Current HEP Computing Models

Ian Bird, WLCG, CERN
ESPPU Open Symposium
Granada, 15th May 2019
Acknowledgement

- Input to this talk has been drawn from many discussions with colleagues from many HEP and closely related experiments over recent years, in particular
  - Preparation of CWP and WLCG strategy documents
  - Recent WLCG and HOW2019 workshops
  - Contributions as input to the ESPPU
Today’s computing models

- HEP has always done distributed computing
- Scale of the challenge for LHC forced a more organized and formal structure
- Built a federated system – grid – to integrate and make easily usable pledged resources
LHC and other HEP experiments

- LHC computing (WLCG) was a significant step in organization and change of computing models
- However, from the beginning of the grid projects
  - Most other HEP experiments have made use of many of the tools and benefitted from the grid infrastructures (national, regional, global)
- All experiments’ computing is distributed to some degree
- Major features and capabilities of today’s HEP computing infrastructures:
  - Networks – international and national, private and public
  - Data management – key to success, data transfers, storage systems, data management tools and data organization
  - Compute – provisioning of resources and workload scheduling; evolution of types of resources
  - AAI (Authentication, Authorization Infrastructure) - the mechanism of federation, single sign on, etc.
  - Operations support – security, incident response, problem tracking, daily operations, upgrade campaigns
  - Other common services – software delivery, databases and db replication/caching, etc.
  - Diverse experiment-specific services and tools – applications

Granada, 15 May 2019
ian.Bird@cern.ch
Networks

- OPN – private connections – managed by WLCG
- LHCOOne – Virtual Overlay on national academic networks – managed by NRENs
  - Also used by many other “HEP & related” experiments

Needs of LHC have helped the networking community deploy state of the art networking

Significant transatlantic capacity
Storage & data services

- **Storage tiers:**
  - Tape – archive and infrequently accessed data
  - Disk – actively processed data
  - SSD – caches for certain services, databases, etc

- **Storage is managed by a software components developed & proven over time**
  - Tape – Castor (CERN), Enstore (FNAL), HPSS or TSM (IBM)
    - Tier 0 and Tier 1s have concentrated on these solutions
  - Disk – EOS (CERN), dCache (DESY/FNAL), DPM (CERN); some others (filesystems like ceph with grid interface, StoRM)
    - All highly reliable, scalable, and robust: needs operational effort to manage
    - All compute sites today require some form of managed storage

- **Databases**
  - Conditions and other volatile data – Oracle as underlying db, with Oracle tools for replication (online→offline, Tier 0 → Tier 1s) and backup
  - Metadata served via a cache infrastructure → Frontier (dev by CMS/FNAL) used by many
    - Replaces original need to have full Oracle replication everywhere

- **Software distribution**
  - Was initially a clumsy problem to distribute complex software (often!) to ~100 sites
  - Solved by cvmfs – publishing, content distribution, caching – now used by many communities
Data management tools

Experiment/common DM tools
Data placement planning – workflow driven

- Dirac
- Rucio
- Alien
- Phedex

Data transfers

- FTS service on gridftp, xroot, https
- xroot

- T1 Disk storage
- T2 Disk storage
- T1 Disk storage

Organised movement

Tools have evolved over many years – experience and changing use patterns
Rucio – becoming a common tool
DIRAC – has broad user community – mainly as workflow

Very robust and tested services
~100 PB/month for 10 years (~10^{11} files!)

