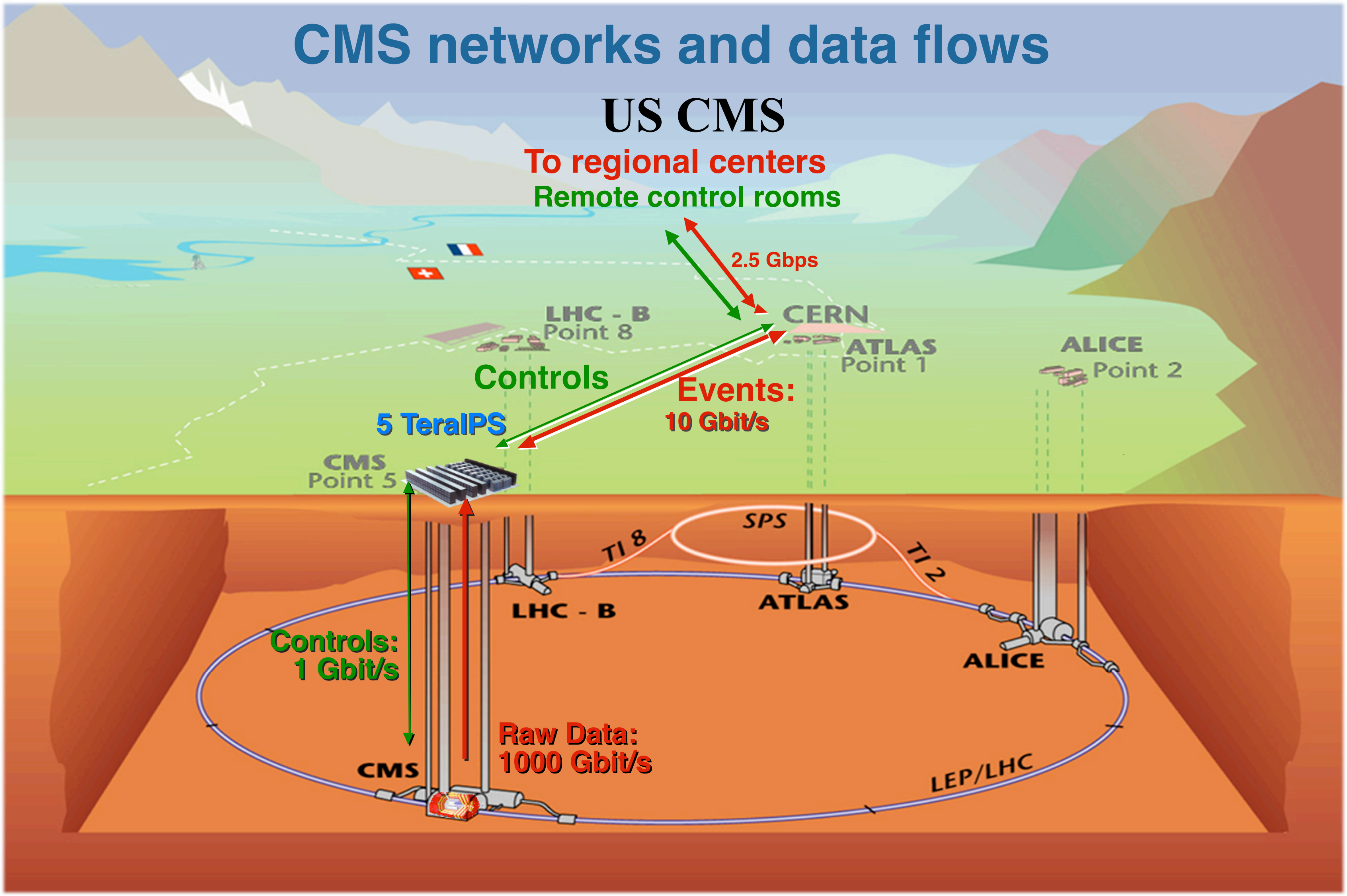
Ruth Pordes

LSU Workshop

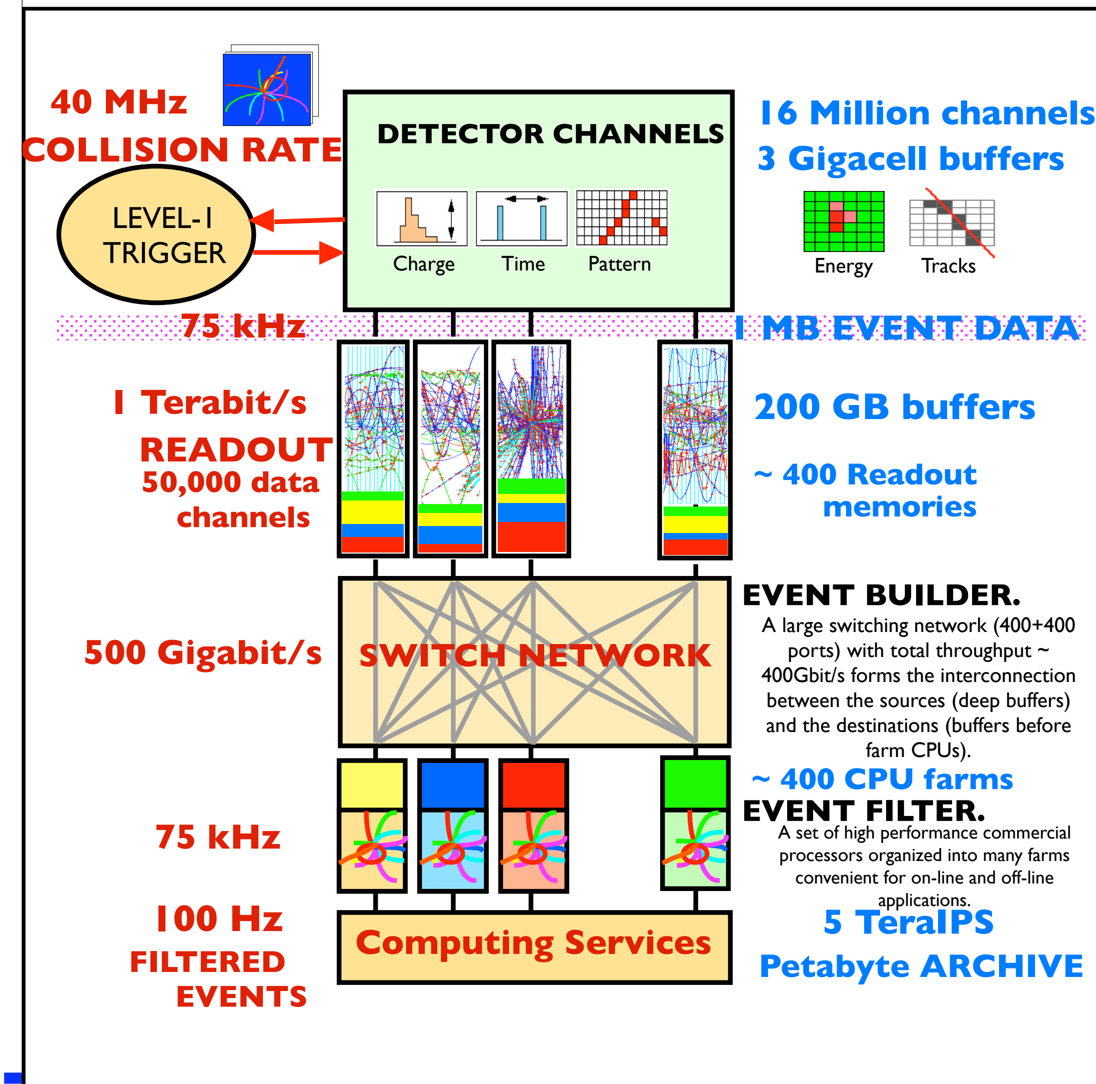
CMS - a High Energy Physics Experiment which will store >50Petabytes of data, spend >$100M on computing over more than a decade. US CMS will rely, in the critical path, on Grid Computing to get its Physics Results.

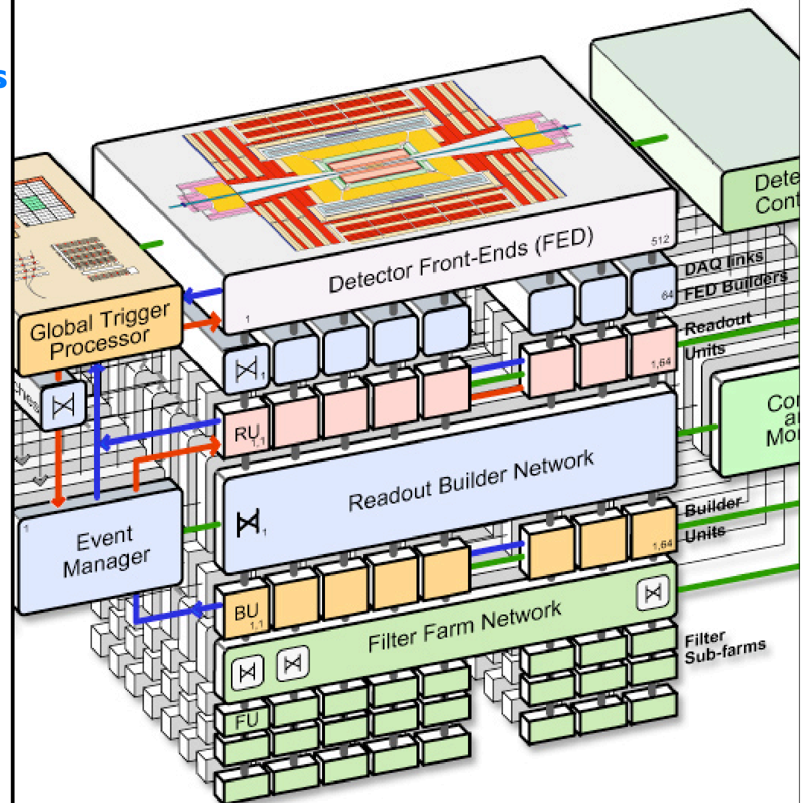
It plans to do this through contributing to and benefiting from a National Grid Infrastructure in the US which will be Open to any Sciences - the Open Science Grid

