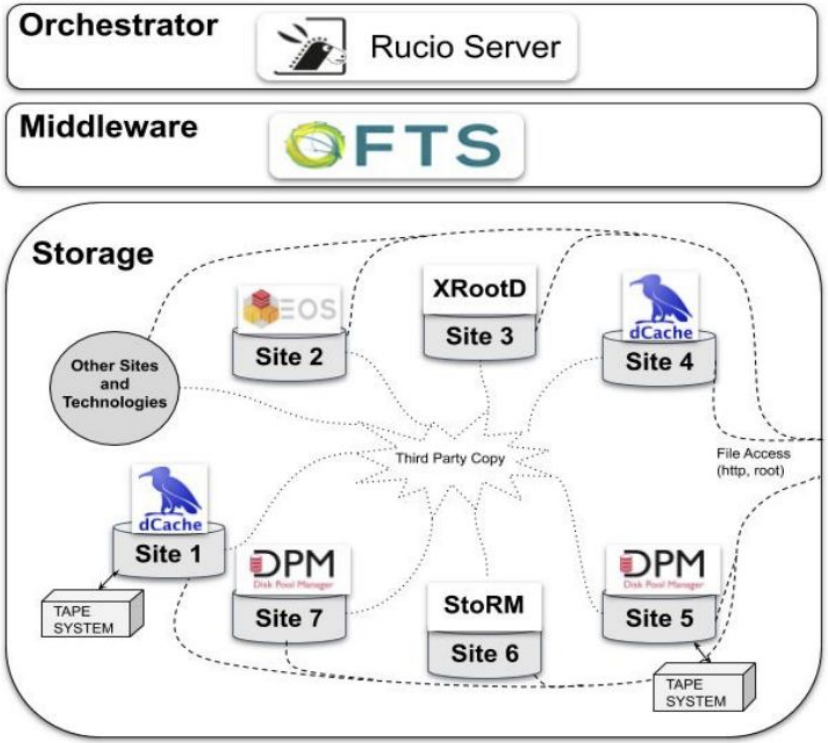
Flexible infrastructure, flexible sites, flexible resources



https://projectescape.eu/

Towards shared infrastructures: ESCAPE project

- Open Science challenge shared by 31 partners including
  - ESFRI facilities: CTA, ELT, EST, FAIR, HL-LHC, KM3NeT, SKA
  - Pan-European research infrastructures: CERN, ESO, JIV-ERIC, EGO-Virgo
- Aiming at delivering solutions to ensure integration of data, tools, services and scientific software
- Started: 1/2/2019 (end date 31/7/2022)

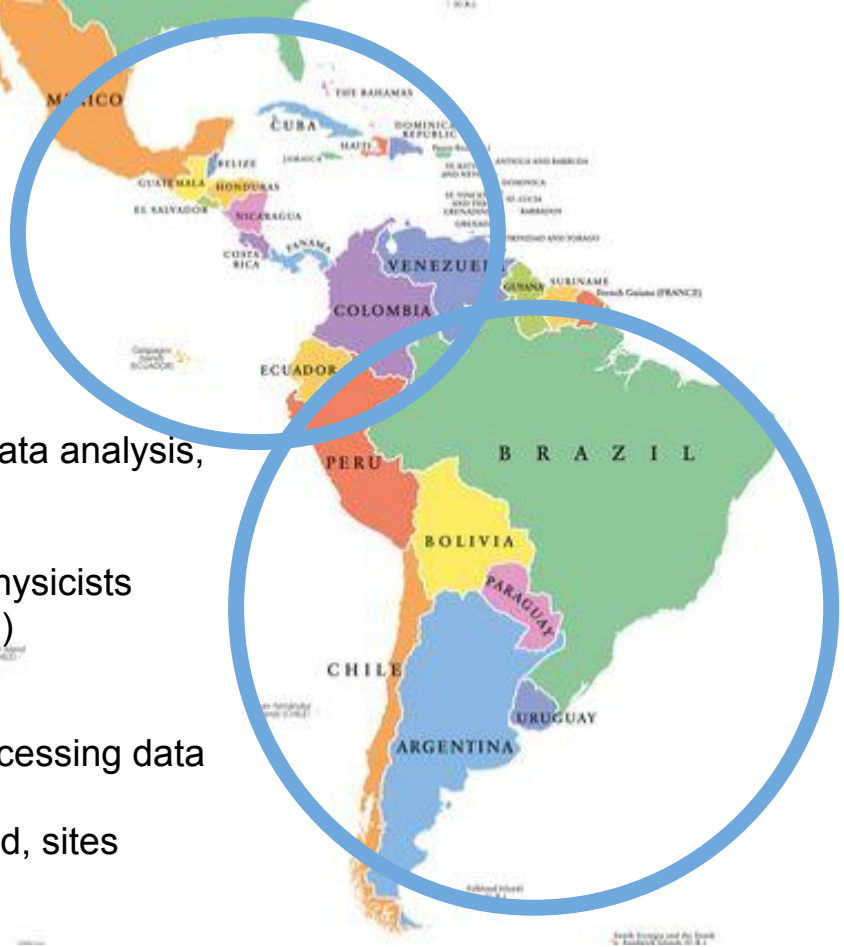


Campana@CHEP2019
https://indico.cern.ch/event/773049/contributions/3474431/

Summary (but not my last slide yet)

