CERN Experience & Plans

The Higgs Boson Years…

Jamie.Shiers@cern.ch

ADMP workshop

June 2016

International Collaboration for **Data Preservation** and **Long Term Analysis** in High Energy Physics

Overview

- From the Worldwide LHC Computing Grid to Data Preservation for long-term Analysis

- What we gained from others: business case / model, certification, DMPs and more

- How we see this fitting together…

- And next steps…

Long Term

- **CERN has existed since 1954: now 21 MS + others in Q**

- The first ideas of a "**Large Hadron Collider**" were mooted in 1978:
  - *"LEP – if it is built – should be housed In a tunnel large enough to accommodate a hadron collider"*

- Council recently approved the **HL-LHC** upgrade: 2025 – 2035 / 40
  - An "**HE-LHC**" will be considered as part of the next European Strategy for Particle Physics update: 2019/2020
  - If approved, will take "LHC" to 2nd half of this century

- A Future (100km) Circular Collider (FCC) (in fact several options) is also under study: more in ESPP update

➢ **LHC data needs to be (re-)usable throughout this period!**

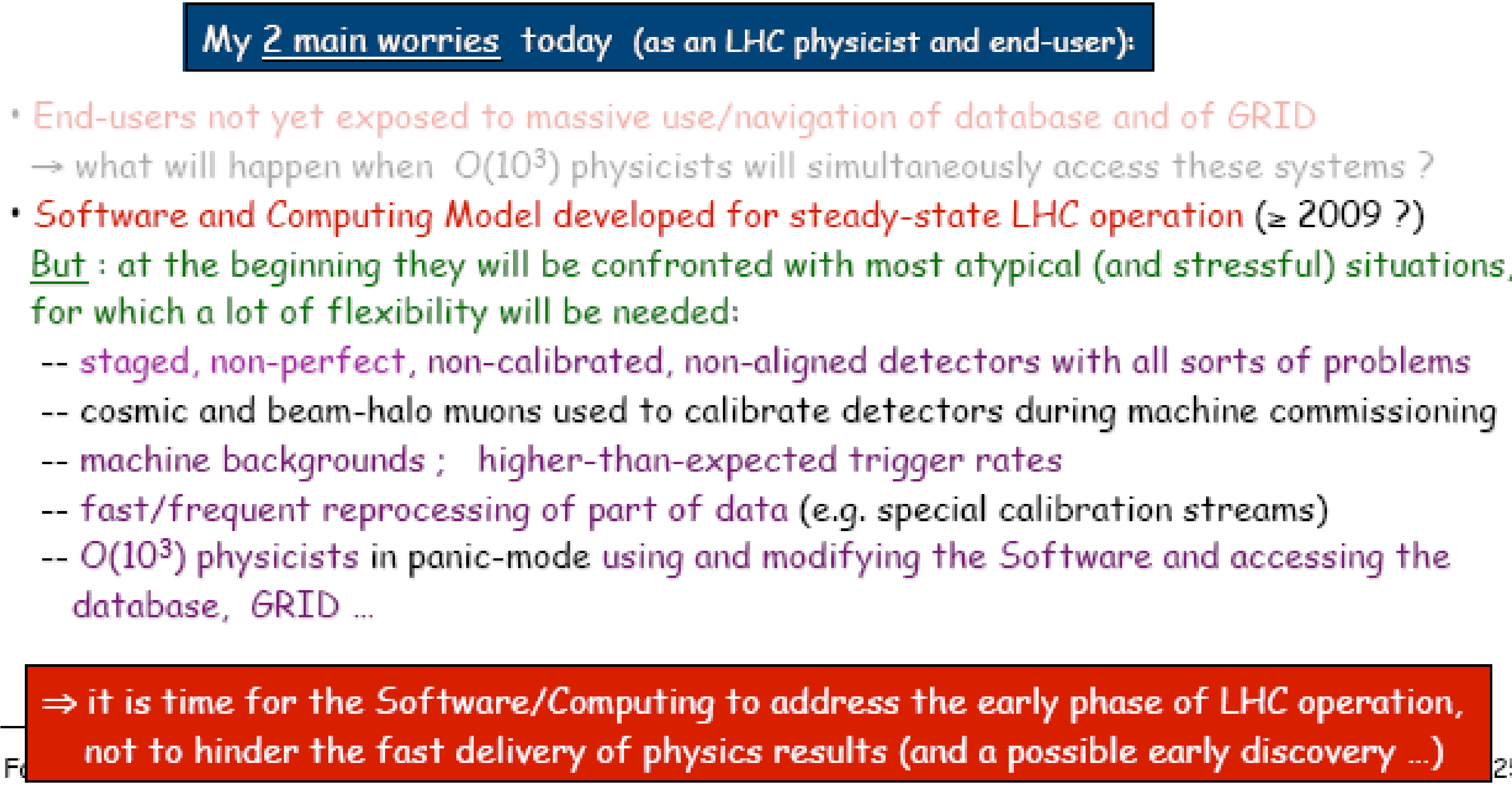
How do we find the Higgs?

