Track 4 summary

Data handling

Overview

A lot of interesting material and discussions
- 46 oral presentations in 7 sessions
- 34 posters

Conveners & session chairs:
- Elizabeth Gallas (Oxford)
- Costin Grigoraş (CERN)
- Tigran Mkrtchyan (Desy)
- Maria Arsuaga Rios (CERN)

Very nice venue

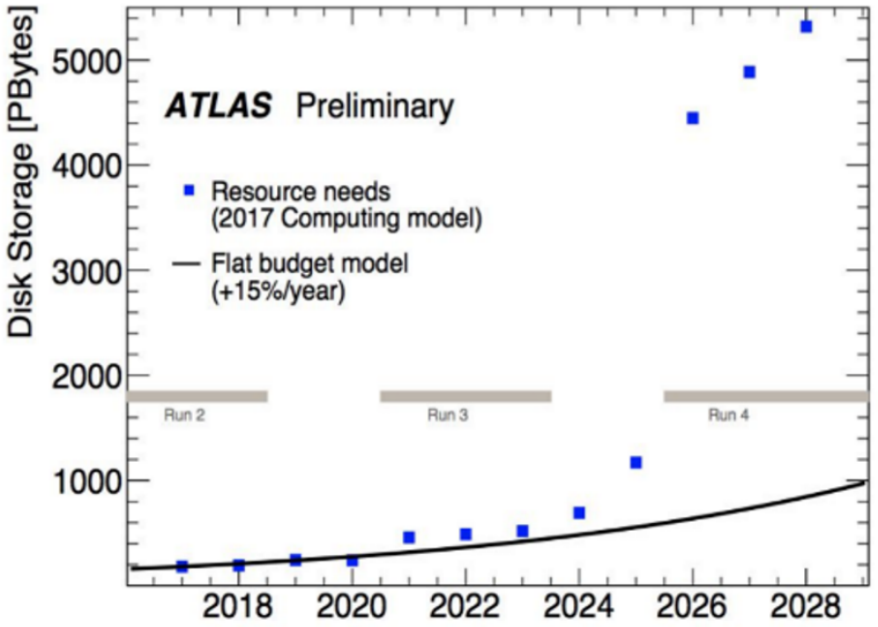
Great interest

Main topics

- Tape & Archival
- Caching
- Analytics
- Conditions data
- Storage solutions
- Data management infrastructure
- Cloud systems
- Event services
- Technology export

Driving force

This innocent plot is the catalyst for much of the work in this field



Driving force

This innocent plot is the catalyst for much of the work in this field

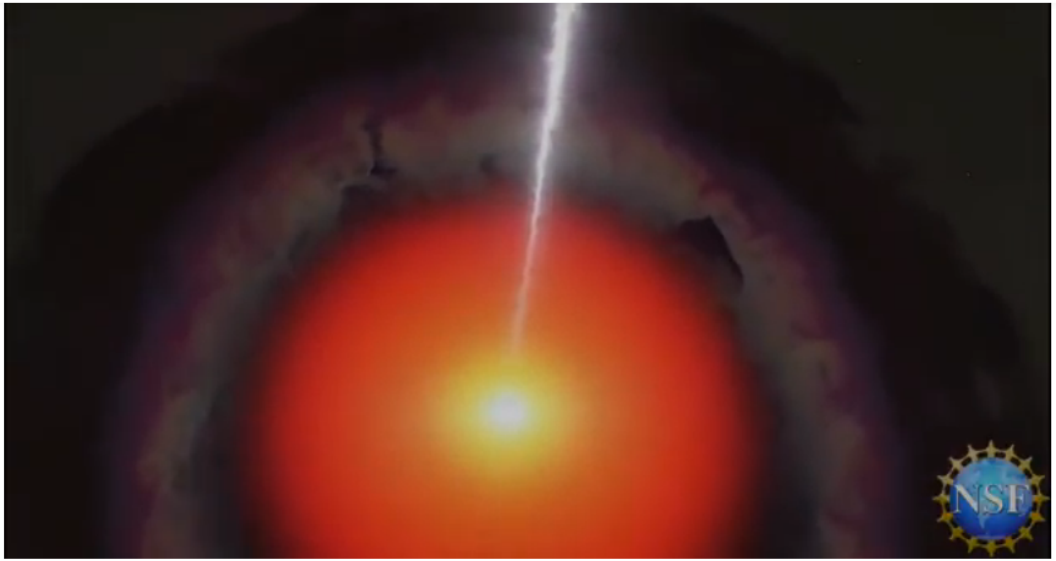
Tape & Archival

Castor → CTA
  Forecasting 4.3 EB by </Run 4>

KIT redesigned the tape pool and increased the tape recall performance 3x while also upgrading the disk pools