- Change of scale in the computing requirements for scientific experiments.
  - Not only in HEP with HL-LHC but in cosmo, astro, neutrino, gw...
- The future of Scientific Computing moving towards collaboration between sciences and its scientific communities
- Time for intense R+D in different areas. WLCCG DOMA R&D project progressing:
  - Deployments of caching initiatives, TPC transfer test machineries for http/xrootd
  - Datalakes, storage consolidation and common infrastructures (ESCAPE)
  - New AAI infrastructure being defined (token based)
  - Activities on active network traffic “shaping” and routing
  - Follow-up new ideas on computing models (compact analysis objects) and analysis models (event streaming services, columnar data files), collaboration with the HSF Analysis WG
- Expect some of the most promising features/ideas to evolve over the next year and start commission on time for HL-LHC

LA DATA LAKE ?



-Easy to start (trivial resources and manpower)
-Multi-national initiative by design
-Interactive analysis platforms with data at hand: trainings, data analysis, etc.
-Multi-science experiment data: PhDs and MSc
- Enhance job Opportunities for Computing engineers and physicists learning widely used Technologies (k8&co, data caches,etc..)

i.e.

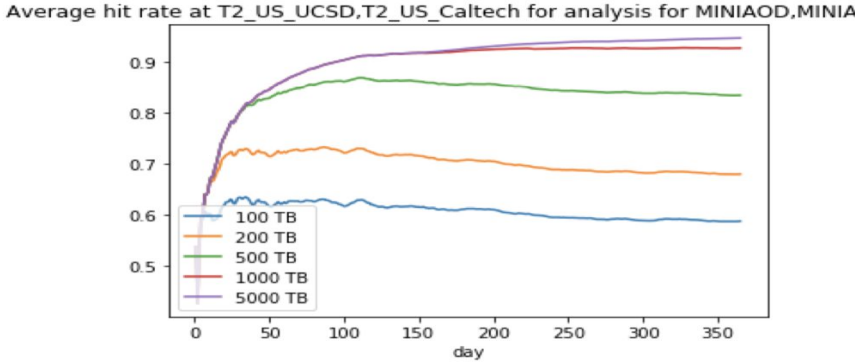
- Interactive analysis platform (=>SWAN+cvmfs) and accessing data through streaming caches?
- Federated data processing clusters? centrally managed, sites providing standar container stack: helm+k8s?
- ...

Backup Slides on Data access

Compact analysis objects

- Compact analysis data (nanos/phys-lites) spread through caches?
  - <1PB expected, keeping N (to N-x?) versions of the datasets
  - CPUs access data from local caches, minimal latency as analysis data objects are replicated in the cache
- Need *engagement* from physics community

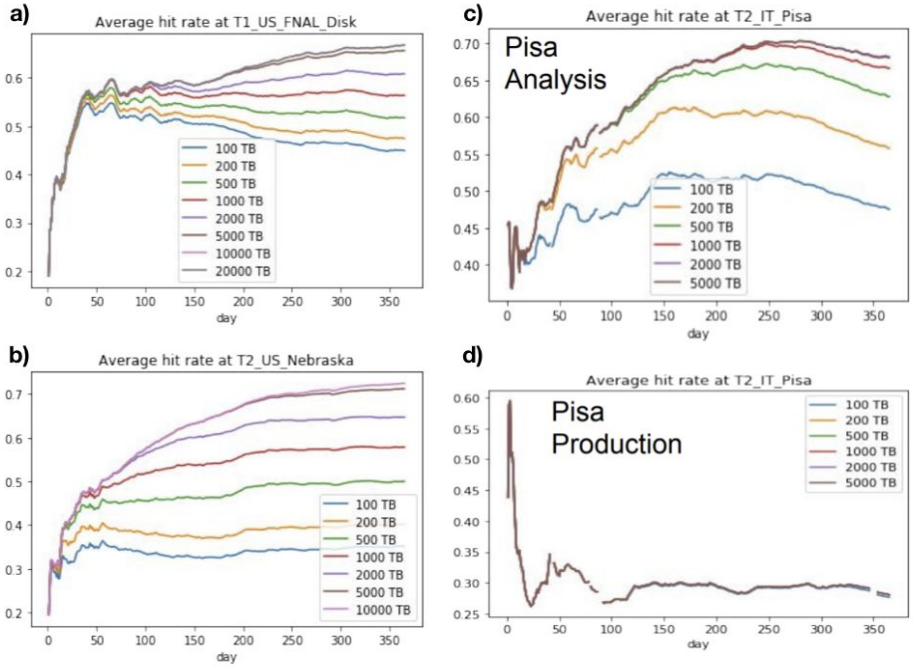
⇒7.4MB x 8e10 ~ 6e11 MB ~ 0.5 Exabytes/year of RAW
⇒2.0MB x 2.4e11 ~ 5e11 MB ~ 0.5 Exabytes/year of AOD
⇒0.2MB x 2.4e11 ~ 0.5e11 MB ~ 50 Petabytes/year of Mini
⇒0.004MB x 2.4e11 ~ 0.01e11 MB ~ 1 Petabyte/year of Nano



