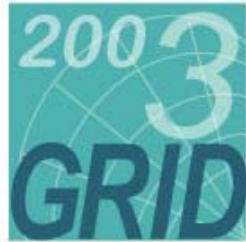
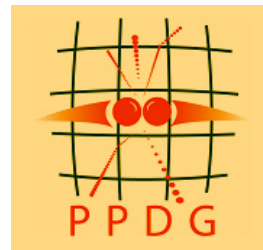


Grid Computing in High Energy Physics

Enabling Data Intensive Global Science



Paul Avery
University of Florida
avery@phys.ufl.edu



Beauty 2003 Conference
Carnegie Mellon University
October 14, 2003

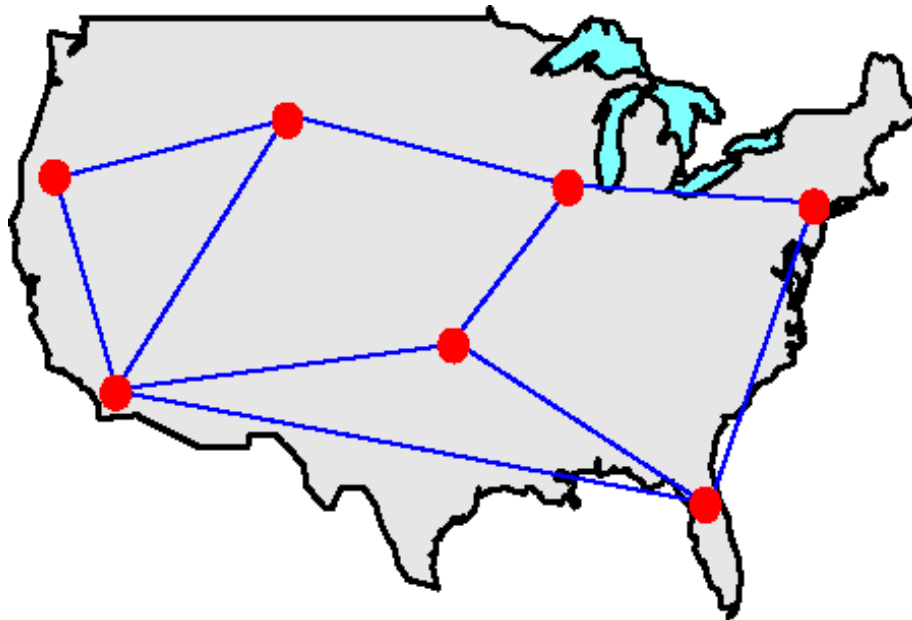


Overview

- Grids, Data Grids and examples
- LHC computing as principal Grid driver
- Grid & network projects
- Promising new directions
- Success story for HEP
 - ◆ Leadership, partnership, collaboration

The Grid Concept

- **Grid:** Geographically distributed computing resources configured for coordinated use
 - ◆ **Fabric:** Physical resources & networks provide raw capability
 - ◆ **Middleware:** Software ties it all together: tools, services, etc.
 - ◆ **Ownership:** Resources controlled by owners and shared w/ others
- **Goal:** Transparent resource sharing





Grids and Resource Sharing

- Resources for complex problems are distributed
 - ◆ Advanced scientific instruments (accelerators, telescopes, ...)
 - ◆ Storage, computing, people, institutions
- Organizations require access to common services
 - ◆ Research collaborations (physics, astronomy, engineering, ...)
 - ◆ Government agencies, health care organizations, corporations, ...
- Grids make possible "Virtual Organizations"
 - ◆ Create a "VO" from geographically separated components
 - ◆ Make all community resources available to any VO member
 - ◆ Leverage strengths at different institutions
- Grids require a foundation of strong networking
 - ◆ Communication tools, visualization
 - ◆ High-speed data transmission, instrument operation



Grid Challenges

- Operate a fundamentally complex entity
 - ◆ Geographically distributed resources
 - ◆ Each resource under different administrative control
 - ◆ Many failure modes
- Manage workflow of 1000s of jobs across Grid
 - ◆ Balance policy vs. instantaneous capability to complete tasks
 - ◆ Balance effective resource use vs. fast turnaround for priority jobs
 - ◆ Match resource usage to policy over the long term
- Maintain a global view of resources and system state
 - ◆ Coherent end-to-end system monitoring
 - ◆ Adaptive learning for execution optimization
- Build managed system & integrated user environment



Data Grids & Data Intensive Sciences

- Scientific discovery increasingly driven by data collection
 - ◆ Computationally intensive analyses
 - ◆ Massive data collections
 - ◆ Data **distributed** across networks of varying capability
 - ◆ Internationally distributed collaborations
- Dominant factor: data growth (1 Petabyte = 1000 TB)
 - ◆ 2000 ~0.5 Petabyte
 - ◆ 2005 ~10 Petabytes
 - ◆ 2010 ~100 Petabytes
 - ◆ 2015 ~1000 Petabytes?

