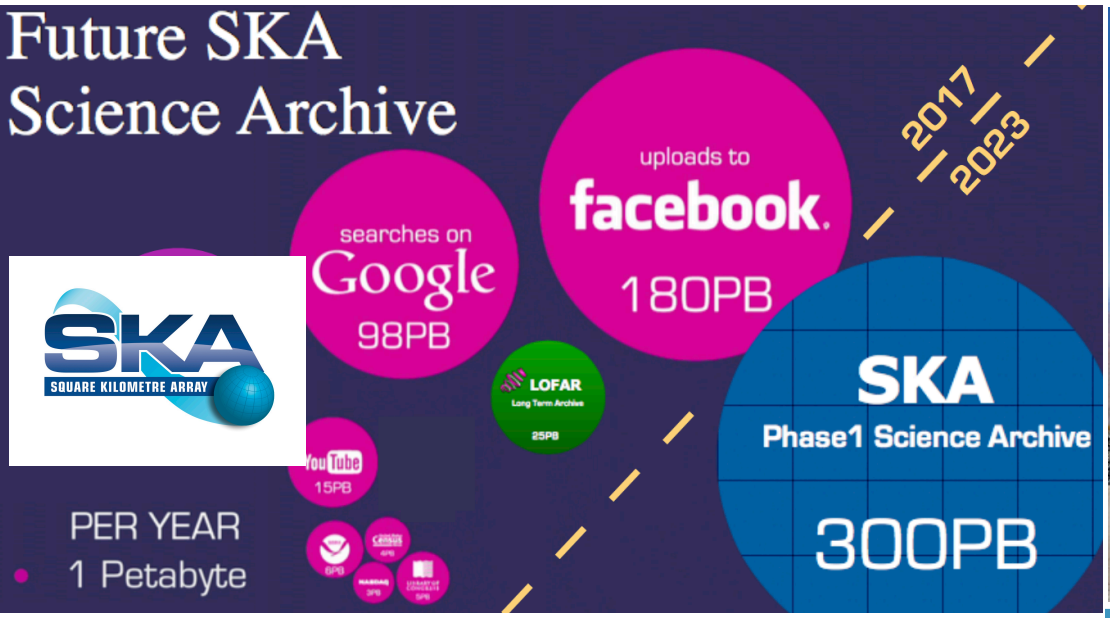

CERN

A Data Lake for WLCG

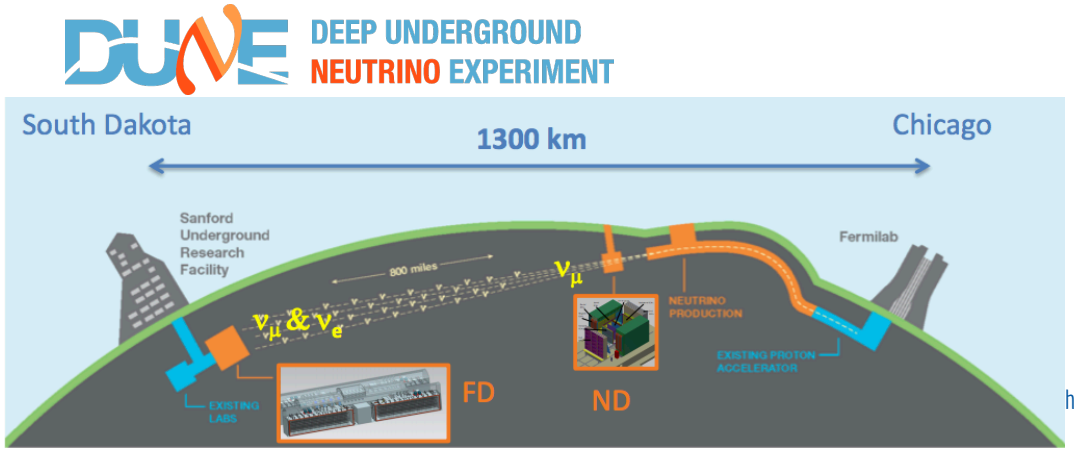
Ian Bird, Simone Campana, Xavier Espinal, Maria Girone, Gavin McCance, Jaroslava Schovancova
(CERN)

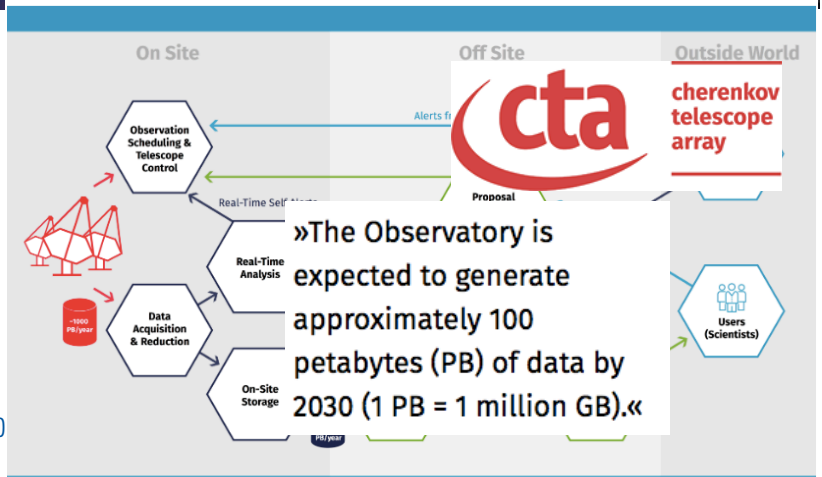
The motivation (1/2)