	MC			DATA			SUM per year
	AOD	DAOD	DAOD_LITE	AOD	DAOD	DAOD_LITE	
2026-2028							
events	6.40E+11			1.50E+11			
DAOD: 5*AOD events, 50 & 100 kB/event, 1 DAOD_LITE 10 kB/event, no replication, no extra versions							
events per year	2.13E+11	1.07E+12	2.13E+11	5.00E+10	2.50E+11	5.00E+10	
size/event [kB]	1000	100	10	700	50	10	
disk [PB/year]	213.33	106.67	2.13	35.00	12.50	0.50	369.63
	MC			**DATA**			**SUM (25-28)**
	AOD	DAOD	DAOD_LITE	AOD	DAOD	DAOD_LITE	
DAOD: 5*AOD events, 50 & 100 kB/event, 1 DAOD_LITE 10 kB/event, no replication, no extra versions							
events (25-28)	6.40E+11	3.20E+12	6.40E+11	1.50E+11	7.50E+11	1.50E+11	
size/event [kB]	1000	100	10	700	50	10	
disk 25-28 [PB]	640.00	320.00	6.40	105.00	37.50	1.50	1,108.90

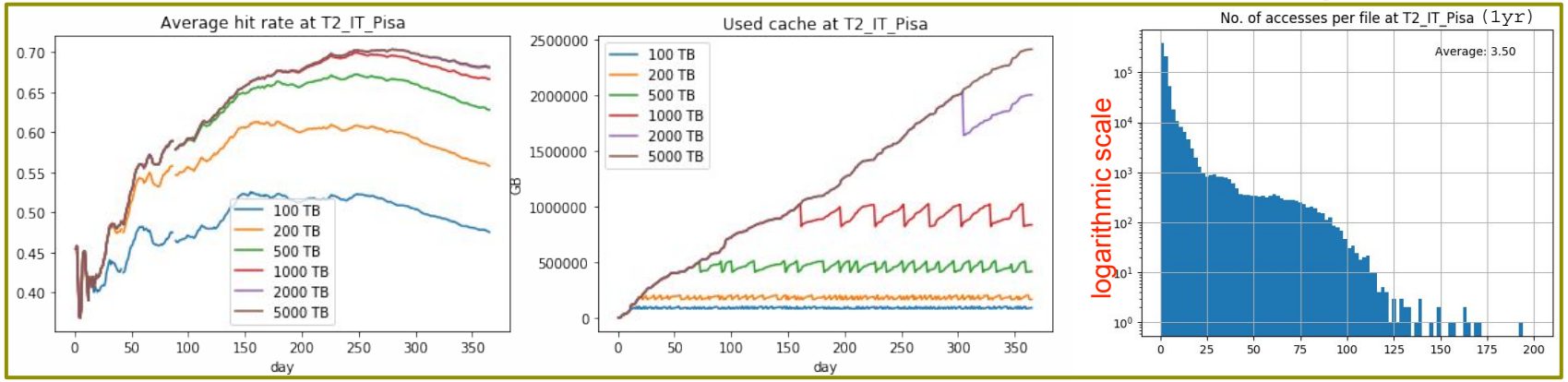
Data access patterns

- Data access frequency after data placement. The two extremes are:
  - **Cold data**: files are WORN (Write Once Read Never)
  - **Hot data**: where files are expected to be accessed continuously
- A large fraction of our files are **neither cold nor hot**.
  - AOD/dAOD files seems to lose popularity with time and the access rate decreases significantly after days/weeks.
- Could this be better handled with caches?
  - Available when popular and replaced when less demanded
  - Less frequently used data might be re-fetch again from the lakes (disk or tape)
    - experiments handling QoS labels to set the best cost/usage ratio

