



The U.S. CMS Grid

Michael Ernst

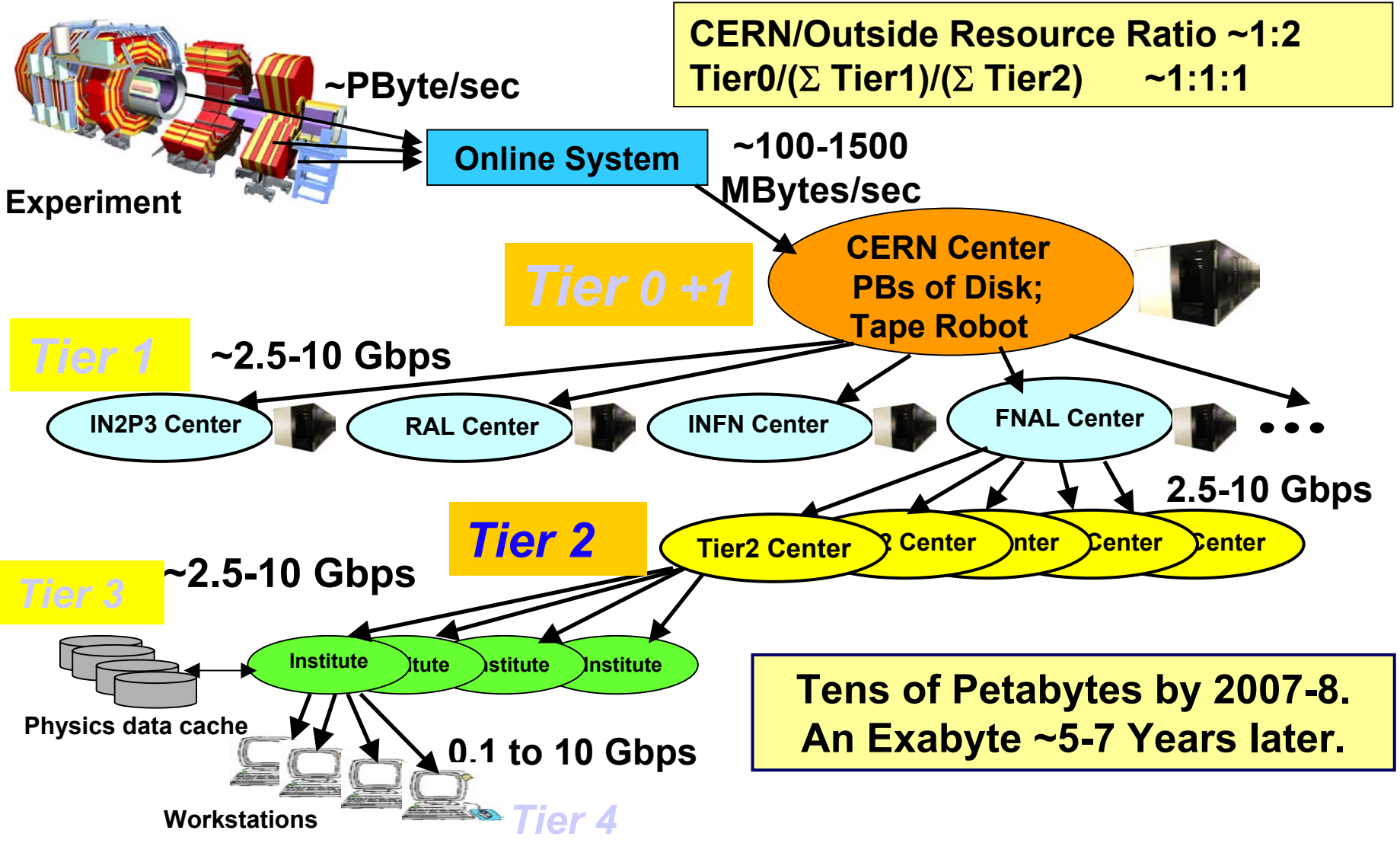
DESY & Fermilab

DESY IT Seminar

April 9, 2003



LHC Computing Hierarchy





US CMS S&C Scope and Deliverables

- ➔ Provide software engineering support for CMS
- ➔ Provide S&C Environment for doing LHC Physics in the U.S.
 - ➔ Develop and build “User Facilities” for CMS physics in the U.S.
 - ◆ A Grid of Tier-1 and Tier-2 Regional Centers connecting to the Universities
 - ◆ A robust infrastructure of computing, storage and networking resources
 - ◆ An environment to do research in the U.S. and in globally connected communities
 - ◆ A support infrastructure for physicists and detector builders doing research
 - ➔ This U.S. infrastructure, together with the U.S. share on developing the framework software is the U.S. contribution to the CMS software and computing needs

Tier-1 center at Fermilab provides computing resources and support

- ➔ User Support for “CMS physics community”, e.g. software distribution, help desk
- ➔ Support for Tier-2 centers, and for Physics Analysis Center at Fermilab

Five Tier-2 centers in the U.S.

- ➔ Together will provide same CPU/Disk resources as Tier-1
- ➔ Universities to “bid” hosting Tier-2 center, take advantage of resources and expertise
- ➔ Tier-2 centers to be funded through NSF program for “empowering Universities”
 - ◆ Proposal to the NSF for 2003 to 2008 was submitted Oct 2002



The US CMS Grid System

The US CMS Grid System of T1 and T2 prototypes and testbeds has a really important function within CMS

- ◆ help develop a truly global and distributed approach to the LHC computing problem
- ◆ ensure full participation of the US physics community in the LHC research program

To succeed requires the ability and ambition for leadership

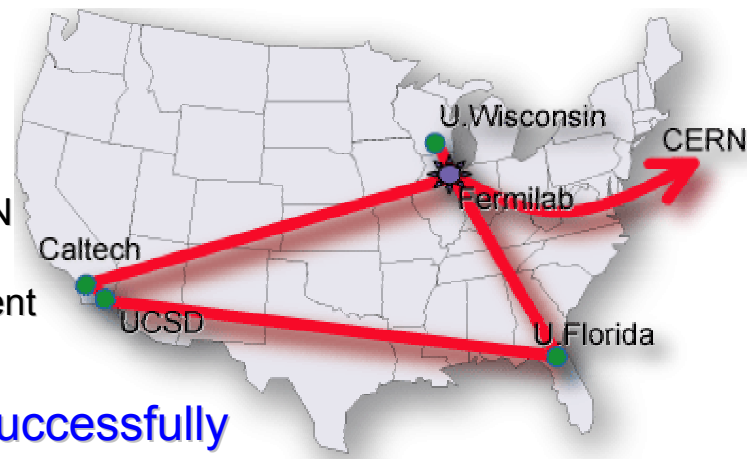
- ➔ US CMS has prototyped Tier-1 and Tier-2 centers for CMS production
- ➔ US CMS has worked with Grids and VDT to harden middleware products
- ➔ US CMS has integrated the VDT middleware in CMS production system
- ➔ US CMS has deployed an “Integration Grid Testbed” and has used it for real productions
- ➔ US CMS will participate in the series of CMS Data Challenges
- ➔ US CMS will take part in the LCG “Production Grid” milestone in 2003



From Facilities to a Grid Fabric

We have deployed a system of a Tier-1 center prototype at Fermilab, and Tier-2 prototype facilities at Caltech, U.Florida and UCSD

- Prototype systems operational and fully functional
US CMS Tier-1/Tier-2 system very successful
 - ◆ R&D, Grid integration and deployment
 - ◆ e.g. high-throughput data transfers Tier-2/Tier-1, CERN data throughput O(1TB/day) achieved!
 - ◆ Storage Management, Grid Monitoring, VO management



Tier-1/Tier-2 distributed User Facility was used very successfully in the large-scale, world-wide production challenge

- — part of a 20TB world-wide effort to produce simulated and reconstructed MC events for HLT studies
- ended on schedule in June 2002

Large data samples (Objectivity and nTuples) have been made available to the physics community ⇨ DAQ TDR

Using Grid technologies, with the help of Grid projects and Grid middleware developers, we have prepared the CMS data production and data analysis framework to work in a Data Grid environment.

- Intense collaboration with US Grid projects
- Grid-enabled MC production system operational

prototyping, early roll out, strong QC&documentation, tracking of external “practices”