LHC Computing: A Long & Winding Road

- Started (for me) in September 1992 – CHEP in Annecy
- R&D projects from 1994 on – led to major migration of media, data format & code (~1 year to plan, ~1 year to do)
- ***Grid-itus*** from ~2000: (W)LCG Service Challenges from 2004
- CHEP 2004: ***"It is time for the Grid to deliver… and not get in the way"*** (Fabiola Gianotti)

➢ **This was a period of "tumultuous change" (but people have already forgotten – a major risk for LTDP…)**

Reminder – one of the conclusions from the plenary talk at CHEP’ 04 by Fabiola Gianotti



My 2 main worries today (as an LHC physicist and end-user):

- End-users not yet exposed to massive use/navigation of database and of GRID
  → what will happen when $O(10^3)$ physicists will simultaneously access these systems ?
- Software and Computing Model developed for steady-state LHC operation (≥ 2009 ?)
  But : at the beginning they will be confronted with most atypical (and stressful) situations,
  for which a lot of flexibility will be needed:
  -- staged, non-perfect, non-calibrated, non-aligned detectors with all sorts of problems
  -- cosmic and beam-halo muons used to calibrate detectors during machine commissioning
  -- machine backgrounds ; higher-than-expected trigger rates
  -- fast/frequent reprocessing of part of data (e.g. special calibration streams)
  -- $O(10^3)$ physicists in panic-mode using and modifying the Software and accessing the database, GRID …

⇒ it is time for the Software/Computing to address the early phase of LHC operation, not to hinder the fast delivery of physics results (and a possible early discovery …)

The LCG Service Challenges:

Rolling out the LCG Service

Jamie Shiers, CERN-IT-GD-SC

http://agenda.cern.ch/fullAgenda.php?ida=a053365

June 2005

LCG Service Hierarchy

Tier-0 – the accelerator centre

- Data acquisition & initial processing
- Long-term data curation
- ➢ Data Distribution to Tier-1 centres

Canada – Triumf (Vancouver)
France – IN2P3 (Lyon)
Germany –Karlsruhe
Italy – CNAF (Bologna)
Netherlands – NIKHEF/SARA (Amsterdam)
Nordic countries – distributed Tier-1
Spain – PIC (Barcelona)
Taiwan – Academia SInica (Taipei)
UK – CLRC (Oxford)
US – FermiLab (Illinois)
– Brookhaven (NY)