- Future change of scale in data volumes common to all scientific communities
  - HL-LHC, SKA, DUNE, LSST, CTA, FAIR, BELLE-II, JUNO,…


Future SKA
Science Archive
2017
2023
uploads to
facebook
180PB
searches on
Google
98PB
SKA
SQUARE KILOMETRE ARRAY
LOFAR
Long Term Archive
25PB
YouTube
15PB
SKA
Phase1 Science Archive
300PB
PER YEAR
1 Petabyte

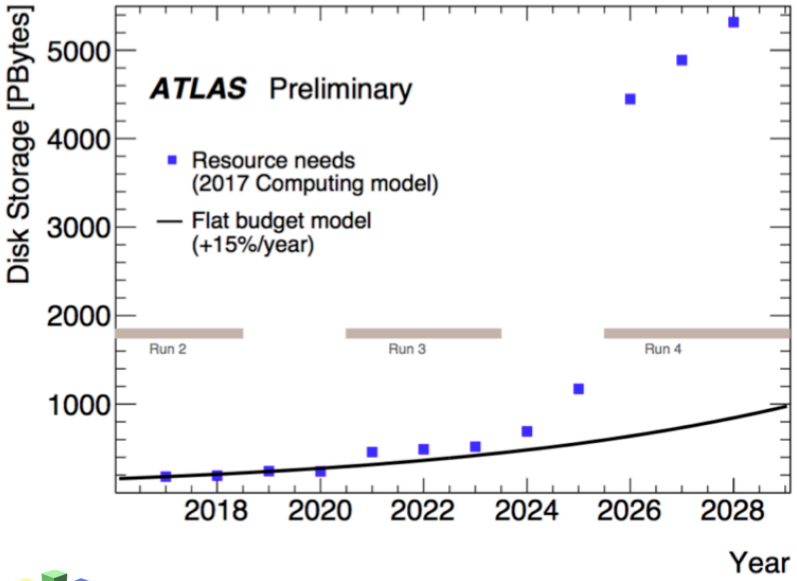

LSST
20PB/night

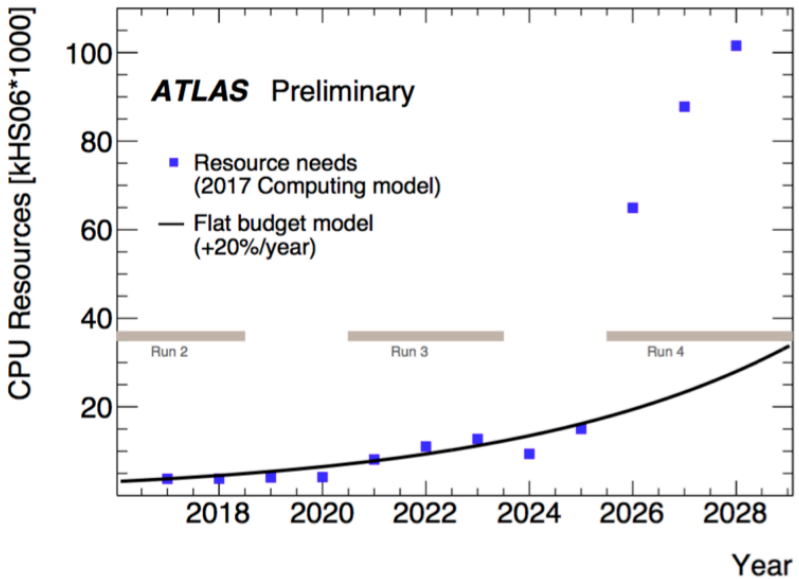

DUNE
DEEP UNDERGROUND
NEUTRINO EXPERIMENT
South Dakota
1300 km
Chicago
Sanford
Underground
Research
Facility
800 miles
Fermilab
NEUTRINO
PRODUCTION
EXISTING PROTON
ACCELERATOR
EXISTING
LABS
FD
ND


On Site
Off Site
Outside World
Observation
Scheduling &
Telescope
Control
Real-Time
Analysis
Data
Acquisition
& Reduction
On-Site
Storage
Users
(Scientists)
cta
cherenkov
telescope
array
»The Observatory is
expected to generate
approximately 100
petabytes (PB) of data by
2030 (1 PB = 1 million GB).«

The motivation (2/2)

- For HL-LHC the future storage needs are above the expected technology evolution (15%/yr) and funding (flat)





