DPHEP Future Activities (from Status Report)

Over the next period, one can expect progress to be made in the following areas:

- The establishment of a formal policy regarding data preservation for CERN experiments (perhaps linked to the approval process through the Research Board);
- At least a "self-audit" for the CERN Tier0 and WLCG Tier1 sites in the context of the WLCG project;
- Further developments in terms of Analysis Capture and Preservation;
- Further releases of Open Data through the CERN Open Data Portal;
- Harmonization of similar activities across various laboratories and projects;
- Extension of DPHEP's activities to consider also those of potential FCCs;
- Clarifications regarding funding – of particular importance to past experiments where resources have already become sub-optimal;
- The continuation of regular meetings and workshops, aligning as much as possible with related events (WLCG, CHEP, HEP Software Foundation etc.);
- Further input to the next round of ESPP – building on concrete experience, results and remaining challenges.

The long-term management of the Collaboration also has to be considered – up to 2020 but also beyond.

Outlook and Conclusions

There are clearly many similarities in the approaches being taken, the technologies deployed and the issues encountered. Regular reporting of results (possibly synchronised with major events such as CHEP) should be sufficient to ensure that coordinated approaches remain and that duplication is minimised.

The following quote[1] is traditionally attributed to Leslie Lamport – the initial author of LaTeX and an expert on distributed computing systems.

> A distributed system is one in which the failure of a computer you didn't even know existed can render your own computer unusable.

This reminds us that data preservation is inherently unstable – with many components and dependencies, constant attention is required to ensure that the entire "system" remains usable. Some changes may be relatively minor, such as a name change in a webserver. Others can be much more disruptive, such as major change in operating system (think VAX/VMS to Unix) or programming language – even a standard-conforming language changes over time, with some constructs being first deprecated, then obsolete and finally unsupported.

Given the cost of today's storage and the likely evolution, there is no inherent cost why "data" cannot be stored more or less indefinitely. What is harder is to capture the

[1] See http://research.microsoft.com/en-us/um/people/lamport/pubs/distributed-system.txt.

necessary knowledge and validation procedures so that it can be used over long periods of time.

The "natural periodicity" of recent collider generations – some twenty years – is perhaps all one can hope for in terms of affordable data preservation. (Most LEP data – that of ALEPH, DELPHI and OPAL – may be usable somewhat longer, perhaps up to 25 / 30 years). Beyond that, re-use of the data will probably still be possible but may require a larger investment to "resuscitate", as has been done on rare (one?) occasion(s), notably for the JADE[2] experiment at the PETRA storage ring in DESY.

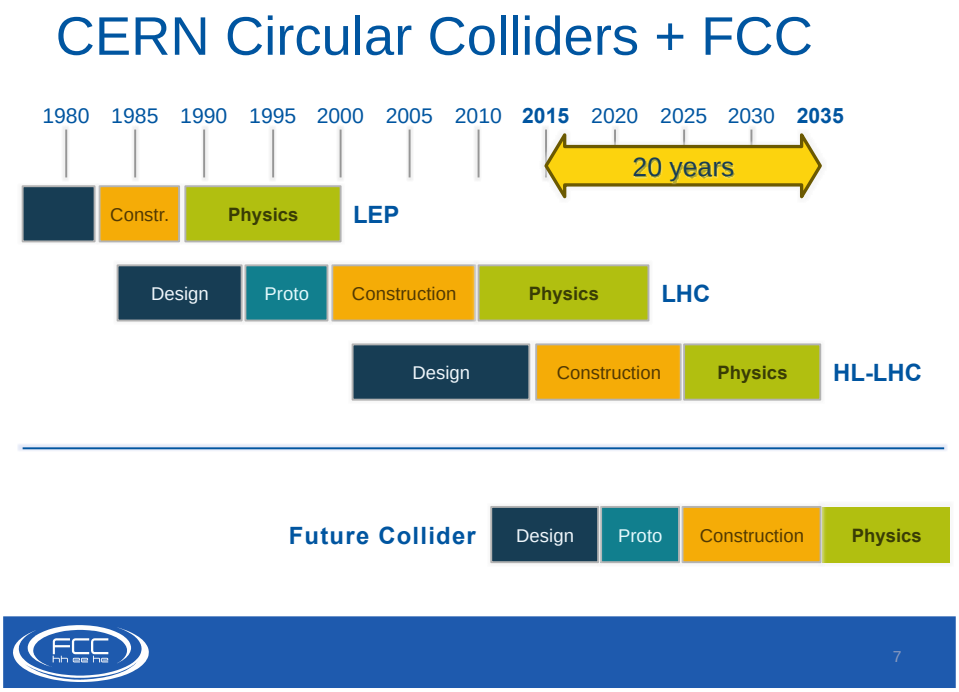

Figure 1 - Timeline of Major Colliders at CERN (+ "FCC")

[2] See https://wwwjade.mpp.mpg.de/ and the DPHEP Blueprint for further information.