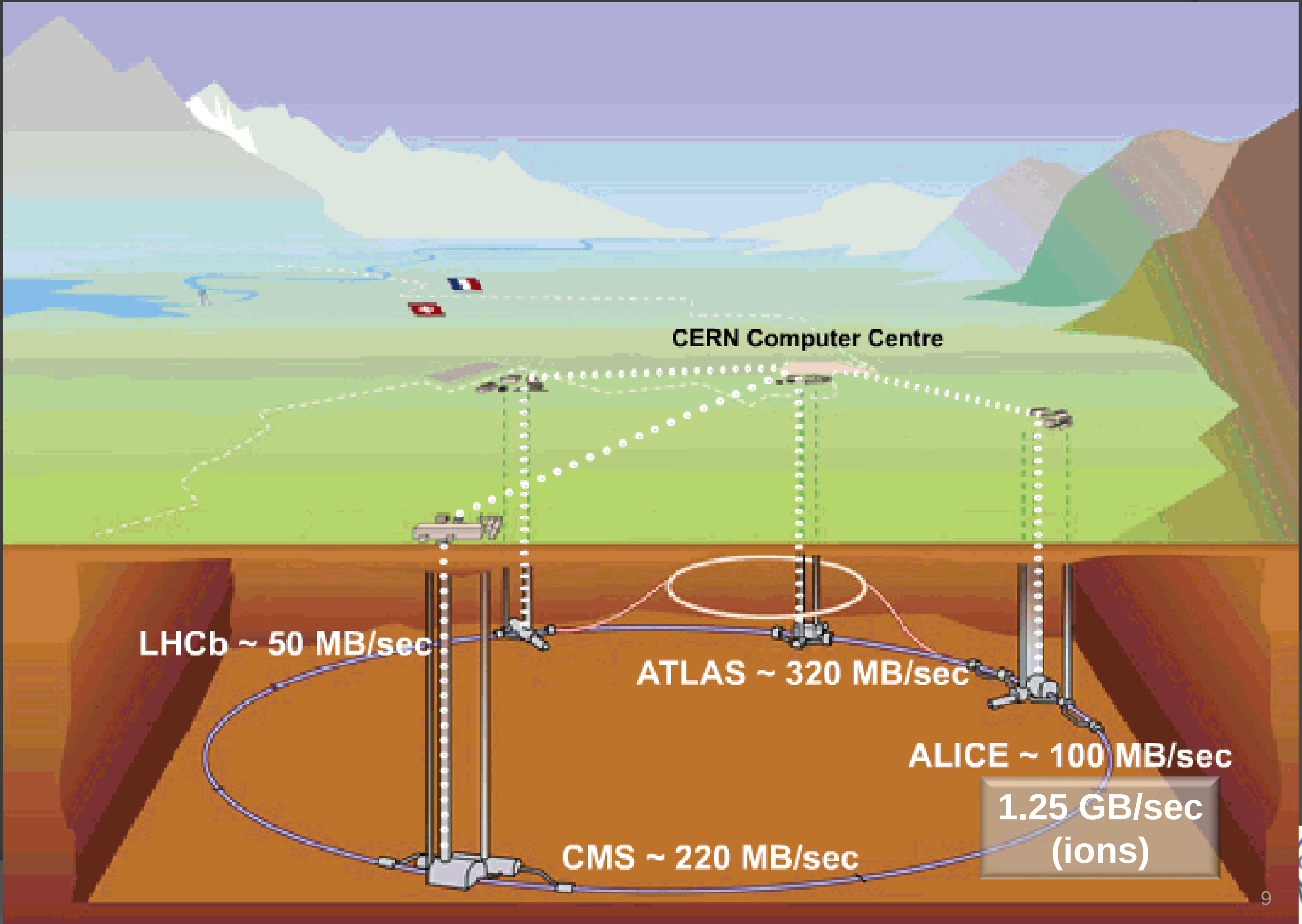
Tier-1 – "online" to data acquisition process → high availability

- Managed Mass Storage – → grid-enabled data service
- ➢ All re-processing passes
- Data-heavy analysis
- National, regional support