CMS Data Flow

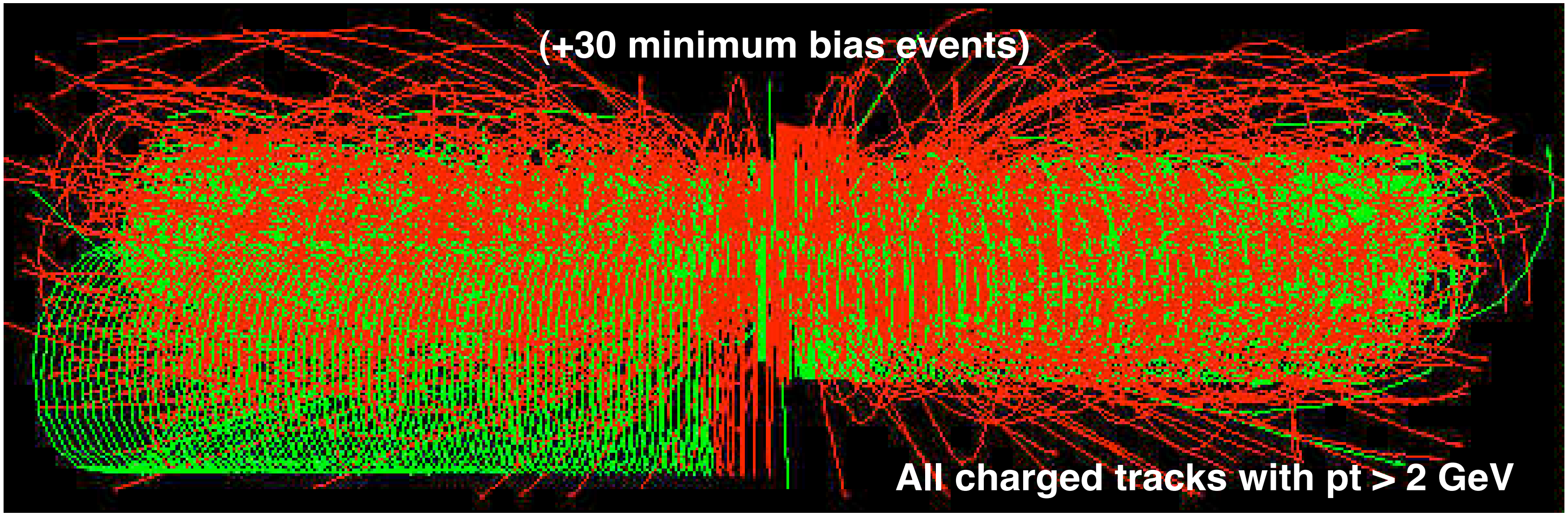


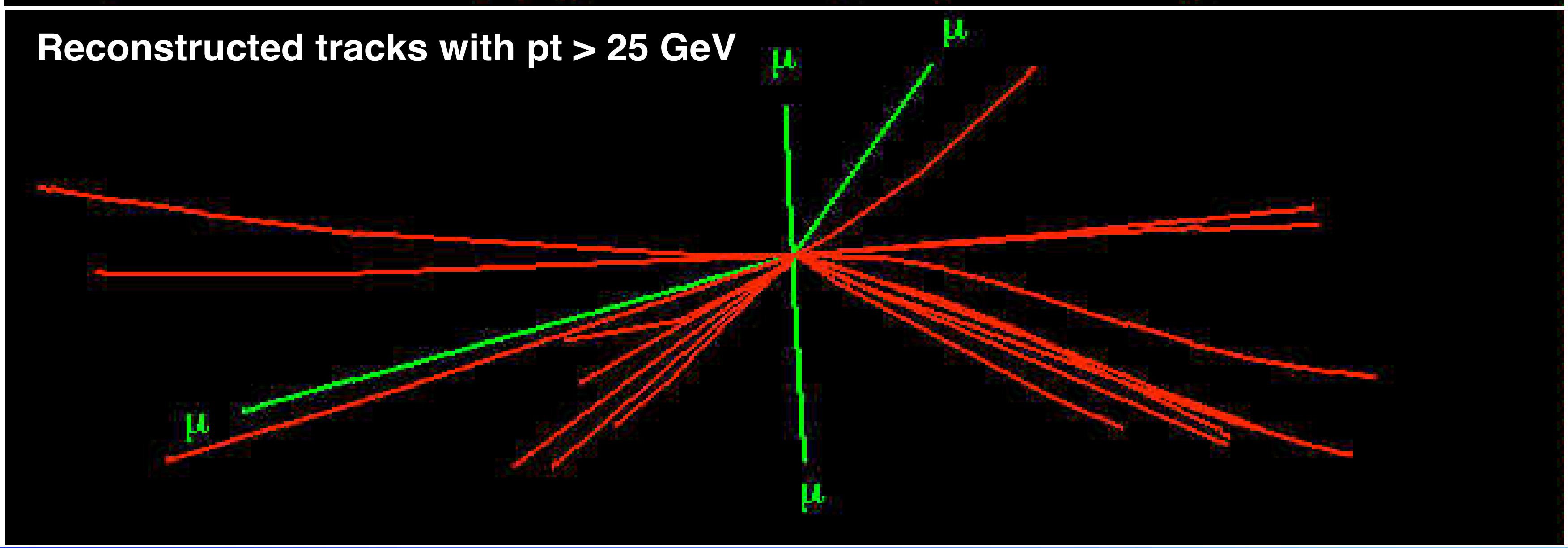
Acquiring the Data





From Data to Physics





CMS Computing Model

Support global access to and analysis of acquired data from 2007.

- Provide framework to support efficient analysis, access to calibrations, simulation production in a widely distributed standard environment.
- Provide services to allow experiment validation and so publication of results.

Develop global network of centers for analysis of the data

- Tier-0 at CERN: one copy of raw data, one reconstruction pass through the data, distribution of output from reconstruction to Tier-1s.
- 6-10 Tier-1s in regional centers worldwide: collective responsibility for 2nd copy of raw data, reprocessing, distribution to Tier-2s,archiving of Tier-2 output.
- 25 Tier-2s dependent on Tier-1s (5 in US):collectively responsible for simulation production and each for analysis needs of 20-100 physicists.

Depends on

- Robust, high performance, managed networks
- Interoperable, ubiquitous, production Grid Services.

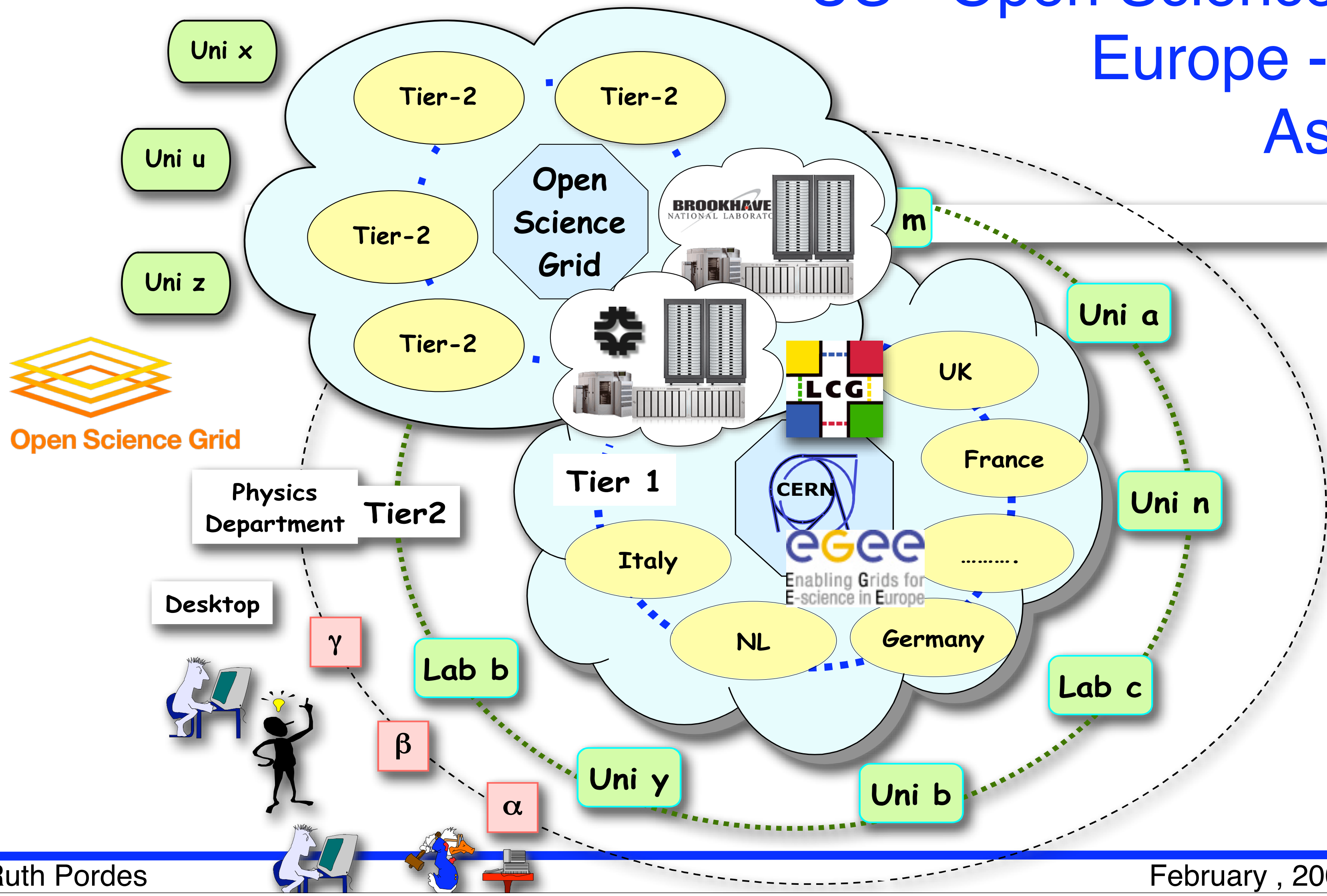
CMS Computing Model

a Global Federation of Grids:

US - Open Science Grid

Europe - LCG

Asia - ?



Plan to Store and Distribute 10s-100Petabytes Worldwide

Event Format	Content	Purpose	Event size (MByte)	Events / year	Data volume (PByte)
DAQ-RAW	Detector data in FED format and the L1 trigger result.	Primary record of physics event. Input to online HLT	1-1.5	1.5×10^9 $= 10^7$ seconds $\times$ 150Hz	–
RAW	Detector data after on-line formatting, the L1 trigger result, the result of the HLT selections ("HLT trigger bits"), potentially some of the higher-level quantities calculated during HLT processing.	Input to Tier-0 reconstruction. Primary archive of events at CERN.	1.5	3.3×10^9 $= 1.5 \times 10^9$ DAQ events $\times 1.1$ (dataset overlaps) $\times 2$ (copies)	5.0
[illegible]	[illegible]	[illegible] ... of tracks, etc.	[illegible]	...0^9 ... 10^9 DAQ events ... (dataset overlaps) ... copies of 1st pass)+ ... eprocessings/year)]	2.1
AOD	Reconstructed objects (tracks, vertices, jets, electrons, muons, etc.). Possible small quantities of very localized hit information.	Physics analysis	0.05	53×10^9 $= 1.5 \times 10^9$ DAQ events $\times 1.1$ (dataset overlaps) $\times 4$ (versions/year) $\times 8$ (copies per Tier $-$ 1)	2.6
TAG	Run/event number, high-level physics objects, e.g. used to index events.	Rapid identification of events for further study (event directory).	0.01	–	–
FEVT	Term used to refer to RAW+RECO together (not a distinct format).		–	–	–

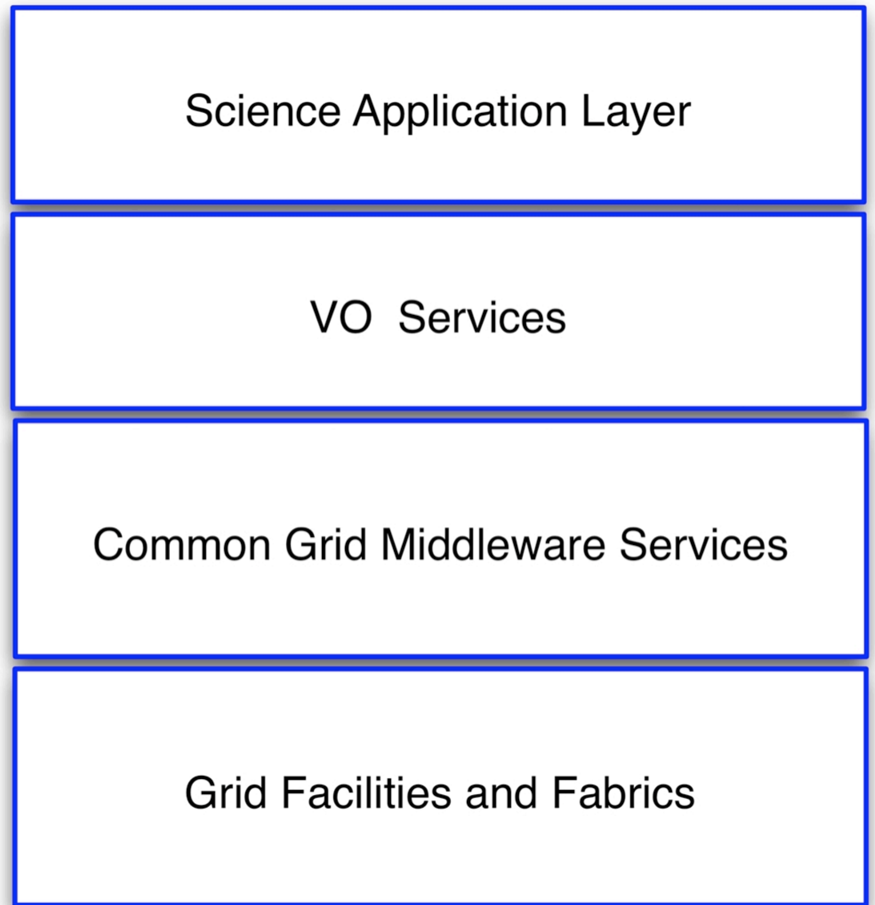
Table 2.2: CMS event formats at LHC startup, assuming a luminosity of $\mathcal{L} = 2 \times 10^{33} \mathrm{cm}^{-2}\mathrm{s}^{-1}$.

