Data and Knowledge Preservation in High Energy Physics

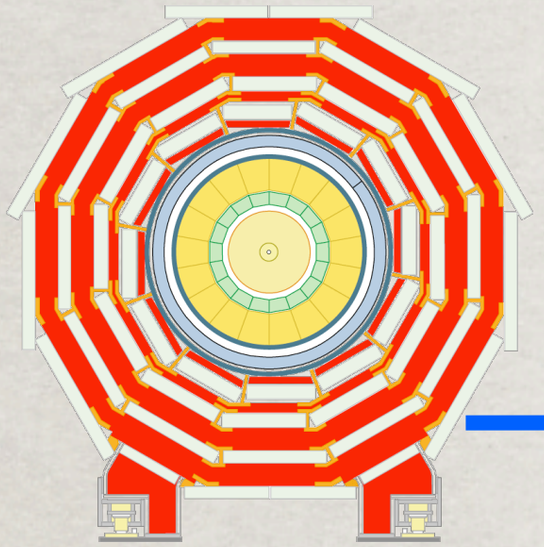
Mike Hildreth
University of Notre Dame
representing the DASPOS Project

Large Hadron Collider (LHC), CERN, Geneva, Switzerland

- Circumference: 27 km
- Current (2012) proton kinetic energy: 4 TeV
  - protons move at 99.999997% of speed of light
  - (8 m/s slower than light)
- Current (2012) total energy in beam: 135 MJ
  - Equivalent to a Nimitz-class aircraft carrier moving at 3.8 MPH
  - or a Subaru Impreza moving at 1045 kilometers/hour (nearly the speed of sound)
  - Or calories in seven 8” Cold Stone Creamery “Cheesecakes Named Desire”
- Superconducting magnet temperature is 2 K (colder than outer space)
- Truly international effort: >10,000 scientists from over 100 countries


LHCb
LHCb
CMS
CMS
ATLAS
ATLAS
ALICE
ALICE

Some Basic Data Facts



Basic facts:

➡Data from detector: ≥ 250 kB/ collision

➡Processing time for analysis: 5 sec (basic)

