

Data Preservation and Long Term Accessibility

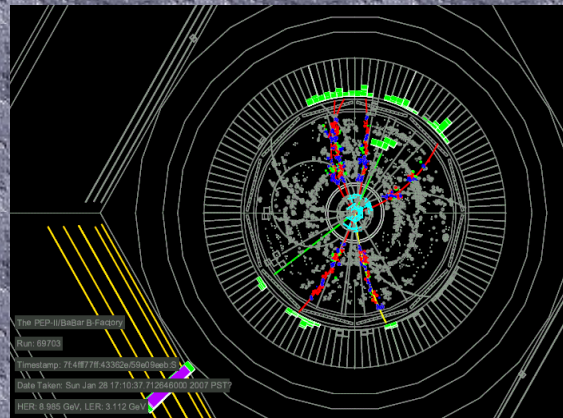
@ First International Conference on Frontier in Diagnostic Technologies

@ Frascati, Italy

26 November 2009

by

Homer Neal ()
BaBar Computing Coordinator



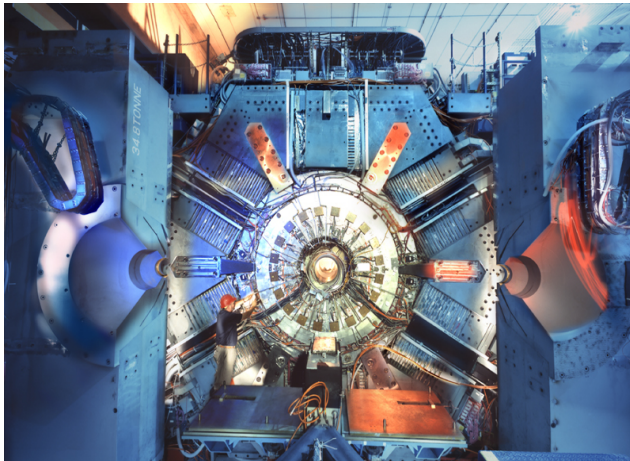
The rise of the importance of data preservation in high-energy physics

- Size, complexity, cost and time-scale of experiments has greatly increased
- Period between an experiment and its successor has become long and uncertain
- Data is too rich and analyses take too long to be fully exploited during the lifetime of the centralized collaborations

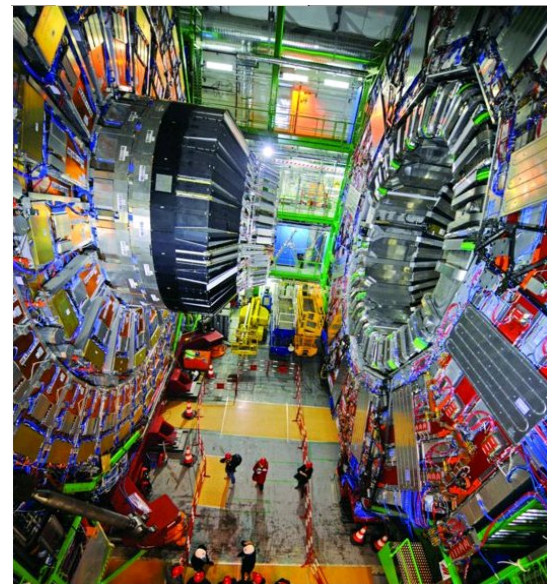
Examples: High-Energy Physics Experiments

- HEP experiments have become very complex, taking years to plan and build and they result in a profusion of data taking decades to fully analyze

BaBar:

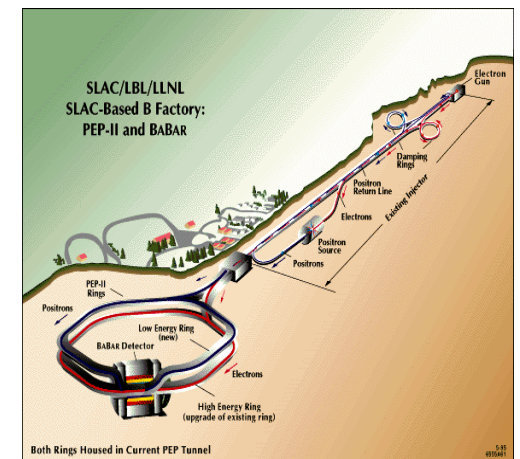


CMS:



The HEP landscape (colliders)

- **HERA:** end of collisions in 2007
 - + No follow-up before at least two decades
- **B-factories**
 - + Next generation around 2017
- **Tevatron**
 - + A majority of the physics program will be taken over at LHC
 - + However: p - \bar{p} is unique, no follow-up foreseen
- **In general:**
 - + Present HEP experiments have “cycles” of more than 5-10 years



An Example: BaBar

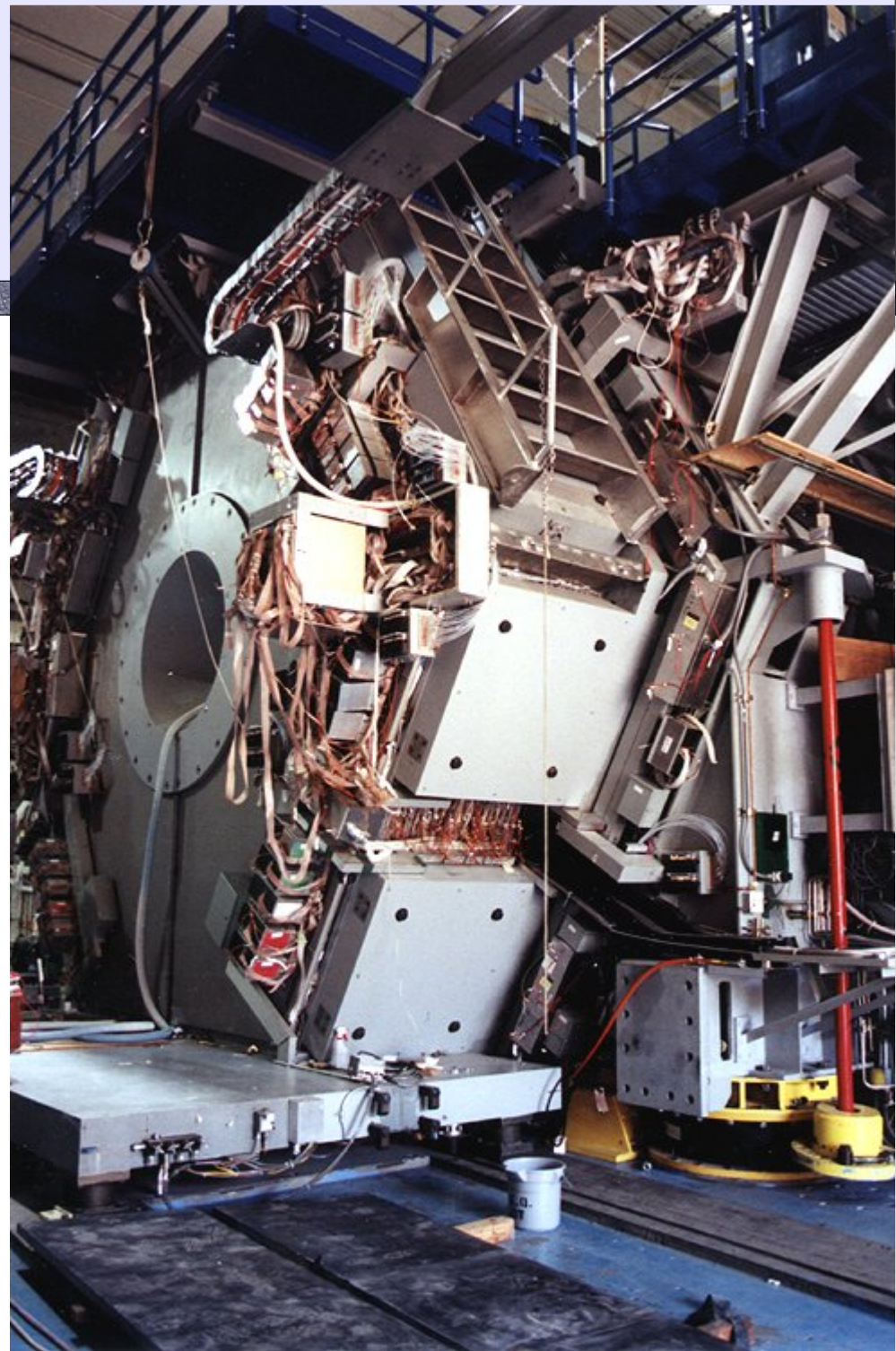
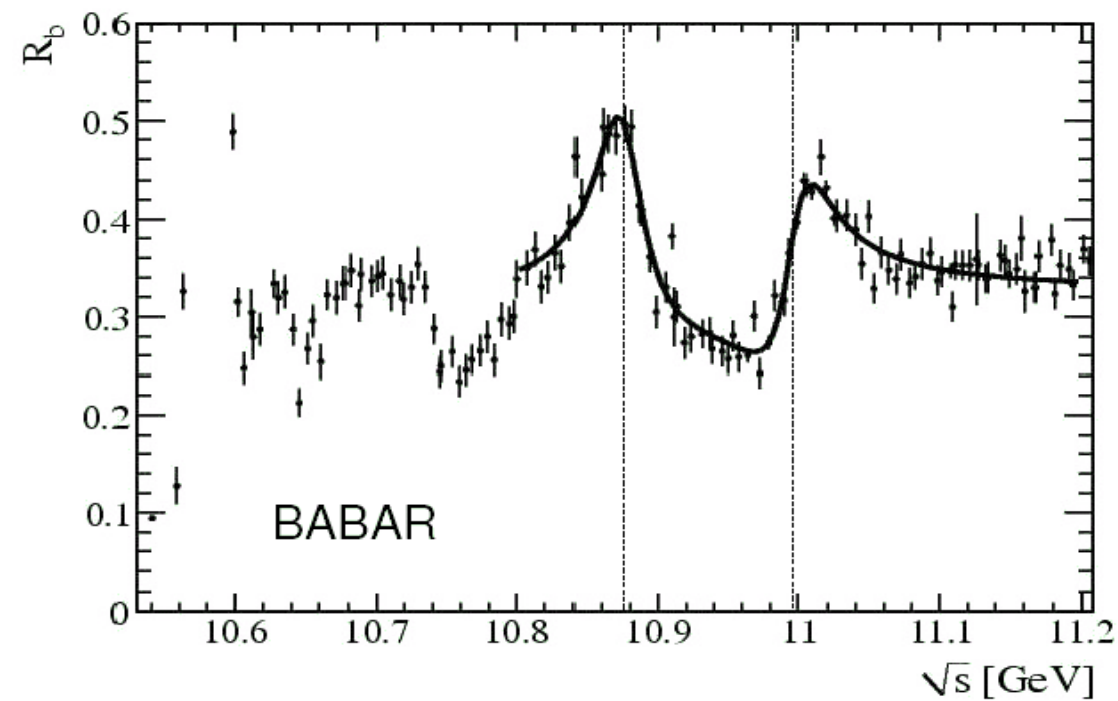
conceived 15 years ago (LOI 1994)