Open Science Grid is the Means

Open Science Grid

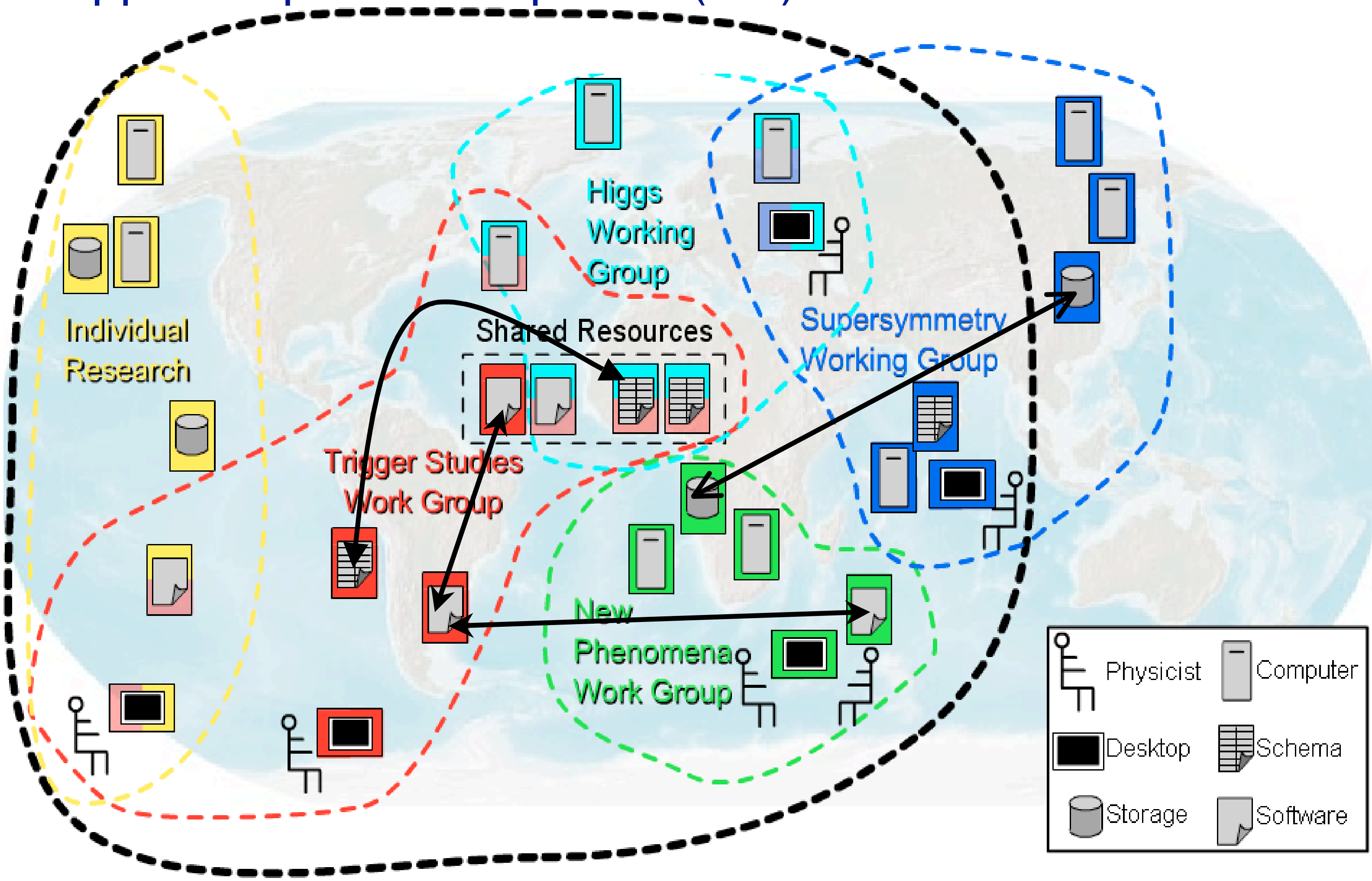
US CMS plans a distributed systems infrastructure and will

- Interface the science applications to the grid.
- Develop Experiment Application Services to live in the grid envrionment and provide common and generalised interfaces and capabilities.
- Adopt and sponsor common middleware services
- Present our Storage and Computing Resources to the common grid fabric and support opportunistic use of those resources by other organizations.



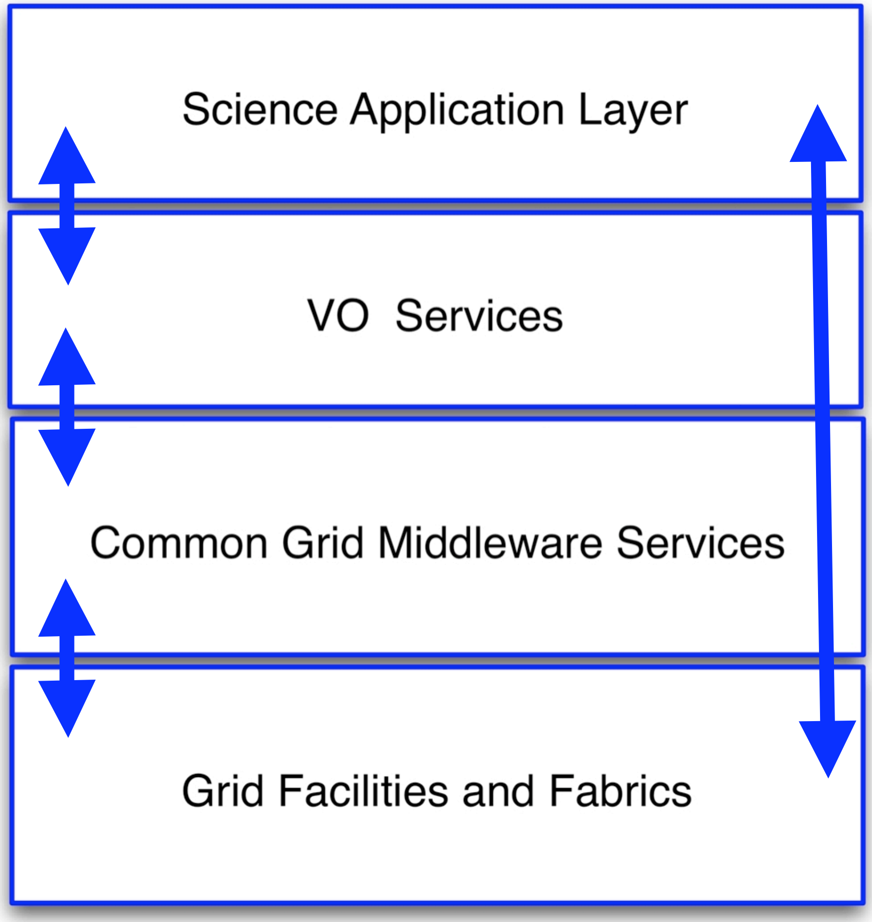
Make grid infrastructure transparent

Support Experiment Specific (VO) Environment on Remote Facilities.



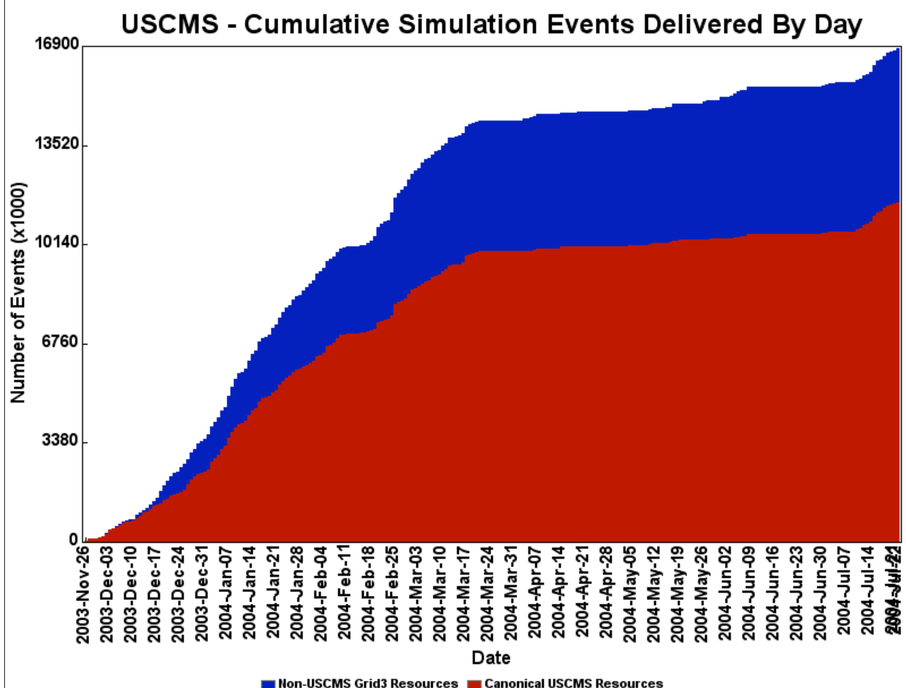
Distributed System Architectural Principles

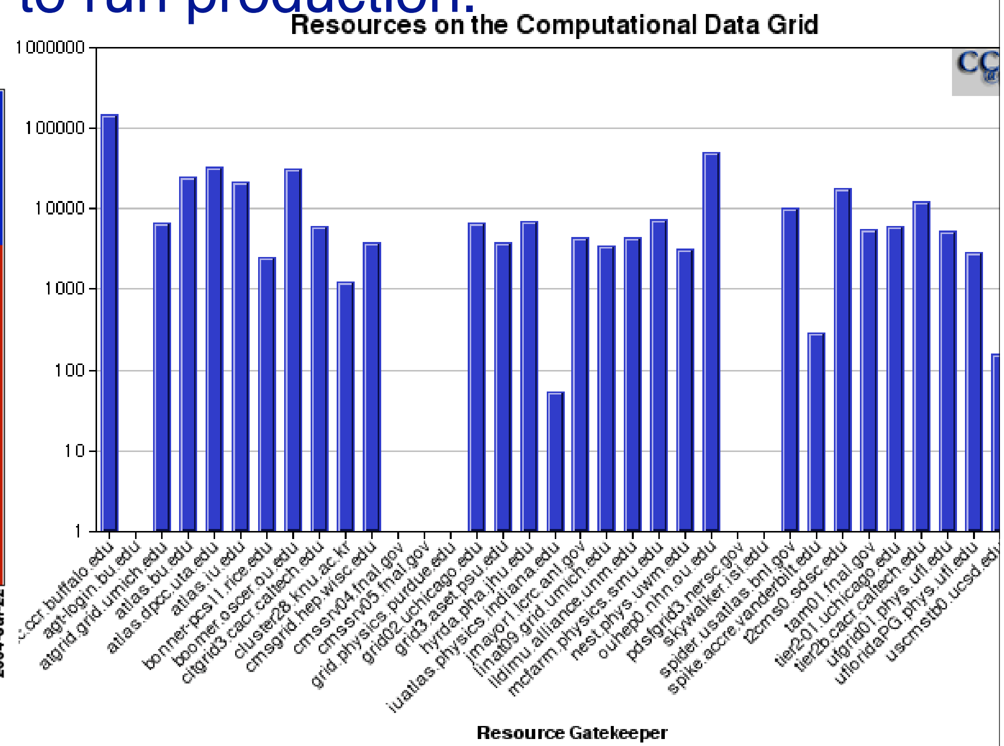
- Deploy end to end systems in support of ongoing physics user goals and iteratively build the final system to the needed performance and scale.
- Develop well specified Services with defined scope and responsibilities where
- Functionality is pushed as down the software stack as far as possible,
- Dependencies between the layers are explicit and
- Interfaces are generalized.



