PIC
port d'informació
científica

CHEP2015
OKINAWA, japan

Architectures and methodologies for future deployment of multi-site Zettabyte-Exascale data handling platforms

V. Acin[5,9], I. Bird[1], T. Boccali[7], G.Cancio[1], I. Collier[10], D. Corney[10], B. Delaunay[6], M. Delfino[9,11], L. Dell'agnello[7], J. Flix[2,9], P. Fuhrmann[4], M. Gasthuber[4], V. Guelzow[4], A. Heiss[8], G. Lamanna[3], P.-E. Macchi[6], M. Maggi[7], B. Matthews[10], C. Neissner[5,9], J.-Y. Nief[6], M. Porto[2,9], A. Sansum[10], M. Schulz[1], J. Shiers[1]

[1]CERN, [2]CIEMAT, [3]CNRS, [4]DESY, [5]IFAE, [6]IN2P3, [7]INFN, [8]KIT, [9]PIC, [10]RAL, [11]UAB

Institut de Física
d'Altes Energies

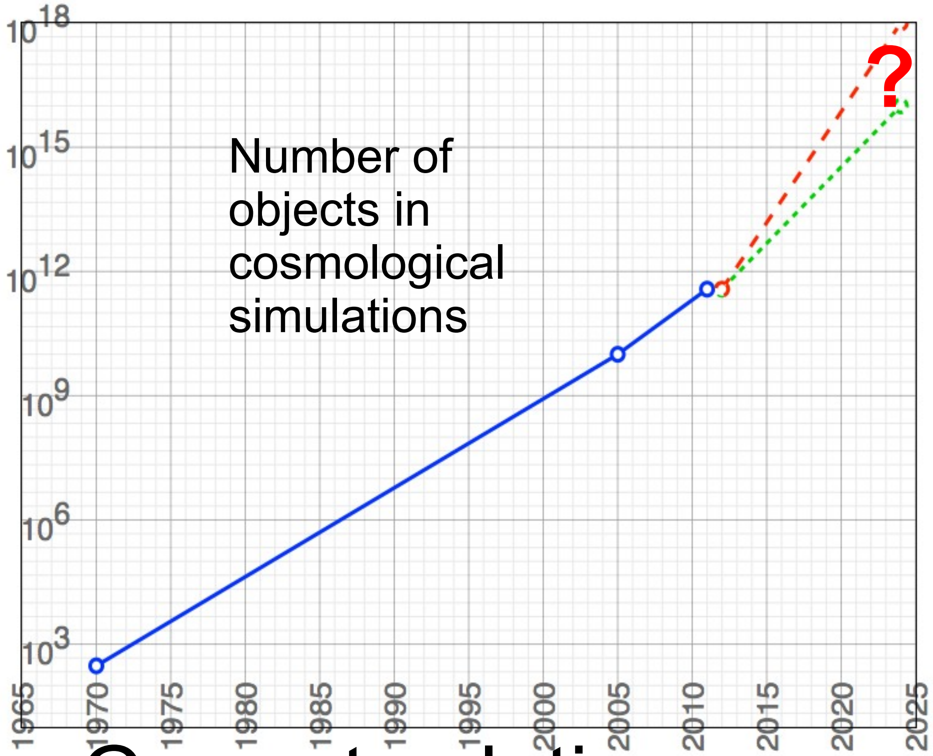
IFAE

Ciemat

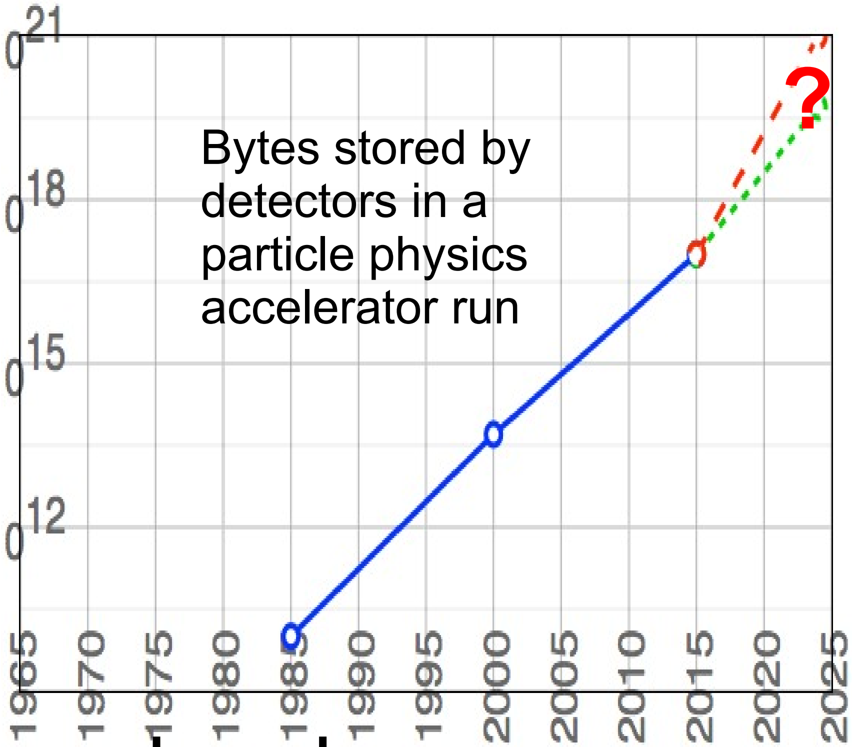
Centro de Investigaciones
Energéticas, Medioambientales
y Tecnológicas

The issue

- Exa = 10^{18} ; Zetta = 10^{21}
- Growth of number of objects <u>and</u> data volume