**collected frontier HEP data over
almost a decade**

**through the efforts of 10 countries
and**

~600 collaborators

April 7th, 2009 marked
the one year anniversary
of the end of BaBar data
taking



Recent publication from the Jade Experiment was a wake-up call

- The Jade HEP experiment last collected data in 1985

- Yet

arXiv.org > hep-ex > arXiv:0810.2933

High Energy Physics - Experiment

Study of moments of event shapes and a determination of α_s using \sqrt{s} annihilation data from \sqrt{s} Jade

Christoph Pahl, Siegfried Bethke, Stefan Kluth, Jochen Schieck, for the JADE collaboration

(Submitted on 16 Oct 2008 (v1), last revised 4 Mar 2009 (this version, v2))

Data from \sqrt{s} annihilation into hadrons, collected by the Jade experiment at centre-of-mass energies between 14 GeV and 44 GeV, are used to study moments of event shape distributions. Models with hadronisation parameters tuned to the LEP 1 precision data provide an adequate description of the low energy data studied here. The NLO QCD calculations, however, show systematic deficiencies for some of the moments. The strong coupling measured from the moments which are reasonably described by NLO QCD, $\alpha_s(m_Z) = 0.1287 \pm 0.0007 \text{ (stat)} \pm 0.0011 \text{ (expt)} \pm 0.0022 \text{ (had)} \pm 0.0075 \text{ (theo)}$ is consistent with the world average.

Comments: 14 pages, 9 figures, EPHJA style, acc. by Eur.Phys.J.C. New version similar to published version

Subjects: **High Energy Physics - Experiment (hep-ex)**

Journal reference: Eur.Phys.J. C60:181-196,2009; Erratum-ibid. C62:451-452,2009

DOI: 10.1140/epjc/s10052-009-0930-5 10.1140/epjc/s10052-009-1032-0

Report number: MPP-2008-135

Cite as: **arXiv:0810.2933v2 [hep-ex]**

- We asked ourselves ... might there still be a need to analyze BaBar data in 2041?

JADE

JADE data and software

- 1995: „private“ (*neither collaboration nor lab*) initiatives to :
 - rescue data from original archive tapes and copy them onto more modern media (IBM cartridges & Exabyte)
(J. Olsson @ DESY)
 - reanalyse data using modern (LEP-like) methods and observables plus improved theoretical calculations
(S. B. and P. Movilla-Fernandez @ RWTH Aachen)
 - revitalise JADE software on modern computer platforms to enable generation of new MC data files
(P. Movilla Fernandez, J. Olsson)
- so far, the only example of reviving and still using 25-30 year old data & software in HEP
- since 1996, O(10) publications, O(10) conf. contributions; no competition in e^+e^- data analysis at $E_{\text{cm}} \sim 14 \dots 200 \text{ GeV}$

BaBar: Some notable events since the end of data collection

July 9, 2008 -
Physicists
Discover New
Particle: The
Bottom-most
"Bottomonium"

The Nobel Prize in Physics 2008
and the B FACTORIES SLAC

The Nobel Prize in Physics 2008 was awarded to
Masatoshi Hagiwara
High Energy Accelerator Research Organization (KEK),
Tsukuba, Japan
&
Terukazu Maeno
Tohoku Gakuin University, Yamaguchi Institute for Theoretical Physics (YITP),
Kyoto University, Kyoto, Japan

"for the discovery of the origin of the broken symmetry which predicts the existence of at least three families of quarks in nature"

Broken Symmetries Predicted Extra Quarks

Matter and antimatter are nearly exact opposites of each other. But this near-perfect symmetry is broken in nature as we observe it. In 1972, Kobayashi and Maskawa discovered that the root of the mystery could be explained by the properties of quarks, the fundamental constituents of protons and neutrons. But only if there were three more types of quarks than had previously been observed. At that time, experimenters had seen the up, down, and strange quarks, but the charm, bottom, and top would not be discovered until later.

B Factory Experiments Confirmed the Predictions

Experiments of the B factories in the United States and Japan in the early 2000s made detailed investigations of billions of high-energy particles containing bottom quarks. International Collaborations at the B factories made numerous measurements of the parameters of the Cabibbo, Kobayashi, and Maskawa (CKM) mixing matrix and confirmed the precise links of these with the observed differences between matter and antimatter. The B factories each consist of an accelerator and a particle detector. At the SLAC National Accelerator Laboratory in California, USA, the PEP-II accelerator provides the collisions observed by the BaBar detector. At KEK in Tsukuba, Japan, the KEK-B accelerator supplies the Belle detector with the particles needed for these studies.

"Please accept our deepest respect and gratitude for the B factory achievements. In particular, the high-precision measurement of CP violation and the determination of the mixing parameters are great accomplishments, without which we would not have been able to earn the Prize."

小林 尊 (Masatoshi Hagiwara)
若川 敏彦 (Terukazu Maeno)

BaBar
PEP-II
Belle
KEK-B



BaBar Collaboration Caps Meeting Week with 400th Scientific Publication

by Lauren Knoche

The BaBar Collaboration reached another milestone Tuesday—just in time for celebration during the [group's meeting](#), which ends today at SLAC. The collaboration published its 400th paper Tuesday, less than nine years after publishing its first in 2001. That's an average of one publication per week, every week, for nearly nine years straight.

"I do not know of any other collaboration that has achieved such a production rate of outstanding quality science in particle physics, it is really something rare," said BaBar spokesperson Francois Le Diberder.

The [milestone paper](#) was published online Tuesday and appears in the November 1, 2009 issue of *Physical Review D* (Volume 80, Number 9). The study examines differences in the rates at which subatomic particles called *B*-mesons and their antiparticle partners, *B*-mesons, decay to related particles called "charm" and "strange"



(Photo by Brad Plummer.)



- We foresee >100 papers being published over the next several years and about 35 papers to be published after the loss of the existing large BaBar computing infrastructure provided by SLAC, other TierA sites and ~20 universities.

There are 35 (20 of these are “MUST DO”) analyses already foreseen to be users of the archival system. Among these are:

- Initial State Radiation Physics These are very delicate analyses aiming at the determination of the cross-sections : $e^+e^- \rightarrow 3\pi^0$, $K_S K_L$, $K_S K_L \pi$, $K_S K_L \pi \pi$, 7(8) pions with the goal of providing a complete set of cross-sections for $e^+e^- \rightarrow hadrons$, at low energies, where the contribution to g-2 is the most sensitive.
- Charm Physics : rare decays, many body final states, etc.
- Y(nS) Physics like $Y(mS) \rightarrow \pi\pi$ (or $\gamma\gamma$) $Y(nS)$
- sin2beta analyses (sic: a few could be done, although they are not high profile) like $Y(4S) \rightarrow J/\psi\rho$

In addition, there is a strong likelihood that new models will need to be tested in the archival period and that checks against the BaBar data will need to be done by new projects such as SuperB.

How soon could BaBar be superceded

It will take ~8 years for SuperB to be approved by the governments and funding agencies, constructed, commissioned and obtain a dataset more significant than BaBar's.

There may also be a need to validate initial results against those of BaBar and Belle

BaBar's approach

- Preserve all capabilities of doing full analyses from inception to publication including some new simulation, access to raw data and latest reprocessings of the detector and simulated data, all databases, documents, simplified/improved interfaces for job submissions and introduction of new models, frozen trusted reconstruction in an

The BaBar Archival System

- A modest many core box with enough storage for the data, databases, documents, releases, and new generators and simulated signals.
- Running a current platform that will be protected by living in virtualization containers



Status of BaBar Effort

- Migration to SL5 and ROOT 5.22 ~ completed
- Simplified job managers ~ completed
- Tests using virtualization and virtual machines in batch on a single node successful
- Working on acquiring a prototype archival system
- Working on finding analysts to test it and provide feedback

BaBar Data Preservation

BaBar & Belle collaborating

In real life: $B^\pm \rightarrow K^\pm \pi^+ \pi^\mp$ decay

Same exercise with the master at Caltech (Los Angeles), one worker at SLAC and the other worker at ccin2p3 (France) with secured connections.



Fit performed in a bit less than 20 minutes. Note that we had slow 32-bits machines, a fit SLAC-SLAC-SLAC took almost 4 minutes

It worked very well

B. Echevard / E. Ben-Haim

BaBar Collaboration Meeting / November 2009

p. 11

Virtualization

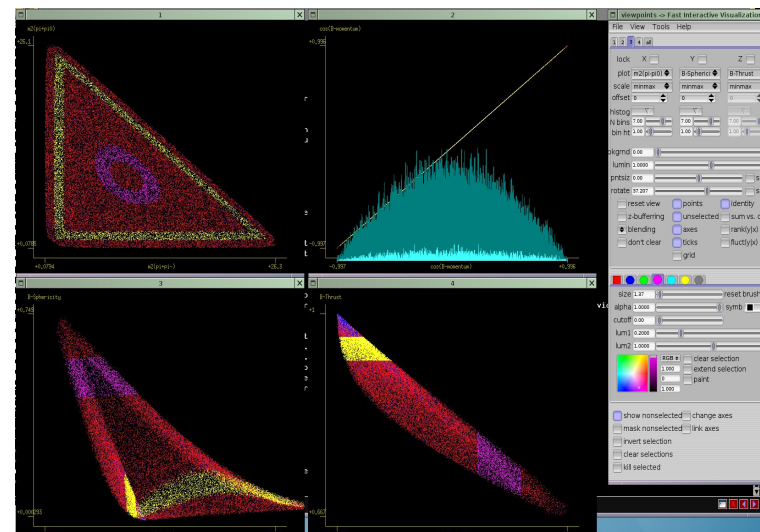
- The status at SLAC: 4 SL5.3 VMs installed on yakut13.
- VMs were added to a special batch queue.
- SL5 migration checks to be done on virtual machines.
- Simultaneously validates the SL5 build and the VM technology.