INTERNATIONAL JOURNAL OF HIGH-ENERGY PHYSICS

CERN COURIER

CERN COURIER APRIL 2018

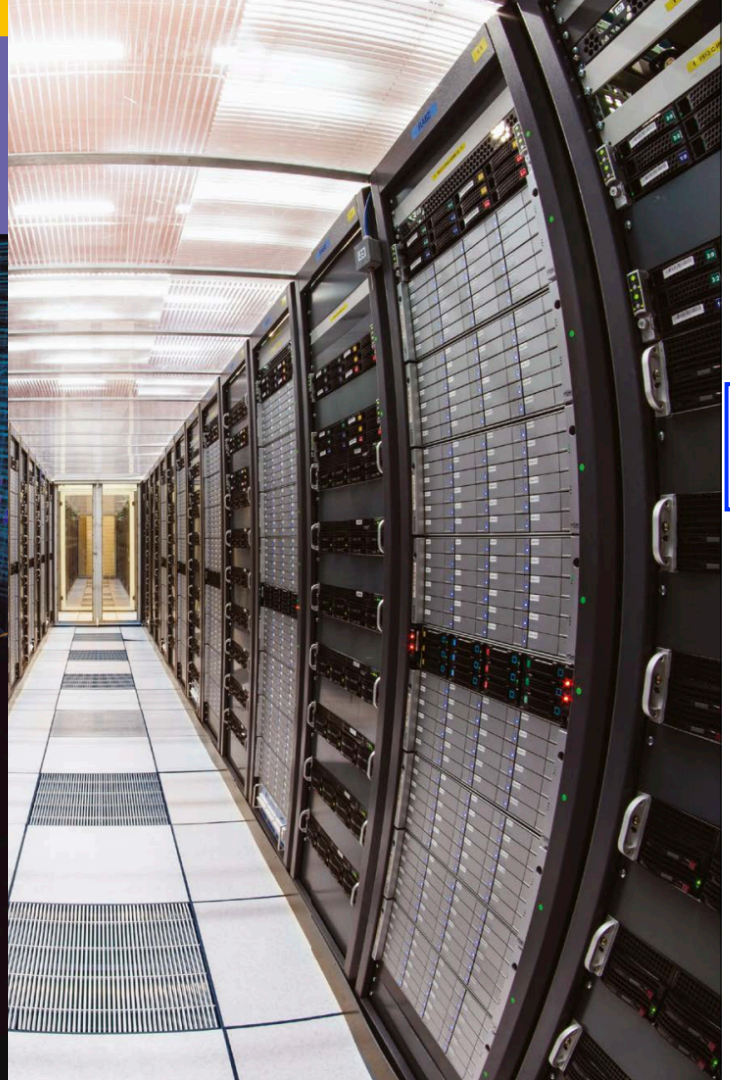
VOLUME 58 NUMBER 3 APRIL 2018

Tracking beauty decays

X-band technology spreads

Computing's radical future

Speaking up for the Higgs



Software and computing

Time to adapt for big data

Radical changes in computing and software are required to ensure the success of the LHC and other high-energy physics experiments into the 2020s, argues a new report.

It would be impossible for anyone to conceive of carrying out a particle-physics experiment today without the use of computers and software. Since the 1960s, high-energy physicists have pioneered the use of computers for data acquisition, simulation and analysis. This hasn't just accelerated progress in the field, but driven computing technology generally – from the development of the World Wide Web at CERN to the massive distributed resources of the Worldwide LHC Computing Grid (WLCG) that supports the LHC experiments. For many years these developments and the increasing complexity of data analysis rode a wave of hardware improvements that saw computers get faster every year. However, those blissful days of relying on Moore's law are now well behind us (see panel overleaf), and this has major ramifications for our field.

The high-luminosity upgrade of the LHC (HL-LHC), due to enter operation in the mid-2020s, will push the frontiers of accelerator and detector technology, bringing enormous challenges to software and computing (*CERN Courier* October 2017 p5). The scale of the HL-LHC data challenge is staggering: the machine will collect almost 25 times more data than the LHC has produced up to now, and the total LHC dataset (which already stands at almost 1 exabyte) will grow many times larger. If the LHC's ATLAS and CMS experiments project their current computing models to Run 4 of the LHC in 2026, the CPU and disk space required will jump by between a factor of 20 to 40 (figures 1 and 2).

Even with optimistic projections of technological improvements there would be a huge shortfall in computing resources. The WLCG hardware budget is already around 100 million Swiss francs per year and, given the changing nature of computing hardware and slowing technological gains, it is out of the question to simply throw

Inside the CERN computer centre in 2017. (Image credit: J Ordan/CERN.)

Evolution of federated storage (1/3)

- Data redundancy re-evaluation
  - Nowadays local storage redundancy is on top of global redundancy
    - An experiment wants two distributed copies but in reality we are storing more:
      - The copy on site A runs a system storing two distributed replicas
      - The copy on site B runs a system configured with local RAID (i.e 4+2)
  - Global redundancy to by-pass local redundancy?
    - Two replicas means two files: one in site A and one in site B (with no extra redundancy)
  - or further question redundancy?
    - Do we need redundancy for the entire life of a file? probably *not.* File level redundancy could evolve with time: file workflows based on QoS
  - or droping redundancy?
    - Can we *recompute* rather than *replicate*? Could this be cost effective and performed within reasonable timing?

Evolution of federated storage (2/3)

- File workflows: *expensive* to *cheap (and back)*
  - Data popularity decreases with time. *Value* drops. Can we follow *datasate market value* with storage media?
    - One can imagine workflows like this:
      1. Dataset first stored as double replica on disk
      2. Transition to RAIN layout after some weeks
      3. End up with a single replica on t*ape* or *tape-equivalent* media
  - This could be leveraged by the storage system at namespace level or by the experiment data management systems as a data(set) metadata