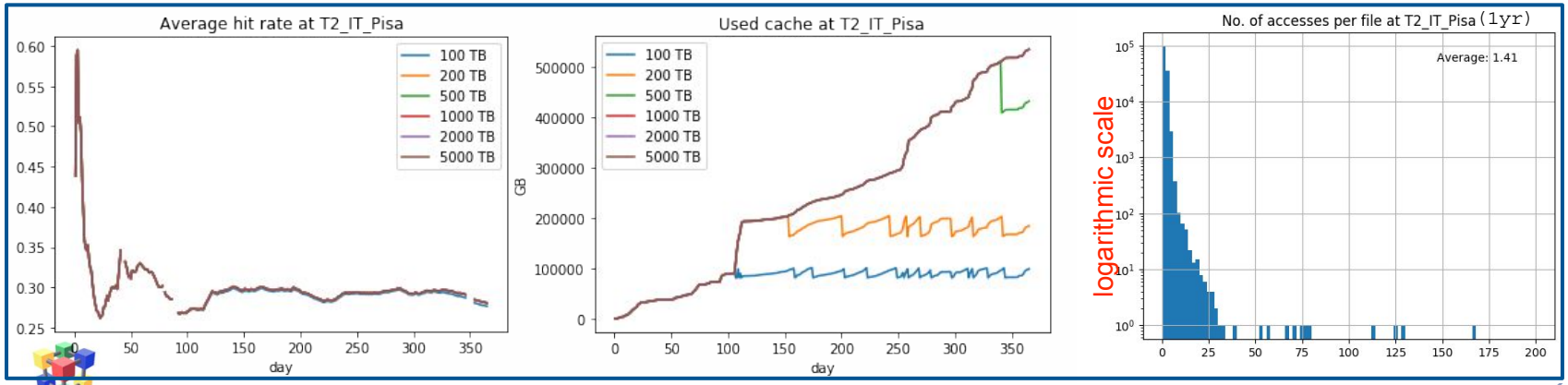


Read-ahead Caches: storage

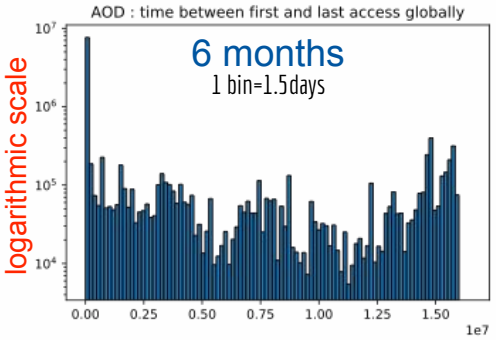
analysis

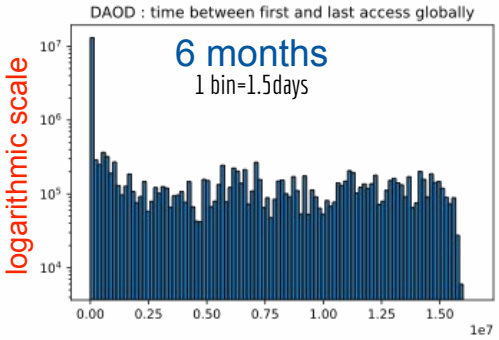


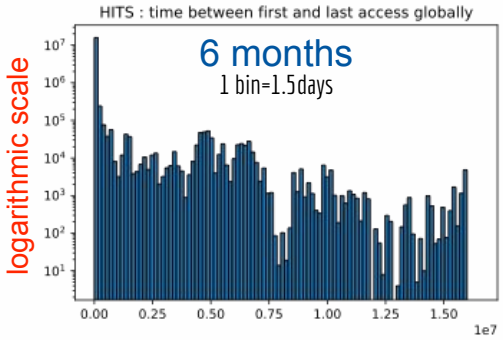
production



Read-ahead Caches: storage



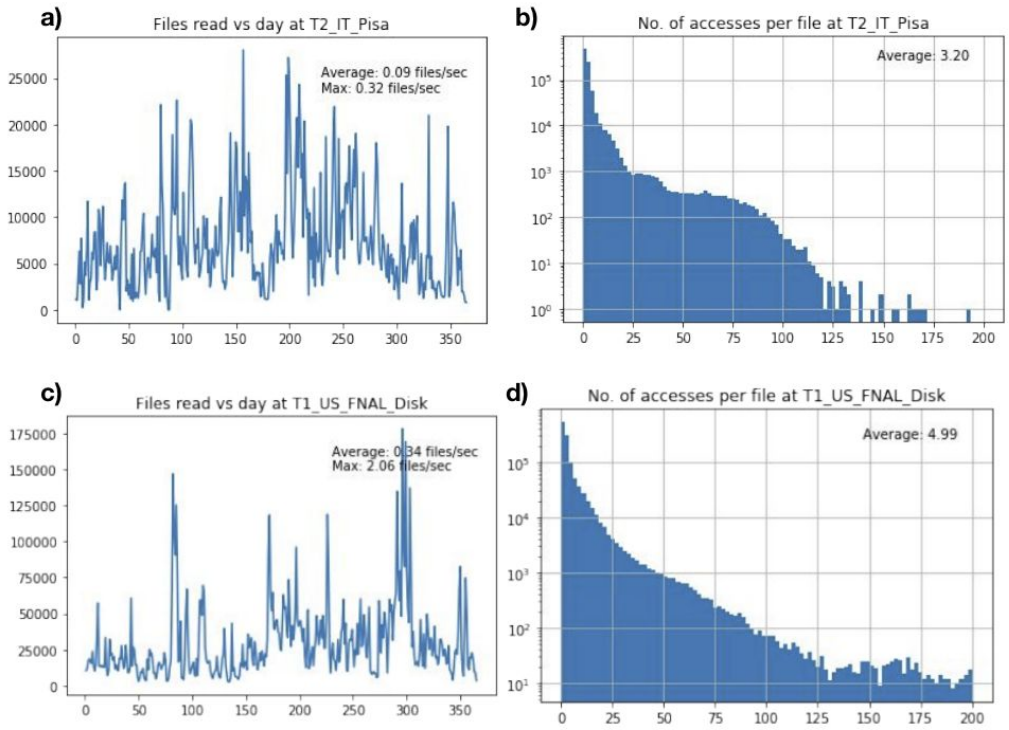




- What could be learn from this?
  - Data isn't accessed very often, most likely to be re-read within days. Then data gets cold.
- This plots provide us with hints. We still miss more substantial analysis to draw conclusions:
  - 6 months cache replay is not enough, need to enlarge the data set
  - Need to add staging and deletion information
  - Looking for access rate versus absolute time
  - Correlate with seasons: conference rushes, holidays, etc.

Analysis vs. Production: changing strategy? (1/2)

- Analysis has **noticeable re-use**
- **Production files have few re-reads**
  - But production files push analysis data out of the caches
- Data isn't accessed very often
  - Most likely to be re-read within days after placement



@A. Sciaba and cost modeling WG

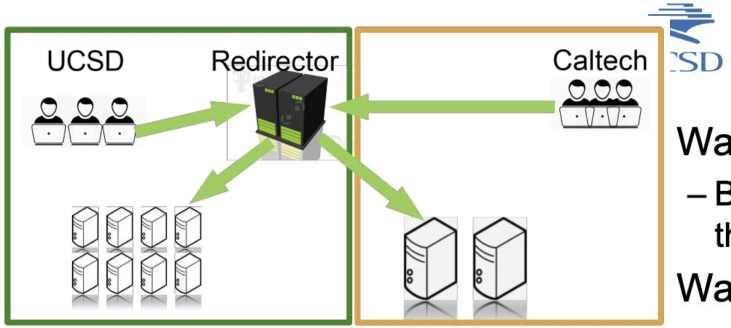
Preliminary studies based on a period of 6 months

Analysis vs. Production: changing strategy? (2/2)