How to collect, manage, access and interpret this quantity of data?

Drives demand for “Data Grids” to handle additional dimension of data access & movement



Data Intensive Physical Sciences

- High energy & nuclear physics
 - ◆ Belle/BaBar, Tevatron, RHIC, JLAB, LHC
- Astronomy
 - ◆ Digital sky surveys: SDSS, VISTA, other Gigapixel arrays
 - ◆ VLBI arrays: multiple- Gbps data streams
 - ◆ “Virtual” Observatories (multi-wavelength astronomy)
- Gravity wave searches
 - ◆ LIGO, GEO, VIRGO, TAMA
- Time-dependent 3-D systems (simulation & data)
 - ◆ Earth Observation
 - ◆ Climate modeling, oceanography, coastal dynamics
 - ◆ Geophysics, earthquake modeling
 - ◆ Fluids, aerodynamic design
 - ◆ Pollutant dispersal



Data Intensive Biology and Medicine

- **Medical data and imaging**
 - ◆ X-Ray, mammography data, etc. (many petabytes)
 - ◆ Radiation Oncology (real-time display of 3-D images)
- **X-ray crystallography**
 - ◆ Bright X-Ray sources, e.g. Argonne Advanced Photon Source
- **Molecular genomics and related disciplines**
 - ◆ Human Genome, other genome databases
 - ◆ Proteomics (protein structure, activities, ...)
 - ◆ Protein interactions, drug delivery
- **High-res brain scans (1-10 μ m, time dependent)**