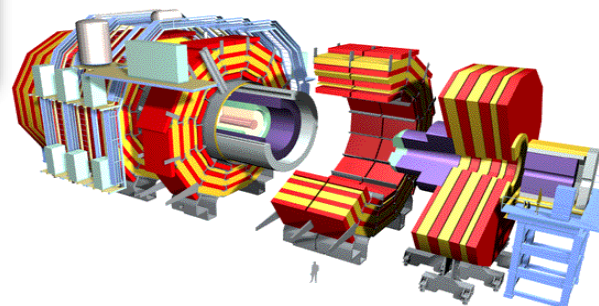
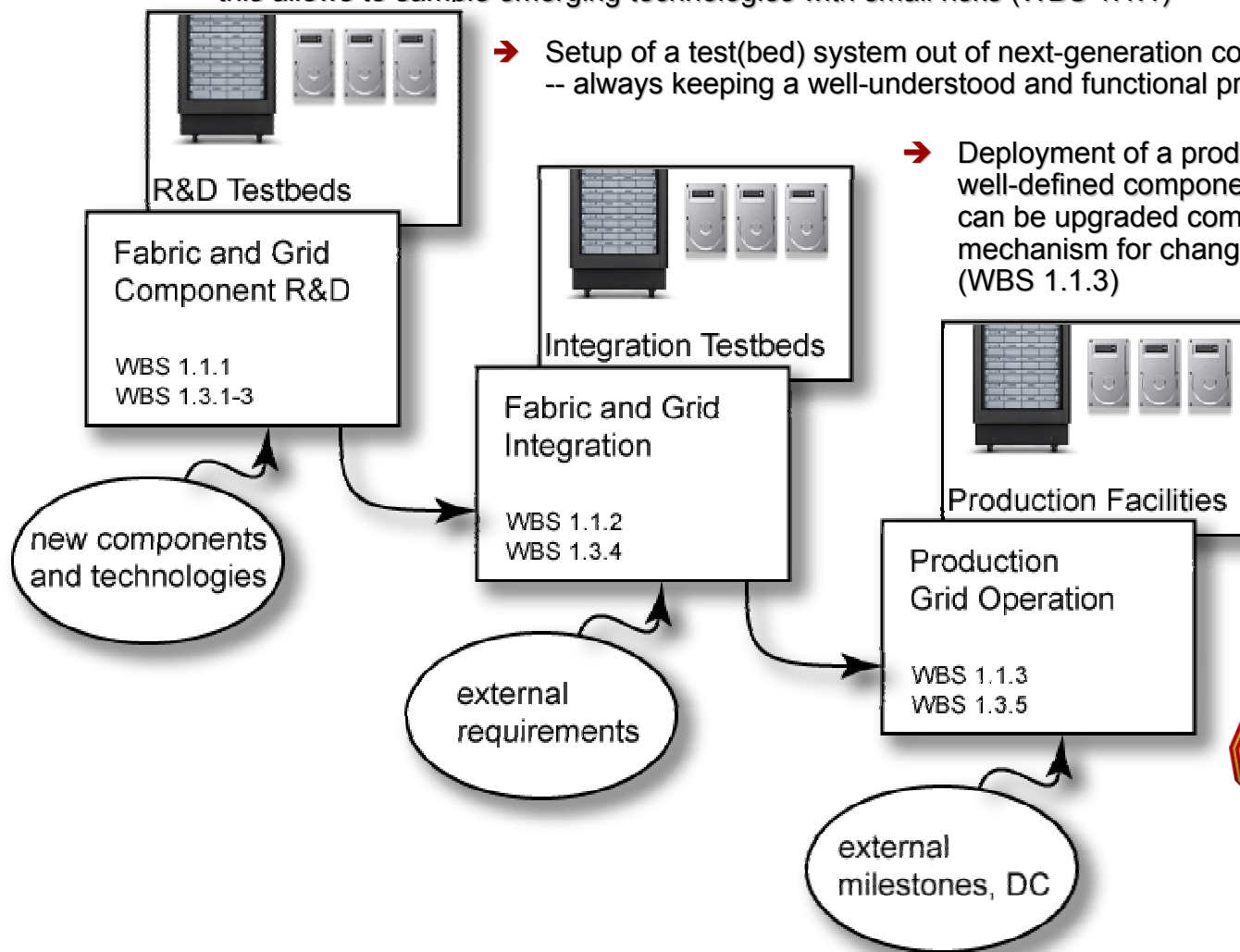
Approach: “Rolling Prototypes”: evolution of the facility and data systems

- ➔ Test stands for various hardware components and (fabric related software components) -- this allows to sample emerging technologies with small risks (WBS 1.1.1)

- ➔ Setup of a test(bed) system out of next-generation components -- always keeping a well-understood and functional production system intact (WBS 1.1.2)

- ➔ Deployment of a production-quality facility --- comprised of well-defined components with well-defined interfaces that can be upgraded component-wise with a well-defined mechanism for changing the components to minimize risks (WBS 1.1.3)

- ➔ This matches to general strategy of “rolling replacements” and thereby upgrading facility capacity making use of Moore’s law





US CMS Grid Technology Cycles

Correspondingly our approach to developing the software systems for the distributed data processing environment adopts “rolling prototyping”

- Analyze current practices in distributed systems processing and of external software, like Grid middleware (WBS 1.3.1, 1.3.2)
- Prototyping of the distributed processing environment (WBS 1.3.3)
- Software Support and Transitioning, including use of testbeds (WBS 1.3.4)
- Servicing external milestones like data challenges to exercise the new functionality and get feedback (WBS 1.3.5)

Next prototype system to be delivered is the US CMS contribution to the LCG Production Grid (June 2003)

- CMS will run a large Data Challenge on that system to prove the computing systems (including new object storage solution)

This scheme will allow us to react flexibly to technology developments AND to changing and developing external requirements

It also requires a set of interesting technologies concerning e.g.

- System architectures, farm configuration and partitioning
- Storage architectures and interfaces
- How to approach information services, configuration management etc



IGT Results

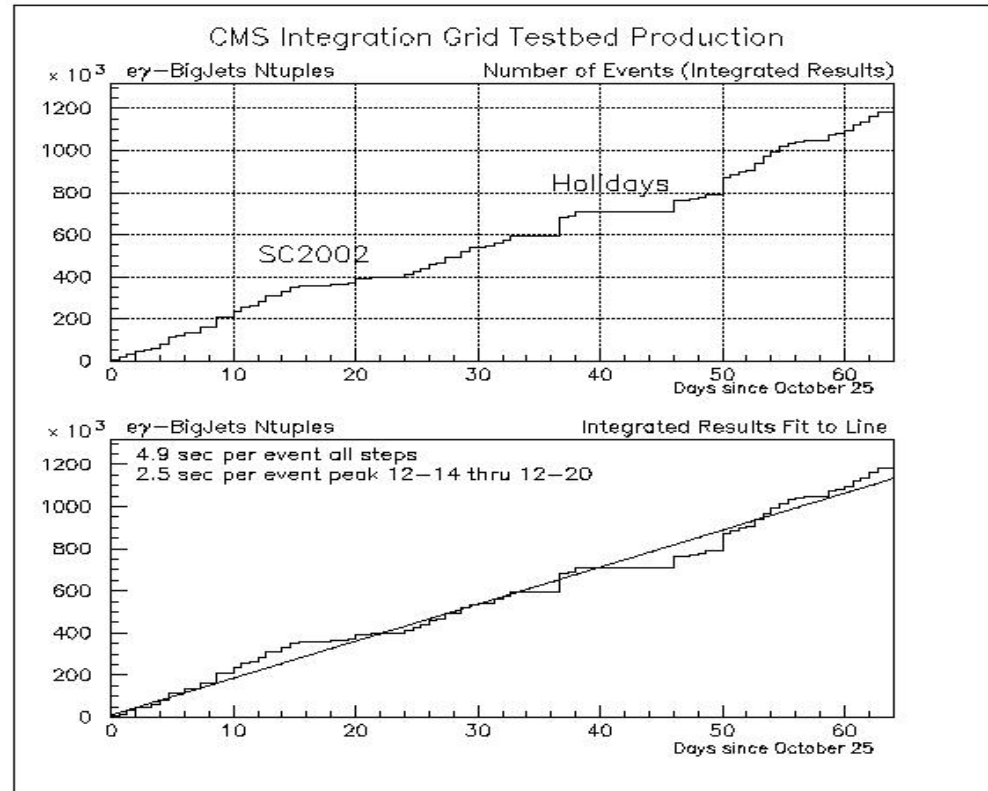
1M fully simulated and reconstructed events!

*Time to process
1 event:
500 sec @ 750 MHz*

*Speedup:
Avg. factor of 100
speedup during current
run*

*Resources:
Approximately 230 CPU
@750 MHz equiv.*

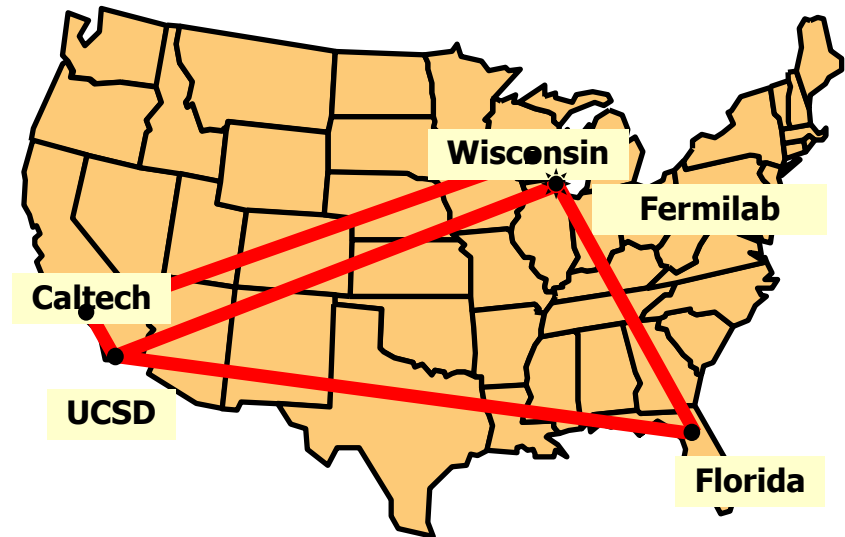
*Sustained efficiency:
about 43.5%*





US CMS Development Grid Testbed Resources (Fall 2002)

- **Fermilab**
 - 1+5 PIII dual 0.700 GHz processor machines
 - 0.2 TB dedicated disk + Mass Storage
- **Caltech**
 - 1+3 AMD dual 1.6 GHz processor machines
 - 0.4 TB dedicated disk
- **San Diego**
 - 1+3 PIV single 1.7 GHz processor machines
 - 0.07 TB dedicated disk
- **Florida**
 - 1+5 PIII dual 1 GHz processor machines
 - 0.5 TB dedicated disk
- **Wisconsin**
 - 5 PIII single 1 GHz processor machines
 - 0.02 TB dedicated disk
- **Total:**
 - ~41 1 GHz dedicated processors
 - ~1 TB dedicated storage