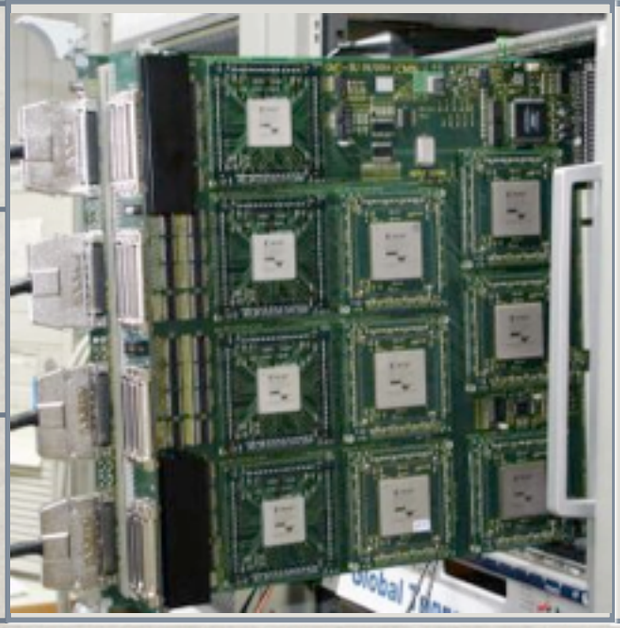
FPGA Chips do very simple analysis

~ μs to analyze data

Simplified analysis code

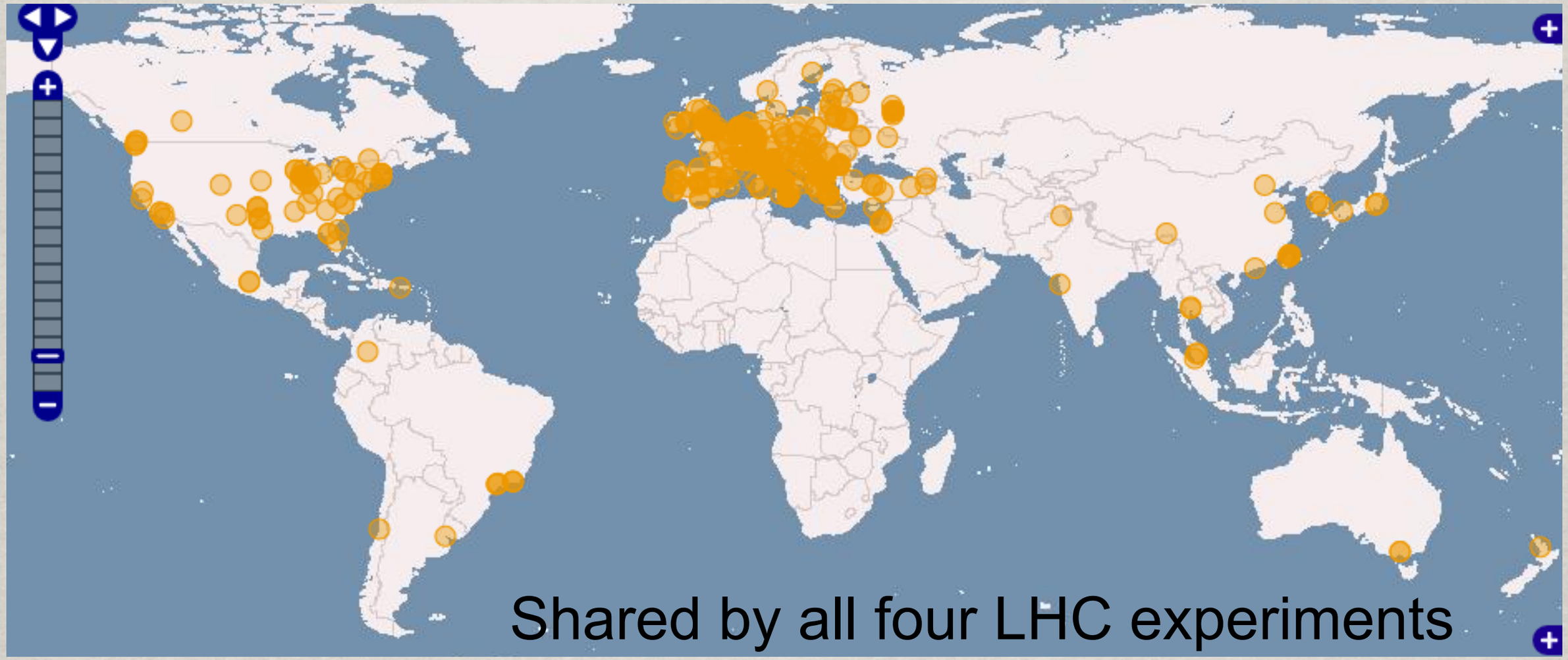
~ ms to analyze data

To physicists

	Proton Collisions in Detector	Level 1 Trigger	High Level Trigger
Data Rate	20 MHz		Computer farm with ~ 1000 CPUs High speed network/switch 
Data Collected	50 EB		
Processing time	45 Million CPU years!		