Tier-2 – ~100 centres in ~40 countries

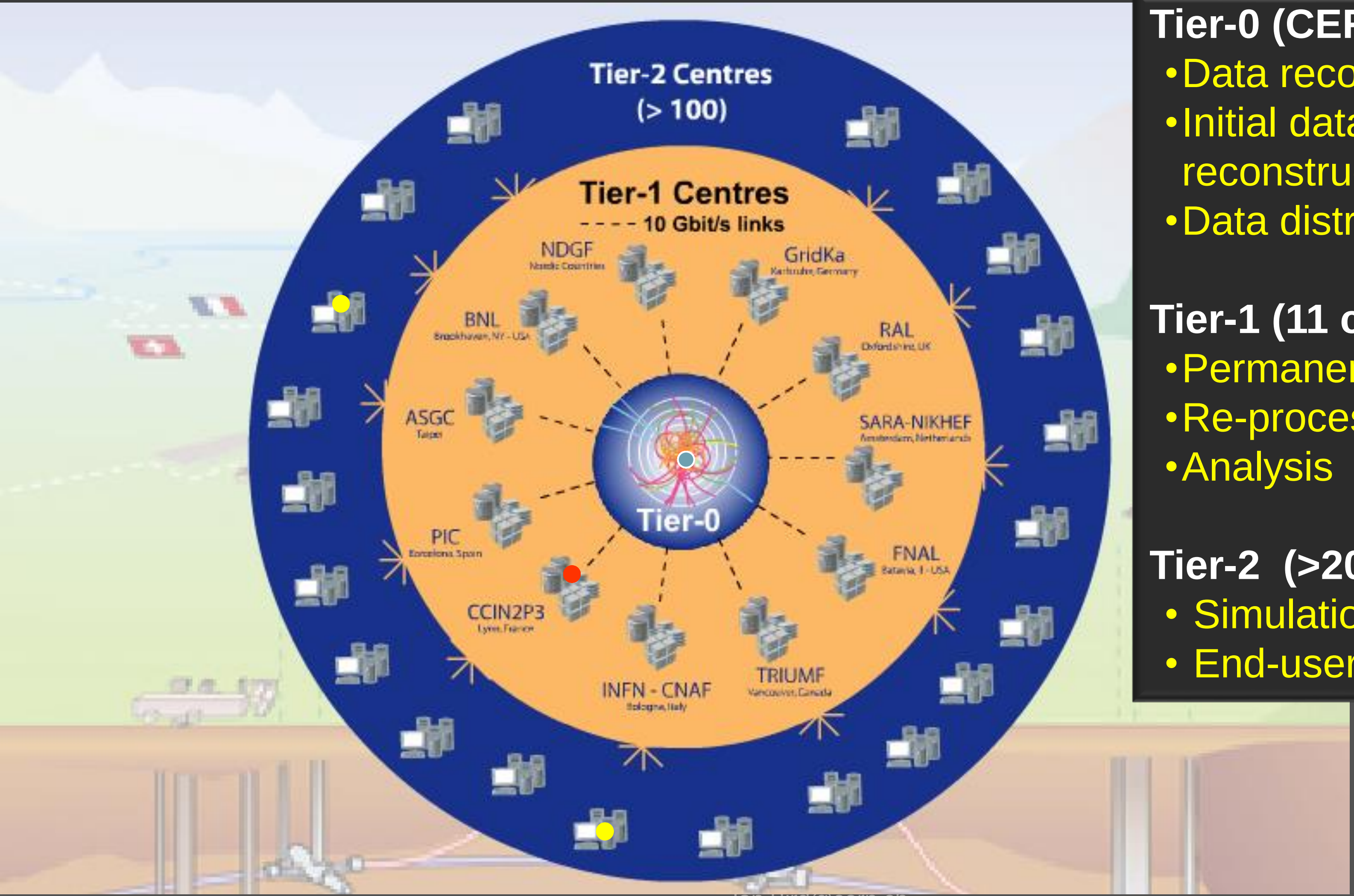
- Simulation
- End-user analysis – batch and interactive
- ➢ Services, including Data Archive and Delivery, from Tier-1s

Tier 0 at CERN: Acquisition, First pass reconstruction, Storage & Distribution



1.25 GB/sec (ions)

Tier 0 – Tier 1 – Tier 2



Tier-0 (CERN):