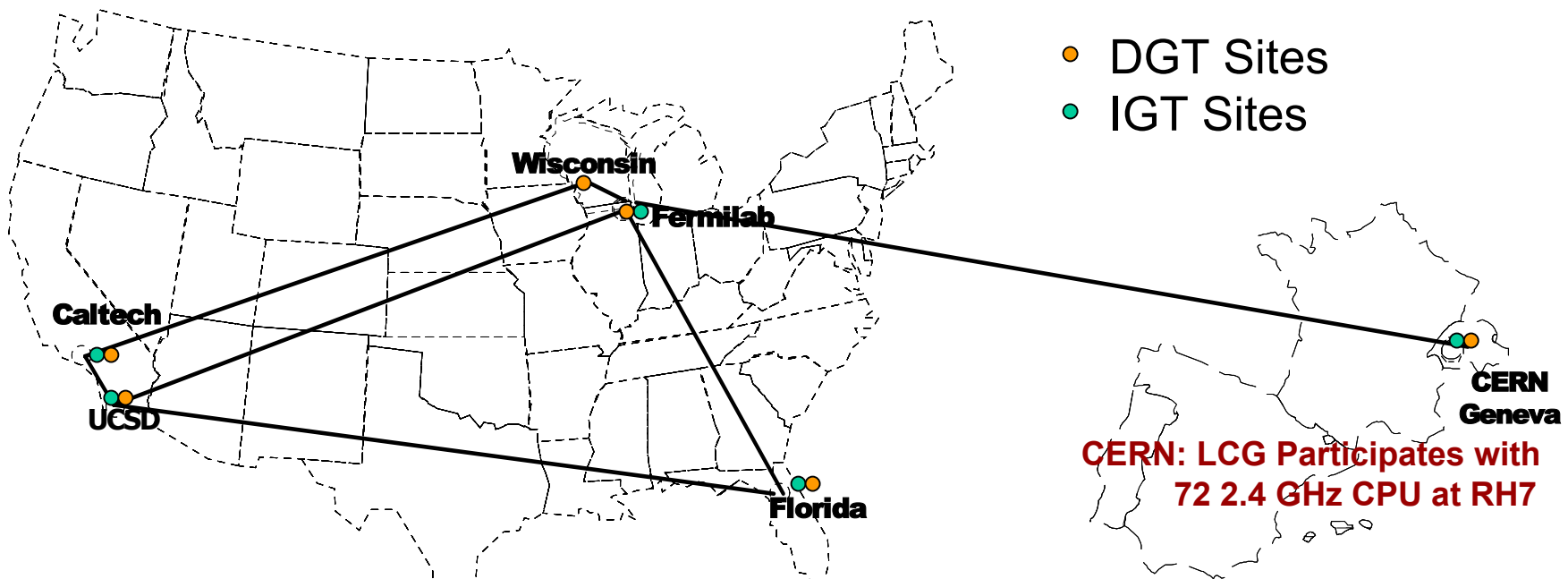


- **Operating System: Red Hat 6**
 - Required for Objectivity

Adding groups at Rice,
MIT, and Princeton
in 2003.



The Current IGT - Hardware



- Fermilab: 40 dual 750 MHz nodes + 2 servers, RH6**
- Florida: 40 dual 1 GHz nodes + 1 server, RH6**
- UCSD: 20 dual 800 MHz nodes + 1 server, RH6**
- New: 20 dual 2.4 GHz nodes + 1 server, RH7**
- Caltech: 20 dual 800 MHz nodes + 1 server, RH6**
- New: 20 dual 2.4 GHz nodes + 1 server, RH7**
- UW Madison: Not a prototype Tier-2 center, support**

**Total: 240 0.8 equiv. RH6 CPU
 152 2.4 GHz RH7 CPU**



Integration Grid Testbed Resources

Resource Allocations (1 GHz equiv. CPU) in 2002/2003 for IGT and Production Grid. (R&D Grid not included.)

	2002(IGT)	2002(PG)	2003(New)	2003(IGT)	2003(PG)
FNAL	60	0	260	10	310
Florida	80	0	175	5	250
Caltech	120	0	88	5	203
UCSD	128	0	88	5	211
Total	388	0	611	25	974

New resources for Tier-2 are from iVDGL.



Grid Efforts Integral Part of US CMS Work

Trillium Grid Projects in the US: PPDG, GriPhyN, iVDGL

- PPDG effort for CMS at Fermilab, UCSD, Caltech, working with US CMS S&C people
- Large influx of expertise and very dedicated effort from U. Wisconsin Madison through the Condor and “Virtual Data Toolkit” (VDT) teams
- We are using VDT for deployment of Grid middleware and infrastructure sponsored by PPDG, iVDGL, now adopted by EDG and LCG
- Investigating use of GriPhyN VDL technology in CMS --- virtual data is “on the map”

US CMS Development Testbed: development and explorative Grid work

- Allows us to explore technologies: MOP, GDMP, VO, integration with EU grids...
- Led by PPDG and GriPhyN staff at U. Florida and Fermilab
- All pT2 sites and Fermilab + Wisconsin involved -- ready to enlarge that effort
- This effort is mostly Grid-sponsored: PPDG, iVDGL
- Direct support from middleware developers: Condor, Globus, EDG, DataTag

Integration Grid Testbed: Grid deployment and integration

- Again using manpower from iVDGL and project funded Tier-2 operations, PPDG, GriPhyN and iVDGL sponsored VDT, PPDG sponsored VO tools, etc.



MOP

MOP is a system for packaging production processing jobs into DAGMAN format

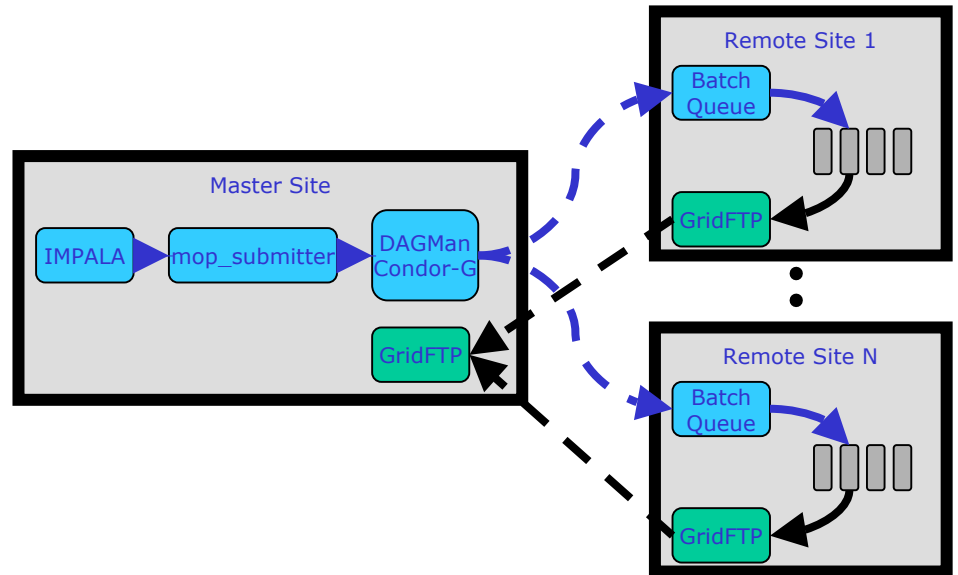
- DAGMAN format is a Directed Acyclic Graph (DAG)
- MOP uses the following DAG Nodes for each job:
 - ◆ Stage-in: Stages in needed application files, scripts, data from the submit host
 - ◆ Run: The application(s) run on the remote host
 - ◆ Stage-out: The produced data is staged out from the execution site back to the submit host
 - ◆ Clean-up: Temporary areas on the remote site are cleansed
 - ◆ Publish: Data is published to a GDMP replica catalogue after it is returned

MOP (cont'd)

Mop_submitter wraps Impala jobs in DAG format at the "MOP master" site

DAGMAN runs DAG jobs through remote sites' Globus JobManagers through Condor-G

Results are returned using GridFTP. Though the results are also returned to the MOP master site in the current IGT running, this does not have to be the case.

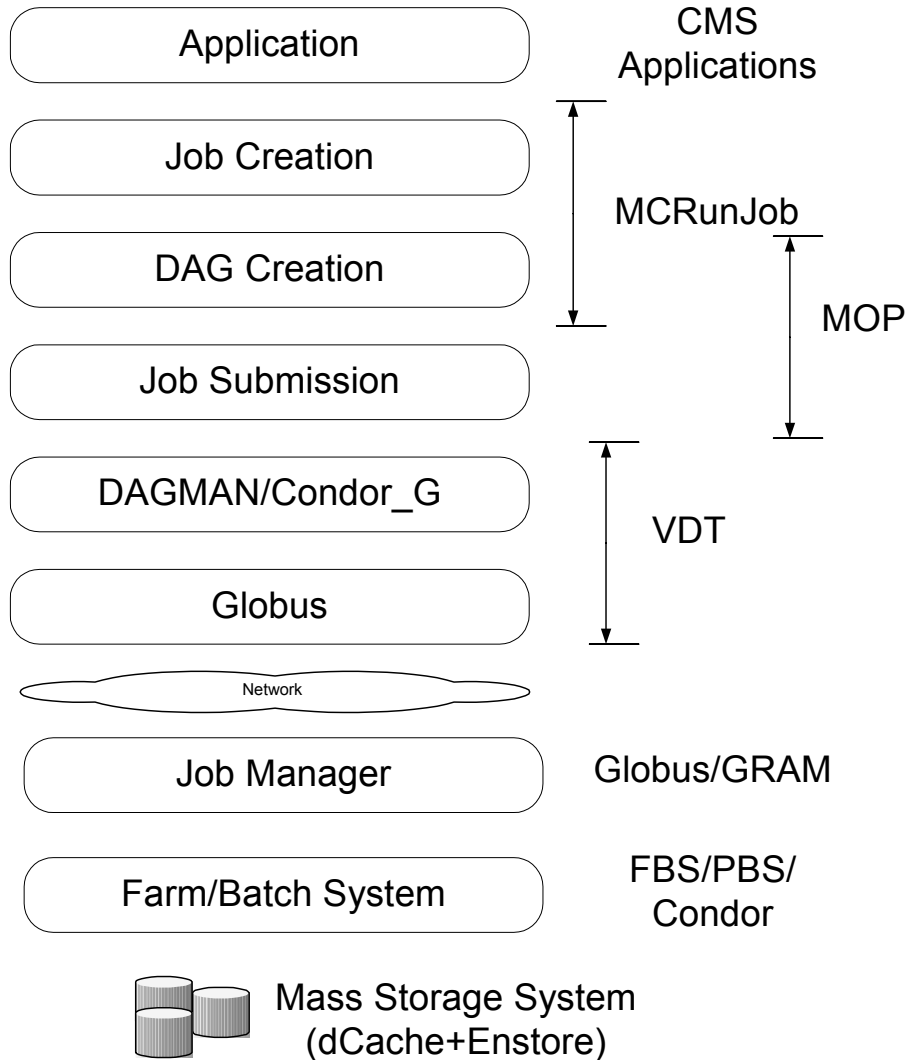


UW Madison is the MOP master for the USCMS Grid Testbed

FNAL is the MOP master for the IGT and the Production Grid



The CMS IGT “Stack”



The CMS IGT “stack” comprises nine layers. The Application layer contains only CMS executables. The Job Creation layer comprises CMS provided tools MCRunJob and Impala. Neither MCRunJob nor Impala are specifically “grid aware.” Then there is a DAG Creation layer and a Job Submission layer. Both functionalities are provided by MOP. Jobs are submitted to DAGMAN which, through Condor-G, manages jobs run on remote Globus Job Managers. Finally, there is a local Farm or Batch System used by Globus GRAM to manage jobs. In the case of the IGT, the local Batch manager was always FBSNG or Condor. Scheduling and Integrated monitoring are not present.