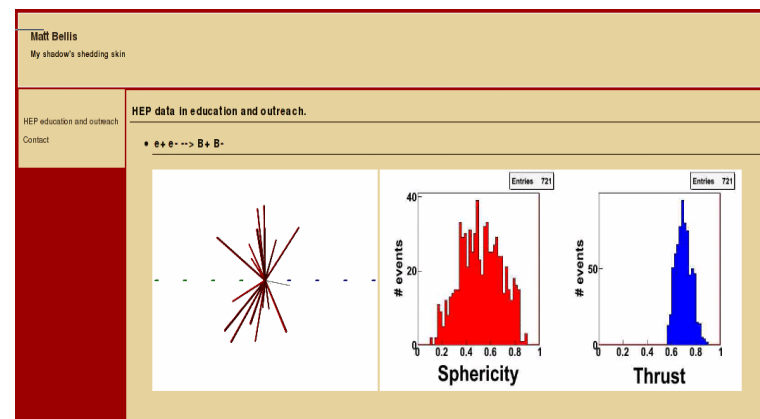
June 22, 2009

Long Term Data Access

6

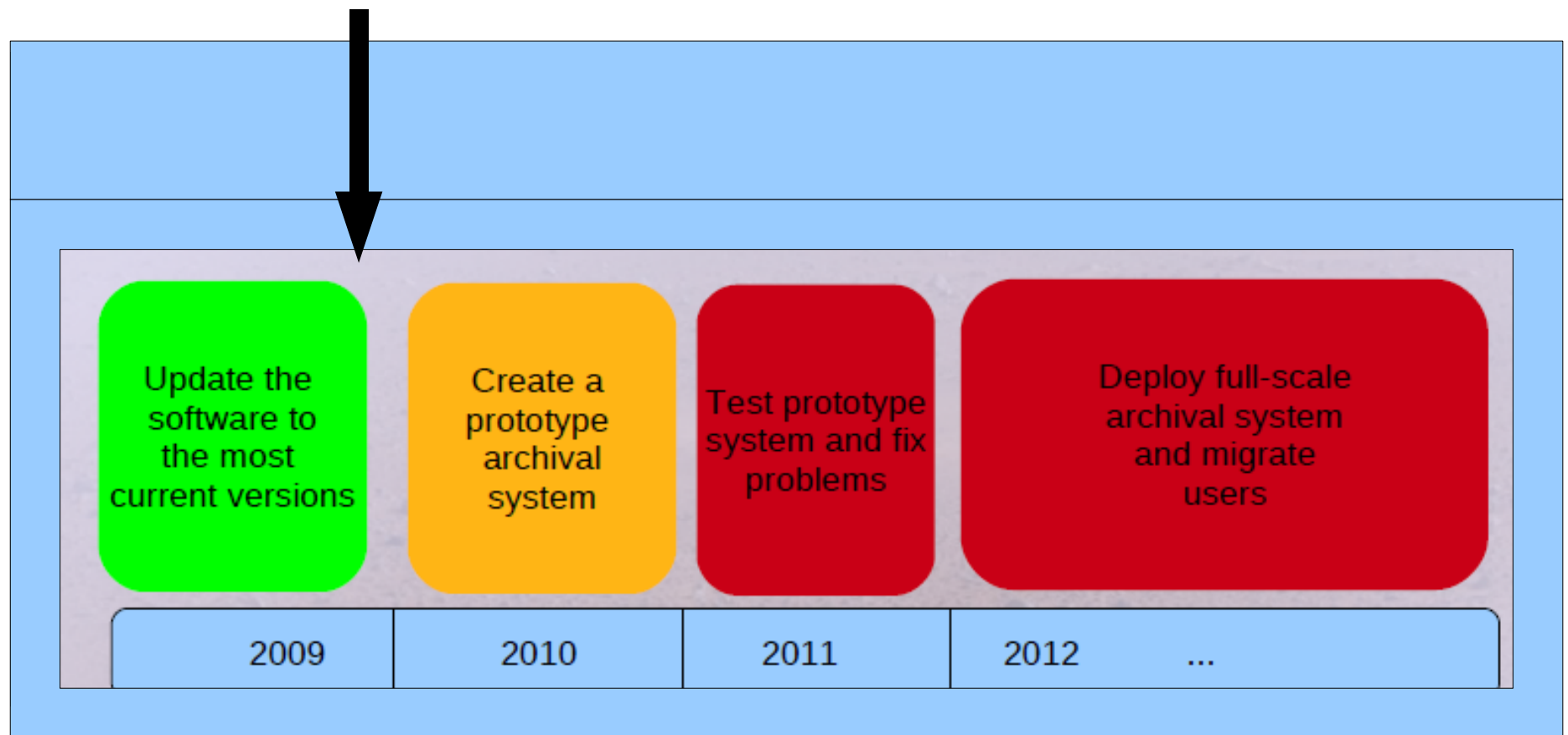


Outreach tools/data already being used in classrooms



Also major advancements in the use of cloud computing

BaBar's Plan



Concerns

- Virtualization technology lifetime (Xen, VMware,...)
- One box as a database, batch, interactive analysis, web server ... will we have a melt down
- Data flow from a many core system – sufficient bandwidth?
- Integrity of results produced from publicly accessible HEP data

IMPORTANT!

- Data preservation was not incorporated into the computing model for BaBar. This is a recent effort.
- BaBar and other experiments would have greatly benefited from guidance on data preservation strategies in the very early stages of the experiment.
- BaBar data store is complex ... too complex
- Other experiments are in the process of making similar decisions that will make it painful later to preserve their data.

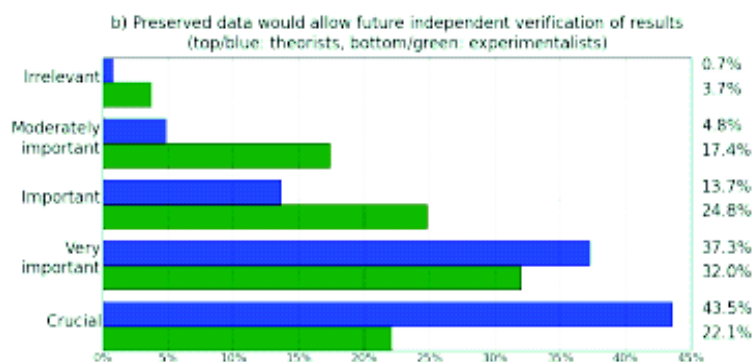
Guidance even more important for the upcoming experiments

- BaBar raw data output rate 530 KB/s
- ATLAS raw data output rate 300 **MB/s**
- LSST calibrated data output rate 60000 MB/s every 40 seconds
- LHC and LSST data will be unique and contain rich new physics that will take decades fully exploit
- The steps beyond LHC, SuperB, ILC are too far in the future for anyone to know when they might occur.

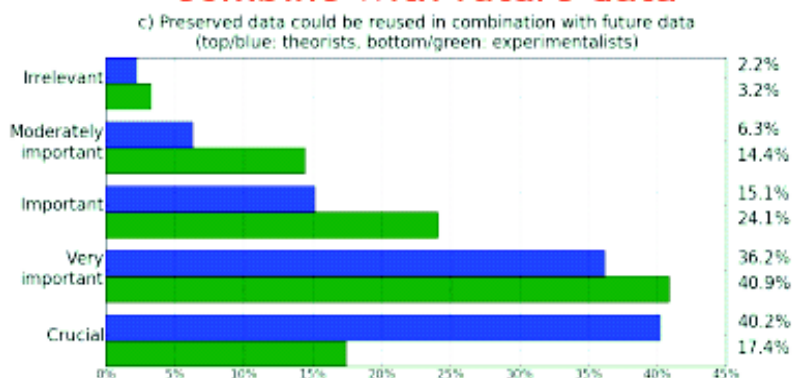
Data Preservation as seen by the HEP World