- User analysis is a big part of the overall grid data processing workload and the most demanding and unpredictable
- Nowadays both processing workloads load are spread all over the infrastructure
- Would a change of strategy in workload management be beneficial?
  - Non-predictable workflows (=analysis) to run at the T1s (and AFs)?
    - Data is close in network terms (datalake)
    - Random access not a killer as data will be (mostly) accessed on the datalake disk storage
  - Predictable workflows (=production) to be run at T2s/T3s/opportunistic?
    - Experiments know how to match data pre-placement and tasks submission, this can be done on a cache-like site or on a full fledge storage site

Data Caching Backup slides

SoCal: XRootD Cache



T2 wish list (I)

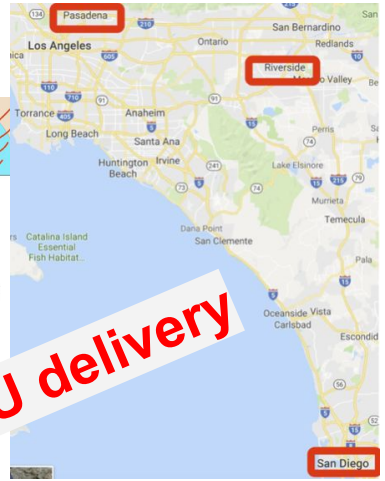
Want CMS to switch to Buffer & Cache mode.

– Buffer that assumes nothing in buffer needs to stay there for longer than a week, to keep buffer small.

Want to operate only JBODs

- Want all CPU that is close to [illegible] of CMS application laten[illegible]ce to benefit from our disks.
  – Why replicate [illegible]ta can be accessed via the network?