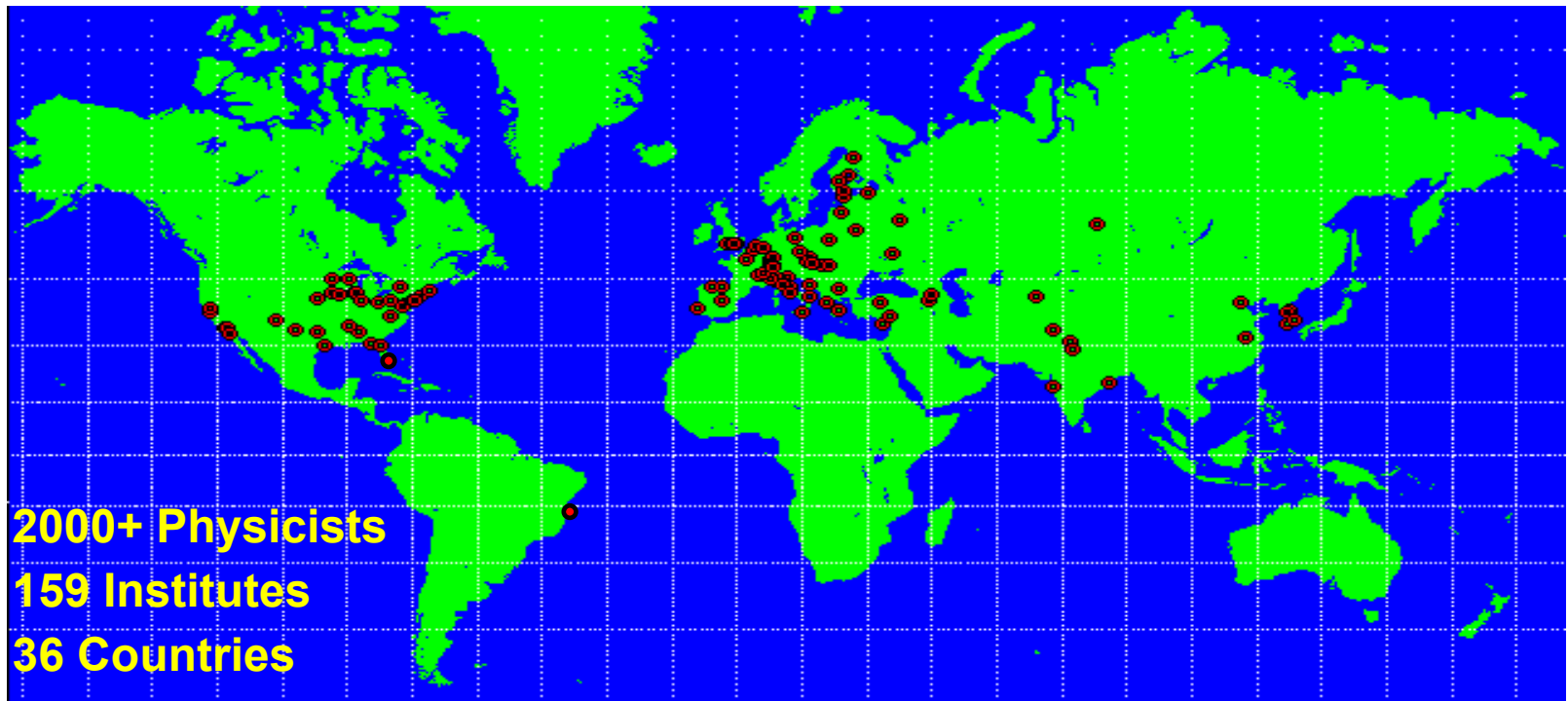


LHC and Data Grids



LHC: Key Driver for Data Grids

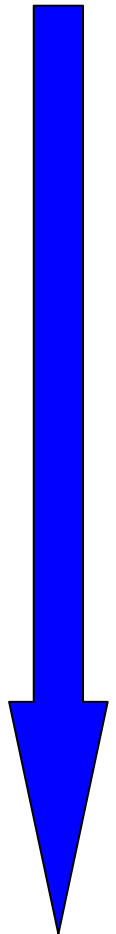
- Complexity: Millions of individual detector channels
- Scale: PetaOps (CPU), Petabytes (Data)
- Distribution: Global distribution of people & resources



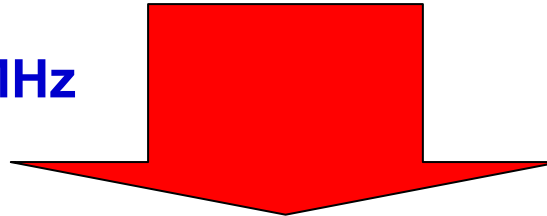


LHC Data Rates: Detector to Storage

Physics filtering



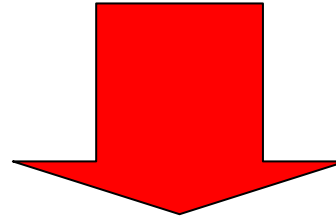
40 MHz



Level 1 Trigger: Special Hardware

75 KHz

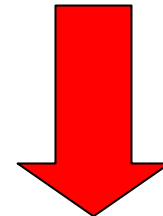
75 GB/sec



Level 2 Trigger: Commodity CPUs

5 KHz

5 GB/sec



Level 3 Trigger: Commodity CPUs

100 Hz

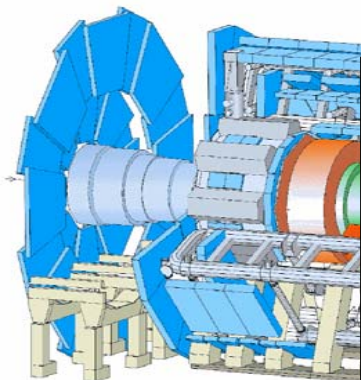
0.1 – 1.5 GB/sec



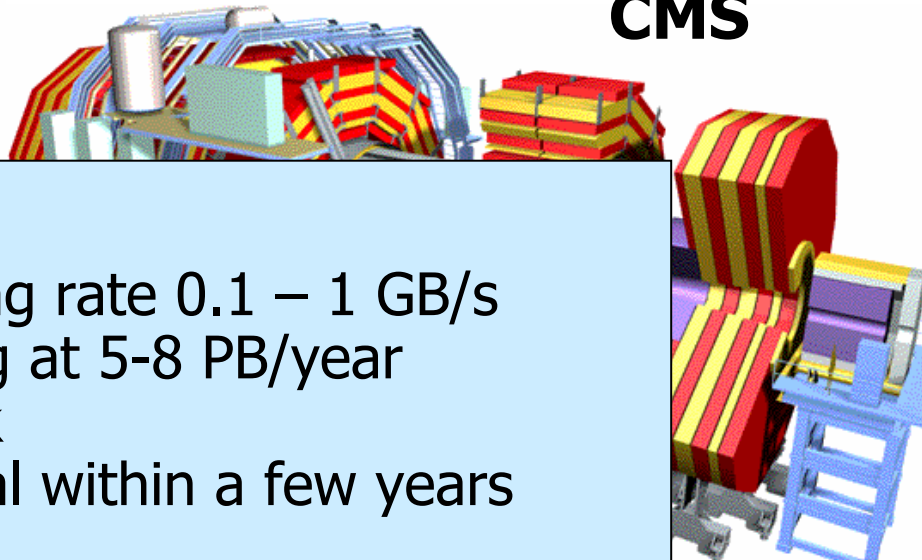
**Raw Data to storage
(+ simulated data)**