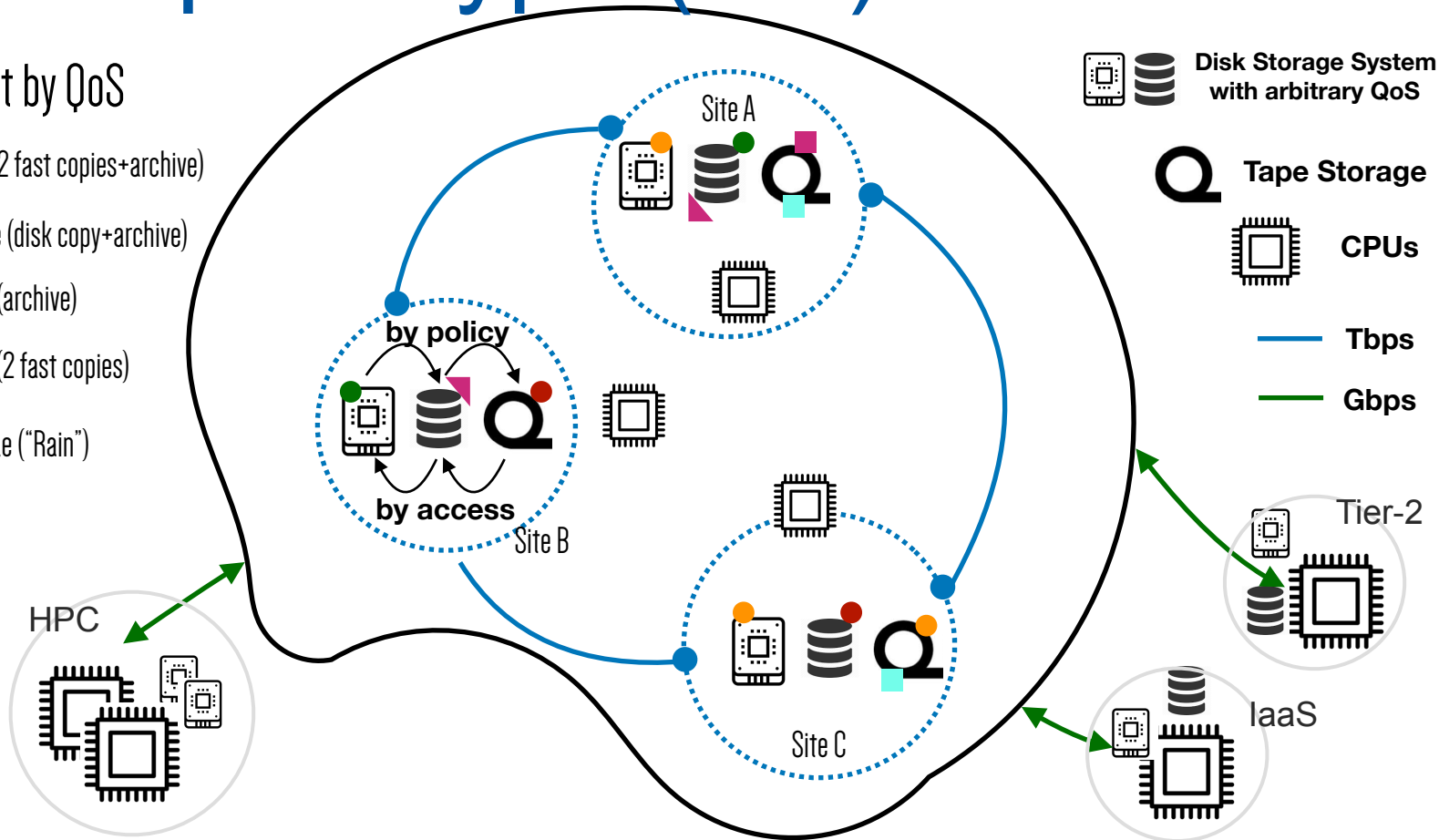
Evolution of federated storage (3/3)

- Co-existence of different QoS (storage media *cost*)
  - Does it makes sense to continue referring to *disk* and *tape* when we want to refer to *qualities* of the underlying storage services
    - Consumer disks vs. Enterprise disks vs. Tape vs. SSDs vs. RAIN
  - Shouldn't we give the flexibility to the sites? and then experiments choose what they need for their files in terms of:
    - Expected reliability (custodial data vs. transient files)
    - Expected access patterns (latency, IOPS)
    - Expected bandwidth
    - Expected cost

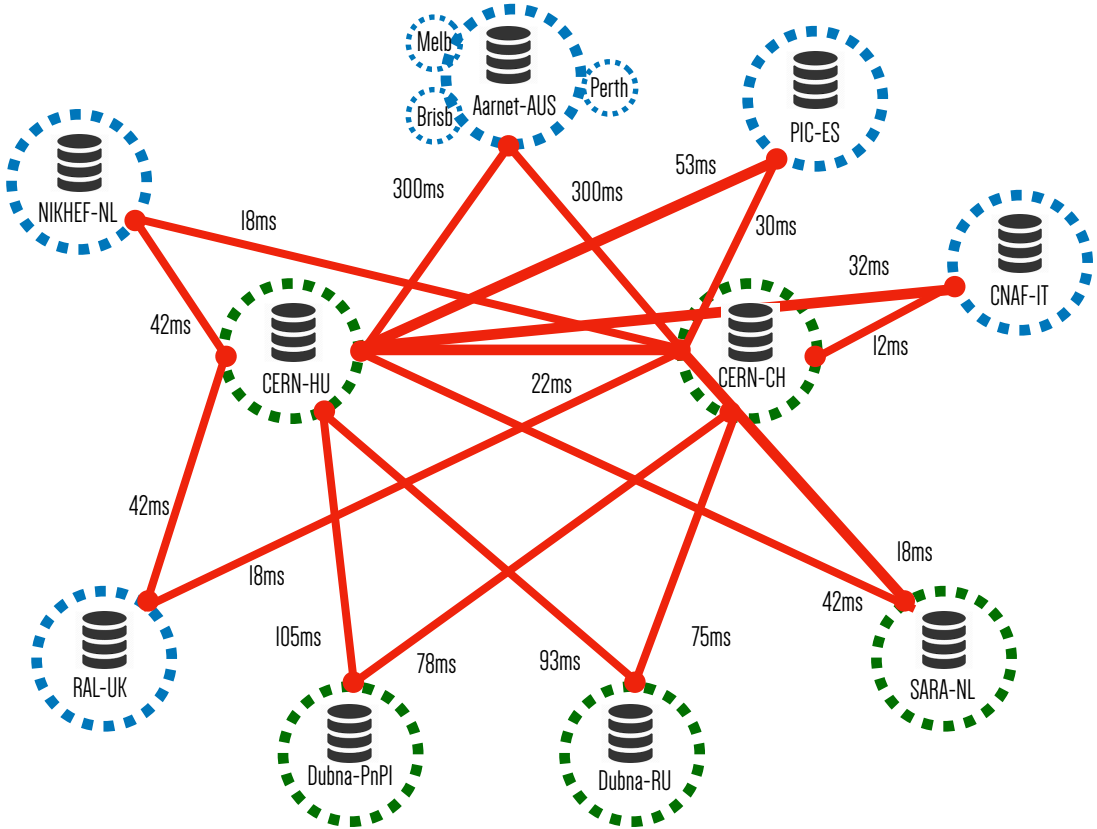
eulake prototype (1/2)