Simple Prototype Built and Used in the US: Grid3

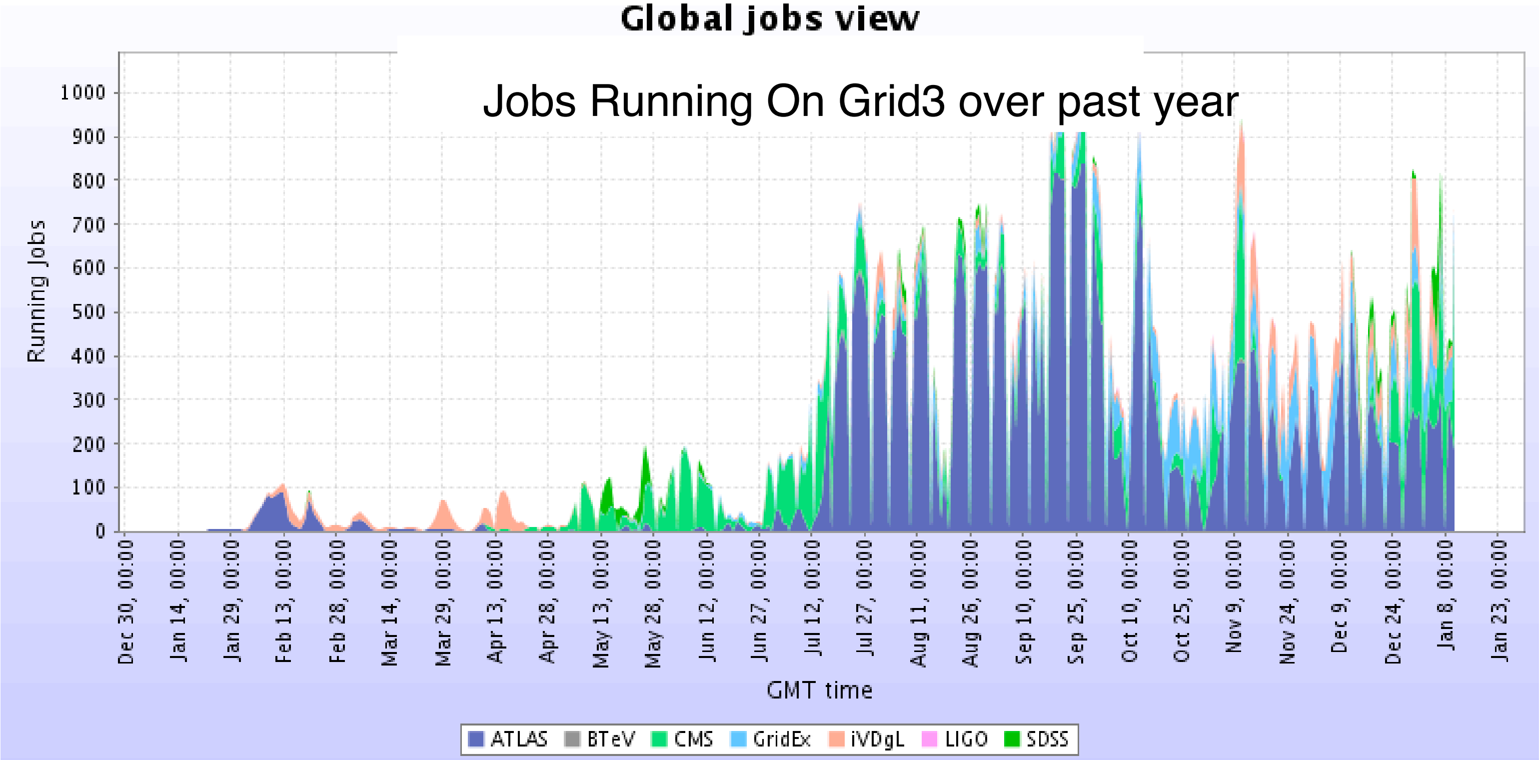
- During 2004 33% of US CMS simulation production was done through opportunistic use of non-CMS resources.
- The jobs ran on 17 sites with an overall efficiency of ~65%.
- The failures were mainly due to hardware problems and filled disk partitions. The overall infrastructure was stable.
- The support load was <1 FTE to run production.





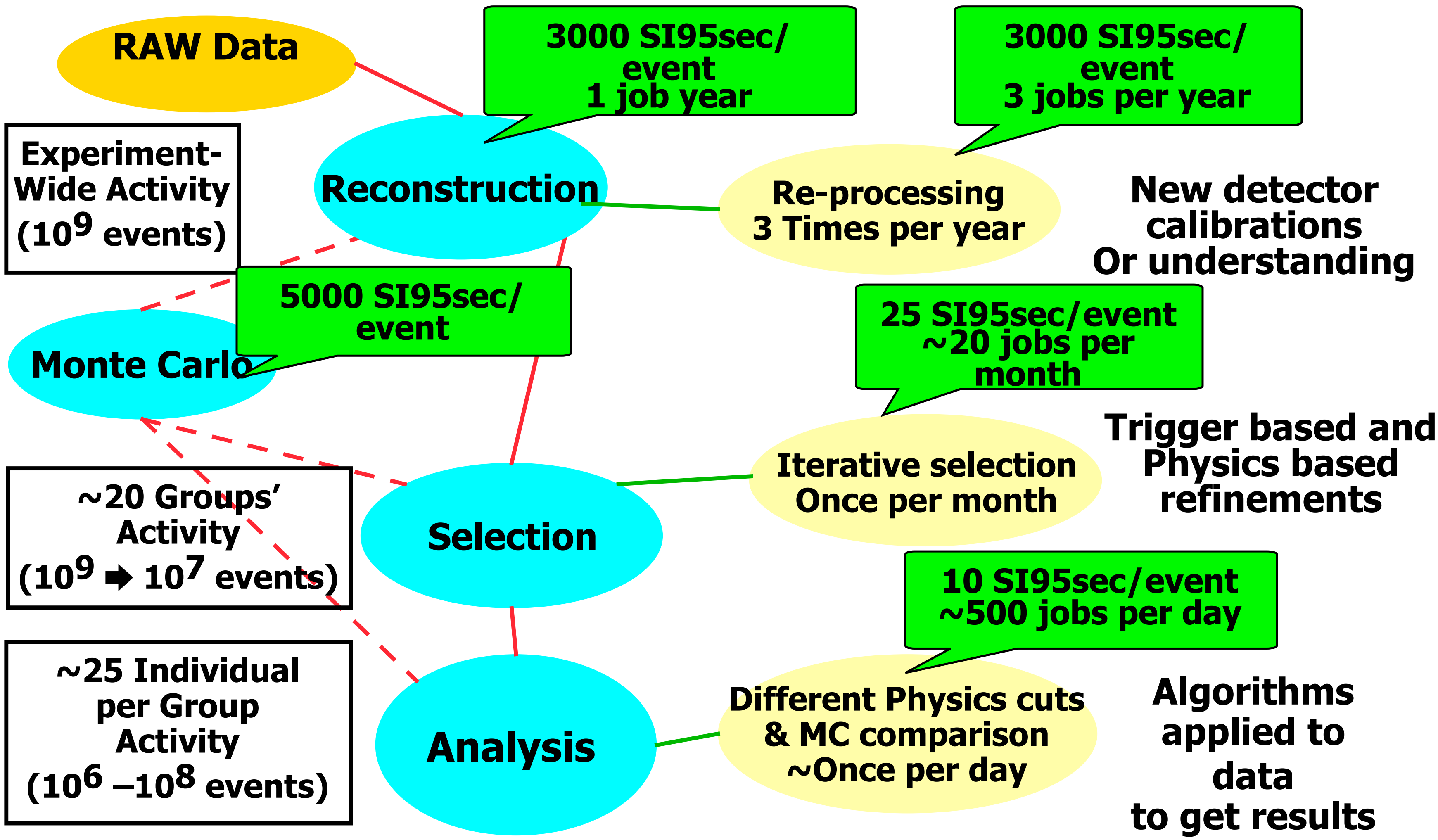
CMS support for Shared Use of Grid3

- In 2nd part of 2004 US CMS allocated shared resources for US ATLAS DC02 and supplied up to ~10% of total throughput.
- Supported opportunistic use for SDSS science analyses.



Analysis Model (ca. 2001)

Hierarchy of Processes (Experiment, Analysis Groups, Individuals)



Distributed Analysis Scenarios

- Persistency of the analysis workspace that a user can later re-connect to, re-submit the analysis with modified parameters or code, check the status, merge results between analyses, share datasets with other users and analysis workspaces, while the system keeps provenance information about jobs and datasets.
- To carry out the analysis tasks users are accessing shared computing resources. To do so, they must be registered with their Virtual Organization (VO), authenticated and their actions must be authorized according to their roles within the VO
- The user specifies the necessary execution environment (software packages, databases, system requirements, etc) and the system insures it on the execution node. In particular, the necessary environment can be installed according to the needs of a particular job
- The execution of the user job may trigger transfers of various datasets between a user interface computer, execution nodes and storage elements. These transfers are transparent for the user

Data Storage and Transfer

Posix I/O to files from the Experiment framework, online system etc

- at present direct access to storage interface; use of common POOL persistency layer.

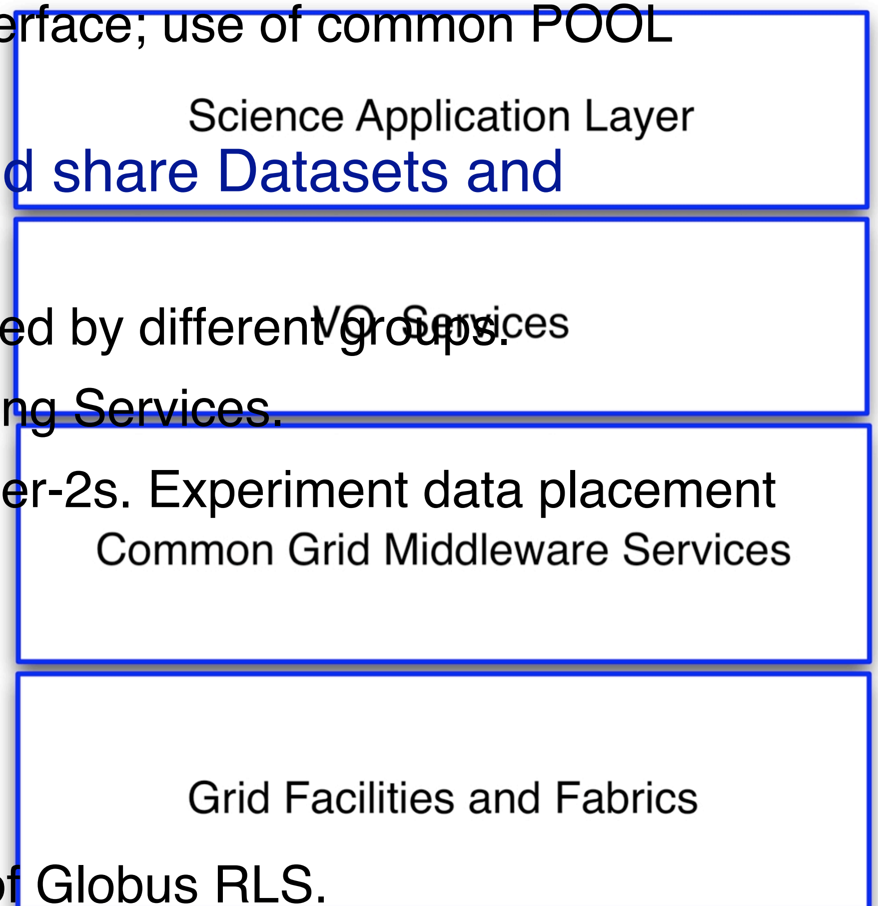
Command line interface to archive and share Datasets and Calibrations

- various prototypes based on mysql used by different groups.
- Experiment Bookkeeping and Publishing Services.
- Managed Data Placement at Tier-1, Tier-2s. Experiment data placement scripts (Phedex).

User of Common Interfaces for

- Storage - SRM V1.1
- Data transfer - GridFTP
- Replica management -prototype use of Globus RLS.

Tier-1 petabye tape store and LAN-distributed, resilient disk pools presented to Grid through SRM



CMS Data Transfer Challenge

Concentrate on full system level issues rather that optimizing single components. A production system was used at FNAL for the challenge

- Service challenge co-existed with normal uses of the system.
- S.C. Could not impede or interfere greatly with normal work.
- Able to study feature interaction and scaling in the most realistic environment.