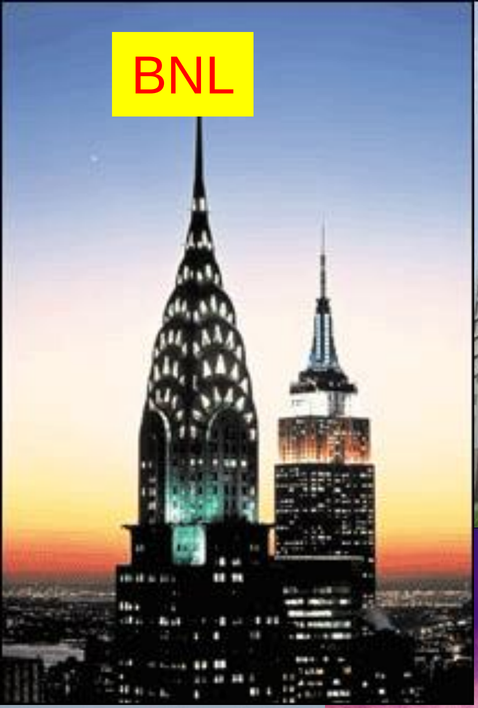
- Data recording
- Initial data reconstruction
- Data distribution

Tier-1 (11 centres):

- Permanent storage
- Re-processing
- Analysis

Tier-2 (>200 centres):

- Simulation
- End-user analysis


BNL

ASGC/Taipei

CCIN2P3/Lyon

TRIUMF/BC

NIKHEF/
SARA

FNAL

NDGF

CNAF

RAL

PIC

TIER2s

CERN

FZK

So What Made it all Work?

- Top-down is fine, but you also need a parallel “bottom-up” approach to explain, convince, motivate
  - To get from A to B, you first have to gather at A – don’t assume everyone is already there
- Keep things simple (e.g. operational procedures), inform users, listen…
- Establish clear and meaningful goals and metrics
- There are no such things as “show-stoppers” – don’t panic: analyse, solve, work-around
- Plan for what you have today: assume major changes and understand how you will accommodate them

Valid for other large scale projects, such as those on ESFRI roadmap – can DMPs help us find synergies?

And that metric?

- To find the Higgs you need 3 things:

1. **The machine;**
2. **The experiments;**
3. **The GRID**

- Rolf-Dieter Heuer, DG, CERN, July 4 2012

Why this WLCG digression?

1. For most people in our community, data management is about large scale, high throughput data movement, caching & popularity and so forth; (WLCG TDR2)
2. Where is the data to be preserved? Raw data is at CERN plus a copy spread across Tier1s. Where is the data behind publications?
3. The complexity of our environment: DMPs required by 21 Member States plus many NMS who contribute to the projects (plus many experiments) – N x M x O

And then?

- In May 2012, the DPHEP Study Group published a “Blueprint document”, a summary of which was fed into the ESPP workshop
- ESPP concluded that LTDP was of strategic importance but how to do it?

➢ Another Road Trip – APA(RSEN), RDA, 4C, iDCC, iPRES, DPC, etc

DPHEP Business Case & Cost Model

- The need for a "business case" was clearly articulated
- A common set of Use Cases was agreed across all major HEP experiments…
- As well as a way to measure the "value" based on publications, PhDs and so forth
- STFC study: Tevatron "cost neutral" without including technology spin-offs; x 10 ROI with
- "Bit preservation" costs tend to zero; LEP now has 3 copies at CERN alone (and many outside)

Requirements from Funding Agencies

- To integrate data management planning into the overall research plan, all proposals submitted to the **Office of Science** for research funding are required to include a **Data Management Plan** (DMP) of no more than two pages that describes how data generated through the course of the proposed research will be **shared and preserved** or explains why data sharing and/or preservation are not possible or scientifically appropriate.

- At a minimum, DMPs must describe how data sharing and preservation will enable **validation of results**, or how results could be validated if data are not shared or preserved.