T2s focussing on CPU delivery

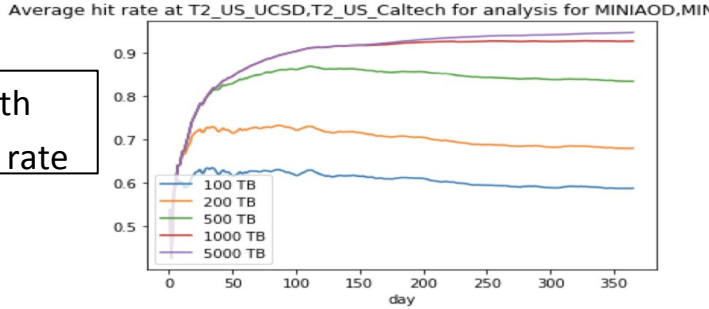


MINIAOD and MINIAODSIM in XCache

	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

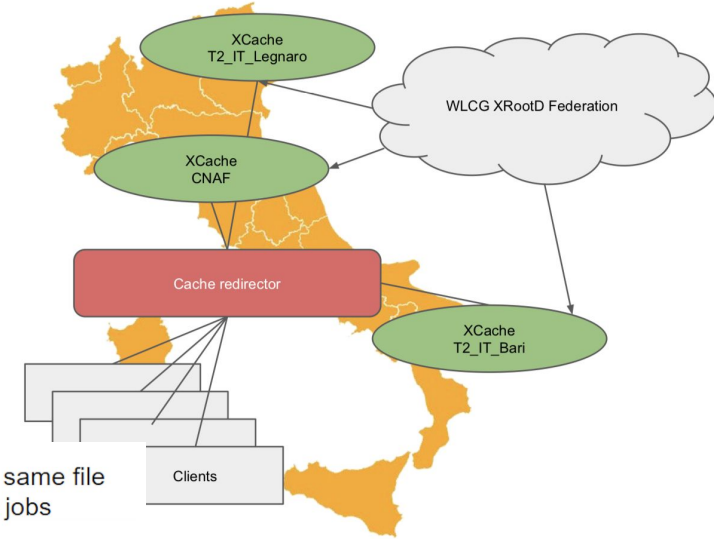
Datasets	Size (TB)
/*/Run2016*-03Feb2017*/MINIAOD	182.8
/*/RunIISummer16MiniAODv2-PUMoriond17_80X_*/MINIAODSIM	502.5
/*/*RunIIFall17MiniAODv2*/MINIAODSIM	211
/*/*-31Mar2018*/MINIAOD	137.9
Total	1041

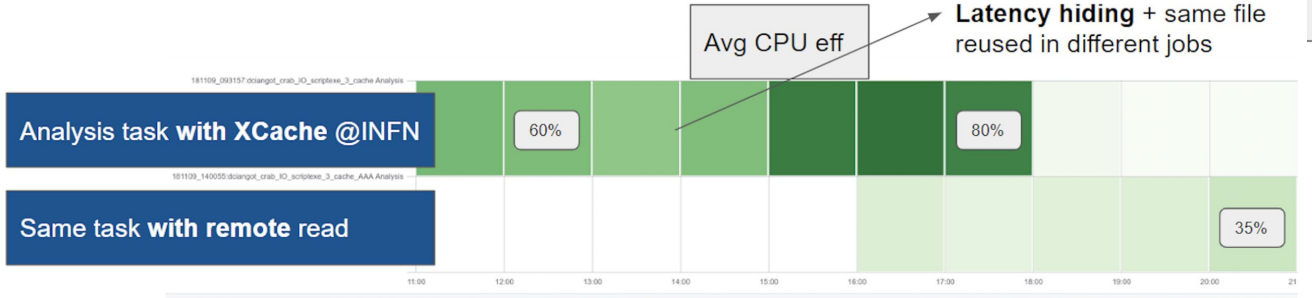
A 1PB cache in SoCal filled with Mini/Nano AODs has 90% hit rate