LHC Data Requirements

ATLAS



CMS

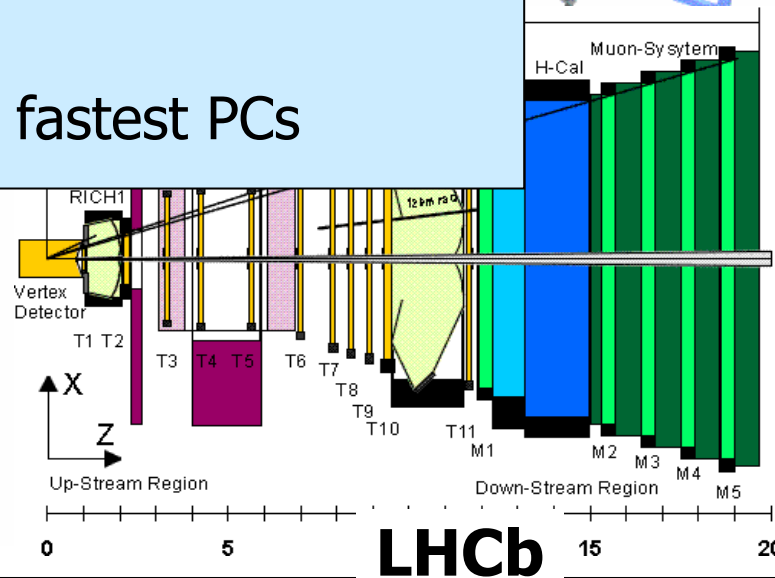
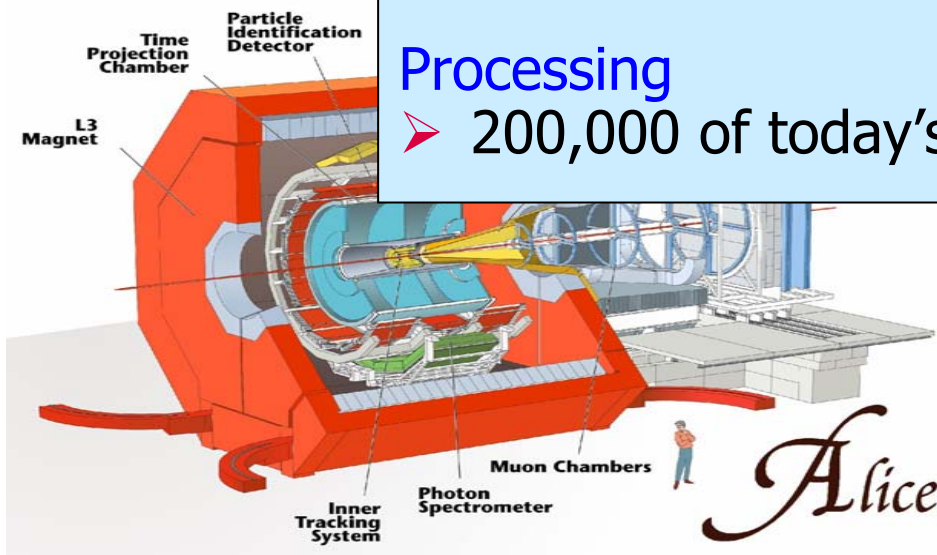


Storage

- Raw recording rate 0.1 – 1 GB/s
- Accumulating at 5-8 PB/year
- 10 PB of disk
- ~100 PB total within a few years

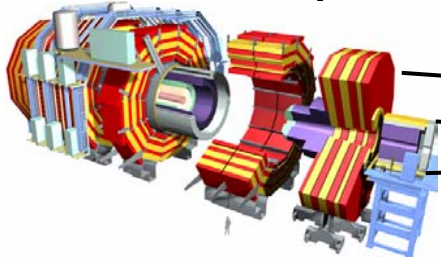
Processing

- 200,000 of today's fastest PCs



Hierarchy of LHC Data Grid Resources

CMS Experiment



$$\text{Tier0}/(\Sigma \text{Tier1})/(\Sigma \text{Tier2}) \sim 1:2:2$$

Online System

100-1500 MBytes/s

Tier 0

CERN Computer Center > 20 TIPS

10-40 Gbps

Tier 1

Korea

UK

Russia

USA

...