Whole files are moved to and from the SE over the GRID w/ grid interfaces.

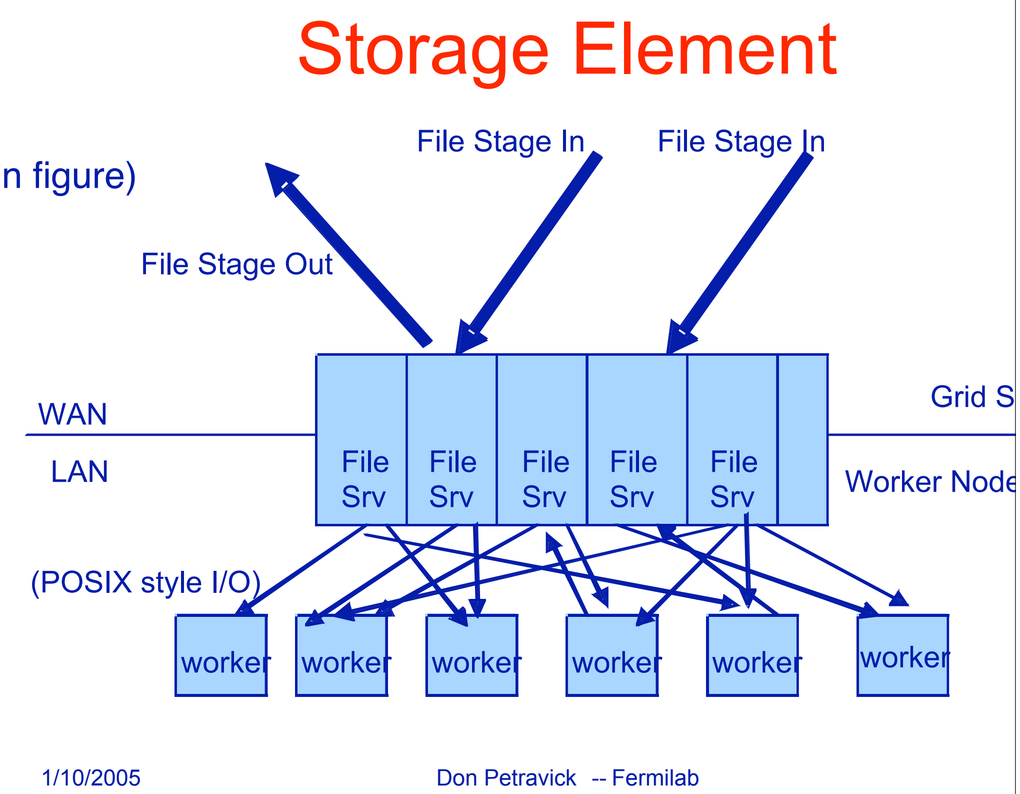
- Large bandwidth*delay
- Grid interfaces (SRM, gFTP)

Local access by WN's is Posix

- Files are accessed bit by bit.

dCache SE's can have more structure (not detailed in figure)

- Can support tape.
- Can move data through firewalls and NAT devices.
- Can add nodes without recycling the system
- Can carry on if nodes fail.



Basic dCache System *dCache.ORG*

Basic Specification

- Single 'rooted' file system name space tree
- Data may be distributed among a huge amount of disk servers.
- Supports multiple internal and external copies of a single file

Scalability

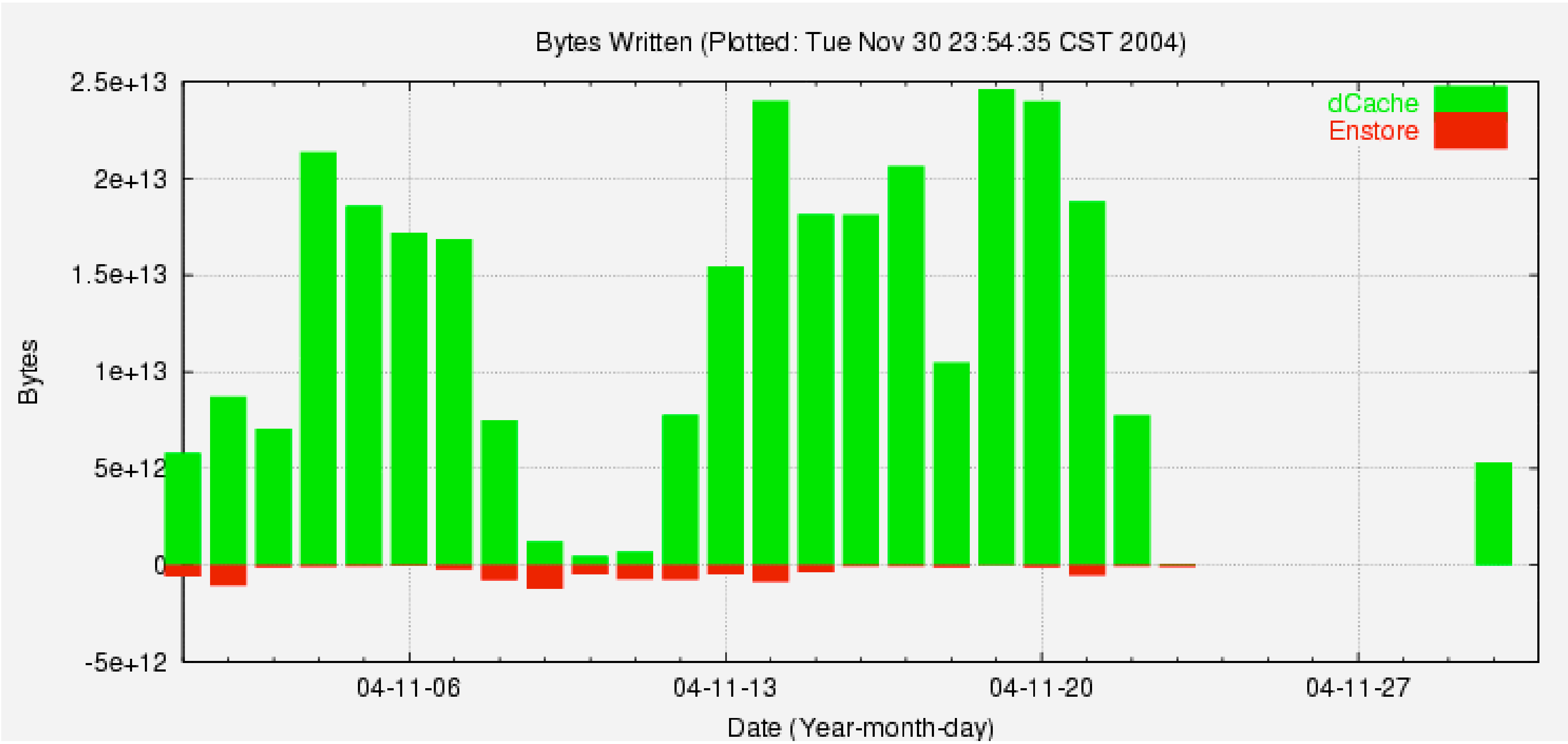
- Distributed Movers AND Access Points (Doors)
- Automatic load balancing using cost metric and inter pool transfers.
- Pool 2 Pool transfers on pool hot spot detection

Patrick Fuhrmann *dCache, Storage Element and HSM optimizer* *Darmstadt, GSI Oct 2004*

Results

November graph GridFTP protocol Throughput:

Many consecutive days of > 10 TB/day 25 TB/day peak (~3 Gbit sustained)



Results

The design used:

- 2 MB TCP buffers and 20 parallel streams for each transfer. 2 MB/sec/stream giving rates of 40 MB/s for each file. Tuned with FNAL production netflow analyzer.

Expectations for this hardware:

- The current set of hardware combined with the dCache IO yields a maximum rate between 50-60 MB/s for each node.
- Unit tests using optimized C code can achieve 70-80 MB/s for each node.
- Therefore, the 40 MB/s per file transfer was deemed acceptable at this point of development

Much was tried and learned,

- This challenge provided more rate than some R&E networks currently carry.
- Did not achieve the best performance levels seen in unit tests
- Did not investigate whole parameter space E.g. Did not use large MTU's

- Discovered Pull gives advantages over Push

Many bugs were identified and fixed over the 1st weeks of the challenge. This was the real goal of the challenge

- Properly clean up when xfers were killed
- Developed a simple system view to understand transfers.
- Applying priorities properly (service challenge uses v.s. production use)
- Regulate # of concurrent FTP's independently of # of local accesses (data movement resources are different)
- Preserve state across crashes, power-downs, failure of pool nodes, Use preserved state to recover where possible.
- Properly scaling monitoring (SPY)
- Configuration issues (gridmap files, corrupted certificates).

As a separate test, a special gridftp only script was written. Only 1 file from each disk at CERN was used, (3 per node), possibly leading to some memory caching at CERN.

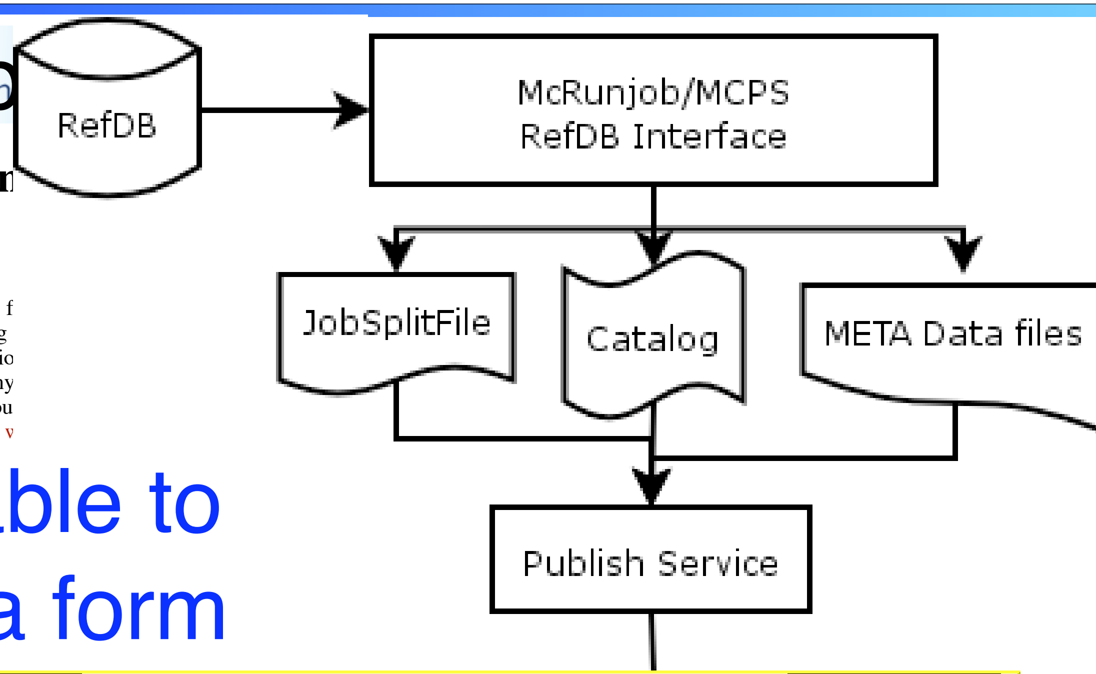
Results:

- Files written to memory at FNAL, provided a rate of 500 MB/s.
- Files were then written to dCache pool disks (by the GLOBUS COG gridftp Client , not through the dCache), and the rate was 400 MB/s.

USCMS
The Compact Muon Solen

Pub

Dataset Building and Validation



Introduction

There are many steps involved in building, publishing and validating datasets f
Some of the steps currently required are due to limitations,or bugs, in existing
become obsolete in the future. However, until the changes make into productio
long since stopped being used, it will be necessary to continue executing many
document describes the steps necessary, given the current circumstances, to bu
Note: These instructions are for building in the production area only. Experts v

Make Data Available to the Physicists in a form they can use.