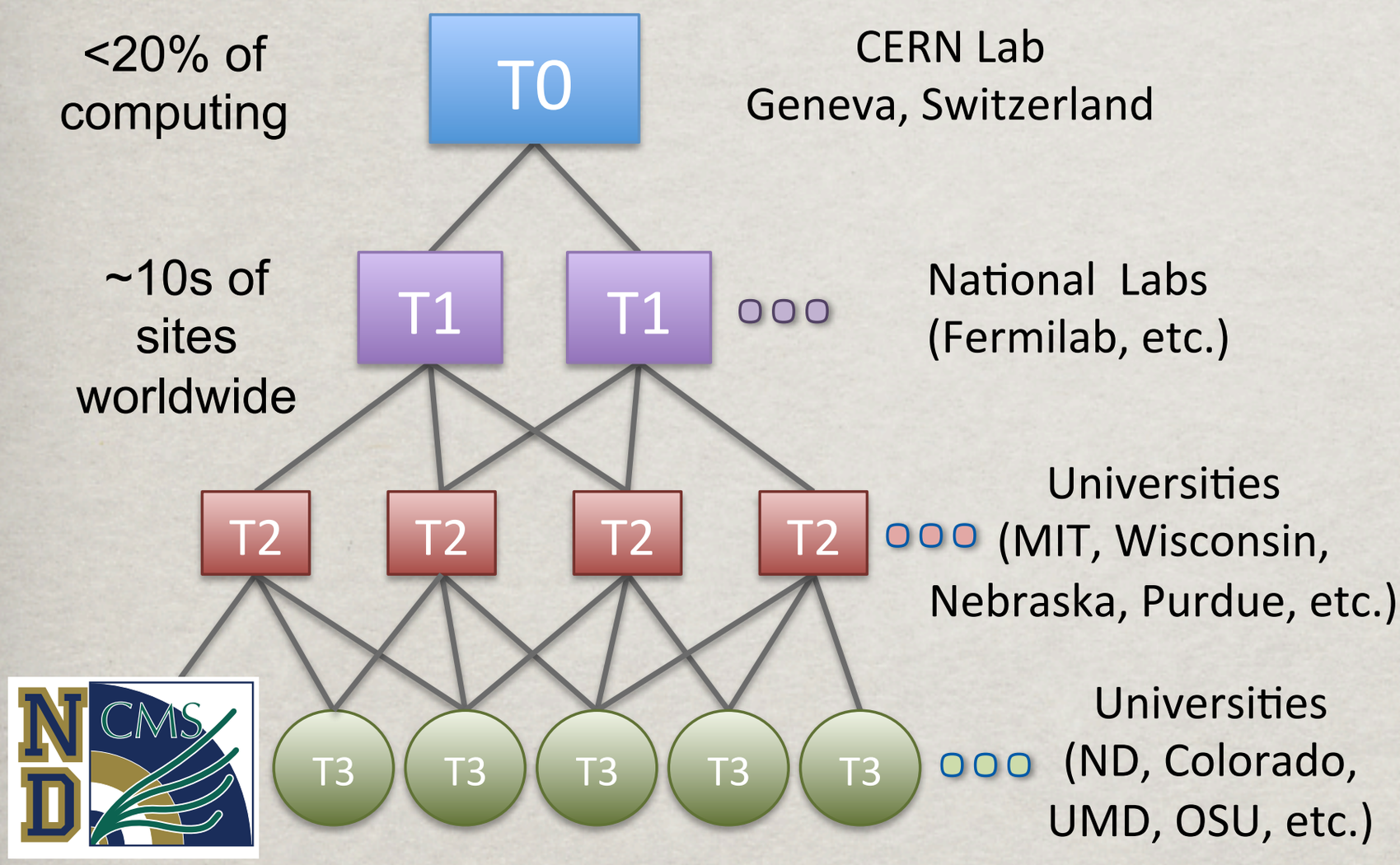
For 1 year's worth of data

Worldwide LHC Computing Grid



- Over 160 sites around world (including OSG sites in US)
- > 250k CPU cores available
- As many as 1 million jobs submitted in a single day
- > 300 PB of total storage available

Computing Tier Model



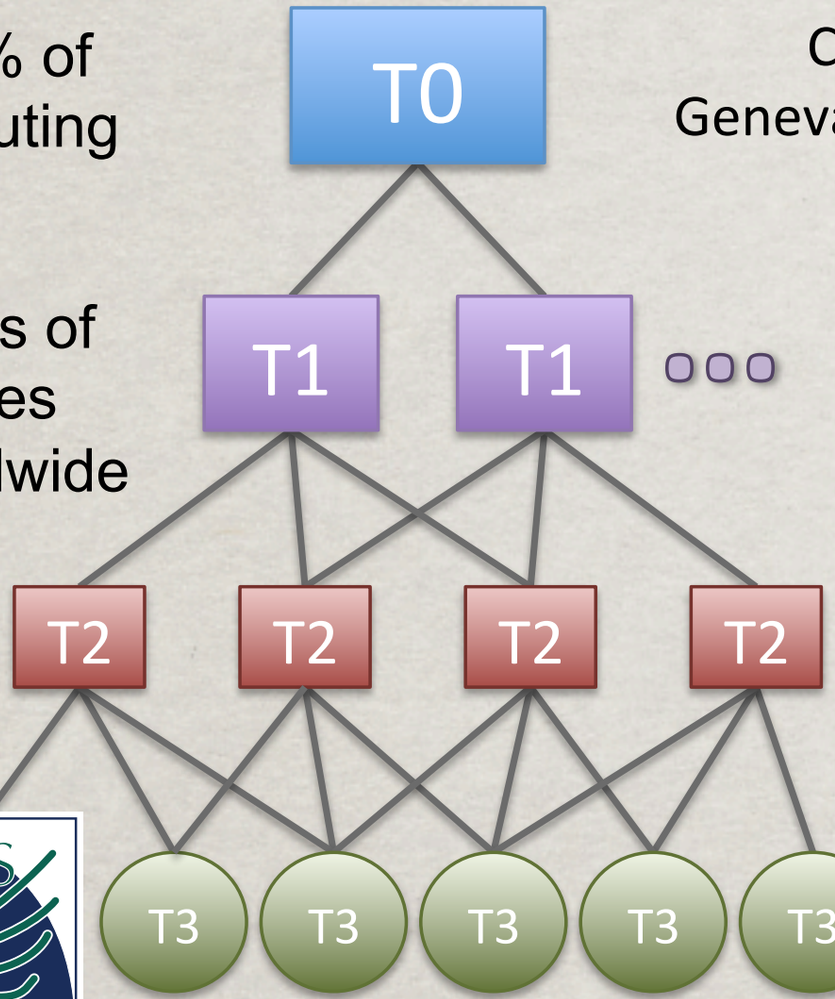
- Over 140 T2s worldwide
  - typically >1000 cores, ~ 1 PB of storage
- T3s provide local research computing
  - most part of global grid infrastructure