- Similar requirements from European FAs and EU (H2020)

H2020: Annex 1 (DMP Template)

The DMP should address the points below…

1. Data set reference and name
   - Identifier for the DS to be produced
2. Data set description
   - Description; origin; nature & scale; to whom useful; underpins publication? similar data?
3. Standards and metadata
   - Reference to standards *of the discipline*
4. **Data sharing**
   - How will it be shared? Embargo periods? Mechanisms for dissemination, s/w and other tools for re-use, access open to restricted to groups, where is repository? Type of repository?
5. **Archiving and preservation**
   - Description of procedures, how long will it be preserved? End volume? Costs? How will these be covered?

HEP LTDP Use Cases

1. Bit preservation as a basic “service” on which higher level components can build;
2. Preserve data, software, and know-how in the collaborations; Basis for reproducibility;
3. Share data and associated software with (wider) scientific community, such as theorists or physicists not part of the original collaboration;
4. Open access to reduced data sets to general public.

➢ **Basically, a reflection of DMP requirements**

LHC Experiments’ Data Policies

- These are basically “extended DMPs” that capture the small variations between each experiment
  - Variations in duration of embargo periods, designated communities, fraction of data released
- A generic “WLCG DMP” exists – just like a generic WLCG TDR (complemented by experiment-specific reports)
- More detail in Thursday’s talk about CMS experience with data releases

Another key “lesson”…

- “Trusted” or “certified” digital repositories
  - (Also cost recovery for repositories)
- Several such standards exist: CERN (WLCG) following ISO 16363 route
  - Some sites start with DSA, then DIN, then ISO
  - This would not work at CERN…
- At CERN, the closest thing to a “mission statement” is an Operational Circular
  - This, and other steps required for “certification” could not realistically be repeated as we moved up the ladder…

Certification – Current Status

- Original idea was to perform Certification in the context of WLCG
- However:
  a) Quite a few of the metrics concern the (CERN) site;
  b) Interest also in an OAIS archive for “CERN’s Digital Memory”;
  c) The two are linked: policies, strategies, mission statements for the former are part of the latter
  d) Some things will be easier in the latter which will in turn help the former ☺

➢ **Current thinking: (self-)certify site-wise; “project-specific details” via “Project DMPs”**

ISO 16363 metrics

Organisational Infrastructure

3.1	Governance & Organisational Viability	Mission Statement, Preservation Policy, Implementation plan(s) etc. [CERN, CERN, project(s)]
3.2	Organisational Structure & Staffing	Duties, staffing, professional development etc. [APT etc.]
3.3	Procedural accountability & preservation policy framework	Designated communities, knowledge bases, policies & reviews, change management, transparency & accountability etc. [At least partially projects]
3.4	Financial sustainability	Business planning processes, financial practices and procedures etc
3.5	Contracts, licenses & liabilities	For the digital materials preserved… [CERN? Projects?]

- **Logical to have an Operational Circular for “Data”**
  - Obviously should include “meta-data” (as per DPHEP SR)
    - Software + environment, documentation etc.
  - Symmetry with OC3 and OC6
    - Archival material and archiving at CERN
    - CERN scientific documents
    - **[CERN scientific data, s/w, doc + meta-data]**