Italy: Caching layer prototype for CMS

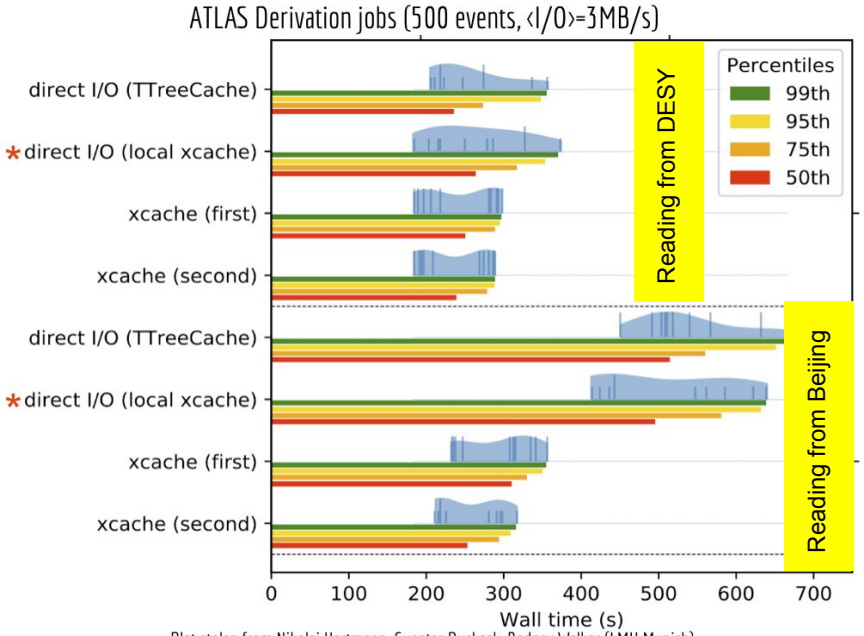
- A distributed XCache model for a national datalake
  - INFN's Legnaro, CNAF and Bari
- Enabling storage-less T2s concept
- Ability to scale out over opportunistic resources
- Improve CPU time for remote I/O





Munich: XCache performance studies

- Analysis jobs performance at LRZ-LMU
- XCache hides latency as well as ROOT TTC with Asynchronous Prefetch
- Modest hardware used (single 2012 disk pool):
  - Successfully served 3.2k
  - Jobs running analysis and ATLAS derivations
    - average I/O 1MB/s and 3MB/s respectively



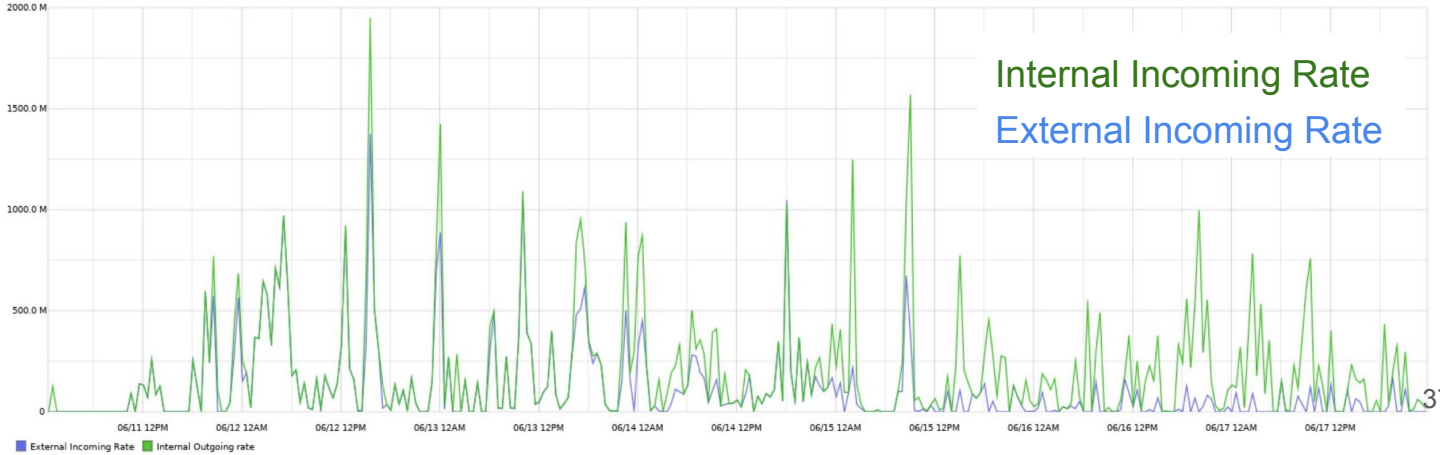
Plot stolen from Nikolai Hartmann, Guenter Duckeck, Rodney Walker (LMU Munich)

* Read-ahead was disable on XCache by mistake

Birmingham: caches in production

- Birmingham has no pledged storage any longer
  - Data source for the BHAM Worker Nodes is Manchester
- **Simple direct read** was **overloading** Manchester SE
  - Deployed xCache in Birmingham
- Conclusion:
  - Caching works as expected
  - Files reused ~3 times
  - Significant saving in network traffic