- Data movement by LHC private optical network
  - Total data flow of >70Gb/s achieved
- Automatic data request/transfer between sites
  - "PhedEx" (CMS)
- xrootd-based I/O redirection:
  - use network to open remote files directly
- Global Database for description, tracking

Computing Tier Model

<20% of computing

~10s of sites worldwide



CERN Lab
Geneva, Switzerland

National Labs
(Fermilab, etc.)

Universities
(MIT, Wisconsin, Nebraska, Purdue, etc.)

Universities
(ND, Colorado, UMD, OSU, etc.)

mostly used for central processing/ reprocessing of data, generation of simulated data

mostly used for user analysis, with some production of simulated data

- Over 140 T2s worldwide
  - typically >1000 cores, ~ 1 PB of storage
- T3s provide local research computing
  - most part of global grid infrastructure

HEP Data is not just one dataset: it is a huge raw dataset plus many smaller datasets produced by unique further processing

Data preservation in HEP

- What to do with 100s of PB of data?
  - Irreplaceable resource
  - should be preserved, somehow, for the future
- DPHEP Working Group
  - Convened by International Committee on Future Accelerators (ICFA)
  - ~ 100 members from different HEP experiments, Labs
  - Two Reports:
    - DPHEP-2009-00, http://arxiv.org/pdf/0912.0255
    - DPHEP-2012-01, May 2012, arXiv:1205.4667v1
  - Conclusions:
    - "an urgent and vigorous action is needed to ensure data preservation in HEP"
    - "A clear and internationally-coherent* policy should be defined and implemented"

* still waiting…

DPHEP Data Tiers & Uses

Preservation Model		Use Case	
1	Provide additional documentation	Publication related info search	Documentation
2	Preserve the data in a simplified format	Outreach, simple analyses	Outreach/Science
3	Preserve the analysis level software and data format	Full scientific analysis, based on the existing reconstruction	Technical Preservation Projects/Science
4	Preserve the reconstruction and simulation software as well as the basic level data	Retain the full potential of the experimental data	Technical Preservation Projects/Science

- Useful as a point of discussion
- Growing recognition that simple data preservation, or even data and software preservation, is insufficient
  - focus on means of more complete “knowledge preservation”

Knowlege Preservation

- Current efforts exist for Tiers 1 and 2:
  - supplementary INSPIRE content gives more complete information for publications (http://inspirehep.net/)
  - outreach efforts using Tier 2 data already
    - Also: **RECAST**: re-run analysis given new Monte Carlo specified by outside queries (JHEP **1104** (2011) 038 [arXiv:1010.2506])
- Serious work needed for Tiers 3 and 4
  - necessary within experiments themselves to preserve their own data for future analysis
  - outreach/public access component could be added in parallel

CMS Data Preservation

- The CMS Experiment has approved a Data Preservation and Access plan
  - first LHC experiment to do so
    - some other LHC experiments have also adopted similar policies
  - prompted by US groups needing to define “Data Management Plans” for the funding agencies
- Under Collaboration Board oversight, calls for:
  - appointment of “Data Preservation Coordinator”
    - Kati Lassila-Perini
  - “prompt” public release of Tier 1 and Tier 2 data
  - delayed release of Tier 3 data (Tier 4 will not be released)
    - hopes to release some fraction of reconstructed 2010 data in 2014
  - Creative Commons CCO waiver for released data

http://creativecommons.org/publicdomain/zero/1.0.