Preparing CMS for the Grid

Making US CMS and CMS fit for working in a Grid Environment

- Production environment and operations
- Deployment, configuration and management of systems, middleware environment
- Monitoring of the Grid fabric and configuration
- Providing information services
- Management of user base on Grid, interfacing to local specifics at Universities and labs (VO)
- Devising a scheme for software distribution and configuration (DAR, PACMAN) of CMS application s/w

In all these areas we have counted on significant contributions from the Grid Projects

Thus these efforts are being tracked in the project through the US CMS S&C WBS



Global LCG and US CMS Grid

We expect that the LCG will address many issues related to running a distributed environment and propose implementations

- This is expected from the Grid Deployment Board Working Groups

A “cookie cutter” approach will not be a useful first step

- We are not interested in setting up identical environments at a smallish set of regional centers
- Nor on defining a minimal environment down to the last version level, etc

With the IGT (and the “EDG CMS stress test”) we should be beyond this

- In the US we already do have working (sub-) Grids: IGT, Atlas Testbed, Worldgrid -- it can be done!
- Note however, a large part of the functionality is either missing or of limited scalability, and/or experiment-specific software

From the start we need to employ a model that allows sub-organizations or sub-Grids to work together

- There will always be site-specifics:
 - ◆ e.g. different procurement procedures, DOE-lab security, time zones, etc
- The whole US CMS project and funding model foresees the Tier-1 center takes care of much of the US-wide support issues, and assumes that half of the resources come from Tier-2 centers with limited local manpower

This is cost-effective and a good way to proceed towards the goal of a distributed LHC research environment

- and on the way broadens the base and buy-in to make sure we are successful
- BTW: US CMS has always de-emphasized the role of the Tier-1 prototype to provide “raw power”, but rather is counting on assembling the efforts from a distributed set of sites in the IGT and production grid



Dealing with the LCG Requirements

We are adapting to work within the LCG approach:

- Grid Use Cases and Scenarios
 - ◆ US participation in the GAG, follow up of the HEPCAL RTAG
- Working Groups in the Grid Deployment Board
 - ◆ “invited” to work on specific issues (Grid File Access)
 - ◆ Work through the US GDB representative (Vicky White)
- Architects Forum for the application area
 - ◆ Proposed and started a sub-group with US participation refining the blueprints of Grid Interfaces

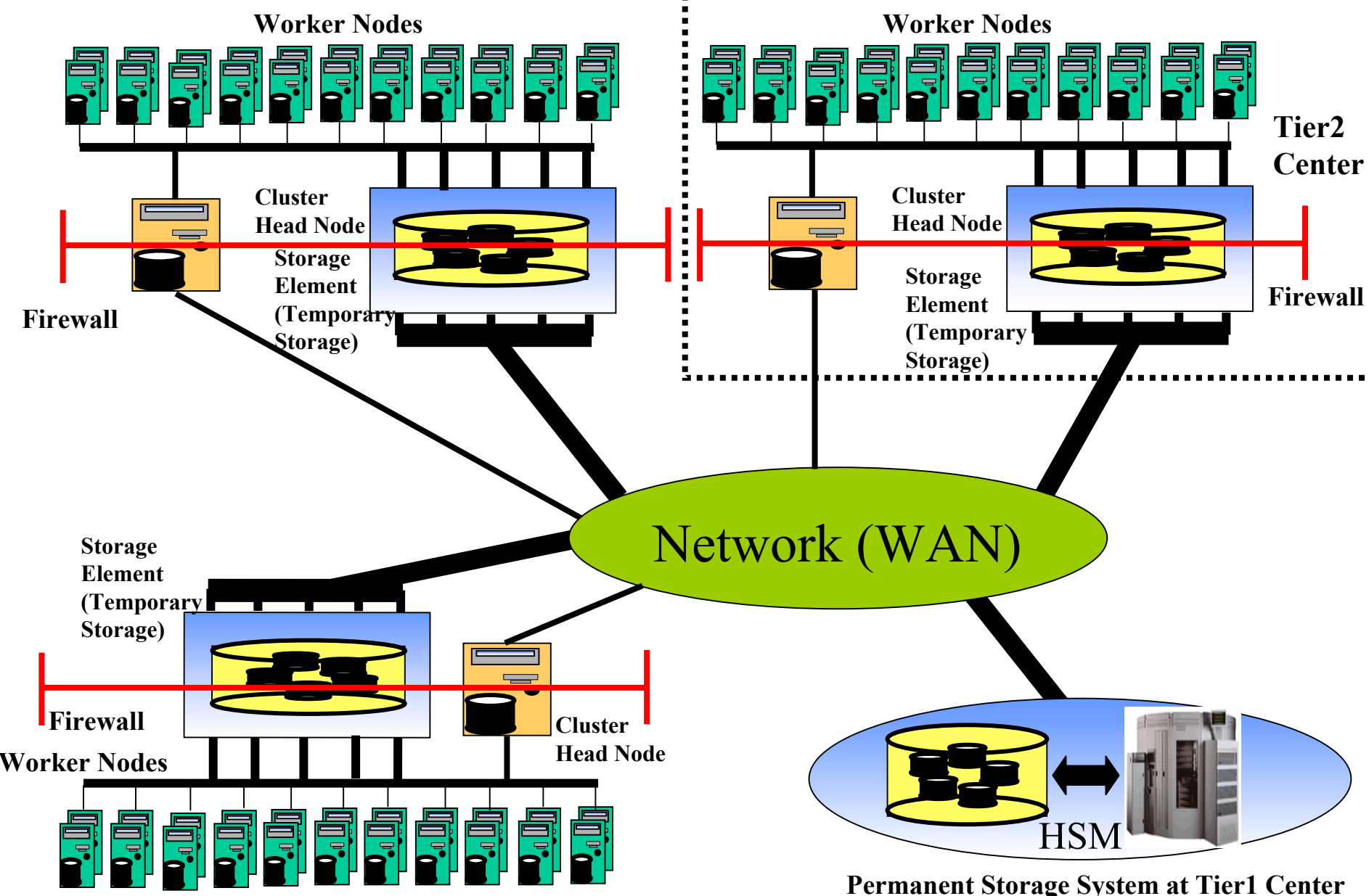
We have to ensure our leadership position for US LHC S&C

- We have to develop a clear understanding of what is workable for the US
- We have to ensure that appropriate priorities are set in the LCG on a flexible distributed environment to support remote physics analysis requirements
- We have to be in a position to propose solutions, and in some cases to propose alternative solutions, that would better meet the requirements of CMS and US CMS

US CMS has setup itself to be able to learn, prototype and develop while providing a production environment to cater to CMS, US CMS and LCG demands