Number of objects in cosmological simulations



Bytes stored by detectors in a particle physics accelerator run

- Current solutions may break or be unaffordable at the Zettabyte-Exascale level

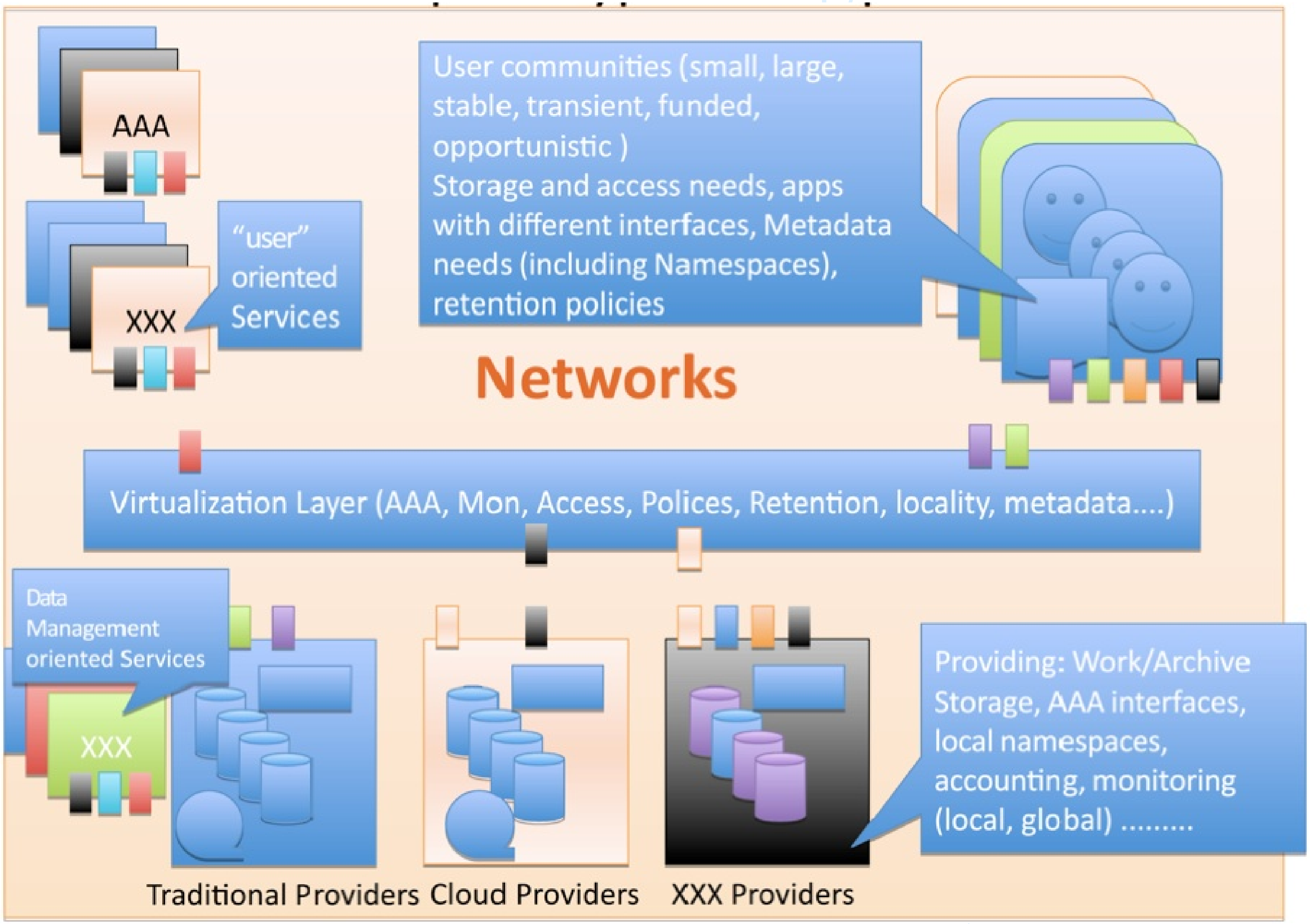
Disclaimer

- Throughout this talk, all statements on activities refer to our opinion that, to get to solutions in 10 years time:
  - Attention needs to be given to those subjects now
  - Cooperation between experiments, data centers and domain experts is needed
  - Simulations and small-scale prototypes may be a good way to start

The ZEPHYR study

- Group of people from institutions which run large data centers in Europe (→ EU-T0)
- Concern that it takes 10 years for “next step”
- Funding opportunities from the EU
- Produce concrete outputs which can be discussed, improved, evolved
    - Collaboration with existing/upcoming experiments
    - Collaboration with data centers in Americas and Asia
- Look at Architecture without forgetting about real, practical “build and operate” aspects
- Use simulation and small prototypes to start

Architectures help organize



"user" oriented Services

User communities (small, large, stable, transient, funded, opportunistic)
Storage and access needs, apps with different interfaces, Metadata needs (including Namespaces), retention policies