The importance of preservation

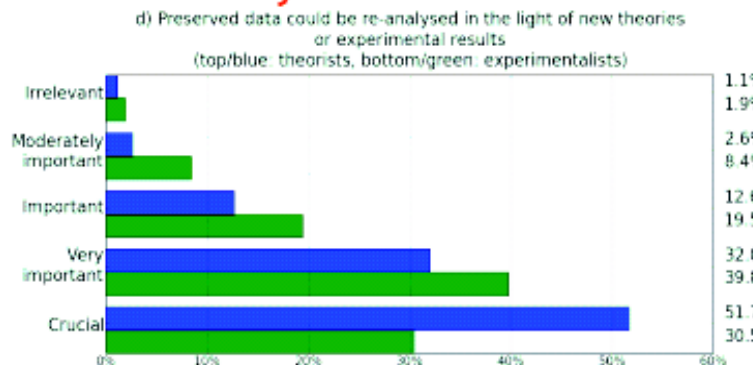
Future independent checks



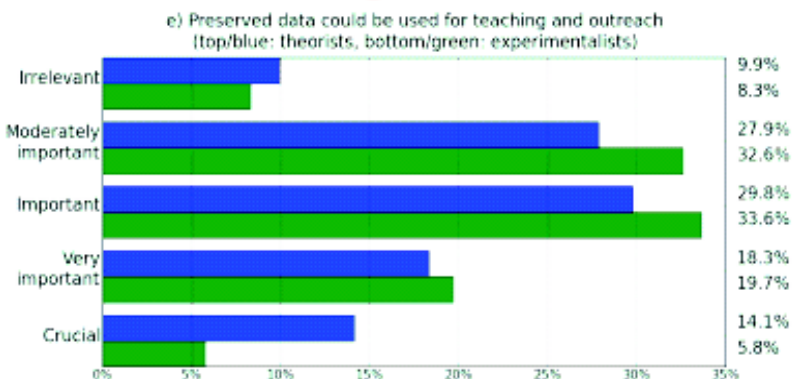
Combine with future data



Re-analyse for future theories



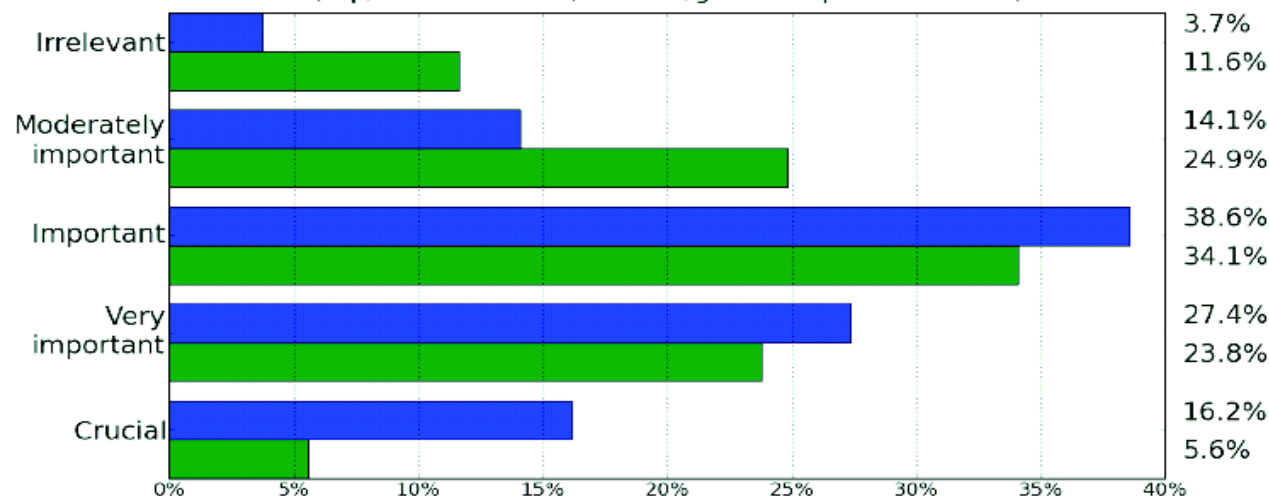
Teaching and outreach



Why to preserve? - Compiling results

How much importance would you attach to the following uses of preserved data ?

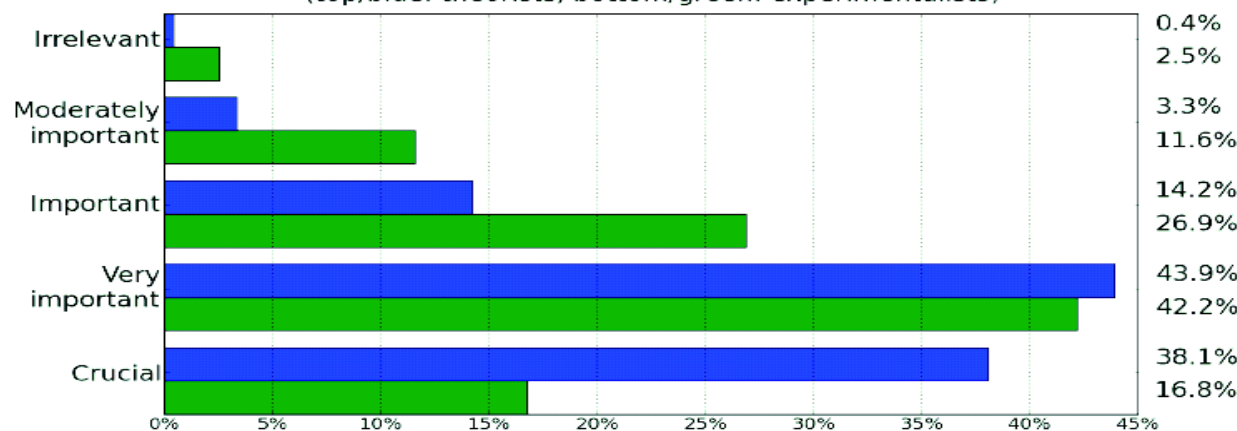
a) Compiling published results on a given subject (e.g. for a review)
(top/blue: theorists, bottom/green: experimentalists)



Why to preserve? - Testing new models

How much importance would you attach to the following uses of preserved data ?

b) Testing new models using preserved data
(top/blue: theorists, bottom/green: experimentalists)



See:

arXiv.org > cs > arXiv:0804.2701

Computer Science > Digital Libraries

Information Resources in High-Energy Physics: Surveying the
Present Landscape and Charting the Future Course

Anne Gentil-Beccot, Salvatore Mele, Annette Holtkamp, Heath B. O'Connell, Travis C. Brooks

(Submitted on 16 Apr 2008 (v1), last revised 22 Apr 2008 (this version, v2))



HEP Data Preservation Progression

First Workshop on Data Preservation and Long Term Analysis in HEP

DESY, Hamburg, Germany
Mon 26th - Wed 28th January 2009

Objectives of the Workshop

Review the physics objectives of data persistency in HEP
Exchange information on the analysis model used by HEP experiments
Address the hardware and software persistency issue
Review the funding programs and other existing international initiatives
Converge to a common set of recommendations for future experiments

<http://indico.cern.ch/conferenceDisplay.py?confId=43722>

Local Organizing Committee

Volker Gülzow (DESY)
Volker Gülzow (DESY-IT)
David South (TU Darmstadt)
Krzysztof Wozniak (DESY)

DESY-IT: Volker Gülzow (DESY)
H1: Cristian Diaconu (CPPM/DESY)
ZEUS: Tobias Haas (DESY)
FNAL/DoE: Amber Boehnlein (DoE)
FNAL-IT: Victoria White (FNAL)
D0: Dmitri Denisov (FNAL), Darien Wood (FNAL)
CDF: Jacobo Konigsberg (FNAL), Robert Roser (FNAL)
IHEP-IT: Gang Chen (IHEP)
BES III: Yifang Wang (IHEP)
KEK-IT: Takashi Sasaki (KEK)
Belle: Masa Yamauchi (KEK), Tom Browder (Hawaii)
SLAC-IT: Richard Mount (SLAC)
BaBar: Francois Le Diberder (LAL/SLAC)
CLEO: David Asner (Carleton)
CERN-IT: Frederic Hemmer (CERN)
CERN/PARIS: Salvatore Mele (CERN)

SLAC
NATIONAL ACCELERATOR LABORATORY

IHEP



KEK
KEK-NSL



DOE



BES III
BELLE

CLEO



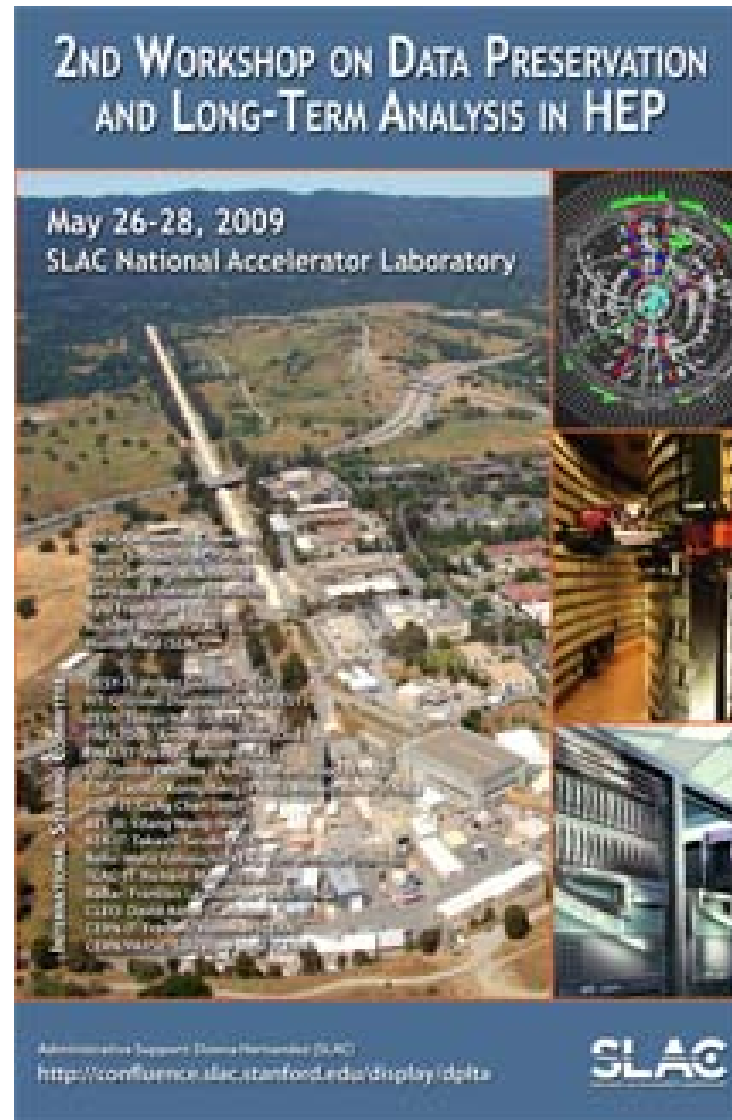
DESY-IT: Volker Gülzow (DESY)
H1: Cristian Diaconu (CPPM/DESY)
ZEUS: Tobias Haas (DESY)
FNAL/DoE: Amber Boehnlein (DoE)
FNAL-IT: Victoria White (FNAL)
D0: Dmitri Denisov (FNAL), Darien Wood (FNAL)
CDF: Jacobo Konigsberg (FNAL), Robert Roser (FNAL)
IHEP-IT: Gang Chen (IHEP)
BES III: Yifang Wang (IHEP)
KEK-IT: Takashi Sasaki (KEK)
Belle: Masa Yamauchi (KEK), Tom Browder (Hawaii)
SLAC-IT: Richard Mount (SLAC)
BaBar: Francois Le Diberder (LAL/SLAC)
CLEO: David Asner (Carleton)
CERN-IT: Frederic Hemmer (CERN)
CERN/PARIS: Salvatore Mele (CERN)

International Steering Committee

The 2nd DPLTA Workshop

26-28 May 2009

@ SLAC




Resulted in initial document detailing the findings/suggestions of the DPHEP group on Data Preservation and Long Term Access in High Energy Physics

DPLTA

- *200908: ICFA Study Group on **D**ata **P**reservation and **L**ong **T**erm **A**nalysis in High Energy Physics. C. Diaconu is the ICFA appointed chair.*
- *200910: reaction of HEPAP members was very positive [to the presentation and report to HEPAP on the DPLTA effort]*
 - Recommendation for strong support from DOE/NSF

3rd DPLTA Workshop

7->9 December 2009 @ CERN



3rd Workshop on Data Preservation and Long Term Analysis in HEP

from Monday 07 December 2009 (08:00)
to Wednesday 09 December 2009 (18:00)
Europe/Zurich at CERN
chaired by:
*Cristinel Diaconu (Faculte des Sciences de Luminy) ,
Salvatore Mele (CERN)*
support:
Andre.Georg.Holzner@cern.ch

Description: This is a follow-up to the first and second workshop held at DESY and SLAC. At this workshop we will have progress reports from the various working groups, status reports of preservation efforts at the various HEP experiments, and will focus on building a blue print for concrete next steps.