File placement by QoS

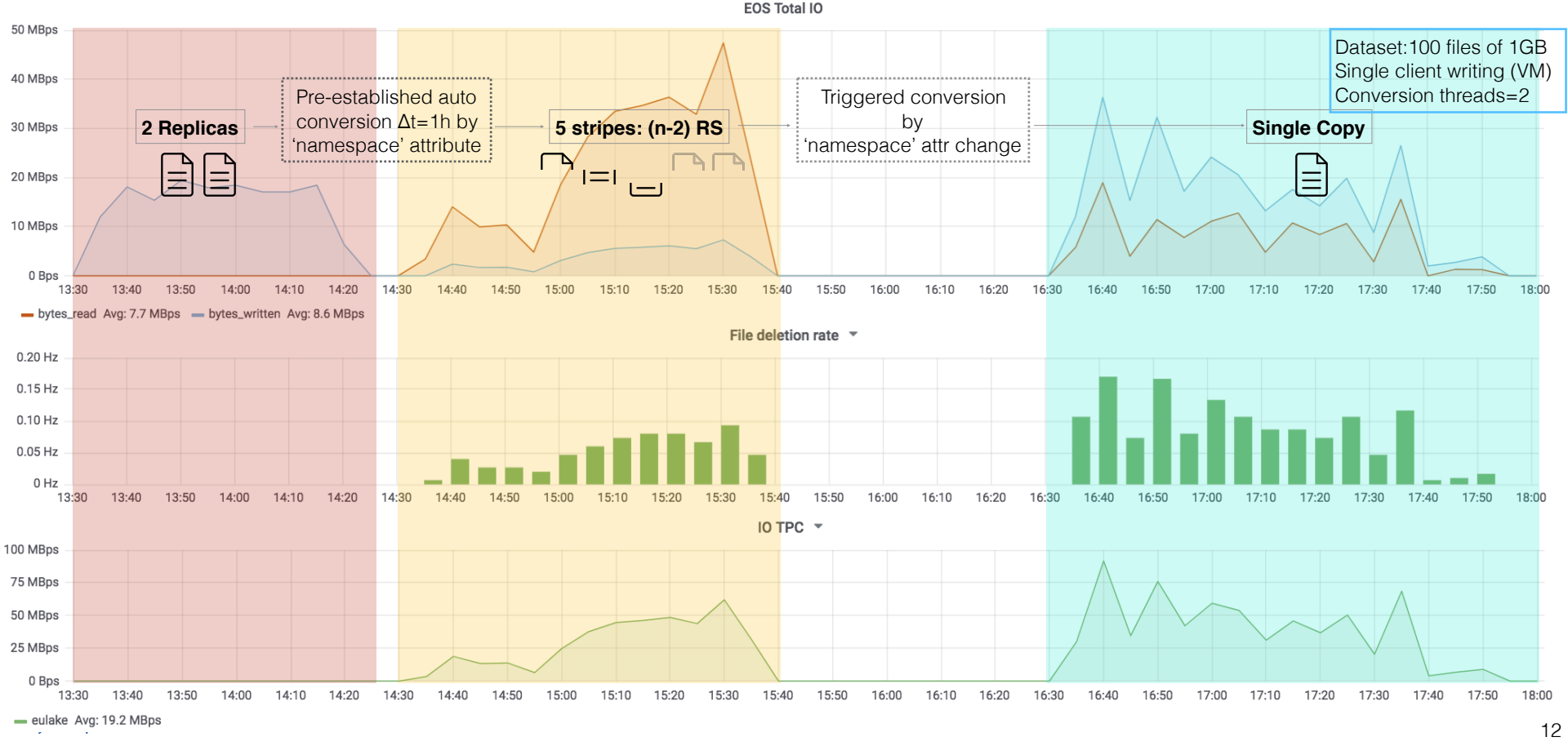
- Hot custodial file (2 fast copies+archive)
- Warm custodial file (disk copy+archive)
- Cold custodial file (archive)
- Hot ephemeral file (2 fast copies)
- Warm ephemeral file ("Rain")