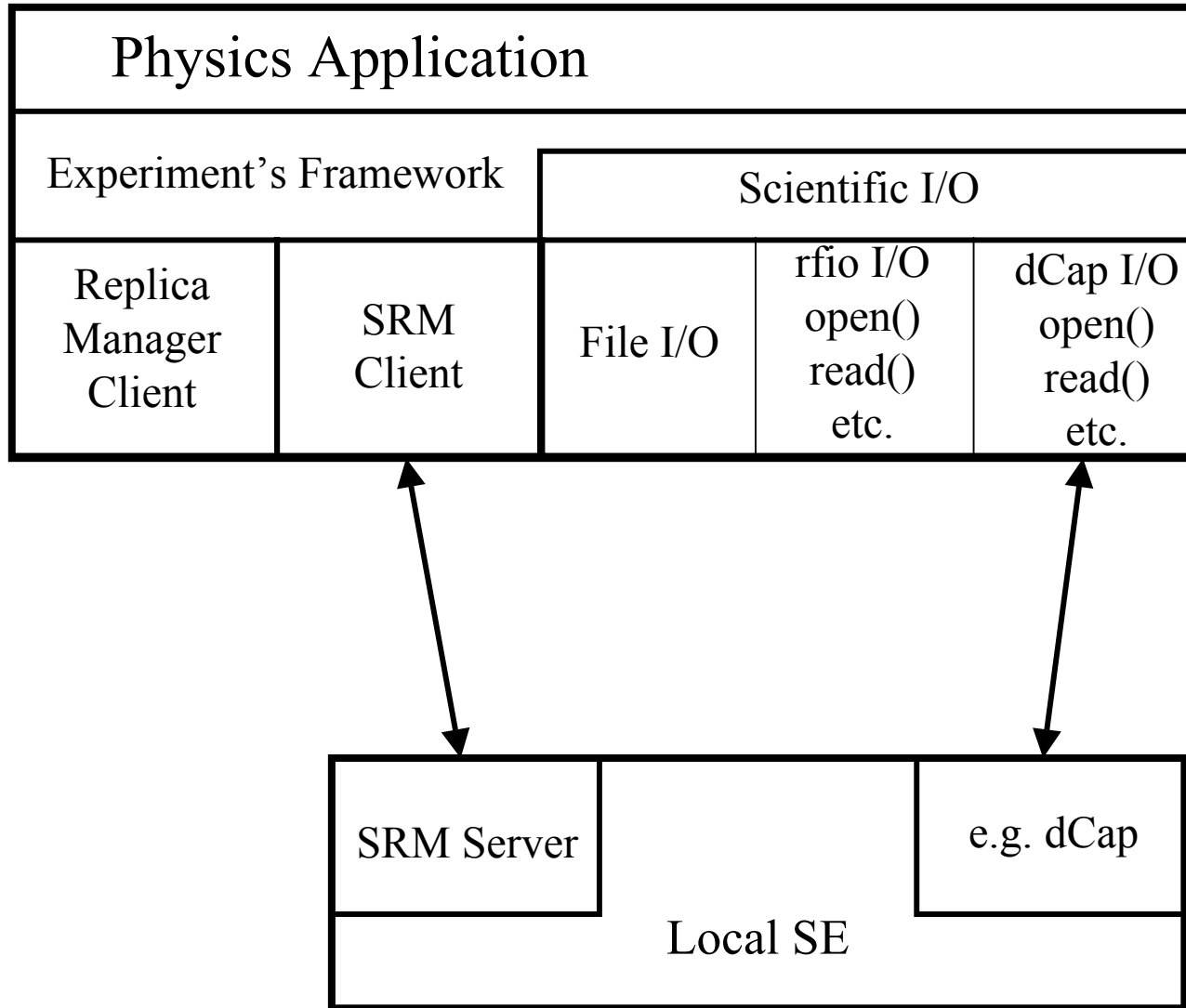


Implementation of the Tier1/Tier2 Architecture



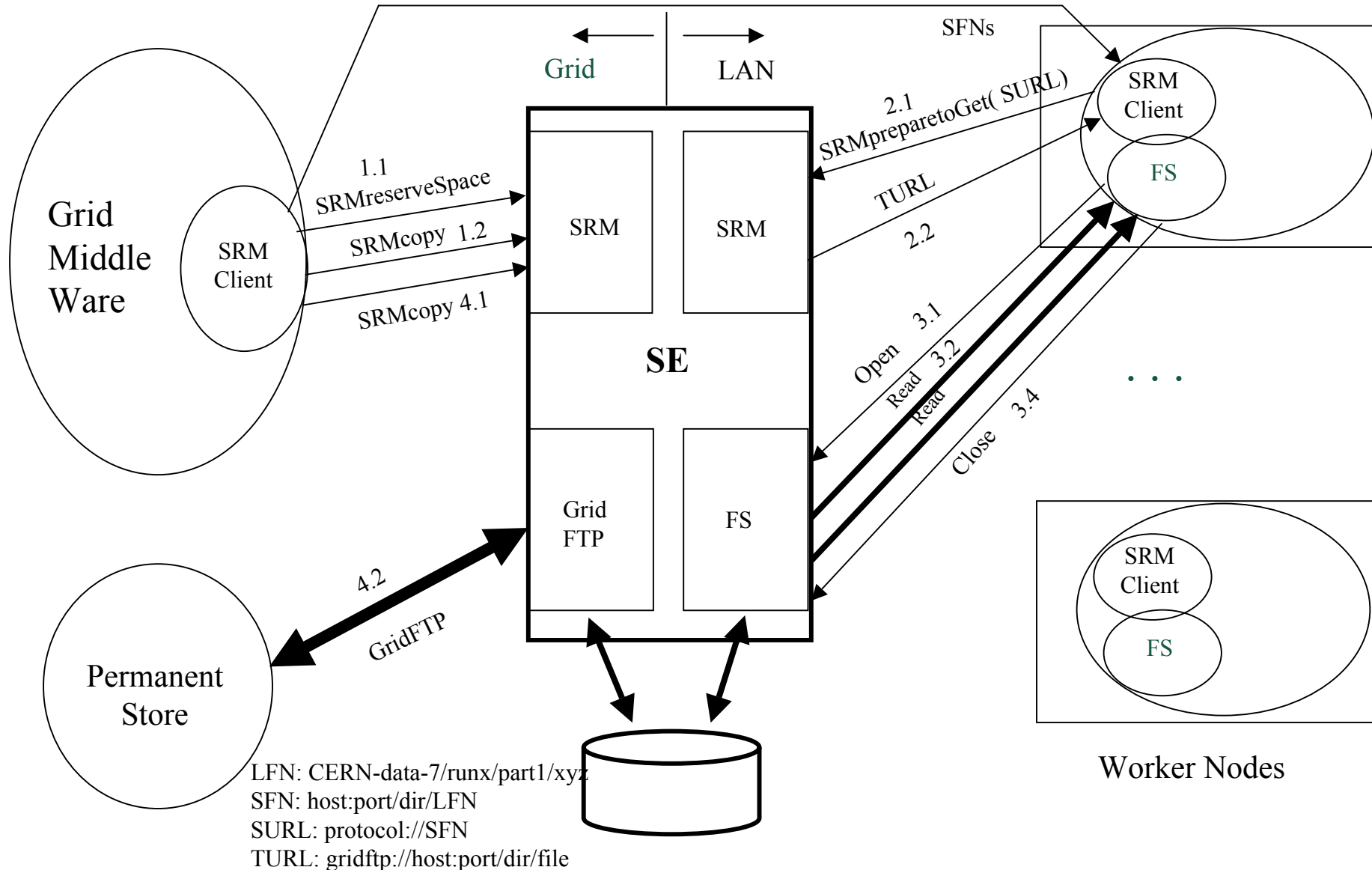


Software Layering of an Application and a Storage Element





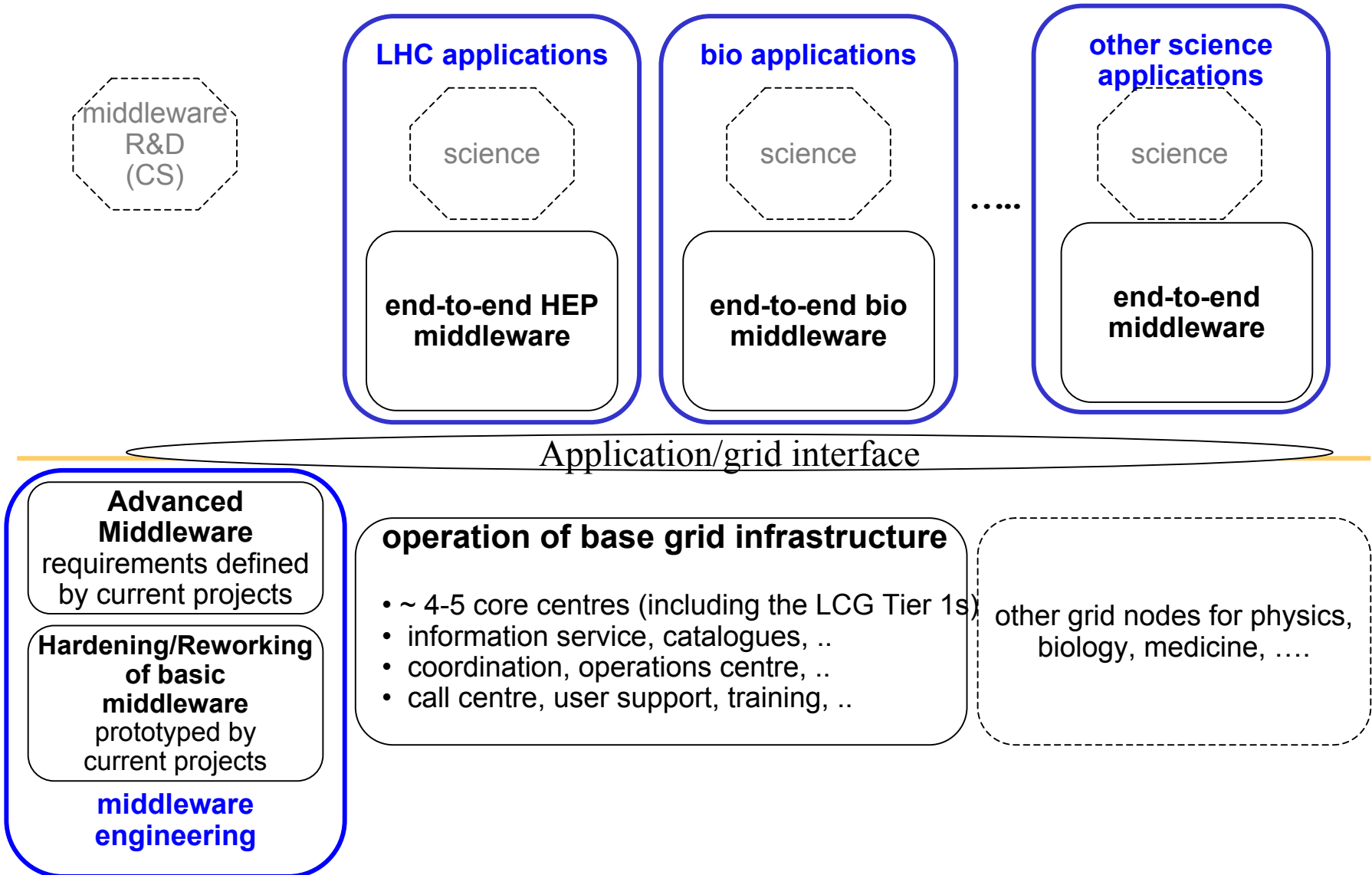
Local and Grid Interactions with a Storage Element





The Global Picture

Development of A Science Grid Infrastructure (L.Robertson)





... and the “missing pieces”

Transition to **Production Level Grids**

- middleware support,
- error recovery,
- robustness,
- 24x7 Grid fabric operations,
- monitoring and system usage optimization,
- strategy and policy for resource allocation,
- authentication and authorization,
- simulation of grid operations,
- tools for optimizing distributed systems
- etc.

Also: much needed functionality of a data handling system is still missing! Even basic functionality

- like global catalogs and location services,
- Storage management,
- High network/end-to-end throughput for Terabyte transfers



Focus Vision on Enabling Science

What does it take to do LHC science in a global setting?

- Avoid focus on setting up big distributed computing facility:
racks of equipment distributed over \llcorner T1 centers, batch jobs running in production

Focus on a global environment to enable science communities:

- How can we achieve that US Universities are full players in LHC science?
- What capabilities and services are needed to do analysis 9 time zones away from CERN?
- (What are the obstacles for remote scientists in existing experiments?)

We are analyzing at a set of scenarios

- “science challenges” as opposed to Grid use cases:
- exotic physics discovery, data validation and trigger modifications etc.

We identify then the capabilities needed from the analysis environment and some of the CS and IT to enable those capabilities

A corresponding project proposal has been submitted to NSF in March and is being followed up in a sub-group of the LCG Architecture Forum



Typical Science Challenge

A physicist at a U.S. university presents a plot at a videoconference of the analysis group she is involved in. The physicists would like to verify the source of all the data points in the plot.

- The detector calibration has changed several times during the year and she would like to verify that all the data has a consistent calibration
- The code used to create the standard cuts has gone through several revisions, only more recent versions are acceptable
- Data from known bad detector runs must be excluded
- An event is at the edge of a background distribution and the event needs to be visualized



Typical Science Challenge

A physicist at a U.S. university presents a plot at a videoconference of the analysis group she is involved in. The physicists would like to verify the source of all the data points in the plot.