These are tools used by and interesting for many HEP and other experiments
Data organization

- Several different formats of data:
  - RAW, ESD, AOD, derived-AOD, other
    new formats of reduced AOD, ntuple
    - Parallel formats for simulated data
- Different formats have different policies for replication, accessibility, lifetimes, distribution
- Initially data was pre-placed according to a policy ready for processing
  - More or less strong hierarchy between Tiers
  - Today much more dynamic – peer-peer and remote access – data federations
  - Each experiment has variations on this
Workload management

Different experiments have different workload management systems, but have converged towards a model of “pilot” jobs

- Late binding – a placeholder job is sent to any free resource, calls home for next (appropriate) priority work
- This has been very effective at filling available resources, and allows dynamic prioritization within an experiment queue of work

LHC Experiments have each developed their own workload management service

- Panda, DIRAC, GlideInWMS/WMAgent, Alien
- Each have broader communities in HEP, NP, astronomy, and other disciplines using these tools
- These tools organize the workflows, dispatch chains of jobs, interact with the DM services, collect output, manage errors and resubmissions
  - They communicate with distributed compute resources

Physicists interact via the experiment frameworks

- Which in turn make use of the workload management and data management systems
Heterogeneous computing

- The majority of today’s HEP processing is performed on dedicated clusters of commodity processors (“x86”)
- Recently: opportunistic use of many types of compute, in particular HPC systems, and HLT
- In future, this heterogeneity will expand; we must be able to make use of all types:
  - Non-x86 (esp GPU), HPC, clouds, HLT farms (inc FPGA?)

HPC Use is challenging
HEP engagement with DOE & NSF in USA and (together with SKA) with PRACE and EuroHPC in Europe and participating in BDEC2 workshops

Granada, 15 May 2019
Ian.Bird@cern.ch
Heterogenous compute

- Requires:
  - Common provisioning mechanisms, transparent to users
  - Facilities able to control access (cost), appropriate use, etc

- HPC, Clouds, HLT will not have (affordable) local storage service (in the way we assume today)
  - Must be able to deliver data to them when they are in active use

Deployed in a hybrid cloud mode:
- Procuers’ data centres
- Commercial cloud service providers
- GEANT network and EduGAIN Federated Identity Management

Granada, 15 May 2019
Ian.Bird@cern.ch
Data delivery “data lake (cloud)”

Idea is to localize bulk data in a cloud service (Tier 1’s ➔ data lake): minimize replication, assure availability.

Serve data to remote (or local) compute – grid, cloud, HPC, ???

Simple caching is all that is needed at compute site (or none, if fast network).

Federate data at national, regional, global scales.
Data management and storage

- Set of R&D projects to prototype such a data management infrastructure – and associated tools

- Aims:
  - Reduce the global cost of storage (hw and operations)
  - Enable a more effective use of existing storage
  - Be able to efficiently and scalably deliver data to large, remote, heterogenous, compute resources (LHC Tier centres or HPC, clouds, other opportunistic)
  - Build a common set of DM tools that can be used by a broad set of scientific experiments
    - Today LHC, DUNE, SKA, Belle-II, GW-3G, and others are all looking at a common set of identified tools

- Also collaboratively (LHC+SKA with GEANT) looking at underlying data transfer and network tools (replace gridftp, network protocols, etc.)
The grid is based on the idea of identity federation:
- “single sign on” in a distributed computing environment
- Credentials issued to a user by a trusted local authority
- Network of trust between all participating institutions and authorities
- Acceptance of the credentials by each resource (storage, compute, service)
- This has been implemented using X.509 certificates as the trusted credential

Authorization:
- Implemented by a service “VOMS” that enables access to resources

The success of this model and the international trust framework has been instrumental in the success of the grid model.

For the future this is moving to a “token-based” system – more common in the rest of the world.
Bit preservation is a solved problem (modulo cost)
Data preservation (reproducibility, accessibility) lacks consistent policies and costing
Open access also lacks consistent policies and is not explicitly funded
  Although required by many funding agencies
There are many use cases of data preservation (for physics) and open access
  Some are coincident – many are not
Long term DP for physics requires evolving the mechanisms and tools used to do analysis
  Cannot afford different solutions
Open access requires policy agreement and appropriate resources
HEP can learn from other sciences who do this well
Evolution of HEP computing

A Roadmap for HEP Software and Computing R&D for the 2020s

HEP Software Foundation

ABSTRACT: Particle physics has an ambitious and broad experimental programme for the coming decades. This programme requires large investments in detector hardware, either to build new facilities and experiments, or to upgrade existing ones. Similarly, it requires commensurate investment in the R&D of software to acquire, manage, process, and analyse the sheer amounts of data to be recorded. In planning for the HL-LHC in particular, it is critical that all of the collaborating stakeholders agree on the software goals and priorities, and that the efforts complement each other. In this white paper, this white paper describes the R&D activities required to prepare for this software upgrade.