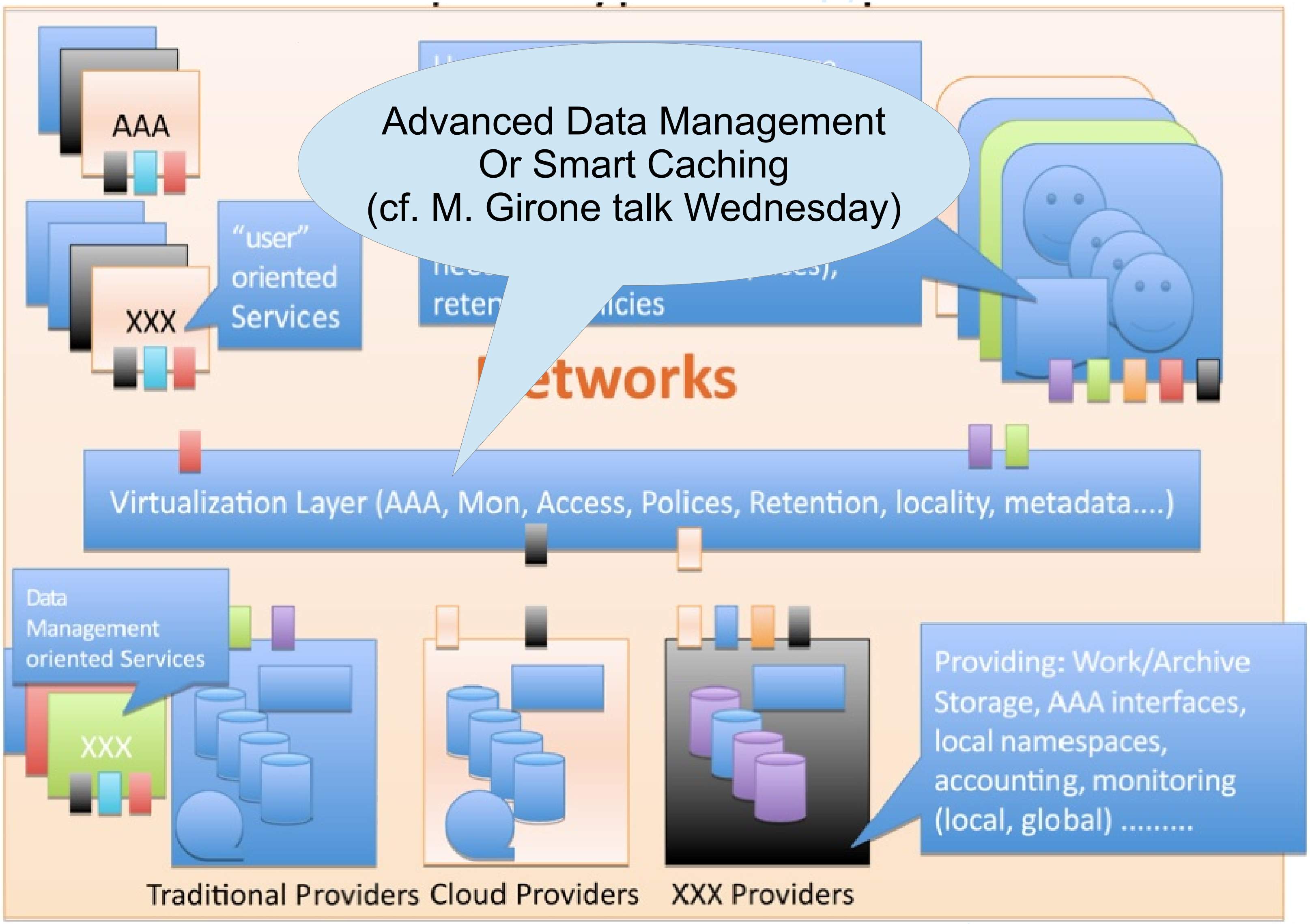
Virtualization Layer (AAA, Mon, Access, Polices, Retention, locality, metadata....)

Data Management oriented Services