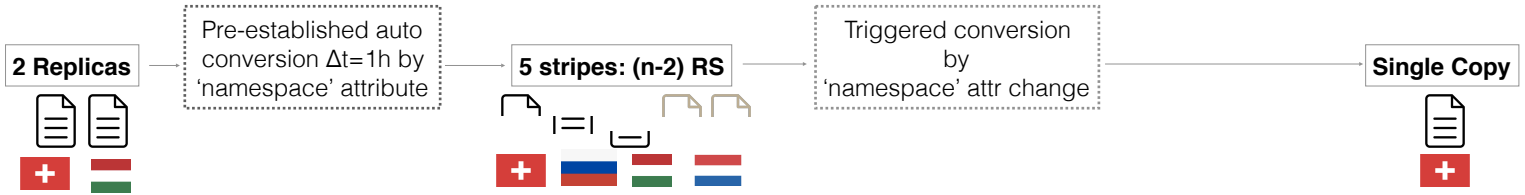
eulake prototype (2/2)



Test: distributed redundancy, file workflows and QoS (1/2)



Test: distributed redundancy, file workflows and QoS (1/2)



180315 14:04:36 func=open path=/eulake/lcg/test/conversion/2replicas-to-rain32/file-workflow-2r-rain32.175.file
op=write target[0]=(**p05799459m56401.cern.ch**,33) target[1]=(**p05798818t49625.cern.ch**,80)

180315 15:04:58 time=1521123718.328306 func=open path=/eulake/lcg/test/conversion/2replicas-to-rain32/file-workflow-2r-rain32.175.file
op=read target[0]=(p05799459m56401.cern.ch,33) target[1]=(p05798818t49625.cern.ch,80)

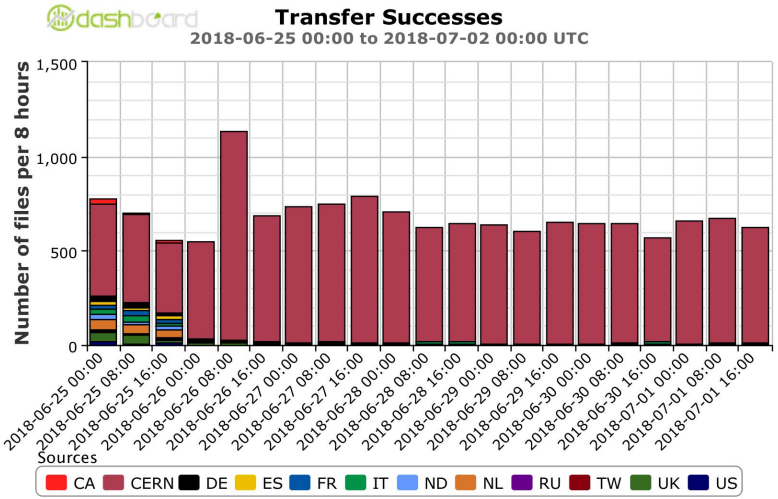
180315 15:04:58 func=open path=/eos/eulake/proc/conversion/0000000000001819:default#20640442
op=write eos.layout.nstripes=5&eos.layout.type=raid6
target[0]=(**fst2.grid.surfsara.nl**,130) target[1]=(**p05496644k62259.cern.ch**,1) target[2]=(**dvl-mb01.jinr.ru**,122) target[3]=(**p05798818t49625.cern.ch**,97)
target[4]=(**fst1.grid.surfsara.nl**,124)

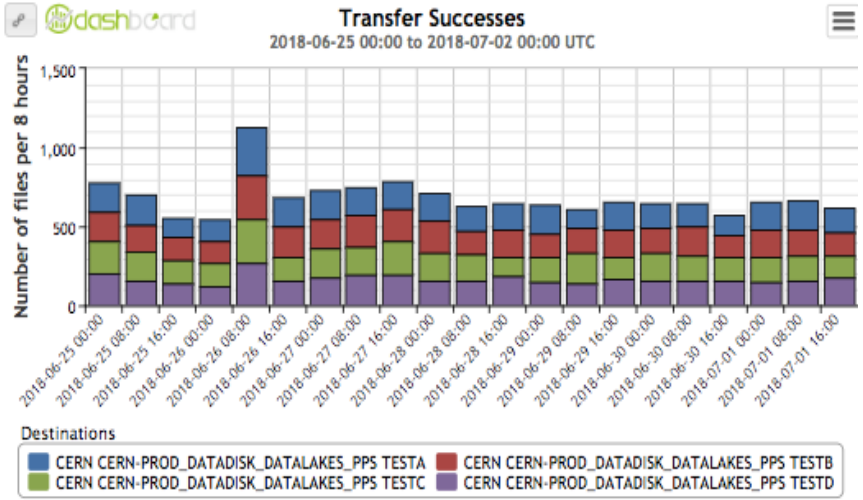
180315 17:22:17 func=open path=/eulake/lcg/test/conversion/2replicas-to-rain32/file-workflow-2r-rain32.175.file
op=read target[0]=(fst2.grid.surfsara.nl,130) target[1]=(p05496644k62259.cern.ch,1) target[2]=(dvl-mb01.jinr.ru,122)
target[3]=(p05798818t49625.cern.ch,97)

180315 17:22:17 func=open path=/eos/eulake/proc/conversion/00000000000018e2:default#00100001
op=write eos.layout.nstripes=1&eos.layout.type=plain tpc.stage=copy redirection=**p05799459m56401.cern.ch**?