[Monday 07 December 2009](#) | [Tuesday 08 December 2009](#) | [Wednesday 09 December 2009](#) | [top](#)

Monday 07 December 2009

09:30->12:30 Symposium on Data Preservation in HEP ([Council Chamber](#))
12:30 Lunch (1h00)

13:30->14:00 Registration ([6-2-004](#))

14:00->14:30 Introduction ([6-2-004](#))

14:00	Presentation of the Workshop Agenda (05')	Salvatore Mele (CERN) , Cristinel Diaconu (Faculte des Sciences de Luminy)
14:05	Welcome (10')	Salvatore Mele (CERN)
14:15	Overview (15')	Cristinel Diaconu (IN2P3)

14:30->19:00 Input from other fields: astrophysics, distributed computing, generic digital preservation programs etc. ([6-2-004](#))

Tuesday 08 December 2009

08:30->12:30 Experiments' data preservation projects: status and plans ([354-1-001](#))
12:30 lunch (1h00)

13:30->18:30 Blue print: Physics Case, Models

Wednesday 09 December 2009

08:30->12:30 Blue print: Technologies, Governance ([6-2-004](#))

Find: **principe** Next Previous Highlight all Match case Phrase not found

Concerns of the Workshop(s)

Steps Towards Long Term Analysis in HEP

- 1) Experiment-wise preparation/organisation for proper conservation of the data/knowledge
 - Proper planning and (new) projects required (**hot topic!**)
- 2) Common framework for similar experiments
 - Similar experiments converge on data release policy/format
 - Enable (further) combined analyses
- 3) Open access to expert community
 - Require sufficient knowledge encapsulation, is a natural and necessary result of the previous steps.
- 4) Open access to a wider community:
 - educational projects, outreach etc.

Steps 2-4 imply a policy for open access to the HEP data (status?)

From Cristinel Diaconu (DPLTA organizer)

Summary of Experimental Status and Data for ee experiments

	BaBar	Belle	BES-III	CLEO
End of Data Taking	07/04/08	~2010	~2017	01/04/08
Collaboration end date	end of 2012	end of 2012	2017-2022	
Type of data to be preserved	raw + sim/recon (ROOT)	raw + MDST	raw+DST (ROOT)	OBJY/PDS (too difficult) preserve analysis data
quantity	2 Pbytes	~4 Pbytes	~6 Pbytes	
desired longevity of long term analysis	unlimited	5 years (until super KEKB)	15 years	superseded by B-Factories and BES-III
Simulation	Geant4	Geant 3	Geant4	Geant3
Platform	SL3,4,5		SLC4	
code	C++	Fortran	C++	Fortran, C, C++

ep experiments

H1 vs ZEUS: A Common Repository?

A comparison of
some H1 and
ZEUS numbers:

	H1	ZEUS
RAW (kB/event)	75	125
POT (kB/event)	200	-
(m)DST (kB/event)	18	75
MC (m)DST (kB/event)	40	200
μ ODS (kB/event)	3	-
HAT (kB/event)	0.4	-
Common ntuple (kB/event)	-	10
Number of data events	1 billion	0.51 billion
Total data to conserve (TB)	100TB (raw + HERA I+II DST + μ ODS + HAT)	30TB (HERA II mDST)
Total MC (TB)	100-200	400
Estimated storage needed (TB)	200-300	430

Same format H1+ZEUS data is an idea, *could do same time as "outreach" format*

- As example: full (searches) analysis in place in one experiment: take 500 pb⁻¹ data from the other experiment and produce improved HERA limit
- But different experimental set ups, resolutions: how to factorise out the detectors?

D0 and CDF

After Run II Data Taking Complete

- Each experiment will have ~10PB data.
- Goal to keep all raw data, reconstructed data and root analysis data.
- Keep migrating to higher technology storage by Fermilab/CD.
- CDF and D0 plan to keep the computing environment and infrastructure live for about 5 years while collaboration active
- Mainly for analyses.
- This is our current thinking.

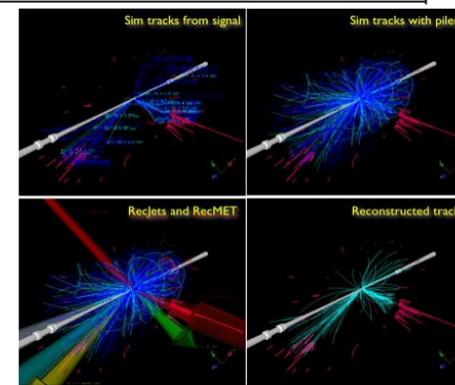
Qizhong Li (Fermilab/D0); Robert Roser (Fermilab/CDF)

DPHIEP: Possible Preservation Models

Preservation Model	Use case
1. Provide additional documentation	Publication-related information search
2. Preserve the data in a simplified format	Outreach, simple training analyses
3. Preserve the analysis level software and data format	Full scientific analysis based on existing reconstruction
4. Preserve the reconstruction and simulation software and basic level data	Full potential of the experimental data

benefits and cost

BaBar, H1, Jade



DPHEP: data preservation timeline

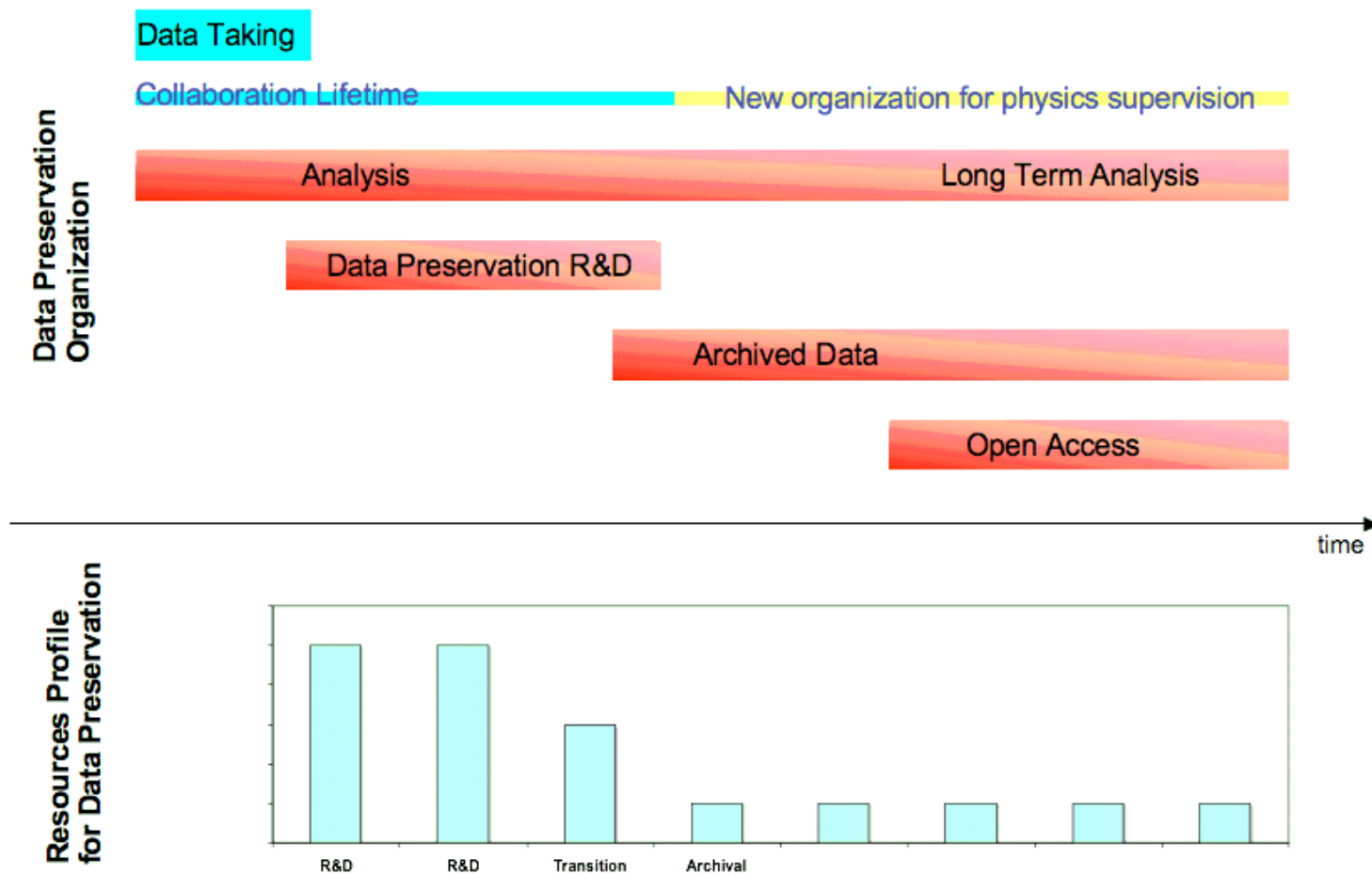


Figure 1: A possible model for data preservation organisation and resources.