Providing: Work/Archive Storage, AAA interfaces, local namespaces, accounting, monitoring (local, global) ....

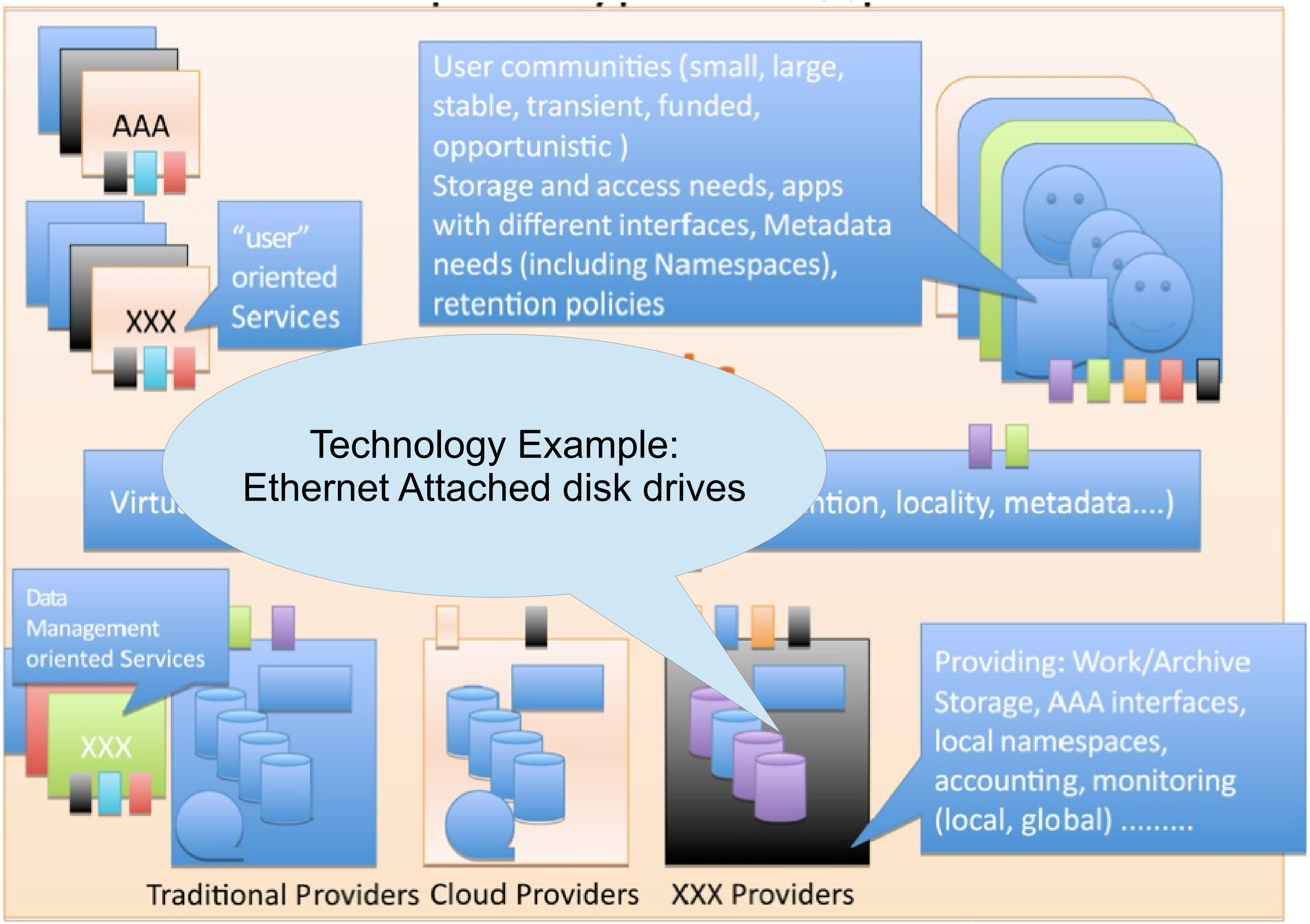
Traditional Providers Cloud Providers XXX Providers