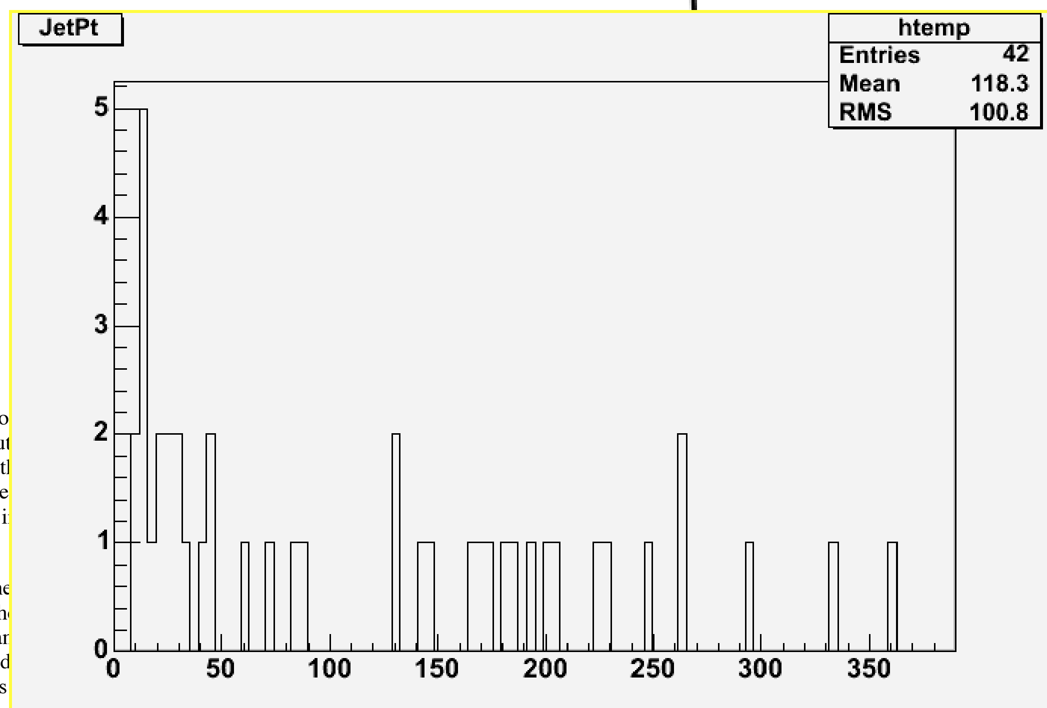
own working directory.

```
cd /storage/data3/pubWork/prod/DatasetBuild
source setup.csh
```

Step 1: Building the Metadata files

Presumably one will receive datasets from different sources and I need to
here. Scripts to automate as much of this as possible will be provided, but
attention and may depend on more manual activities than is desirable. If t
then one may already have a POOL XML file catalog to start with and re
However, if the catalog is not available then it will need to be generated i

1. **Identify Dataset Name and All Owner Names:**
The name of the dataset and all owners that will be grouped togethe
last production stage will serve as the primary owner and will be th
validation stages. For example, for a DST datast we might also war
DST owner becomes the primary and we also need the SimHits and
moment we assume an expert has provided us with a list of owners



Job Submission and Execution

Users submit Production Simulation Requests to central team at CERN who allocate assignments round the world.

- Request database based on mysql
- Application and job logging at local submission site
- Record successful jobs back at central database at CERN

Job Placement and Planning

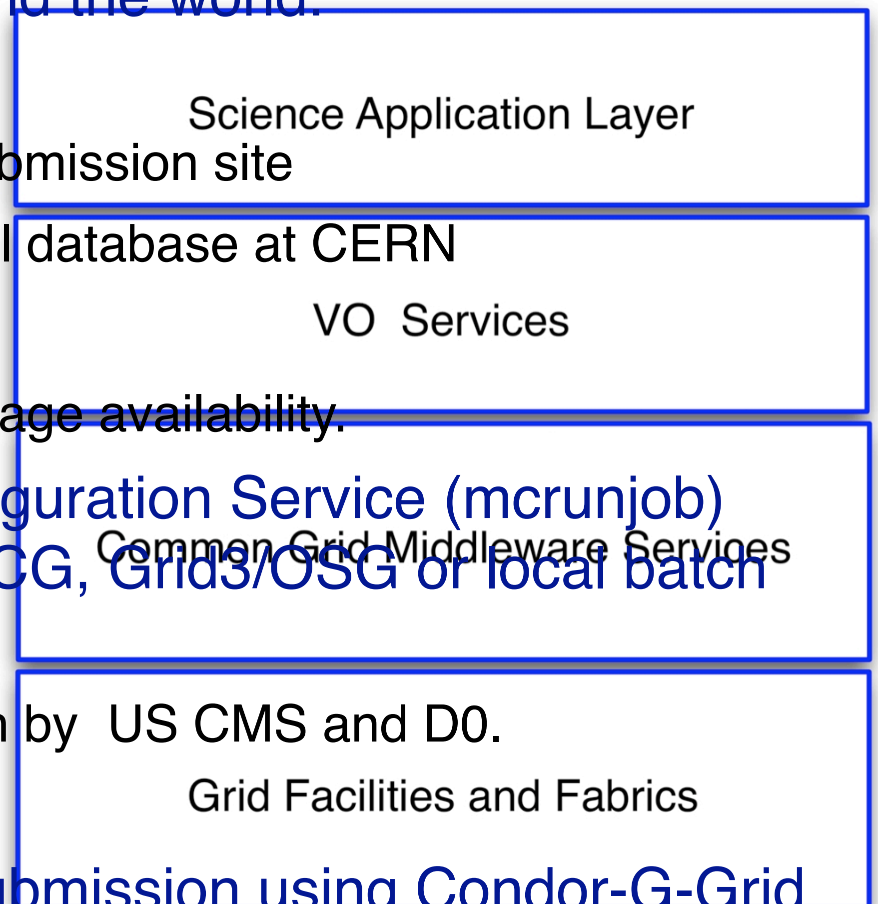
- Poll sites for job queue depth and storage availability.

Job and execution environment Configuration Service (mcrunjob) generates scripts to submit jobs on LCG, Grid3/OSG or local batch queue.

- RUNJOB python framework used both by US CMS and D0.
- Makes environment opaque to users

Workflow using Condor-DAGMAN, submission using Condor-G-Grid monitor, Gram-job managers.

- Use local batch system; have run across 20 Grid3 sites;



Installation, Configuration, Packaging and Distribution

Usual code, dependency & release management for >10million lines code. Still dependent on AFS at central CERN.

On the fly CMS software “collection and distribution” (DAR) + update capability.

VDT + Pacman (ivdgl/usatlas) well provisioned build, test, parameter driven distribution and configuration services.

ROCKS installs and configures “rich OS” including Grid services.

- Reinstall rather than “patch” or “reconfigure” works ~hundreds CPUs.

Science Application Layer

VO Services

Common Grid Middleware Services

Grid Facilities and Fabrics

Security - User Identity

Well developed User administrative interface to register and define Experiment roles in one location. (VOMRS)

- Experiment controls and monitors roles and access rights.

VOMS repository of DNs and roles + automated scripts to populate Sites with identity and role information and policies.

- Developed by EU Grid projects. Tested by Grid3. Included in VDT for OSG.
- Automated scripts to include all LCG and Grid3 VOs for compute and storage access.

Site Identity Mapping - common use of GUMS developed by US ATLAS/PPDG.

- Use of standard Global GSI infrastructure and PKI certificates.