- The detector calibration has changed several times during the year and she would like to verify that all the data has a consistent calibration
- The code used to create the standard cuts was gone through several revisions, only more recent versions are acceptable
- Data from known bad detector runs must be excluded
- An event is at the edge of a background distribution and the event needs to be visualized

**Metadata
Data Provenance
Data Equivalence
Collaboratory Tools
User Interfaces**



Science Challenges

A small group of University physicists are searching for a specific “exotic” physics signal, as the LHC event sample increases over the years. Instrumental for this search is a specific detector component that those University groups have been involved in building. Out of their local detector expertise they develop a revolutionary new detector calibration method that indeed significantly increased the discovery reach. They obtain permission to use a local University compute center for Monte Carlo generation of their exotic signal. Producing the required sample and tuning the new algorithm takes many months.

After analyzing 10% of the available LHC dataset of 10 Petabytes with the new method they indeed find signals suggesting a discovery! The collaboration asks another group of researchers to verify the results and to perform simulations to increase the confidence by a factor three. There is a major conference in few weeks – will they be able to publish in time?

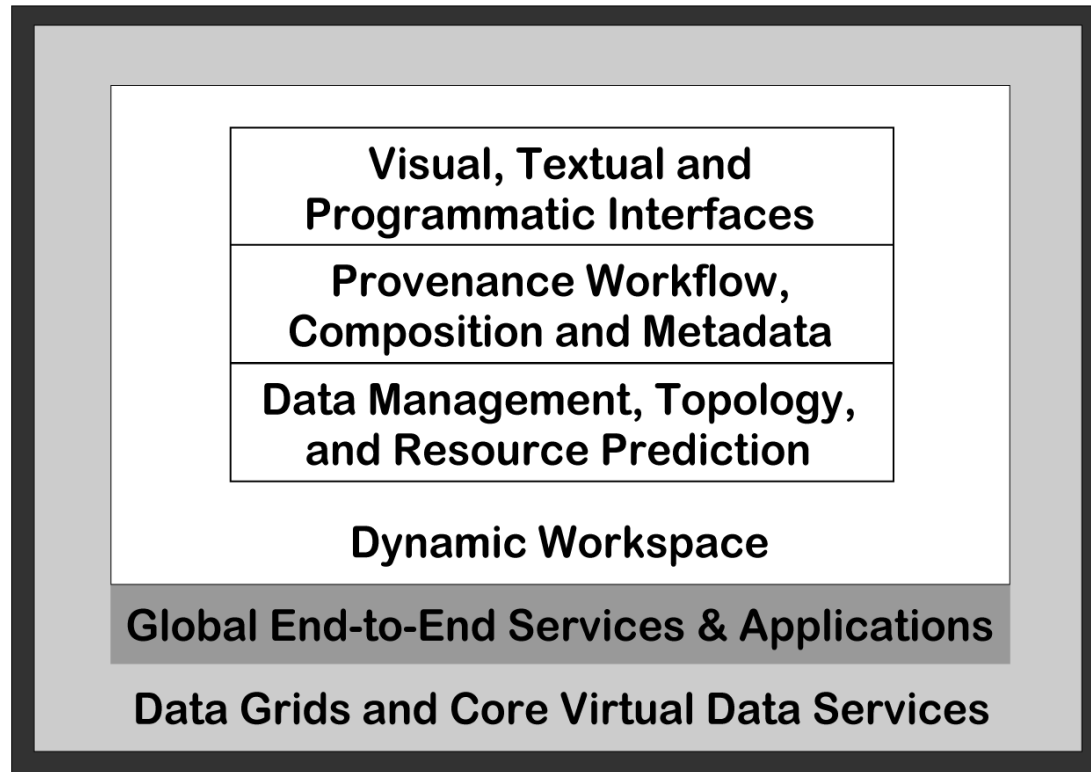
- (a) access the meta-data, share the data and transfer the algorithms used to perform the analysis;
- (b) quickly have access to the maximum available physical resources to execute the expanded simulations, stopping other less important calculations if need be;
- (c) decide to run their analyses and simulations on non-collaboration physical resources to the extent possible – depending on cost, effort and other overheads;
- (d) completely track all new processing and results;
- (e) verify and compare all details of their results;
- (f) provide partial results to the eager researchers to allow them to track progress towards a result and/or discovery;
- (g) provide complete and up to the minute information to the publication decision committee to allow them to quickly take the necessary decisions.

Create and manage dynamic temporary private grids; provide complete provenance and meta-data tracking and management for analysis communities; enable community based data validation and comparison; enable rapid response to new requests; provide usable and complete user interaction and control facilities



The Global Environment

Globally Enabled **Analysis Communities** (a proposal was submitted to NSF)



Enabling **Global Collaboration** (a medium-sized ITR proposal)



Goals of the Proposal

Provide individual physicists and groups of scientists capabilities from the desktop that allow them:

- To participate as an equal in one or more “Analysis Communities”
- Full representation in the Global Experiment Enterprise
- To on-demand receive whatever resources and information they need to explore their science interest while respecting the collaboration wide priorities and needs.

Environment for CMS (LHC) Distributed Analysis on the Grid

- **Dynamic Workspaces** - provide capability for individual and community to request and receive expanded, contracted or otherwise modified resources, while maintaining the integrity and policies of the Global Enterprise.
- **Private Grids** - provide capability for individual and community to request, control and use a heterogeneous mix of Enterprise wide and community specific software, data, meta-data, resources.



Physics Analysis in CMS

The Experiment controls and maintains the global enterprise:

Hardware: Computers, Storage (permanent and temporary)

Software Packages: physics, framework, data management, build and distribution mechanisms; base infrastructure (operating systems, compilers, network, grid);

Event and Physics **Data and Datasets**

Schema which define: meta-data, provenance, ancillary information (run, luminosity, trigger, Monte-Carlo parameters, calibration etc)

Organization, Policy and Practice

Analysis Groups - Communities - are of 1 to many individuals

Each community is part of the Enterprise :

- Is assigned to or shares the total Computation and Storage
- Can access and modify software, data, schema (meta-data)
- is subject to the overall organization and management

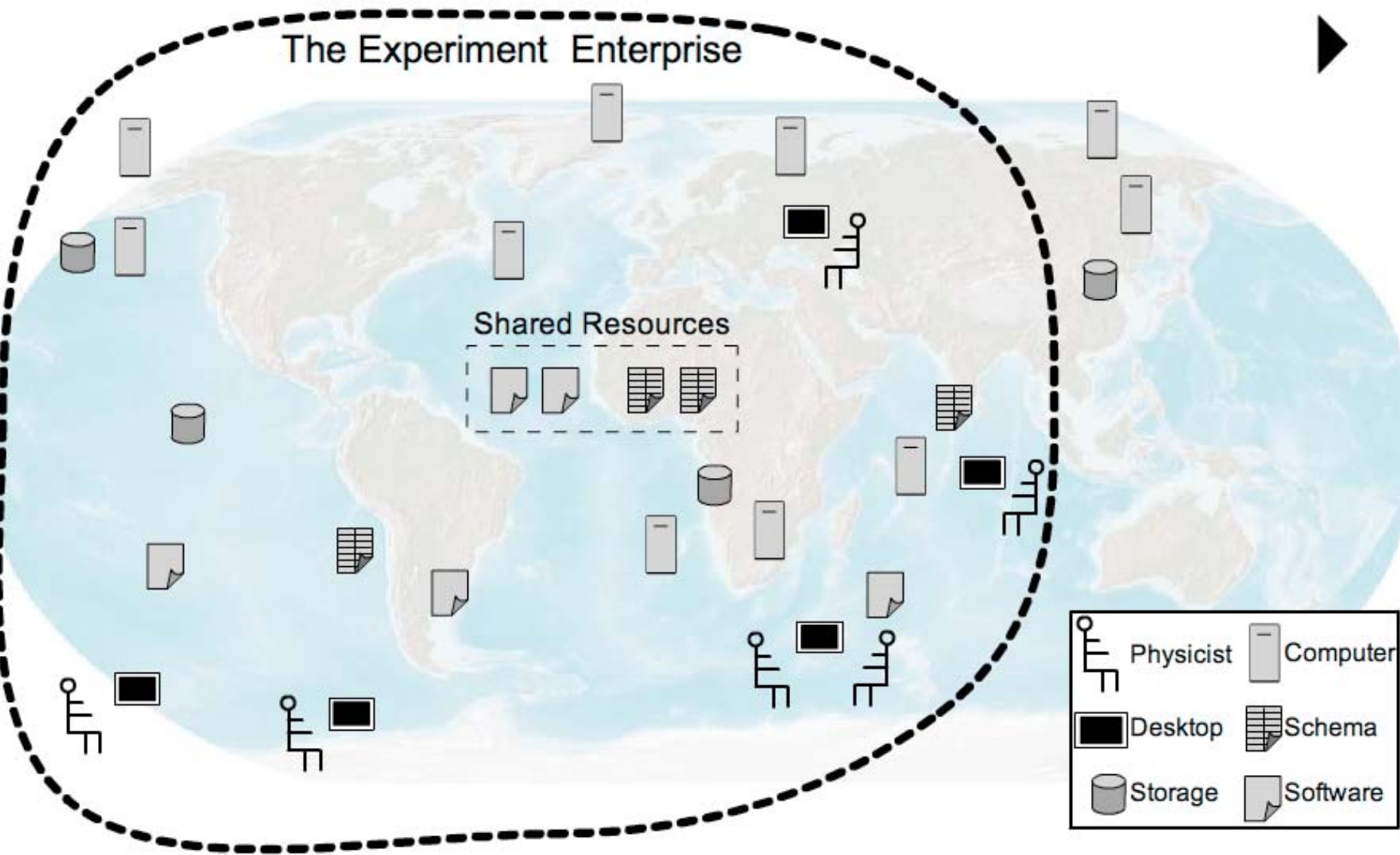
Each community has local (private) control of

- Use of outside resources e.g. local institution computing centers
- Special versions of software, datasets, schema, compilers
- Organization, policy and practice

We must be able to reliably and consistently move resources & information in both directions between the Global Collaboration and the Analysis Communities.

Communities can share among themselves.