The proposed international HEP data preservation organization

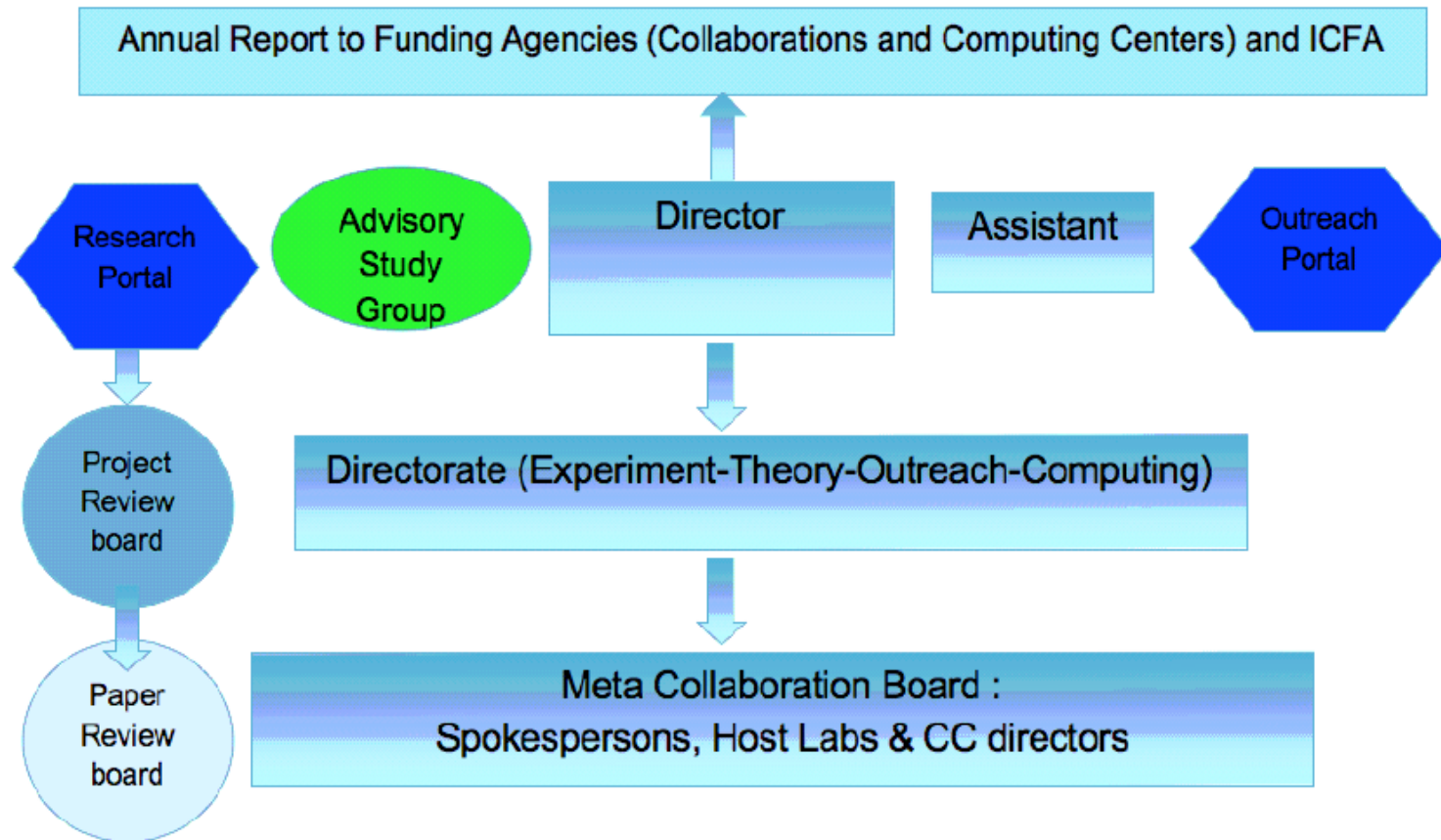


Figure 2: Organisation of an international forum for data preservation in high-energy physics.

Technology Issues

Storage and CPU power of archival systems requires some investigation into the technologies that will be available at the time the archival systems are finalized ...
today's solutions are not necessarily the best for future archiving

Storage Systems

storage technology, contd.



▪ Storage Class Memory/SSD (flash ...)

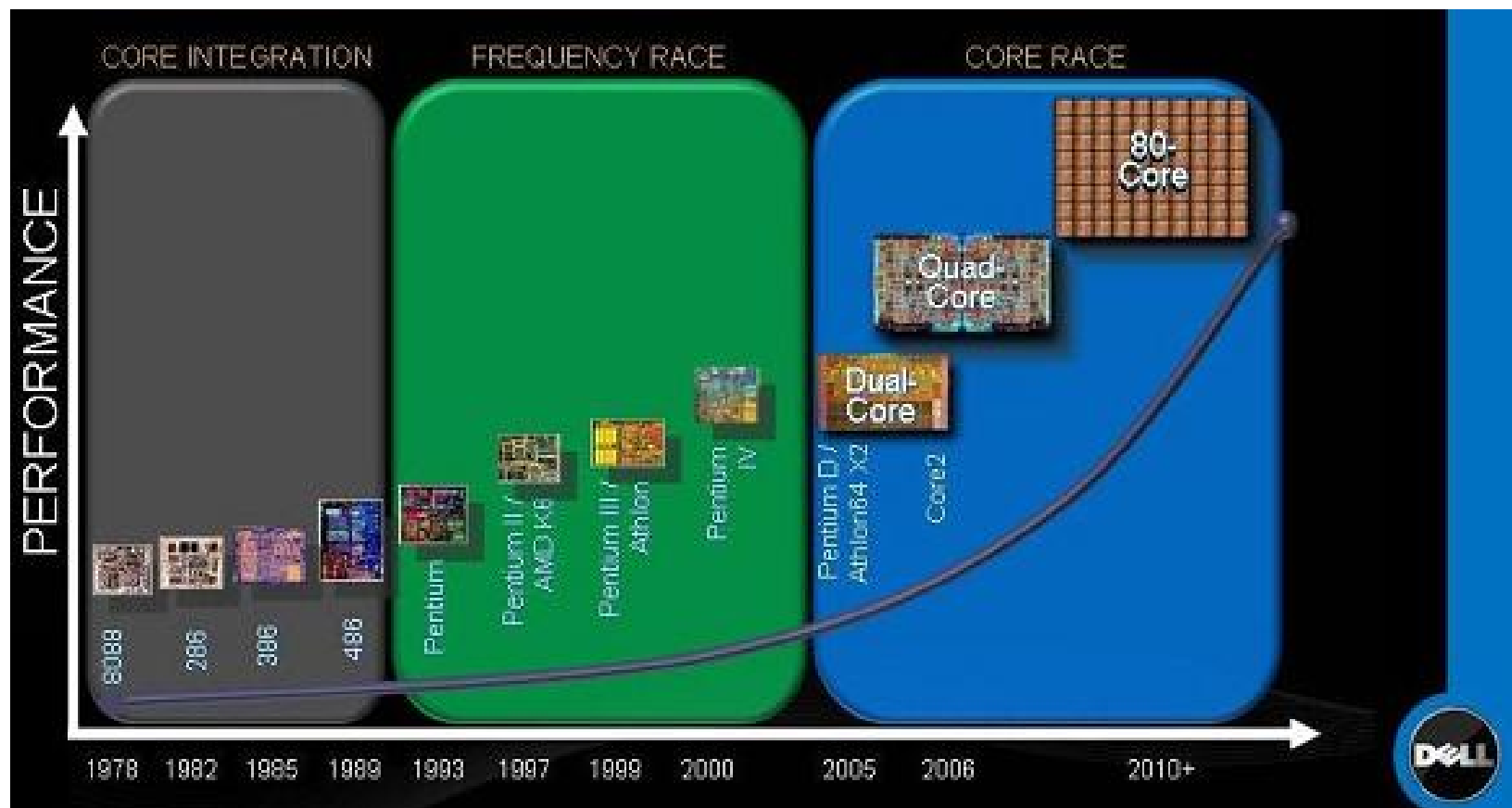
- very reliable (better than tape & disk)
- higher costs (~ x20) to disk
 - but coming down fast
- less power than disk (x10 less)
- getting higher density than disks (footprint)
 - based on conventional form factors (i.e. 3.5", 2.5" disks)
 - expected this year