Coordination

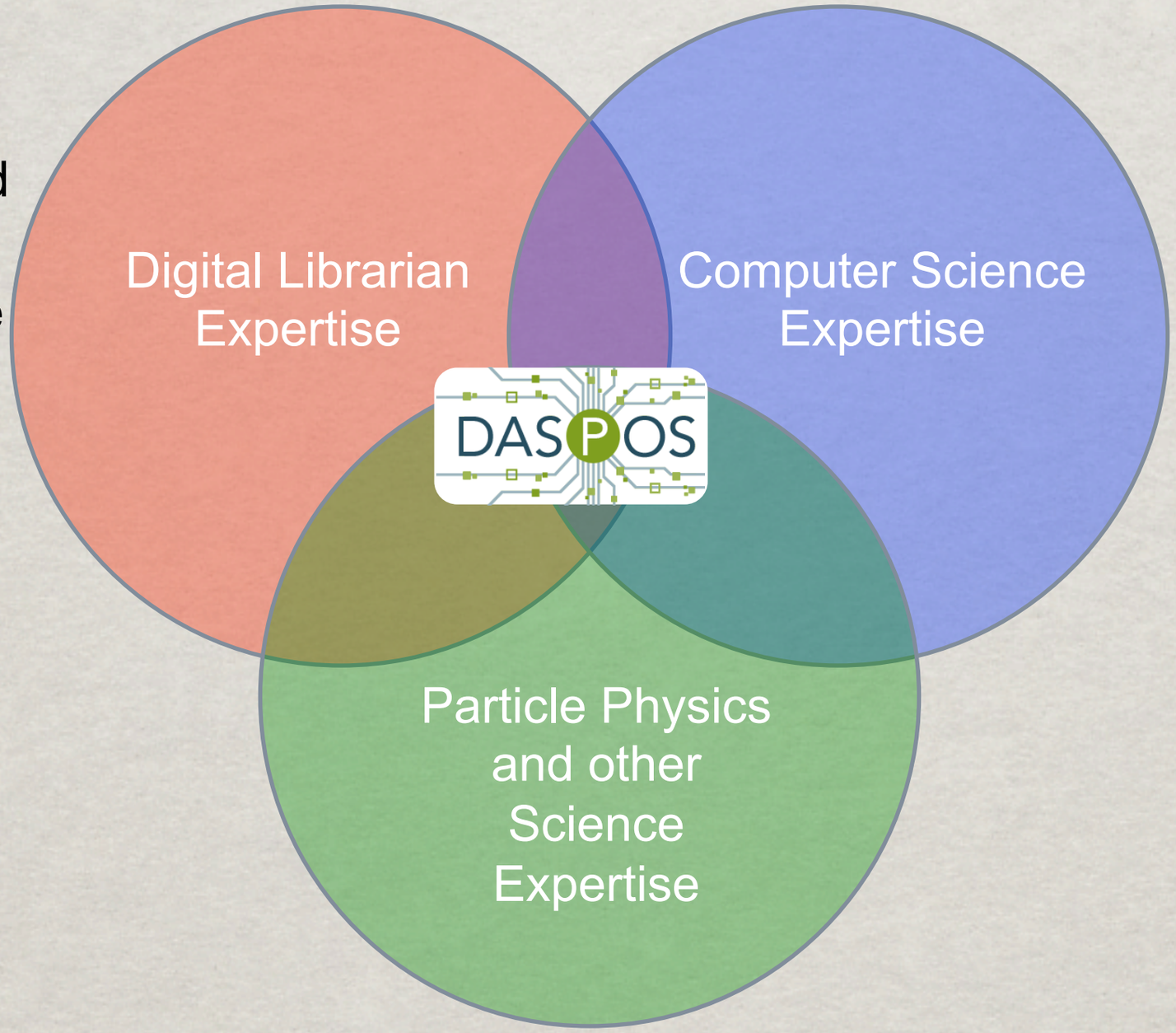
- Next: implementation of technical infrastructure, policy, etc. to make data available
  - guidance, but no FTEs (yet) from DPHEP
    - suggestions of overall structure, but no concrete implementation plans
  - CMS will rely on internal expertise, coordinate with external agencies
    - would be most efficient to build infrastructure that is re-useable by other experiments, or even other disciplines
- Several efforts in this area exist or are in the pipeline
  - e.g. CMS/CERN IT development of analysis preservation portal underway

DASPOS

- Data And Software Preservation for Open Science
  - multi-disciplinary effort recently funded by NSF
  - Notre Dame, Chicago, UIUC, Washington, Nebraska, NYU, (Fermilab, BNL)
- Links HEP effort (DPHEP+experiments) to Biology, Astrophysics, Digital Curation
  - includes physicists, digital librarians, computer scientists
  - aim to achieve some commonality across disciplines in
    - meta-data descriptions of archived data
      - What's in the data, how can it be used?
    - computational description (ontology development)
      - how was the data processed?
      - can computation replication be automated?
    - impact of access policies on preservation infrastructure

DASPOS Overview

- How to catalogue and share data
- How to curate and archive large digital collections
- Ontology/ Metadata expertise



- How to build databases and query infrastructure
- How to preserve software and functionality
- How to develop distributed storage networks