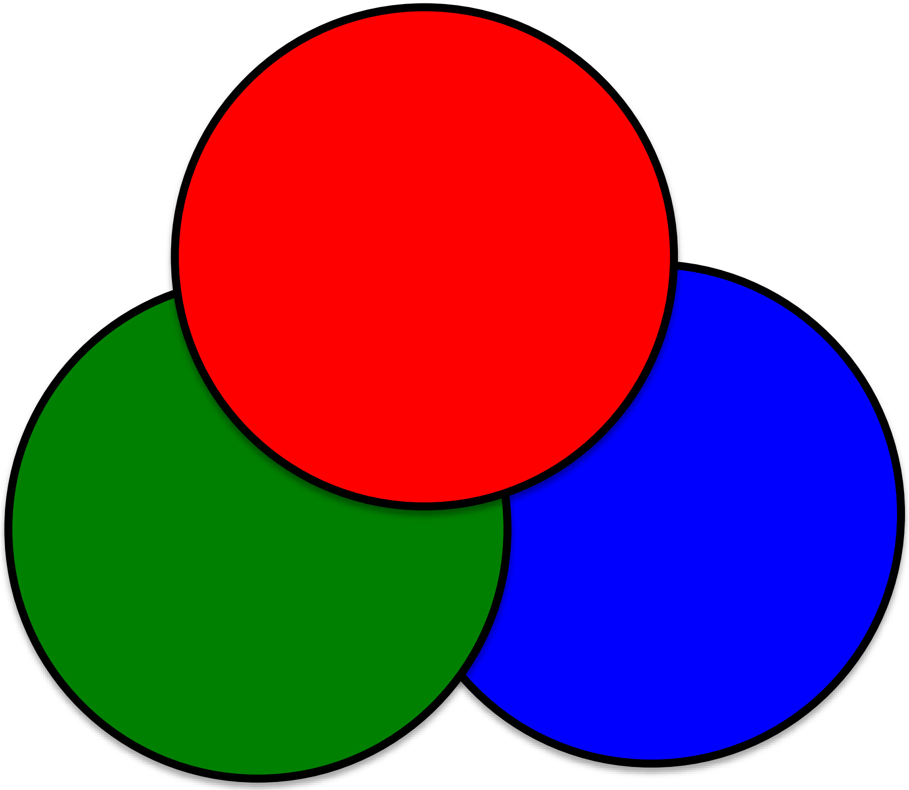
- **This could address “Mission Statement” and “DP Policy” in ISO 16363**

- Complemented by:
  - **Data Preservation Plan (inter-departmental) with ~3 year outlook**
    - Include also experiment plans or as part of their DMPs?
  - Experiment / Project Data Management Plan
  - Data Policy (extended DMP – à la LHC)

Infrastructure & Security Risk Management

5.1	Technical Infrastructure Risk Management **[We do all of this, but is it documented?]**	Technology watches, h/w & s/w changes, detection of bit corruption or loss, reporting, security updates, storage media refreshing, change management, critical processes, handling of multiple data copies etc **OC5, …**
5.2	Security Risk Management **[Do we do all of this, and is it documented?]**	Security risks (data, systems, personnel, physical plant), disaster preparedness and recovery plans … **OC2, …**

Data Preservation &
Certification of Trusted Digital Repositories:
Helps Address the Goals Below.