Architectures help organize


AAA
XXX
"user" oriented Services
reten icies
etworks
Virtualization Layer (AAA, Mon, Access, Polices, Retention, locality, metadata....)
Data Management oriented Services
XXX
Providing: Work/Archive Storage, AAA interfaces, local namespaces, accounting, monitoring (local, global) ....
Traditional Providers
Cloud Providers
XXX Providers

Advanced Data Management
Or Smart Caching
(cf. M. Girone talk Wednesday)

Architectures help organize



“user” oriented Services

User communities (small, large, stable, transient, funded, opportunistic)
Storage and access needs, apps with different interfaces, Metadata needs (including Namespaces), retention policies

Data Management oriented Services

Providing: Work/Archive Storage, AAA interfaces, local namespaces, accounting, monitoring (local, global) ....

Technology Example:
Ethernet Attached disk drives

The scientists' environment

- Analysis during years by a relatively small number of scientists distributed globally
  - {CPU, I/O}/person is huge compared to “Big Data”
- How data will be analyzed not known *a priori*.
- “Summary data” → “Science Metadata”
- Need to “zoom back” all the way to raw data
- Provenance: Who created this item of data ?
- Data Integrity: Ensure analysis input is correct
- Data Preservation: Reproduce results in future

The datacenter environment

- Increase throughput as data volume grows
- Handle growth in “Science Metadata”
  - Increasing fraction of analysis done on metadata
  - Metadata volume increasing to >PB
  - N_objects in metadata = N_objects in data → Exa
- Merge-in “Technical Metadata” management
  - When was this replica of this object last read, by whom?
  - Automation support for “smart” data caching
  - Decision support for cost optimization (inc. energy)
- Make hardware “invisible”. Downtime at object level.
- Data life-cycle management→Data Preservation

The science data handling funding environment

- Data handling costs come after construction
  - Agencies want to minimize operations-type funding
  - Competition with "next project" or "upgrade"
  - Would you donate $10.000/year for preserving data from UA1/UA2, Mark-I and Gargamelle?
- Constant pressure to lower unit costs
- Constant pressure to lower personnel
- Peaks of enthusiastic postdocs and grad students funded in experiments followed by valleys of scarce personnel