IceCube breaking news

(yesterday evening)



NSF Live Press Conference on Astrophysics Breakthrough

4,252 watching now

188 8 SHARE

National Science Foundation
Started streaming 1 hour ago

SUBSCRIBE 55K

An announcement of recent multi-messenger astrophysics findings led by the National Science Foundation's IceCube Neutrino Observatory at the South Pole. Hosted by NSF Director France Córdova, the briefing will feature leading astrophysicists from across the globe.

Panelists:

Francis Halzen, IceCube principal investigator, University of Wisconsin-Madison

Regina Caputo, Fermi-LAT analysis coordinator, University of Maryland/NASA Goddard Space Flight Center

Tape & Archival

Never underestimate the bandwidth of a sleigh full of disks !

Caching

Attracted a lot of attention, a hot topic in light of smaler sites loosing their storage

Two main options are explored

- Transparent (or not) caching of transiting data
  - With the option of an Xrootd plugin
- Pre-filling by the VO (in opportunistic storage mode)

Explored in both popular protocols nowdays: XRootD and HTTP

Caching

Expanding the Frontier & CVMFS caching with Cloudfare
- Part of the fallback mechanism in CVMFS

While data caches are very attractive it is not clear yet what the impact on real life setups is
- Is data actually being reused?

Analytics

Hadoop @ CERN
- Growing ecosystem of components
- Expert knowledge ready to share with the community

HPC: convert data in HDF5 format, both for current parallel analysis and for ML future use
- Also mentioned by dCache as reason for development

CMS WMAgent logs with Mongo, Spark, EOS

Conditions data

Using logs and analytics to identify places to optimize (needless queries, failing sites, caching problems)

Push for REST APIs to decouple clients while profiting from the http caching infrastructure
ATLAS, Belle II

and for using JSON to serialize data & geometry

Conditions data

Radically different approach from LHCb to storing and distributing conditions data as a Git repository on CVMFS

Storage solutions

ECHO entered production at RAL one year ago
- Swift/S3 favoured over legacy protocols

CERN is also using Ceph for CASTOR and for the HPC

dCache active development on supporting Posix for mutable files, auth tokens, storage events

DPM streamlines DB operations by deprecating legacy & SRM-related services

Storage solutions

EOS needs to shed the in-memory namespace for a persistent one to keep up with the growth

- providing an interesting source for hdd failure analysis
- and using *consul* to identify problem nodes

CERNBox is one of the reasons to push EOS development further

- In collaboration with industry for sync & share solutions

Data mgmt infrastructure

Dynamic data placement for ATLAS
  Using ML to predict future use of datasets

SciTokens: replacing GSI (proxy) authentication with OAuth2 + JSON WebTokens
  3rd party HTTP transfers too

Data lake exploratory work across many remote sites

BESIII Data management solution

Cloud systems

Interesting collaboration between ATLAS and Google to integrate both storage and computing

CERN IT involvement in the Horizon2020 eXtreme DataCloud initiative

Cloud Services for Synchronization and Sharing (CS3) conference

Remote data access in ATLAS

Event services

BESIII & ATLAS, aiming at reducing the amount of data transferred to the client and optimizing CPU usage on HPCs

- And doing async IO to prefetch

Event indexing – how to scale up from 350 Hz, x5 for Run 3 and then x3 more for Run 4

Technology export

DIRAC, Rucio, CVMFS, Dynafed and many others are adopted by an expanding community outside CERN

- SoLid, Belle II, LIGO

Feed back from the external use cases

- Generic metadata support in Rucio

Thank you!