Science Application Layer

VO. Services

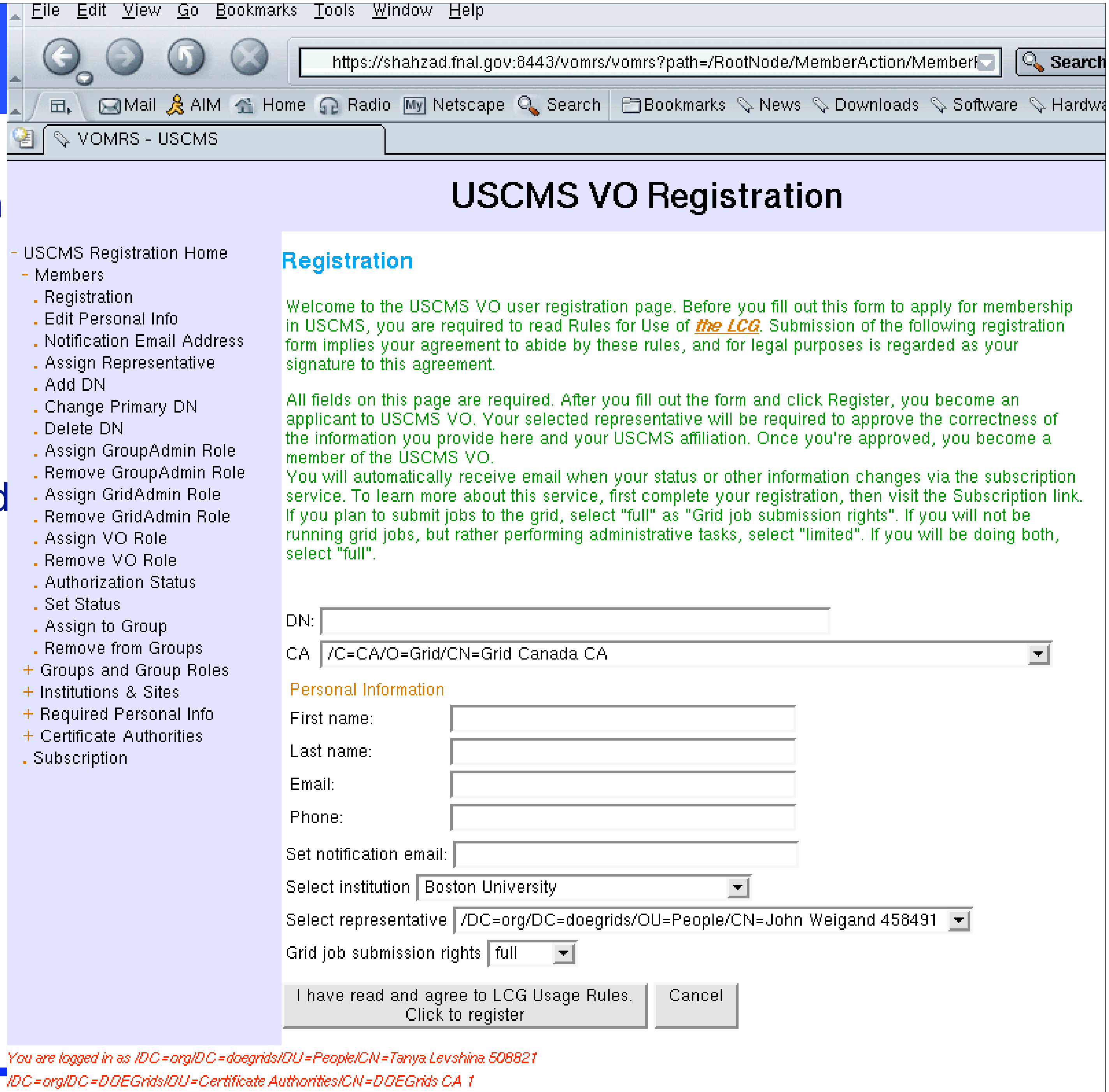
Common Grid Middleware Services

Grid Facilities and Fabrics

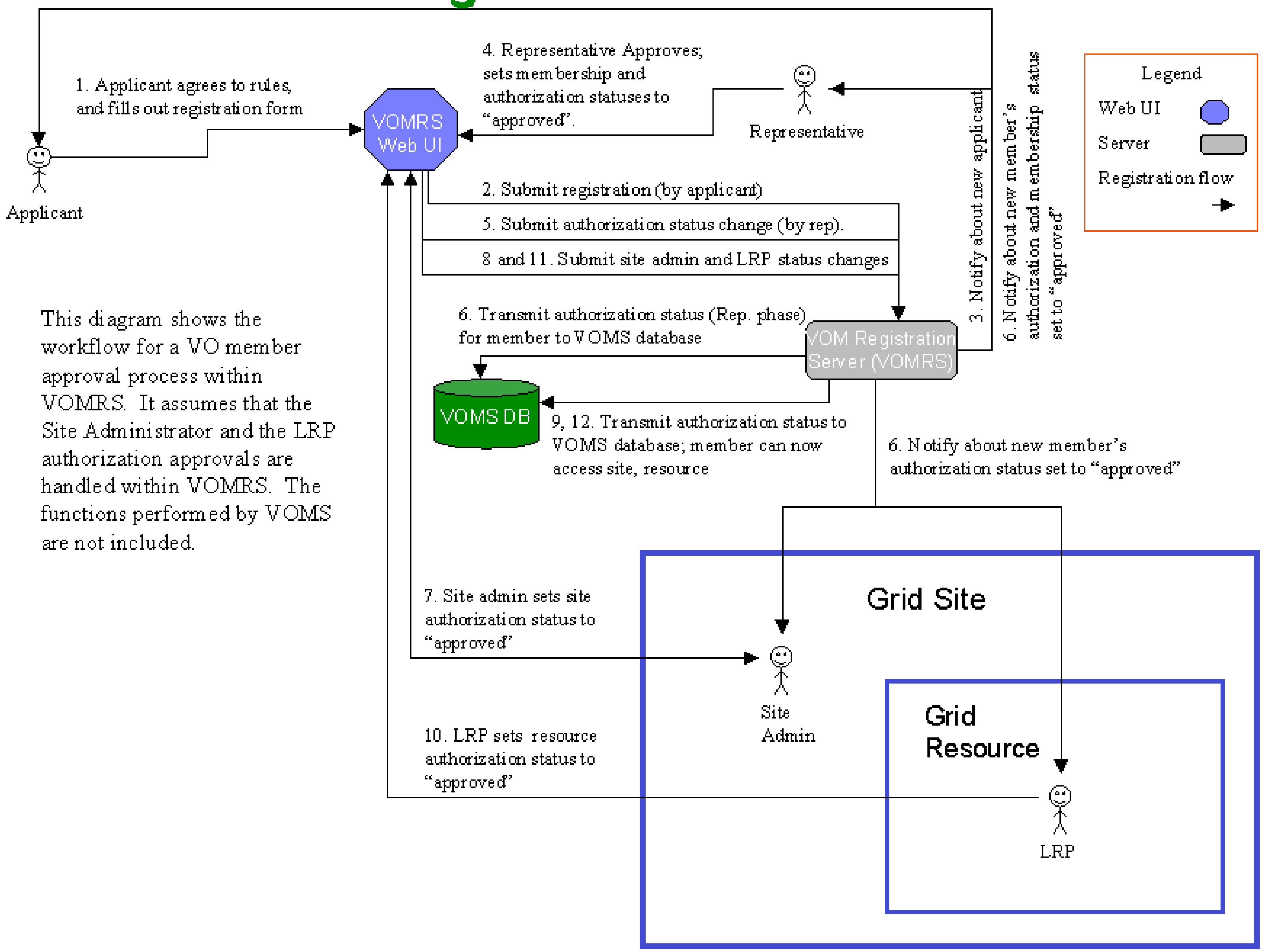
VOMS developed in Europe.

First production use in US.

Now being deployed by LCG/EGEE



VO Registration Workflow



This diagram shows the workflow for a VO member approval process within VOMRS. It assumes that the Site Administrator and the LRP authorization approvals are handled within VOMRS. The functions performed by VOMS are not included.

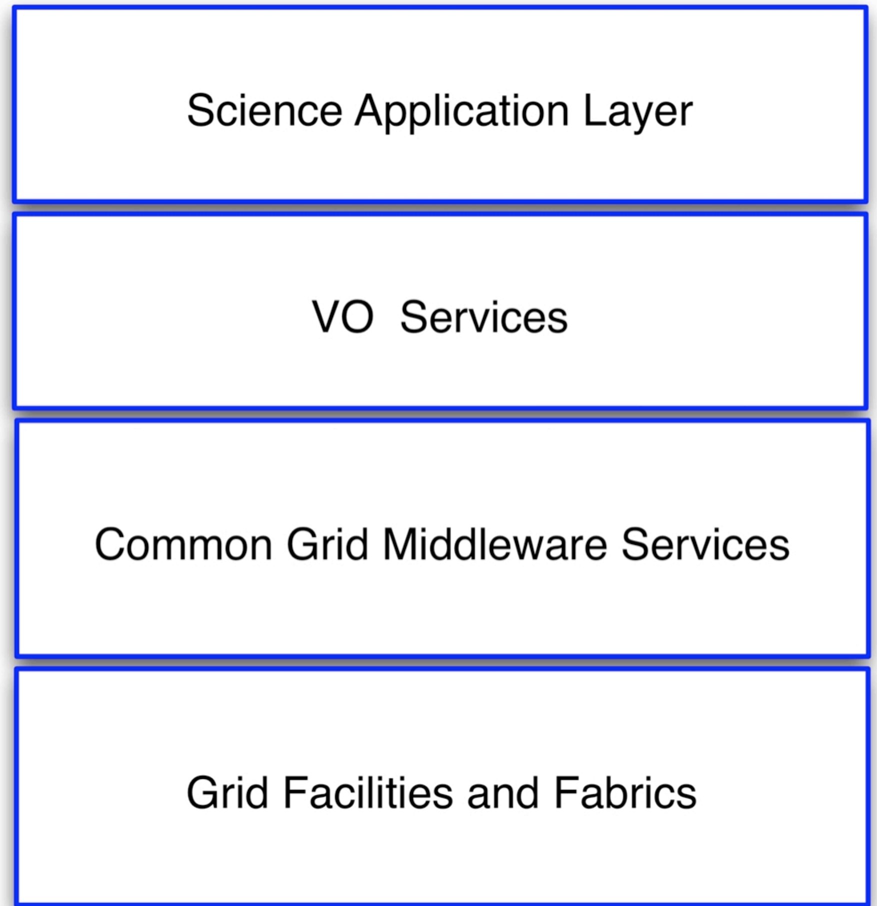
Access Control: Authorization & Policy

Users indicate role or groups they are

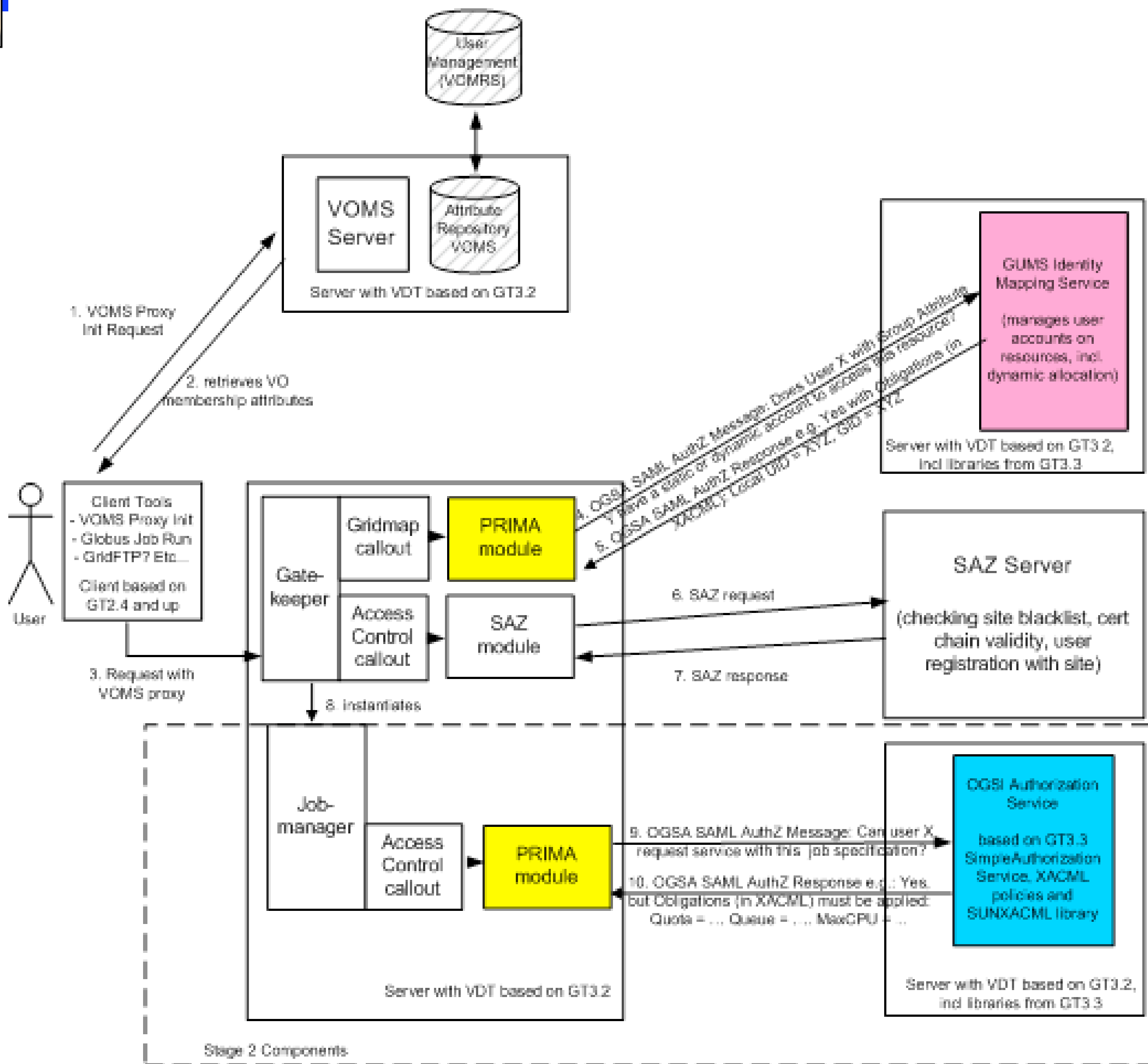
Experiment administrator specifies p
groups share of the resources.

Authorization, Policy and Accounting
information.

Sites apply their own policies and pri



Version 3 - 2004-08-05
mlorch@fnal.gov



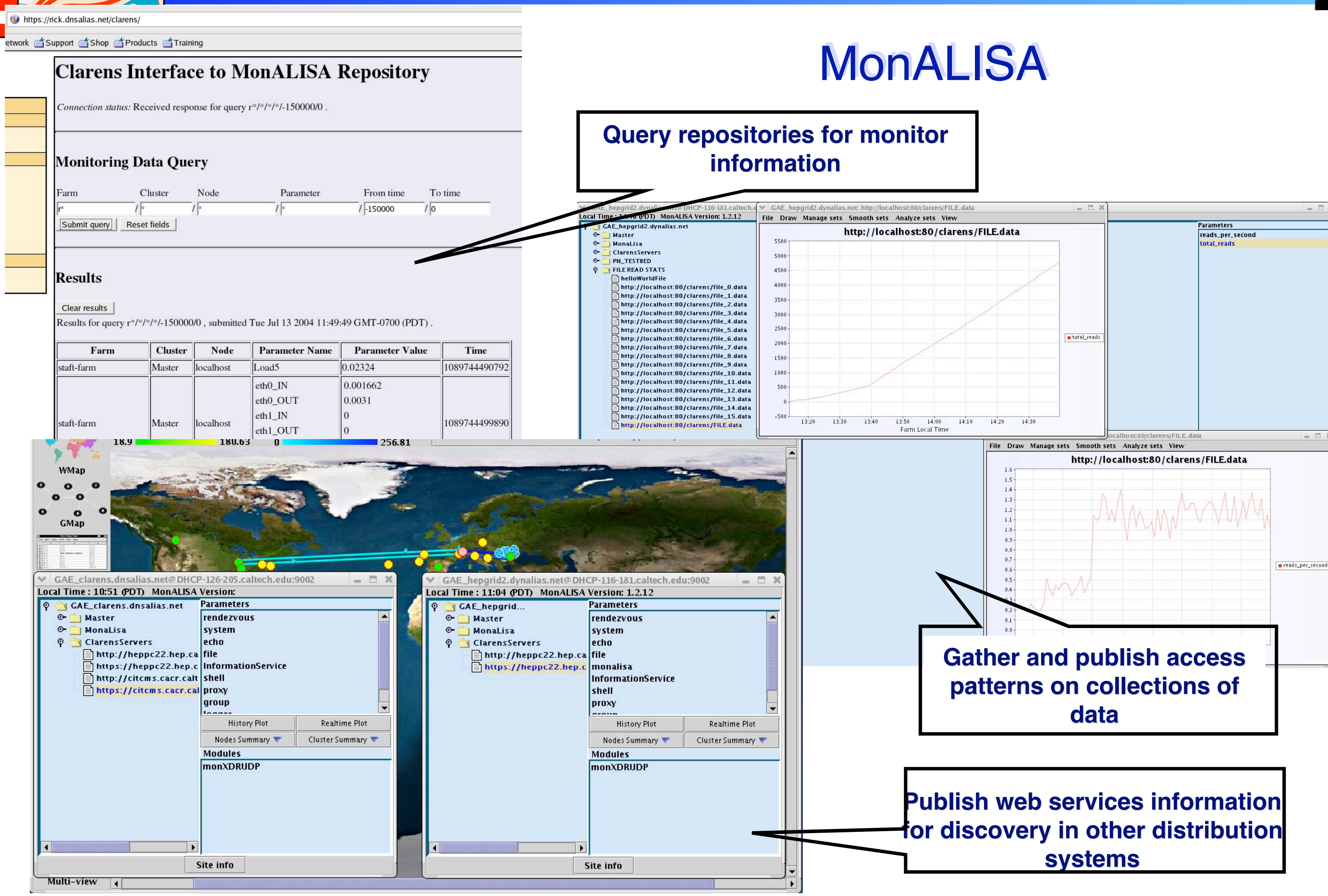
Protect and Respond

- OSG/iVDGL Security Incident Response and Handling Plan/
- Benefits from participating in the OSG Consortium - not having to go it alone.
- Will do a validation test this month - will involve more than 10 people.