1Arhivs are listed at the end of this report.

https://doi.org/10.1007/s41781-018-0018-8

WLCG Strategy towards HL-LHC

Executive Summary

The goal of this document is to set out the path towards computing for HL-LHC in 2020/7. Initial estimates of the data volumes and computing requirements show that this will be a major step up from the current needs, even those anticipated at the end of Run 3. There is a strong desire to maximise the physics possibilities with HL-LHC, while at the same time maintaining a realistic and affordable budget envelope. The past 15 years of WLCG operation, from initial prototyping through to the significant requirements of Run 2, show that the community is very capable of building an adaptable and performant service, building on and integrating national and international structures. The WLCG and its stakeholders have continually delivered to the needs of the LHC during that time, such that computing has not been a limiting factor. However, in the HL-LHC era that could be very different unless there are significant changes that will help to moderate computing and storage needs, while maintaining physics goals. The aim of this document is to point out where we see the main opportunities for improvement and the work that will be necessary to achieve them.

During 2017, the global HEP community has produced a white paper - the Community White Paper (CWP), under the aegis of the HEP Software Foundation (HSF). The CWP is a ground-up gathering of input from the HEP community on opportunities for improving computing models, computing and storage infrastructures, software, and technologies. It covers the entire spectrum of activities that are part of HEP computing. While not specific to LHC, the WLCG gave a chance to the CWP activity to address the needs for HL-LHC along the lines noted above. The CWP is a compendium of ideas that can help to address the concerns for HL-LHC, but by construction the directions set out are not all mutually consistent, not are they prioritized. That is the role of the present document - to prioritise a 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, and the experience of the past.

At a high level there are a few areas that clearly must be addressed, that we believe will improve the performance and cost effectiveness of the WLCG and experiments:

- **Software.** With today’s code the performance is often far from what modern CPUs can deliver. This is due to a number of factors, ranging from the construction of the code, not being able to use vector or other hardware units, layout of data in memory, and end-end IO performance. While some level of code re-engineering, it might be expected to gain a moderate factor (x2) in overall performance. This activity was the driver behind setting up the HSF, and remains one of the highest priority activities. It also requires the appropriate support and tools, for example to satisfy the need to fully automate the ability to often perform software validation of software. This is essential as we must be adaptable to many hardware types and frequent changes and optimisations to make the best use of opportunities. It also requires that the community develops a level of understanding of how to best write code for performance, again a function of the HSF.

https://cds.cern.ch/record/2621698
DUNE will leverage the WLCG for its computing infrastructure

First formal non-LHC

Common theme in many contributions is the desire to collaborate with and benefit from LHC R&D work

implies governance evolution
Governance / organization
Conclusions

- The LHC computing models and WLCG have been very successful, and have evolved significantly over 15 years.

- Many other HEP experiments have collaborated and used the infrastructures and tools (WLCG, EGI, OSG, others).

- Strong desire to strengthen the synergies and collaborations for the future – infrastructure, tools, and software.
Backup
Recent Evolution - LHCb

40 Tb/s

- DETECTOR READOUT
- PARTIAL RECONSTRUCTION

1-2 Tb/s

- REAL-TIME ALIGNMENT & CALIBRATION

80 Gb/s

- 26% FULL
- 68% TURBO & real-time analysis
- 6% CALIB

Buffer

Offline

Online

Event Rate
(events/s)

100%

100%

100%

Throughput
(GB/s)

High Level Trigger

Tape Storage

Disk Storage

High Level Trigger

3.5 GB/s

13%

10 GB/s

00%