Many core systems are the future

- From SC08:

http://news.cnet.com/8301-13924_3-10101987-64.html

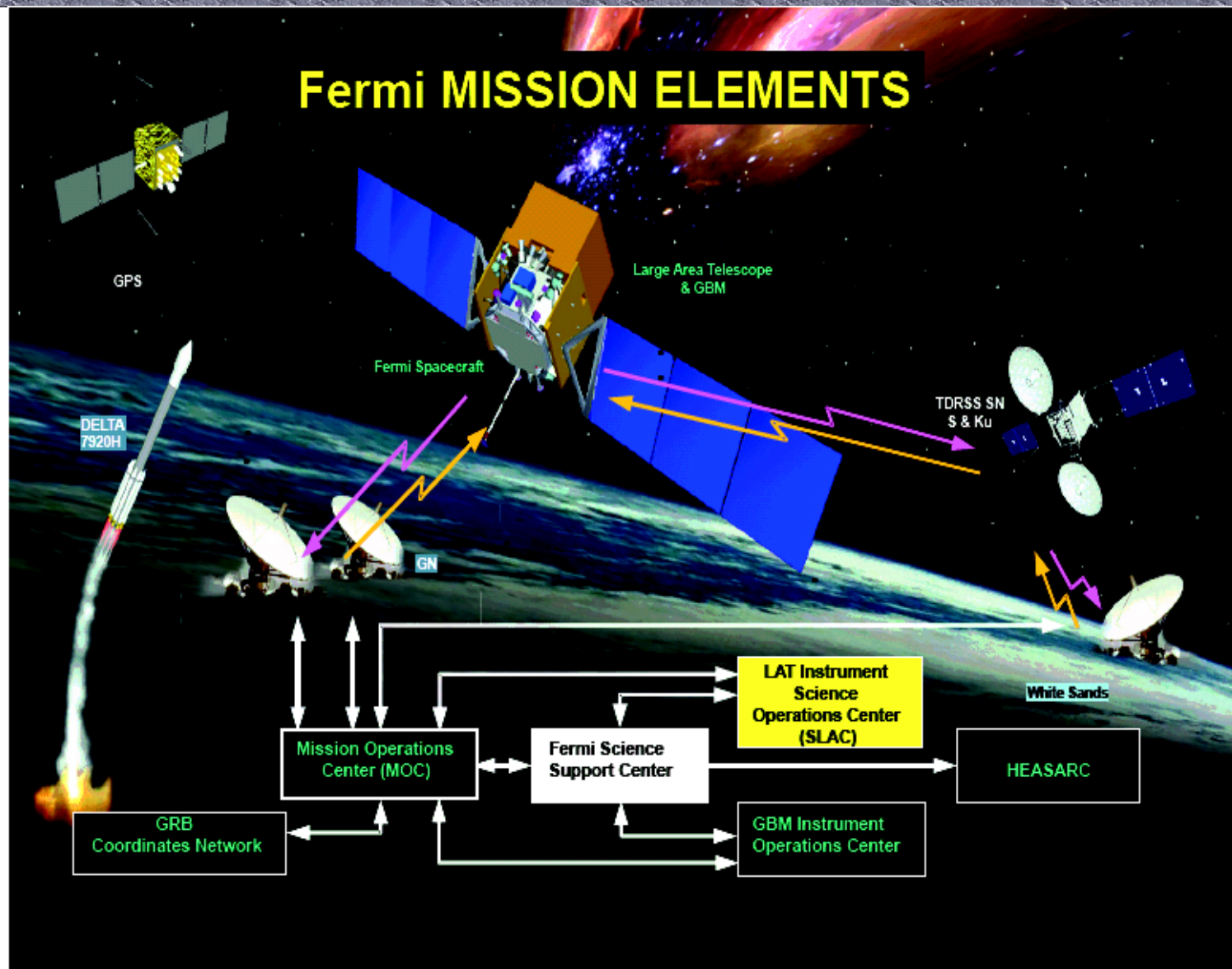


How well will they handle high data output jobs???

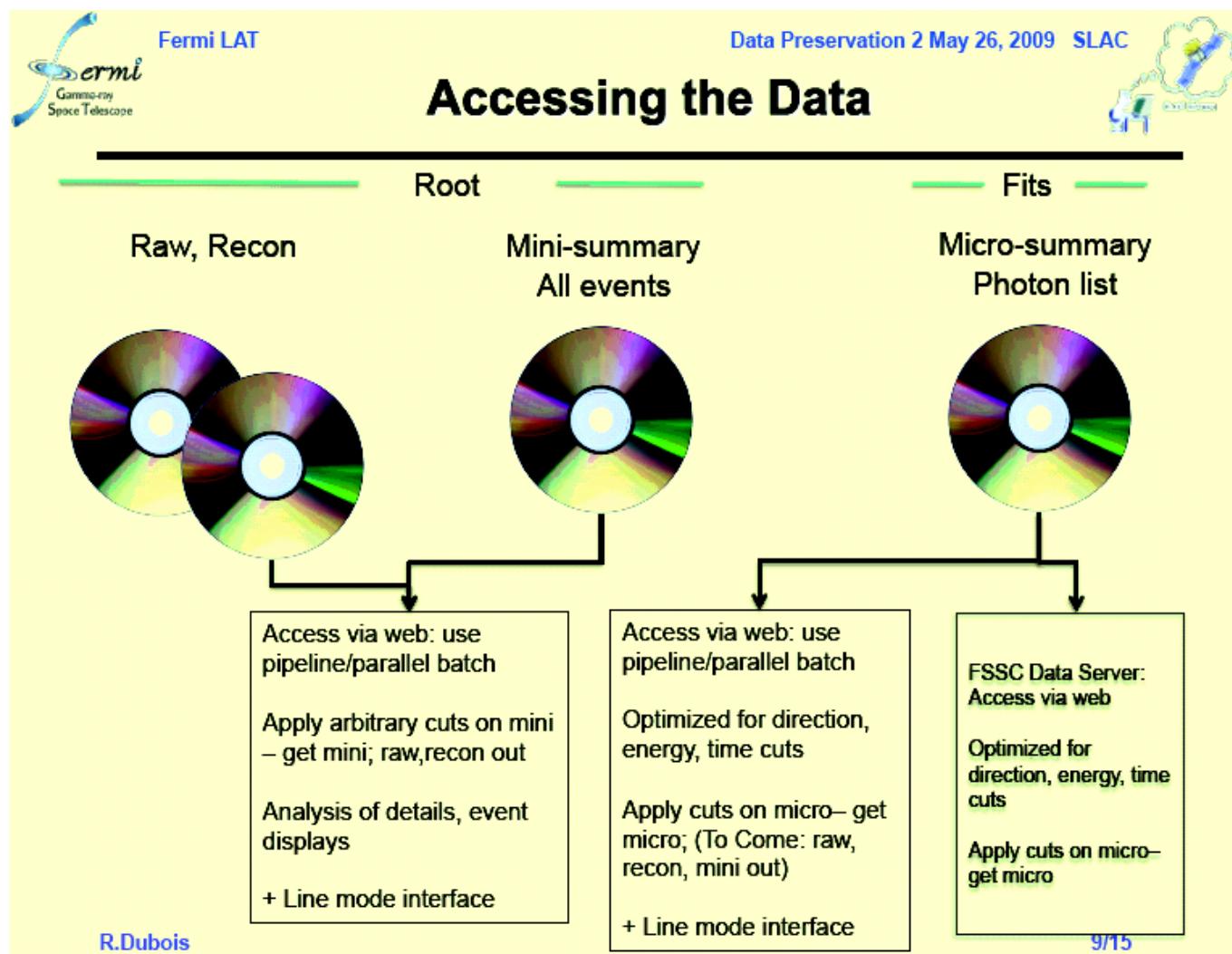
Lessons to be learned from other fields

- Data preservation and data sharing
« common » in astro physics
 - BUT is the level of preservation appropriate for HEP and can we have as much confidence in the results produced from this public data
- Data preservation in biology also well established

Fermi Data Flow



Fermi

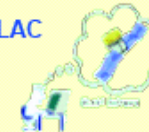


NASA Obliges the Preservation of Mission Data



Fermi LAT

Data Preservation 2 May 26, 2009 SLAC



High Level Analysis

- **Output of a telescope (for “event” data) is:**
 - Location on sky, time, energy, quality
 - Very simple output tuple!
 - Implemented in Root and FITS
 - Where instrument and celestial analysis overlap
 - Public data makes sense in astrophysics
 - NEED data from multiple missions to understand celestial sources
- **NASA mandates that all its space missions use FITS as a data format**
 - In use for 25+ years
 - Format fully documented and files self documenting
 - File headers are an integral part of the format
 - Interface library supplied for popular languages
 - And that the data be made public
 - Fermi negotiated one year hiatus on doing this. Expires Aug 11. Then all existing & ongoing data goes public
 - Funds a Science Support Center to interface to the public
 - Instrument teams not asked to do this
 - 10+ FTEs for Fermi (LAT+GBM)
 - Charter is to support the Fermi data “forever”

Large Area Telescope First Year Data Released

<http://today.slac.stanford.edu/feature/2009/fgst-data-year1.asp>

This all-sky view from the Fermi telescope reveals bright emission in the plane of the Milky Way (center), bright pulsars and super-massive black holes. (Image: NASA/DOE/International LAT Team.)

Ever since the Large Area Telescope launched aboard the Fermi Gamma-ray Space Telescope in June 2008, the LAT team has been analyzing data, searching for answers to some of the most pressing questions in astrophysics. Now everyone else can join in.

Today, the collaboration and the Fermi mission makes the first year of LAT gamma-ray data publicly available.

"This is a way of maximizing the scientific return from the mission," said Fermi Project Scientist Julie McEnery. "There is a very large number of scientists in the community with very good ideas of what to do with this data. By sharing it among a large group of people, we really get a lot more."

To ensure that others in the astrophysics community can take full advantage of the data, the LAT collaboration, working with the Fermi Science Support Center at NASA Goddard Space Flight Center, has spent a considerable amount of time preparing for the release.

"It took significant effort both on our side and the Goddard side to both get the data out and to get it out in a form that's usable by the whole community," said Astrophysicist Jim Chiang, LAT Collaboration member who works on the analysis software for FGST.

The data set released today includes more than 150 million detected gamma rays. In contrast, in the more than nine years that the LAT's predecessor, EGRET, operated, it collected 1.4 million gamma rays. In all, the LAT has collected more than 100 times as many photons in about one-tenth the time.

...

—Kelen Tuttle

SLAC Today, August 25, 2009

A new concept for HEP

Large Area Telescope First Year Data Released

<http://today.slac.stanford.edu/feature/2009/fgst-data-year1.asp>

...

As in all particle physics experiments, Chiang said, LAT data are unique to the instrument and require unique software. With this in mind, the collaboration will also make available high-level software that other researchers will need in order to analyze the data. In addition, NASA is offering further resources and funds to guest investigators who successfully submit proposals.

"We can see both from the large number proposals submitted to the guest investigator program and the large number of references in papers that the community is excited about the data," McEnery said.