Version 1.0 20 Nov. 2004

Monitoring

MonALISA



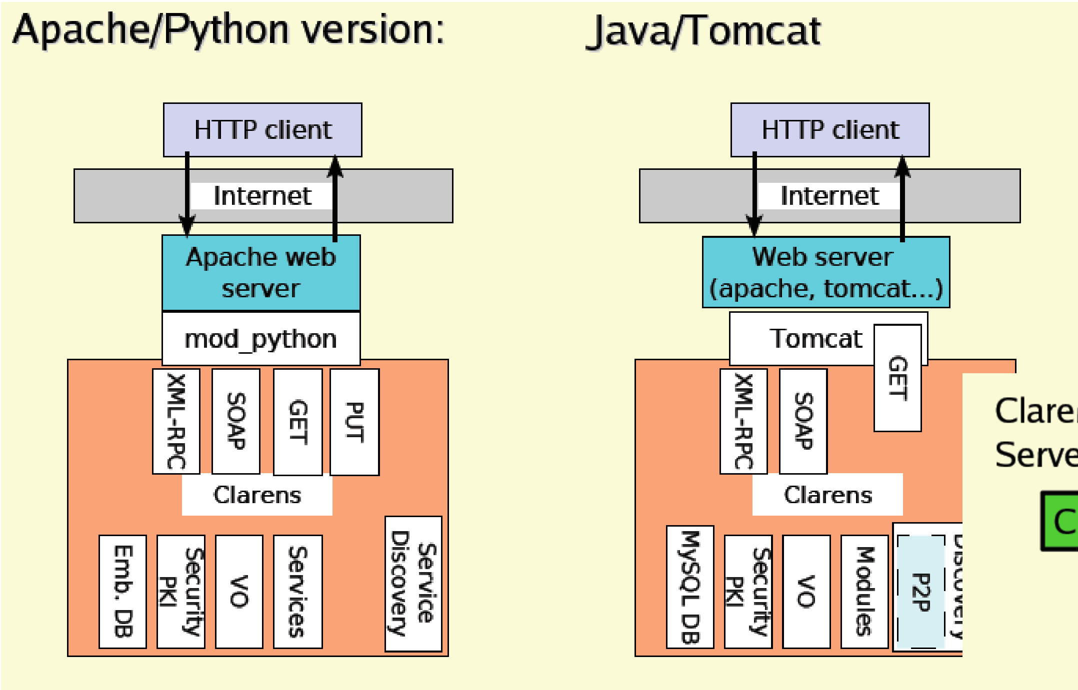
Query repositories for monitor information

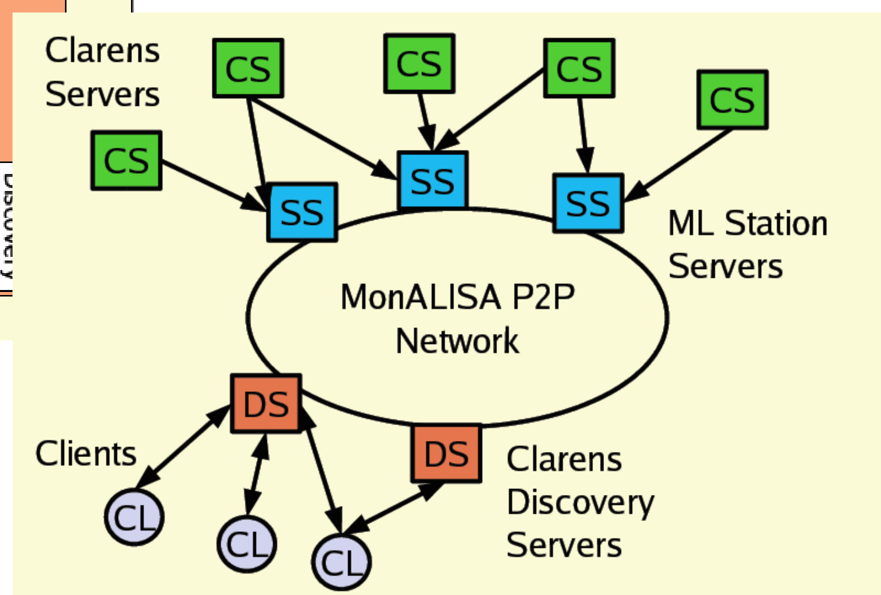
Gather and publish access patterns on collections of data

Publish web services information for discovery in other distribution systems

Service & Discovery Infrastructure

US CMS developing Clarens “web service bus” which will conform to WSRF standards





Virtual Data Toolkit

The Virtual Data Toolkit (VDT) continues to provide the packaging and support for the core middleware for US CMS and US common grid infrastructure.

New components get added through a request, validation and release process overseen by the stakeholders.

VDT Team collaborating with LCG and EGEE. Contributing to new gLITE middleware packaging and deployment.

Nightly builds and test infrastructure for robustness.

VDT

Home
What's New?
Download
FAQ
Documentation
Security
Testing
Support
Working Group

VDT Test Result

Run Info

Package	VDT
VDT Version	1.3.1
OS	RHEL 3
Install User	root
Time	01:41:13 02/03/2005
Test Type	nightly
Install Root	/scratch/vdt-install-test/new-installs/12710

Run Results

Test File	Test Name	Test Result
condor_g.t	-	expand
condor.t	-	expand
edg.t	-	expand
globus.t	-	expand
httpd.t	-	expand
install.t	-	expand
myproxy.t	-	expand

VDT

Home
What's New?
Download
FAQ
Documentation
Security
Testing
Support
Working Group

What is in VDT 1.3.0?

VDT is layered on NMI

VDT

- Apache Tomcat, v4.1.31
- ClassAds, v0.9.7
- Condor/Condor-G, v6.7.3
  VDT Condor configuration script
- DOE and LCG CA Certificates, vv3 (includes LCG 0.25 CAs)
- DRM, v1.2.2
- EDG CRL Update, v1.2.5
- EDG Make Gridmap, v2.1.0
- Fault Tolerant Shell (ftsh), v2.0.5
- GLUE Information Providers, vCVS version 1.79, 4-April-2004
- GLUE Schema, v1.1, extended version 1
- GSI-Enabled OpenSSH, v3.4
- Globus Toolkit, pre web-services, v3.2.1 + patches
- GriPhyN Virtual Data System (containing Chimera and Pegasus), v1.3.5a
- Java SDK, v1.4.2_06
- KX509, v2031111
- Monalisa, v1.2.20
- MyProxy, v1.11
- MySQL, v4.0.22
- Netlogger, v2.2
- PyGlobus, v1.0.5
- RLS, v2.1.5
- UberFTP, v1.3
- VOMS, v1.2.19
- VOMS Admin Server, v0.7.5

Other optional packages (actual Pacman package names specified in **bold**)

- **Globus-Core** - necessary for building against Globus
- **Globus-LSF-Setup** - configures GRAM jobmanager/reporter for LSF
- **Globus-PBS-Setup** - configures GRAM jobmanager/reporter for PBS
- **Globus-LoadLeveler-Setup** - configures GRAM jobmanager/reporter for IBM LoadLeveler
- **Globus-FBSNG-Setup** - Globus jobmanager for Farms Batch System Next Generation
- **Globus-RLS-Server-Setup-MySQL** - installs MySQL, IODBC and sets up MySQL databases for RLS
- **VDT-Test** - VDT Certification Tests

US CMS Grid Services pending

Data and Namespace Management

Auditing, Logging and Bookkeeping

Accounting

Rich Execution Environments controlled by and accessible to users and administrators.

User Interfaces with higher level functionality and robustness.

:

:

Open Science Grid (US CMS) roadmap includes Interoperation with other grid infrastructures

with the LCG

- Jobs submitted across Grid3 and LCG by both LHC experiments.
- Consistent interfaces being maintained for Storage - SRM, Job Submission - GRAM/Condor-G, Information - GlueSchema;

with TeraGrid

- Work at Purdue to provide Grid3 job manager to submit jobs to TeraGrid.

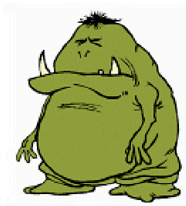
with Campus Grids

- University of Buffalo GRASE
- Emerging “FermiGrid”

Including New Students

Helping Develop America's Technological Workforce

QuarkNet CMS Test Beam e-Lab



OGRE is an Online Graphical ROOT Enviornment

Visit the Root Homepage. (Creates a new window.)

CMS HCal Testbeam 04 Data

Select Variable	Selection Criteria	Plot Style	Histogram Fill Color
☐ Total Energy (name = c3x3.3)		☐ logx ☐ logy	None ☐ Use Dark Colors
☐ Ecal Energy (name = c3x3.ee)		☐ logx ☐ logy	None ☐ Use Dark Colors
☐ Hcal Energy (name = c3x3.eh)		☐ logx ☐ logy	None ☐ Use Dark Colors
☐ Eta (name = peak_eta)		☐ logx ☐ logy	None ☐ Use Dark Colors
☐ Phi (name = peak_phi)		☐ logx ☐ logy	None ☐ Use Dark Colors

Deployment on the OSG in 2005

Applications:

- Robust data movement challenge across Tier-1 and 7 Tier-2s (cf between CERN and Tier-1).
- Data movement between US CMS sites and non-CMS sites on OSG.
- (Multi-)User requested simulation production.
- Data analysis with some interactive features across Tier-1 and 4 Tier-2 sites.

Infrastructure & Middleware:

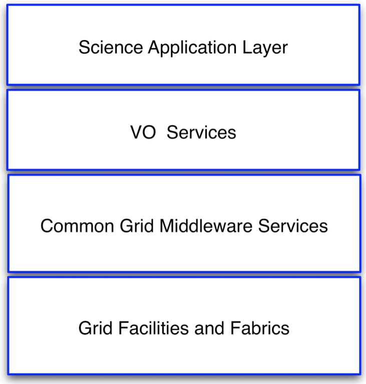
- Multi-VO Managed transient storage and disk caches
- Discovery of and "on demand" managed use of resources.
- Automated planning and scheduling of jobs.

US CMS Roadmap

50% of final system working and delivering to users in 2006.

Enabling of full-scale analysis for up to 1000 users using integrated system of 7-10 Universities, Fermilab Physics Analysis Center and Cern in 2007.

Continued commitment to develop needed capabilities with well defined and described interfaces to contribute to the common infrastructure and participate fully in the Open Science Grid Consortium.



Open Science Grid