Integration with ATLAS and CMS Data Management

- We exposed eulake to the ATLAS and CMS data management system as storage endpoint
- Data can be transferred from any site into eulake (see ATLAS below)
- We imported 4 input samples in different eulake areas for the next tests



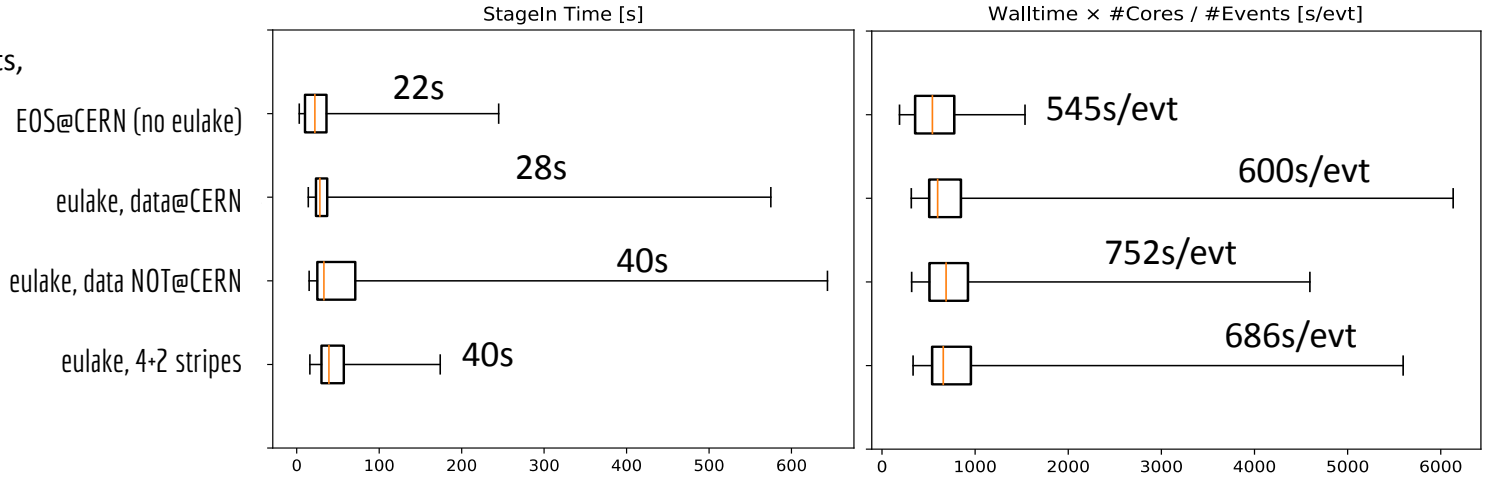


Integration with the Hammercloud framework

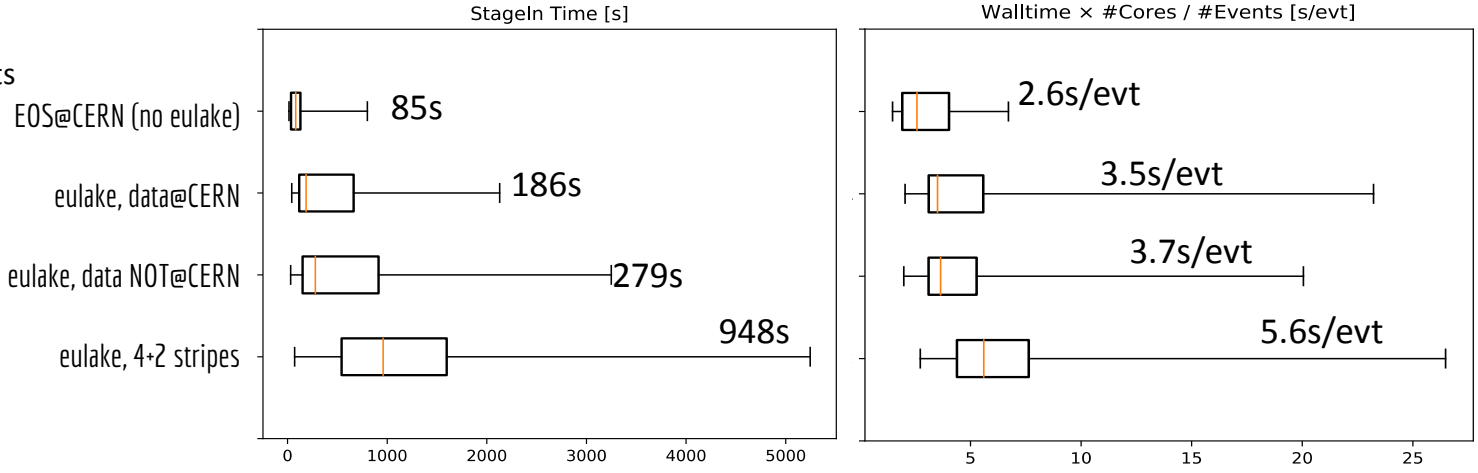
- Allows test real workflows and data access patterns
- Four test scenarios where data is copied to the WN from:
  - Current production EOS instance (no eulake)
  - eulake, data@CERN
  - eulake, data NOT@CERN
  - eulake, 4+2 stripes

Low I/O intensity workflow
(simulation)
~40MB input (1 file), 2 events,
~5 mins/event