ZEPHYR study hints for things to work on

- Life-cycle management of users and their roles
- Extensible metadata management frameworks, handling Scientific and Technical metadata
- Site-independent “Data Virtualization” layer: metadata query with redirection to data objects
- Clearly define what site-storage should do and what it doesn't need to do – relation to costs
- Smarter use of network capacity and capability
- Gatekeepers for Data Provenance/Preservation
- Security and cost-containment in architecture

Users and their roles

- User Authentication as a Service (i.e. use home institution username/password *à la eduGAIN*)
- Enable (*à la grouper*) the flow of information on roles between
  - Project secretariats, data management coordinators
  - Data processing services
- Propagate/map info deep down into the operating systems hosting key services
- Leverage setup to help automate
  - Access rights (who can read, write, delete)
  - Accounting, data provenance and preservation

Extensible metadata management

- Recognition that each project has specific needs for data management which will generate project-specific metadata
- Data "management" is currently a huge sink of human resources
- Lots of ad-hoc patches to merge queries across project-specific and technical metadata
- In addition, more and more science information will be enconded as metadata (c.f. genomics)
- Need to identify candidates and build large-scale prototypes of extensible metadata management services

"Data Virtualization" layer

- Global, high reliability and availability service, probably to be provided cooperatively by n-sites
- Challenge: Reliability/Availability per object
- Dynamic repository of information about objects
  - Information (project+science metadata) on newly created objects
  - Updates of attributes of existing objects
  - Updates on technical metadata (status of objects)
- Responds to metadata queries *à la Big Data*
- Provides I/O redirection to access data blobs
- Has bulk operation capabilities

Site Storage

- Another huge sink of human resources
- Part of the problem is the incoherent piling up of filesystems on top of pseudo-filesystems on top of Grid filesystems on top of project namespace
- Need to clearly/cleanly define its roles. Maybe:
  - Key-value object storage
  - Key-indexed technical metadata reporting
  - End-to-end network optimization
  - Making hardware failures invisible
- Careful: Must avoid dependencies and preserve parallelism in order to achieve throughput

Network

- Wide Area Networks have evolved to have features which we are not using
  - Are we under-utilizing dynamic network capabilities?
  - Or is the NREN model more static than advertised?
- Work with NRENs on WAN for data Exascale

Network

- Wide Area Networks have evolved to have features which we are not using
  - Are we under-utilizing (WAN) network capabilities?
  - Or is the NREN model more static than advertised?
- Work with NRENs on WAN for data Exascale
- Local Area Networks have evolved to have features which we are not using
  - Are we under-utilizing (LAN) network capabilities?
- Virtualization+Data Intensive → re-think LAN
- And of course …. IPv6.

Data Provenance/Preservation

- Lots of work: DPHEP, other scientific disciplines
- Current (few) implementations: mostly manual filling of metadata when "depositing data"
- Encourage development of tools for "batch" data deposit with provenance/preservation information according to international standards
  - Will all future project data be "tagged" using international standards ?
  - Or will there be "internal" data which has non-standard metadata and "external" data with standard tags ?
- Need to build prototypes and understand issues

Security, Access Control, Reliability, Cost-containment

- All of these things should be invisible to the user
- All of these things can generate high costs, particularly when 10^{18} objects are involved
- Must build the handling of these issues into the right architectural layers
- A random example:
  - Hypothetical future storage systems built from hard drives directly connected to Ethernet (cf. CERN openlab talk in this conference)
  - If confidentiality is implemented with an incompatible scheme, advantages may be completely lost

Prototypes and Simulation

- Implementing even a 1% prototype with dedicated resources has prohibitive costs, and anyway what we need at first is practical investigations into specific “slices”.
- Simulations are an alternative which can help evaluate various alternatives
- Once alternatives are reduced, Datacenters can help to setup tests using temporary resources
- This also avoids locking-in on a single solution too early and without considering sufficiently the various alternatives

Conclusions

- Zettabyte volumes and Exascale objects can be expected within 10 years
- It takes 10 years to prototype, develop and implement solutions
- Projects / User Communities and Datacenter experts need to start working on prototypes
- Need to identify a few possible architectures to be able to work on concrete prototypes/tests
- Many projects can provide components which can be assembled into alternatives to be tested
- Evaluation through Simulation+"slice"Prototypes