Network traffic to an XCache Pool node



XCache R&D activities

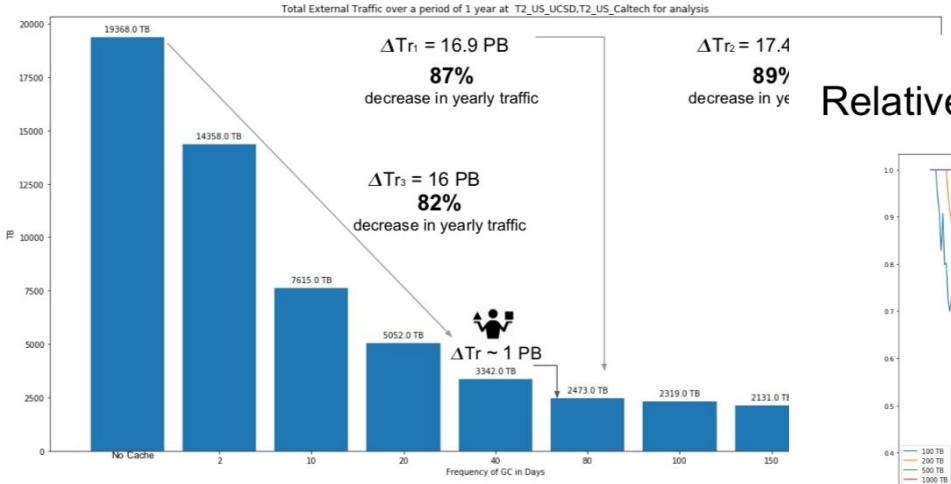
- DOMA ACCESS pre-GDB on Caching (July 2019, https://indico.cern.ch/event/827556)
  - Opportunity to confront directly users and developers
  - Status report on development based in US (W. Yang) an developments at CERN (D. Smith)
  - Feedback from site admins on deployment and operation
    - Security requirements on SLATE (follow-up on a dedicated working group)
- World-wide XCache working group
  - Common place to follow-up the different XCache initiatives
  - Common place to follow-up on developments
  - Common place to discuss further requirements and prioritization, ie.
    - Optimization for **HPCs** shared file systems
    - Xrootd over Remote Direct Memory Access (**RDMA**) to deliver data blocks for partially cached files
    - XCache data ingest from **HTTP** data source

Caching and network implications

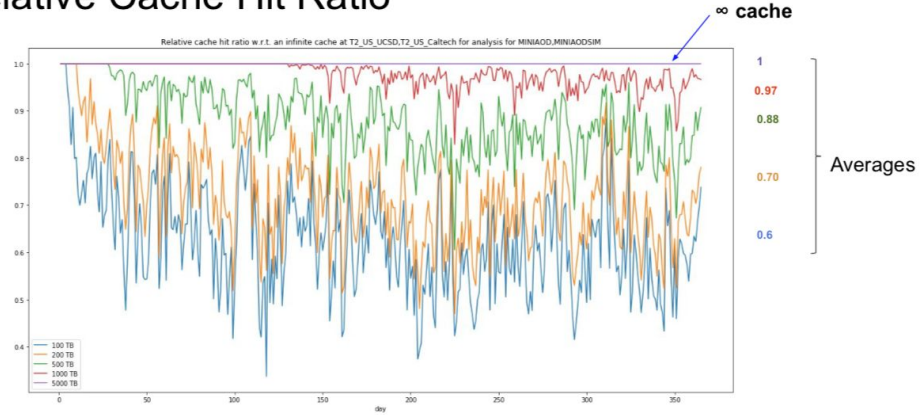
- Cache purging strategy (N> 80 days) equivalent to infinite cache with no deletion
  Large proportion of files never accessed after 80 days

Yearly External Traffic at SoCal

ΔTr = Difference in External Traffic



Relative Cache Hit Ratio



#Note: High water mark algorithm was used here.

Shreya@DOMA ACCESS
https://indico.cern.ch/event/834197/

DOMA QoS: Sites' diversity

- Concerns on storage-less orientation
  - Funding
  - Manpower
  - Local storage needs
- Caches <u>are not</u> a replacement of storage
  - Can be useful as a solution for sites wanting to be CPU oriented
- Caches and storage can co-exist and
  - Help to increase efficiency of the storage size (avoid keeping unused data on disk)
  - Help to increase cpu efficiency accessing remote data
- This is the main focus in the WLCG DOMA ACCESS WG studies

Q4 – Effort

- Inconclusive
  - No clear mapping of "efficiency" onto system
- Difficult to quantify effort
  - But then how will we know if we save?
- T1s report ~2.5 FTE on storage
- T2s report ~0.6 FTE on storage

Q5 - "Storageless sites"

- The vast majority of T2s (who responded) are neither planning nor wanting to move to storageless setups
  - Local storage is needed
  - This can cut off independent lines of funding
  - Mixed/diverging opinions regarding caching solutions
    - we are interested / we are not interested / depends on which kind of cache you're talking about
- T1s indicate uncertainty about the impact of storageless sites on their systems

Site storage survey - results, conclusions, actions
https://indico.cern.ch/event/739883/ (Oliver Keeble)