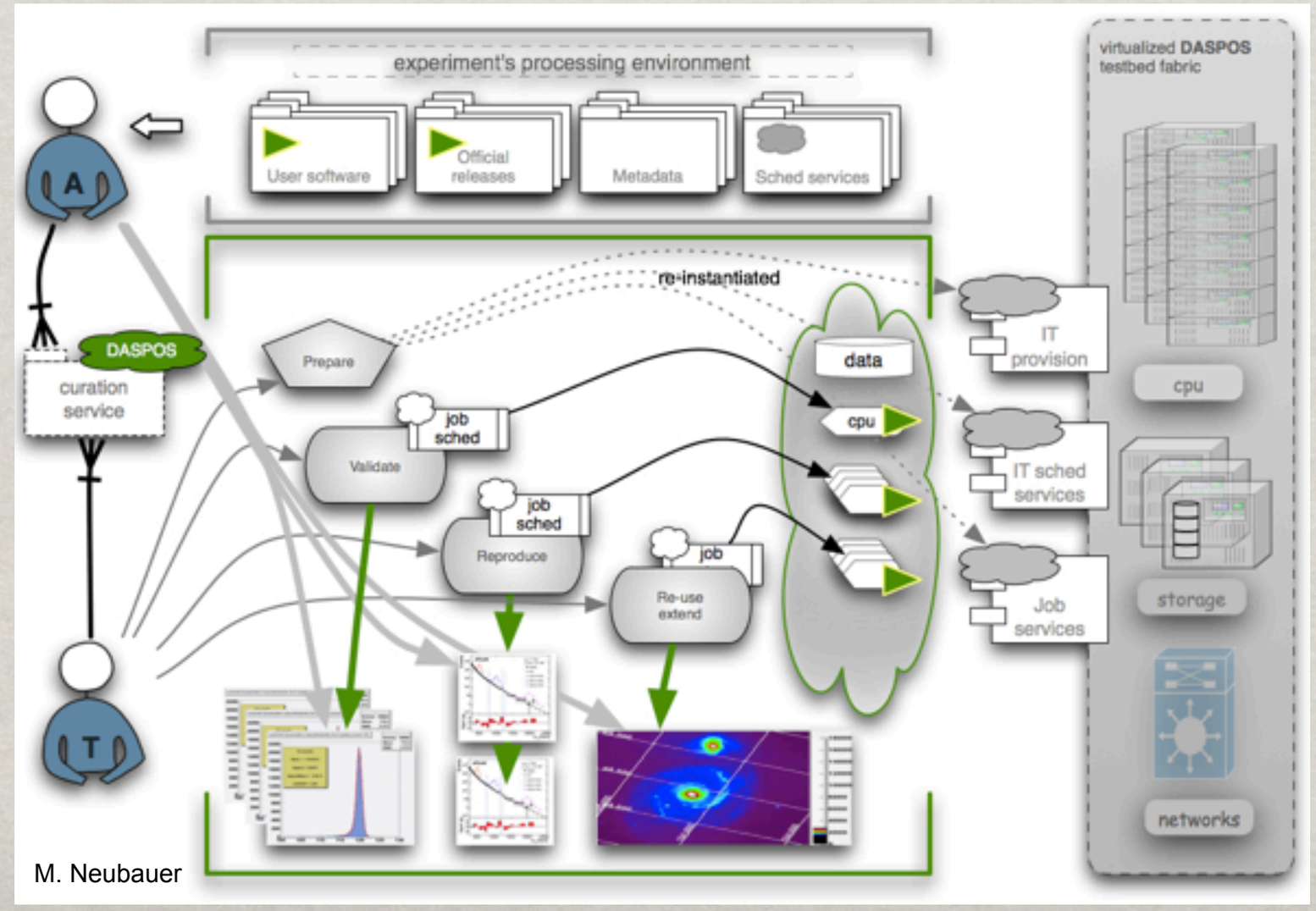
- What does the data mean?
- How was it processed?
- How will it be re-used

DASPOS Activities

- In parallel, will build test technical infrastructure to implement a knowledge preservation system
  - "Scouting party" to figure out where the most pressing problems lie, and some solutions
    - incorporate input from multi-disciplinary dialogue, use-case definitions, policy discussions
  - Will translate needs of analysts into a technical implementation of meta-data specification
  - Will develop means of specifying processing steps and the requirements of external infrastructure (databases, etc.)
    - automatic instantiation of workflows?
  - Will implement "physics query" infrastructure across small-scale distributed network
  - End result: "template architecture" for data/software/knowledge preservation systems

DASPOS

- Final Milestone: “Curation Challenge”
  - an analyst will reproduce some physics result using only curated information
  - success defined by external auditing team



M. Neubauer

DASPOS Activities

- Two “Fact-Gathering” Workshops in 2013:
  - HEP-centric (March, at CERN, joint with DPHEP):
    - Can experiments agree on the types of data they would like to preserve?
      - software and analysis preservation, in addition
    - Can we begin to define some global metadata?
  - Multi-Disciplinary (July, Indianapolis, at JCDL):
    - What are problems, use cases in other fields? (Astro, Bio, etc.)
    - What is the commonality between these and HEP?
      - can we think about common infrastructure?