F.A.I.R. and Open Data:
Requires effort & Resources

Data Management Plans:
Sharing, Re-Use;
Reproducibility of Results

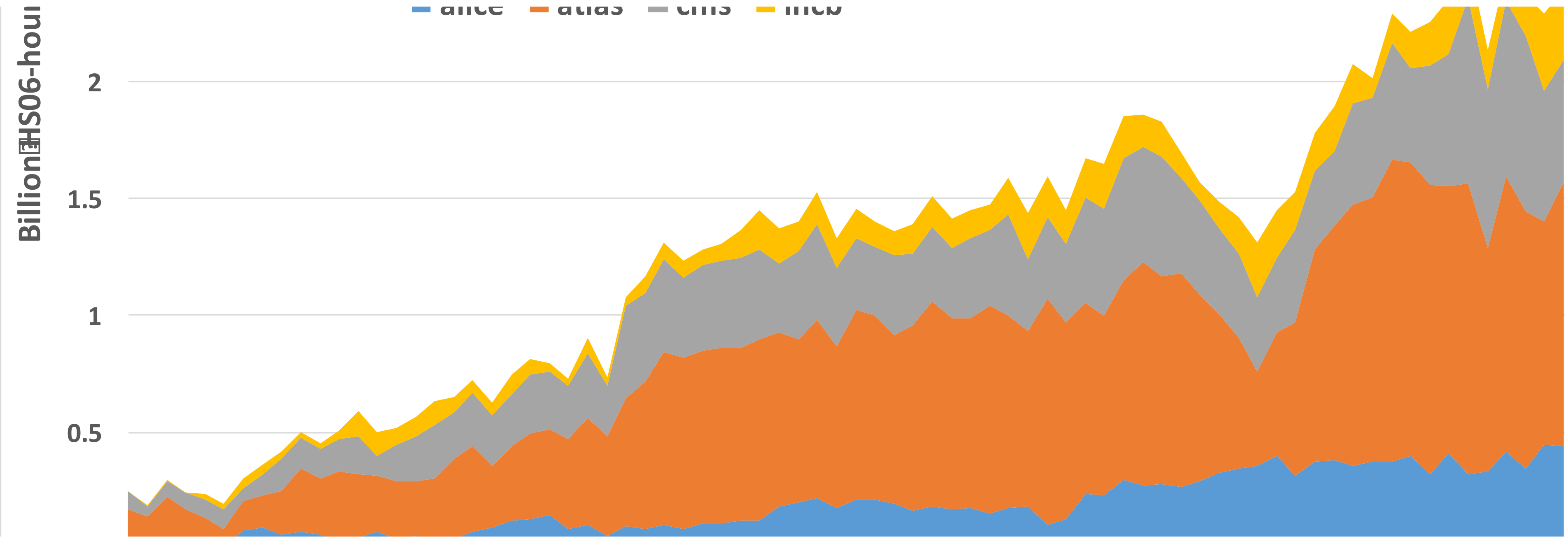
Concluding Remarks

- Data Preservation is a Journey – Not a Destination
  - “Once you stop pedalling, you stop & fall off”

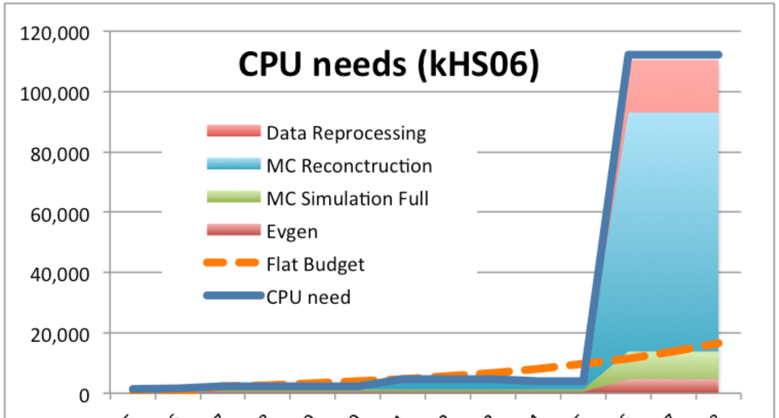
- Data Preservation is not an Island – it is part of a much bigger picture, including the full data lifecycle
  - You can’t share or re-use data, nor reproduce results, if you haven’t first preserved it

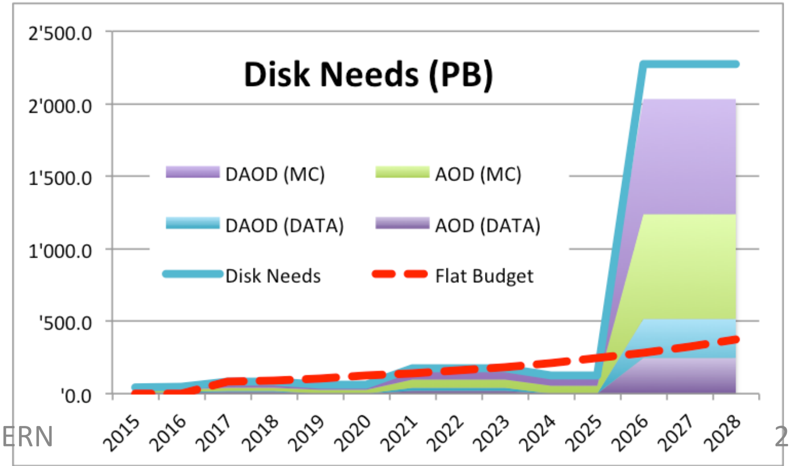
FUTURE NEEDS – NOW TO HL-LHC

Current Ramp-up of CPU



Initial studies on Computing for HL-LHC





TIME FOR ANOTHER PERIOD OF TUMULTUOUS CHANGE?