LAT Principal Investigator **Peter Michelson** added: **"The LAT team has made significant discoveries and significant progress in many areas. I expect that the collaboration will continue to come out with the most results, but I also expect others to make discoveries. Releasing this data is good for the project, good for the collaboration, and good for science."**

—Kelen Tuttle

SLAC Today, August 25, 2009

Words from the Archivists: The government demands the preservation of data

Scientific Data:

- Raw data (all levels)
 - 10 year retention (N1-434-07-01, item 4c(12))
- Evaluated or Summarized data
 - Level 1: permanent retention (N1-434-96-9, item1B13a)
 - Level 2: 25-year retention (N1-434-96-9, item1B13b)
 - Level 3: 10-year retention (N1-434-96-9, item1B13c)

iRODS User Community (1 of 2)

By Wayne Schroeder (SDSC) at the 2nd DPLTA workshop at SLAC

- iRODS Development Collaborations
 - NARA TPAP Transcontinental Persistent Archive Prototype (NARA funded)
 - NSF SDCI Research in Adaptive Middleware Architecture Systems
 - SHAMAN Sustaining Heritage Access through Multivalent ArchiviNg
 - UK e-Science data grid
- Communities Using DICE Technologies, including Biology, Environment, Psychology, Human Subjects
 - BIRN Biomedical Informatics Research Network (NIH funded)
 - ROADNet Real-time Observatories, Applications, and Data management Network (NSF funded)
 - SEEK Science Environment for Ecological Knowledge (NSF funded)
 - TDLC Temporal Dynamics of Learning Center (NSF funded)
- Physical Sciences Uses
 - CADAC Computational Astrophysics Data Analysis Center (NSF funded)
 - BaBar high energy physics data grid (DOE funded)

iRODS User Community (2 of 2)

By Wayne Schroeder (SDSC) at the 2nd DPLTA workshop at SLAC

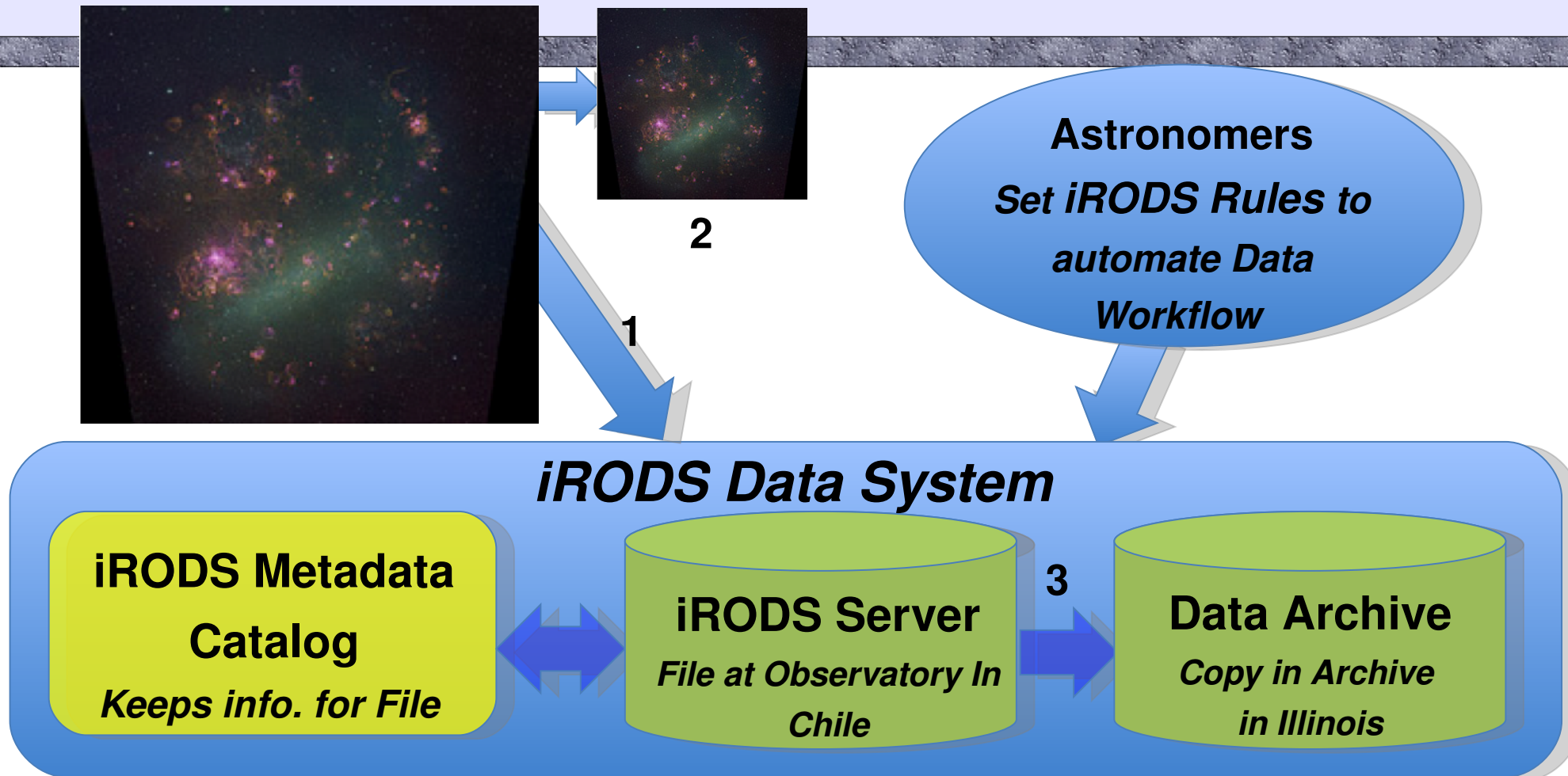
- Physical Sciences (continued)
 - NOAO National Optical Astronomy Observatories data grid (NSF funded)
 - NVO National Virtual Observatory (NSF funded)
 - Observatoire de Strasbourg, France, VOSpace Interface
- Persistent Archives and Digital Preservation / Humanities Uses
 - NARA TPAP Transcontinental Persistent Archive Prototype
 - e-Legacy Preserving the Geospatial Data of the State of California
 - DCPC Distributed Custodial Preservation Center (NHPRC funded)
 - DIGARCH UCTV NSF Digital Archiving and Long-Term Preservation (LoC)
 - T-RACES Testbed for Redlining Archives of CA Exclusionary Spaces (IMLS)
- Geosciences Uses
 - OOI Ocean Observatories Initiative (NSF funded)
 - SCEC Southern California Earthquake Center (NSF funded)
- High Performance and Grid Computing
 - NSF TeraGrid
- Plus many international users.
- And growing all the time...

Some Preservation Challenges

By Wayne Schroeder (SDSC) at the 2nd DPLTA workshop at SLAC

- Maintaining Provenance –history of ownership/changes
 - Audit log, ACLs, checksums
 - Metadata
- Keeping the bits safe/secure
 - Replication (multi-location), checksum, system and user metadata, catalog backup, ACLs, secure authentication(password, GSI, Kerberos, Shibboleth (soon))
- Access/Usability –Storage Format Evolution and Obsolescence
 - Difficult
 - NARA Admin email example; XMS metadata
 - Metadata
- Ontological Continuity
 - Metadata (user defined)
- Search
 - Metadata, queries
- New Storage Systems
 - Irods resources, storage drivers, replication

Adding Data to iRODS Data System



Rules can tell iRODS to automatically 1) Register photo from telescope in Chile into Metadata Catalog and write to Data Server, 2) Make thumbnail, 3) Make copy in US Archive.

By Wayne Schroeder (SDSC) at the 2nd DPLTA workshop at SLAC

On Demand Data Analysis

- Use virtual machines on clouds like Amazon EC2 to simulate/analyze archived data as needed
- Currently being investigated for BaBar
- Used when extra processing power is needed for Belle

Summary

- Data preservation has become a major concern of HEP experiments
- Should be integral part of the infrastructure development
- Must act now to preserve data whose physics potential has not yet been exhausted and to prepare upcoming experiments for the best means of fully extracting the physics potential of their data
- Much to learn from other scientific fields concerning data preservation

Extra Material

Preservation Tools

Welcome to an [inspire](#) test server. Please go to [SPRES](#) if you are here by mistake.



user login

HEP :: PERSONALIZE :: HELP :: HEPNAMES :: INST :: CONF :: EXP :: JOBS

[Home](#) > [Record#735595](#): Axions In String Theory

[Information](#) [References](#) [Citations](#) [Discussion](#) [Usage statistics](#) [Fulltext](#)

Axions In String Theory.

Peter Svrcek (Stanford U., Phys. Dept. & SLAC), Edward Witten (Princeton, Inst. Advanced Study).
May 22, 2006

Published in: JHEP 0606: 051, 2006
e-Print: [hep-th/0605206](#)

Abstract: In the context of string theory, axions appear to provide the most plausible solution of the strong CP problem. However, as has been known for a long time, in many string-based models, the axion coupling parameter F_a is several orders of magnitude higher than the standard cosmological bounds. We re-examine this problem in a variety of models, showing that F_a is close to the GUT scale or above in many models that have GUT-like phenomenology, as well as some that do not. On the other hand, in some models with Standard Model gauge fields supported on vanishing cycles, it is possible for F_a to be well below the GUT scale.

Keyword(s): [string model](#) ; [heterotic](#) ; [gauge field theory](#) ; [SU\(3\)](#) ; [instanton](#) ; [axion](#) ; [violation](#) ; [CP](#) ; [dimensional reduction](#) ; [anomaly](#) ; [membrane model](#) ; [D-brane](#) ; [bibliography](#)

Record created 2008-05-04, last modified 2008-05-04

[Similar records](#)

[Abstract and Postscript and PDF from arXiv.org](#)
[@JHEP Electronic Journal Server](#)
[@SLAC Document Server](#)

Rate this document:



(Not yet reviewed)

⇒ [Add to personal basket](#)
⇒ [Export reference](#)
[BibTeX](#), [EndNote](#), [LaTeXUS](#), [LaTeXEU](#),
⇒ [Export data](#)
[MARCM](#), [HLM](#), [DC](#), [MARC](#)

Words from the Archivists: The government demands the preservation of data

Context @ SLAC: General Definitions -- Record



“..all books, papers, maps, photographs, machine readable materials, or other documentary materials, regardless of physical form or characteristics, made or received by an agency of the United States Government under Federal law or in connection with the transaction of public business and preserved or appropriate for preservation by that agency or its legitimate successor as evidence of the organization, functions, policies, decisions, procedures, operations, or other activities of the Government or because of the informational value of data in them...” (44 USC 3301)