Jun 2018

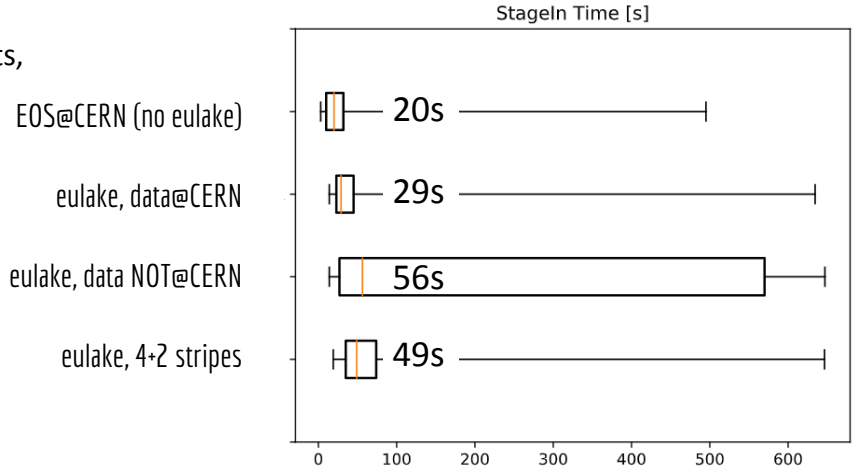


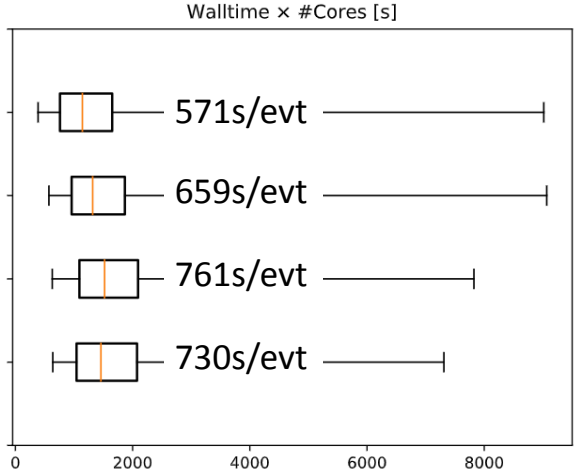
High I/O intensity workflow
(DigiReco)
~6GB input (1 file), 1000 events
~2 seconds/event



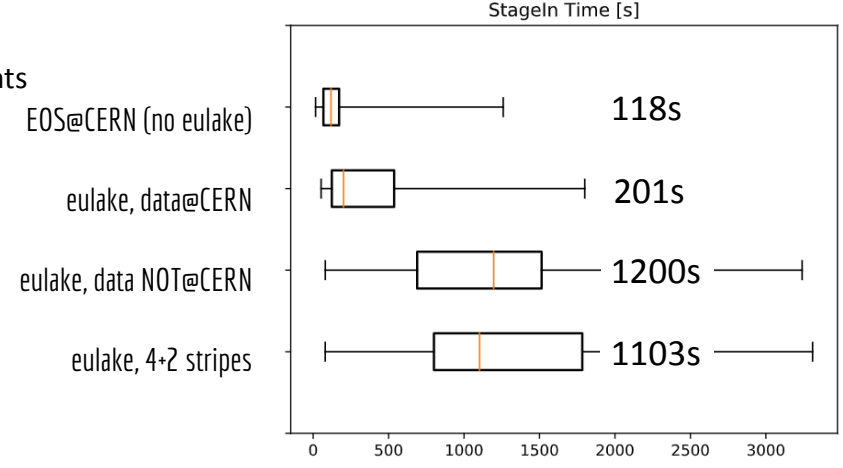
Low I/O intensity workflow
(simulation)
~40MB input (1 file), 2 events,
~5 mins/event

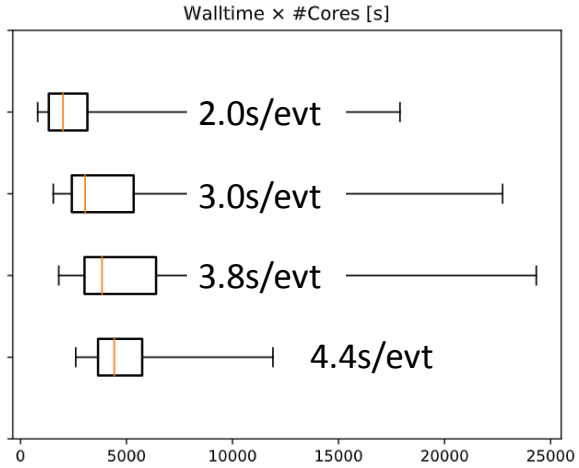
Sept 2018





High I/O intensity workflow
(DigiReco)
~6GB input (1 file), 1000 events
~2 seconds/event





Conclusions (1/2): eulake

- We set up a federated storage prototype to implement some of the concepts to address cost optimization
- This protoype is at the level of proof of concept
  - Very modest in space and available bandwidth but with enough geographic participation to validate some of the main ideas behind
  - Measurements and results to be taken only as a feasibility proof
- We integrated eulake instance with the ATLAS and CMS distributed computing services and HammerCloud
- Next steps towards having a minimal amount of resources to start evaluating performance

Conclusions (2/2): storage evolution and DOMA

- Presented concepts cater for the main storage layer: a data lake or evolved federated storage which mainly impact storage oriented sites
- A parallel evolution is to adapt computing oriented sites to this new topology to improve costs and performance.
- Specific DOMA Working Groups are starting to coordinate and address these challenges:
  - Latency hidding and application data caching is a big leap towards improving performance and costs: DOMA-ACCESS WG
  - Revise and improve the data distribution/aggregation protocols is instrumental: TPC over http, questioning TCP/UDP, DTNs: DOMA-TPC WG
  - Homogeneous concept for Quality of Service across storage services and storage interoperability: DOMA-QoS WG