The Experiment Enterprise

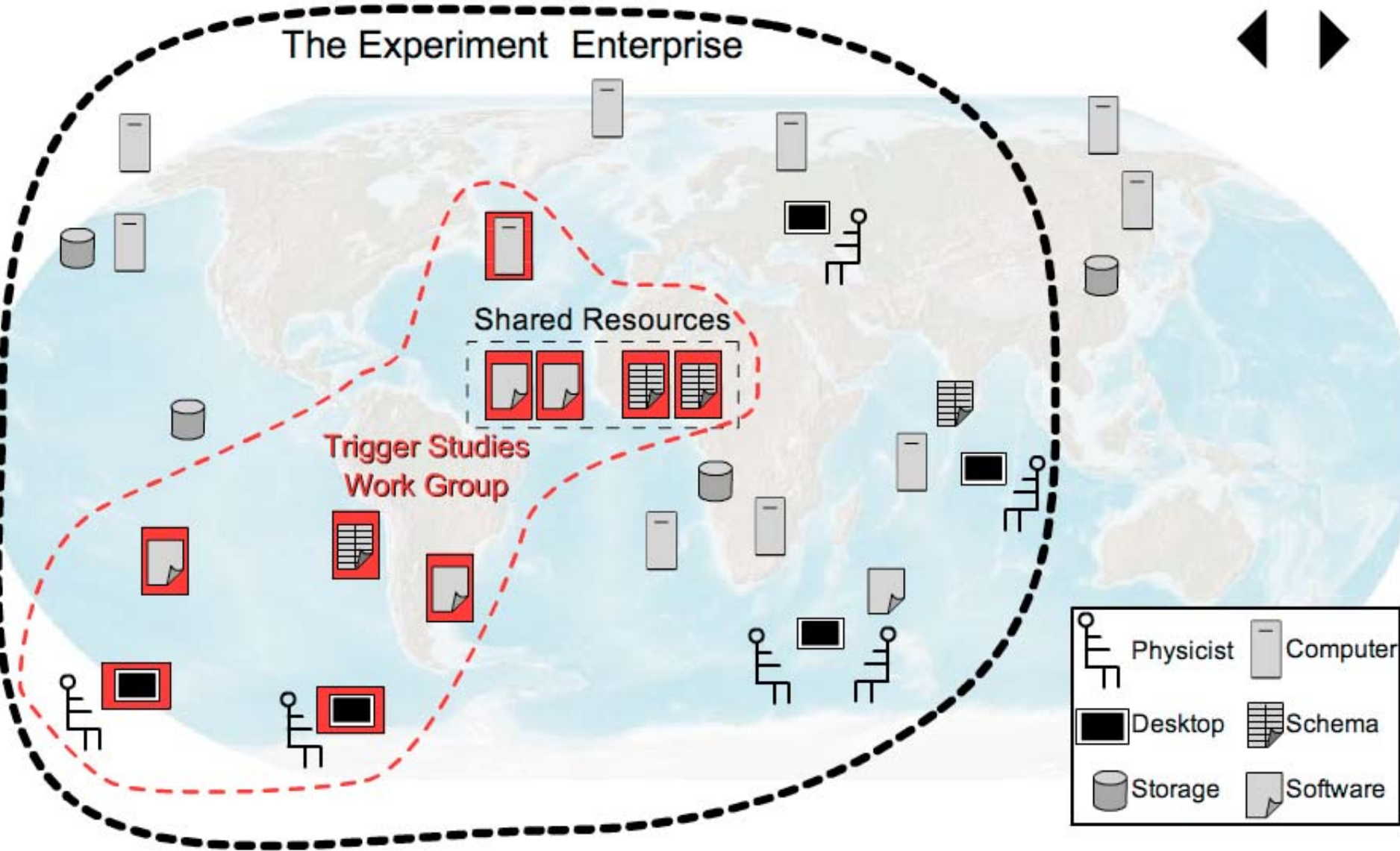


	Physicist		Computer
	Desktop		Schema
	Storage		Software



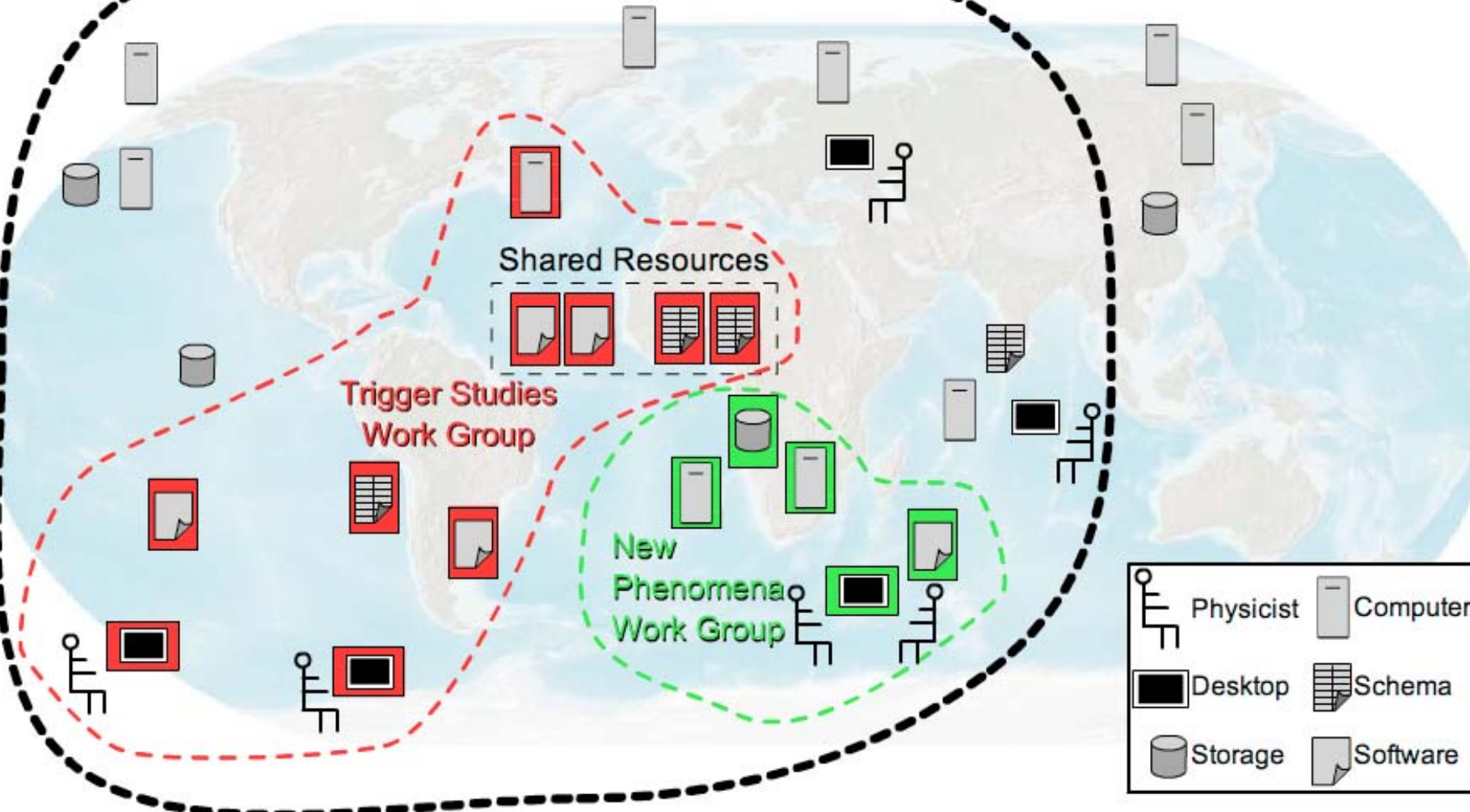


The Experiment Enterprise

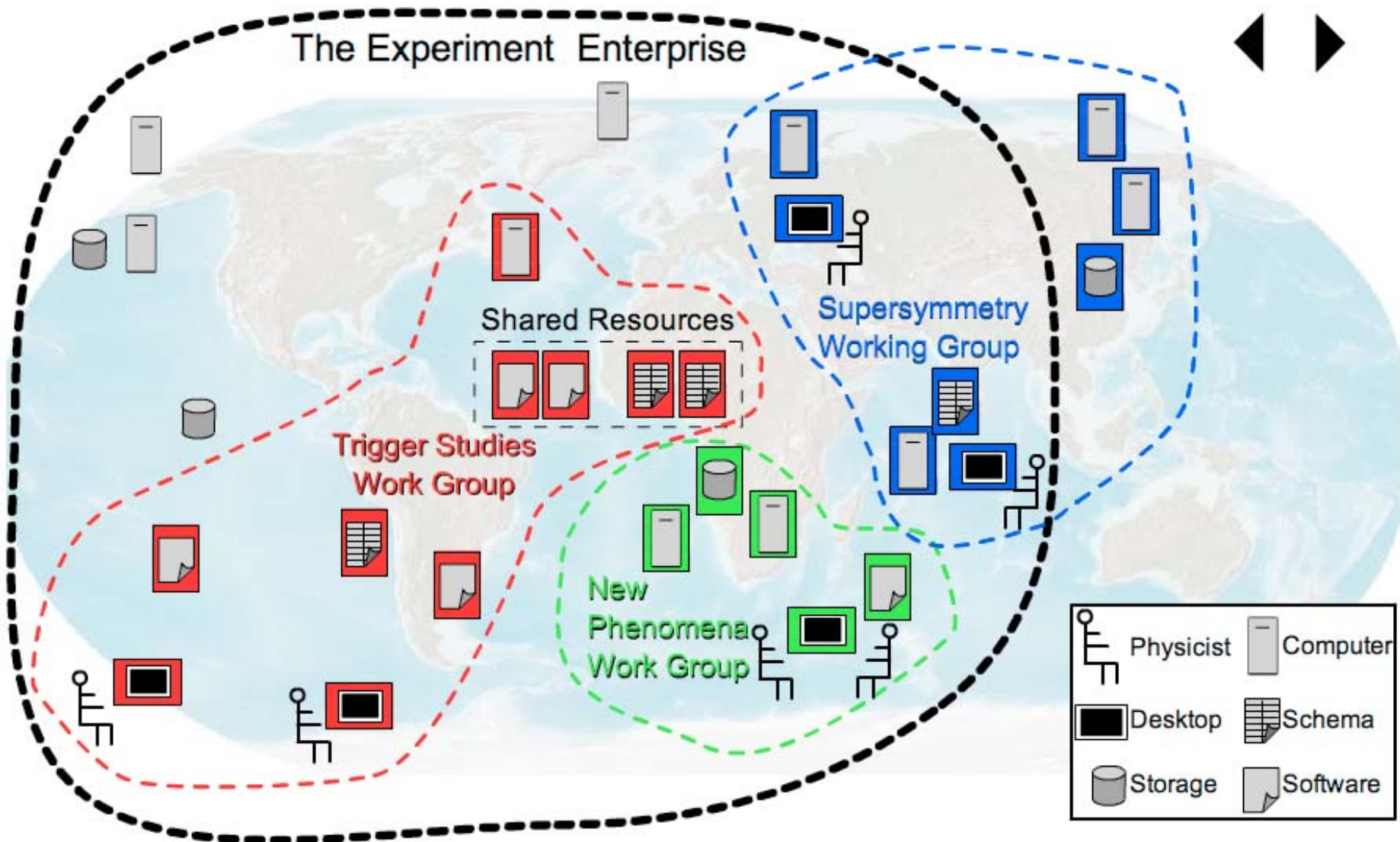


	Physicist		Computer
	Desktop		Schema
	Storage		Software

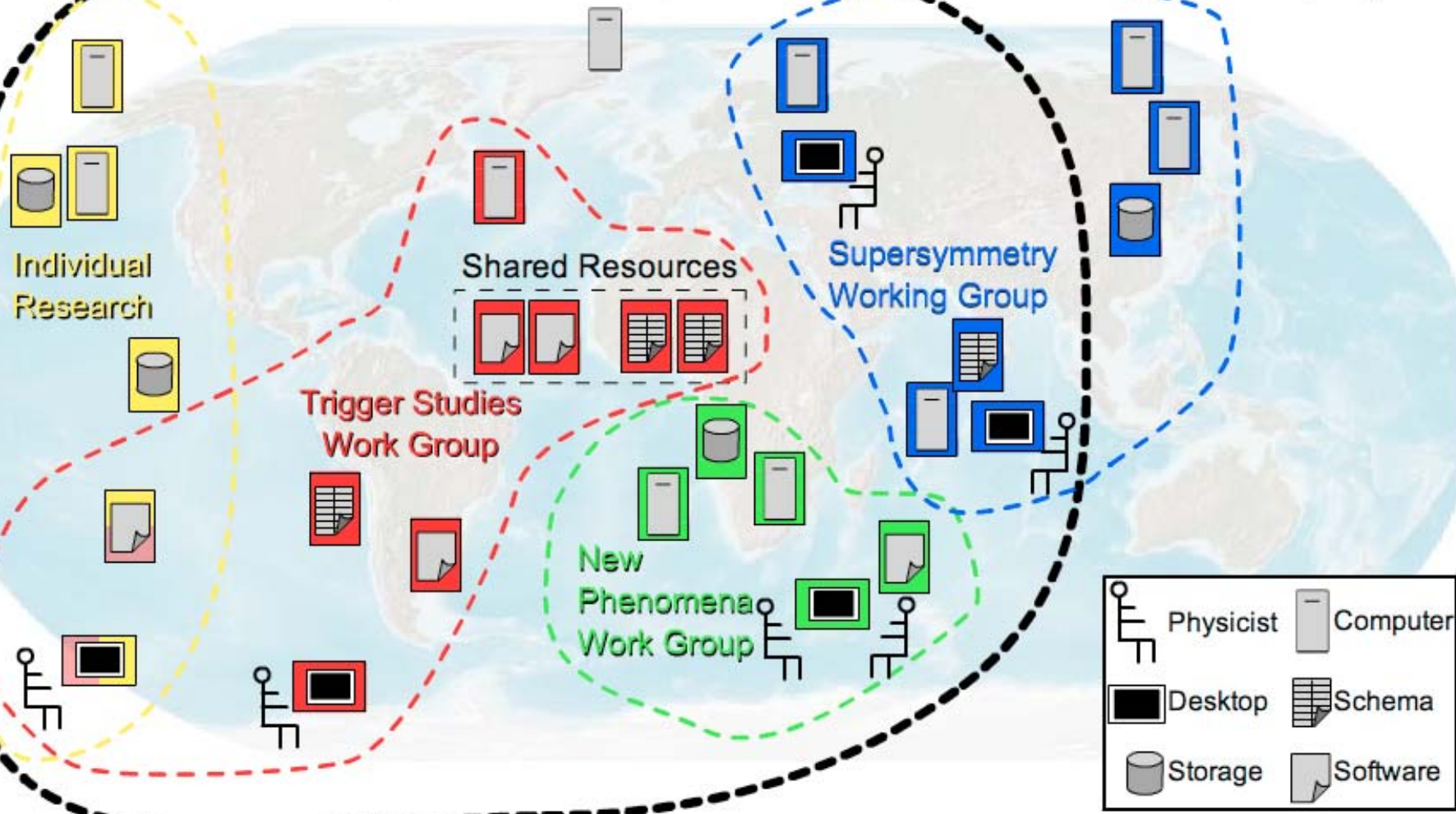
The Experiment Enterprise



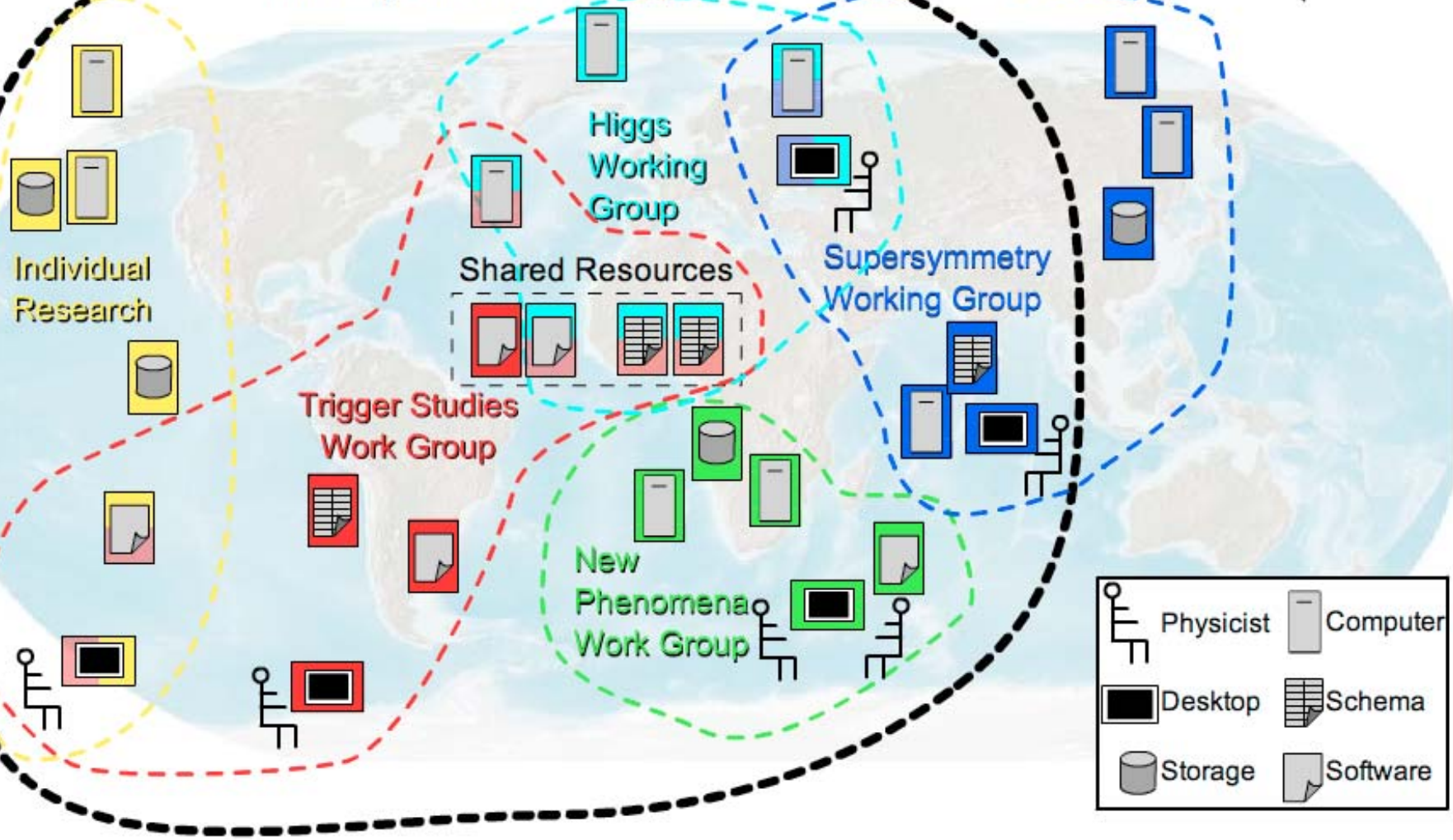
	Physicist		Computer
	Desktop		Schema
	Storage		Software



The Experiment Enterprise



The Experiment Enterprise





This Proposal Addresses Key Issues

- remote analysis groups and individual physicists will be enabled to perform local/remote analyses
- reliable and quick validation, trusted by the collaboration
- demonstrate and compare methods and results reliably and improve the turnaround time to physics publications
- quickly respond to and decide upon resource requests from analysis groups/physicists, minimizing impact to the rest of the collaboration
- established infrastructure for evolution and extension for its long life time
- enable small groups to perform reliable exploratory analyses on their own
- increased potential for individual/small community analyses and discovery
- analysis communities will be assured they are using a well defined set of software and data

This looks obvious and is clearly required for the success of LHC RP

This looks daunting and scarily difficult and is indeed far from what has been achieved in existing experiments

We do need the intellectual involvement and engagement of CS and IT!



Conclusions on US CMS Grids

The Grid approach to US CMS S&C is technically sound,
and enjoys strong support and participation from
U.S. Universities and Grid Projects

→ We need large intellectual input and involvement, and significant R&D to build the system

US CMS is driving US Grid integration and deployment
US CMS has proven that the US Tier-1/Tier-2 User Facility (Grid-) system
can indeed work to deliver effort and resources to CMS and US CMS!

We have a unique opportunity of proposing our ideas to others,
of doing our science in global, open, and international collaboration

→ That goes beyond the LHC and beyond HEP