Workshop I

Establishment of Use Cases for Archived Data and Software in HEP

- **Attendees:** Participants from all of the HEP experiments considering long-term data preservation and access issues (4 LHC experiments, BaBar, D0/CDF); joint with DPHEP
- **Organizers:** A team consisting of the digital librarians from University of Chicago and Notre Dame and HEP physicists from Notre Dame, University of Chicago and University of Illinois at Urbana-Champaign
- **Location:** CERN
- **Purpose:** (i) Establish use cases for data access and re-use, especially for the larger DPHEP data tiers, since this will be a primary driver of the preservation architecture, (ii) define what data and associated information supports the use cases, and (iii) identify a preliminary set of metadata that would serve the needs of the HEP community in accessing the various forms of archived data/algorithms.
- **Inputs**: Questionnaire based on Data Curation Toolkit addressing: use-case scenarios for data re-use and archiving including intended audiences; current practices and policies for data use; data types; and high-level description and rights metadata necessary for discovery and access.

Workshop I

Other Agenda Topics

- prospects for commonality in outreach formats
  - issues being driven by CMS outreach program in Finland that will release "real" data
  - open question, limited by available people
- high-level analysis preservation
  - HepData ⇔ Rivet ⇔ Theorists' Analysis Archive
    - Can an extension of the HepData/Rivet infrastructure serve as a common platform for high-level analysis preservation?
    - Now discussing Rivet back-end interface for RECAST (NYU/UNL)

Outcomes

- Analysis of commonality in data processing, analysis chains
  - first look at abstraction of workflow steps
    - still looking at how best to represent this so that it will be useful for other disciplines

Workshop II

Survey of Commonality with other Disciplines

- **Attendees**: Broad participation from many NSF supported science efforts.
- **Location:** Satellite workshop at IEEE/JCDL (Joint Conference on Digital Libraries) in Indianapolis, IN, July 2013.
- **Organizers:** Digital librarians, HEP, and Computer Scientists
- **Purpose:** (i) Explore areas of commonality and difference, (ii) identify common metadata standards that could be designed to allow generic access and indexing of cross-disciplinary research data, and (iii) identify cross-disciplinary services that would support data preservation (e.g. software repositories).
- **Inputs:** A panel discussion with many cross-disciplinary participants; break out sessions targeting sub-topics of interest within data preservation.

Workshop II

DASPOS Data and Software Preservation for Open Science

www.daspos.org

ABOUT | PEOPLE | WORKSHOPS | RESEARCH | REPORTS

DASPOS Workshops

- Use Cases for Archived Data and Software in HEP
- **Commonality with other Disciplines**
- Data Model and Query Semantics
- Software Sustainability
- Preservation Policy
- Technical Storage Architectures

WORKSHOP 2

Survey of Commonality with other Disciplines

Date: Thursday, July 25, 2013

Agenda: Click here for schedule of events

Purpose: (i) Explore areas of commonality and difference, (ii) identify common metadata standards that could be designed to allow generic access and indexing of cross-disciplinary research data, and (iii) identify cross-disciplinary services that would support data preservation (e.g. software repositories).

Inputs: A discussion framework similar to that of Workshop 1 will be developed and will also be used to conduct individual or small group discussions with targeted colleagues not available for the workshop (e.g. the research staff involved in archiving the Sloan Digital Sky Survey-II).

Outcomes: Provide extensive information about preservation efforts in other disciplines.

Panel Discussion Video: Click here for video recording of panel discussions

Round Table Discussion Video: Click here for video recording of the round table discussions

Discussion: Round Table Discussion Notes

Workshop II

Panel Participants

Dr. George O. Strawn
Director, National Coordination Office (NCO)
National Science Foundation

Dr. Reagan W. Moore
Director, Data Intensive Cyber-Environments
Chief Scientist, RENCI
Professor, School of Information and Library Science
University of North Carolina at Chapel Hill

Dr. Chris Mattmann
Senior Computer Scientist
NASA Jet Propulsion Laboratory

Dr. Don Petravick
Principal Investigator
Dark Energy Survey Data Management System
National Center for Super Computing Applications
University of Illinois at Urbana-Champaign

Dr. Matthew Mayernik
Research Data Services Specialist
NCAR Library / UCAR Integrated Information Services
National Center for Atmospheric Research (NCAR)
University Corporation for Atmospheric Research (UCAR)

Prof. Michael Witt
Associate Professor of Library Science
Interdisciplinary Research Librarian
Purdue University

Dr. Micah Altman
Director of Research
Head/Scientist, Program on Information Science
MIT Libraries

Dr. Clifford Lynch
Executive Director
Coalition for Networked Information

Dr. Line Pouchard
Information Scientist
Scientific Data Group
Oak Ridge National Laboratory
US Department of Energy

Workshop II

Round-Table Topics:

- Policy based Data Management
- Reuse of Big Data, complex digital objects & Scientific Workflows
- Software & Algorithmic Preservation for Open Science

Outcomes:

- Common themes:
  - Provenance of data, Workflows, definition of workflow, reproducibility
  - Software preservation
  - Policy based data management
  - Metrics, citations
  - Economics
- Other tools/concepts to explore (understand uses in other disciplines):
  - Taverna, MyExperiment, iRODS, etc.

Technical Advances

Software/Workflow/Environment Preservation (Chicago)

- Prototype Workflow capture/validation on one step in the analysis chain
  - "Slim/Skim" step in ATLAS environment
  - Start with actual ATLAS analysis workflow and data stored at a Tier 2
  - capture user workflow process in service database, validate against existing results
  - Next:
    - Execute using existing virtualized execution infrastructure at the ATLAS Midwest Tier 2 Center
    - Deploy captured code, data and environment on OpenStack managed preservation platform
    - Develop an "Is the data still alive" monitoring infrastructure

DØ Effort (Washington)

- Establishing analysis platform on virtual machines outside of FNAL infrastructure
- have one VM running SL6, second as a cvmfs server to provide DØ software
- experience will translate to LHC efforts

Analysis Preservation Case Study

- Resource Survey for analysis code in various preservation scenarios
  - (H. Meng, M. Wolf, P. Ivie, A. Woodard, M.H., D. Thain, Notre Dame)
  - Analysis run inside instrumented version of Parrot* to provide complete accounting of necessary external resources
  - Example of source/data usage:

Name	Location	Total	Used
CMSSW code	CVS	88.1GB	6.3MB
Tau source	Git	73.7MB	6.7MB
Configuration	CVMFS	7.4GB	103MB
Linux commands	localFS	110GB	68.4MB
HOME dir	AFS	12GB	32MB
Misc commands	PanFS	155TB	1.6MB
Ntuples data	HDFS	11.6TB	20GB

  - Independent preservation “Package” has been run on different cloud platforms (ND, Amazon) with validated results

* D. Thain, P. Livny, Scalable Computing: Practice and Experience, 6(3):9–18, 2005

Workshops for 2014

- "Virtual Workshop" on Ontologies and Provenance-Driven Workflows
  - ~too complicated to get all of the experts in one place simulataneously
  - meet with/interview smaller groups over the course of a few weeks/ months
- DPHEP: June software preservation workshop? Joint with DASPOS?
- Data Query Model Workshop
  - More CS-driven
  - ~Fall

Challenges

Internal:

- Many, many things to learn
  - merely just figuring out a common vocabulary took a couple of months
  - translates to slower-than-predicted progress

External:

- Coordination
  - tens of data preservation efforts scattered across the globe in different disciplines; growing realization that “knowledge preservation” is the real issue
    - at the highest levels, all have different requirements
    - at the most basic levels, we are all doing the same thing
      - arguably, most other disciplines are “less complicated” than HEP
      - a framework with sufficiently generic pieces should be able to be adopted in a simplified manner by other disciplines
- Would like to see joint workshops (DoE/NSF/NASA -sponsored) on these issues

Conclusions

- Data Preservation and Access will be major issues
  - merely preserving data and analysis knowledge for re-use within the experiments will be a major challenge
  - preserving just the data without all of the associated software, and documentation of expertise is not useful
  - No technical infrastructure in place to handle public release, access to data
- DASPOS project could help guide development of full-scale systems
- Commonality between disciplines must be exploited lest we drown in our own data streams!

Other Global Initiatives

(not an exhaustive list)

- RDA: Research Data Alliance
  - partially funded by NSF
  - kick-off meeting next week in Gothenburg, Sweden http://rd-alliance.org/invitation/
- SCIDIP-ES
  - EU-funded (7th Framework) Earth Sciences project
  - http://www.scidip-es.eu/
- Horizon 2020 Funding for European Projects
  - now coming into focus
  - will almost certainly include data/software preservation
  - also meeting next week in Brussels