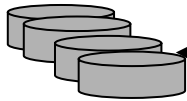
2.5-10 Gbps

Tier 2

Tier2 Center 2 Center 2 Center 2 Center

1-2.5 Gbps

Tier 3



Physics cache

Institute Institute Institute Institute

1-10 Gbps

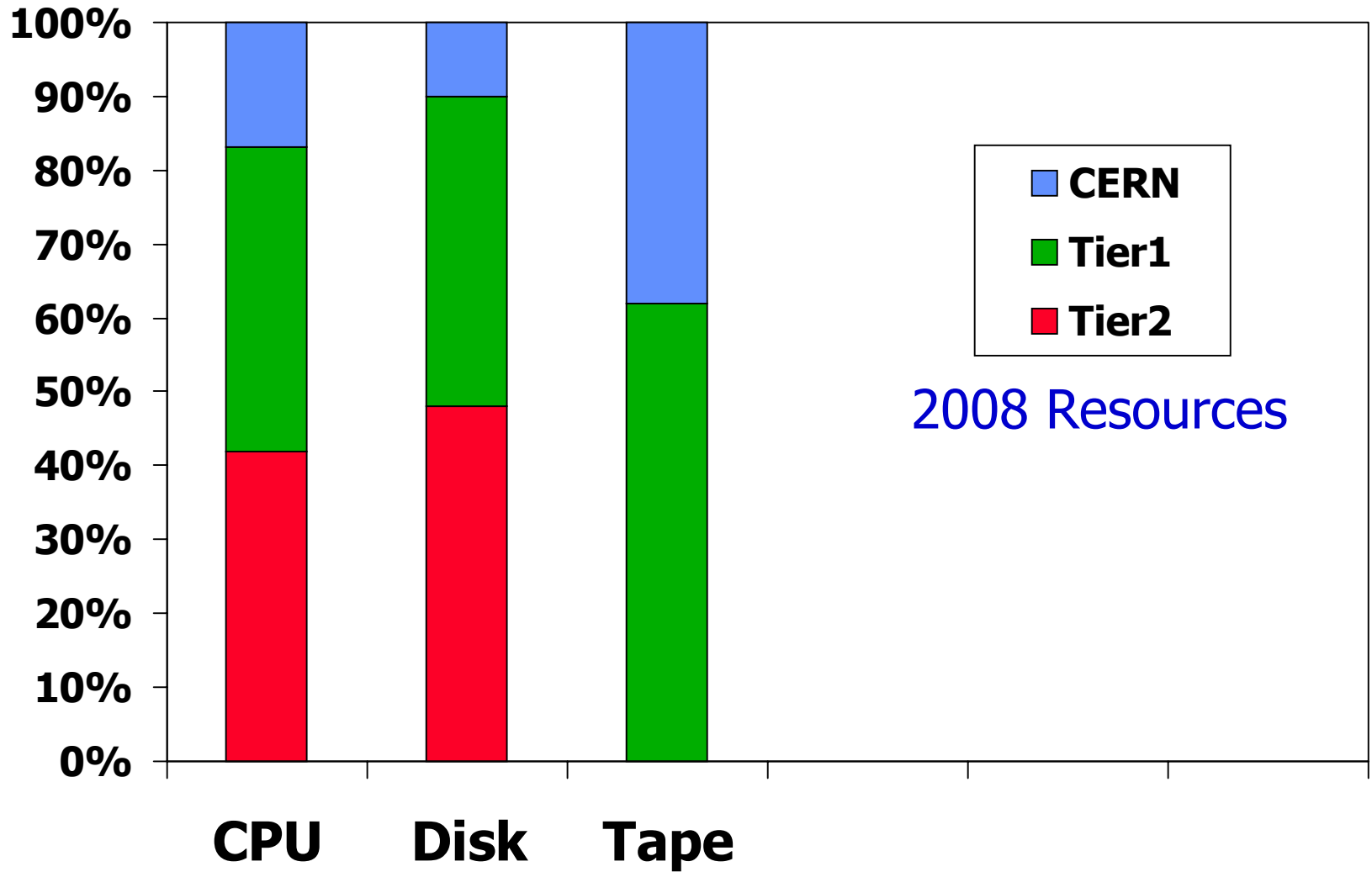


Tier 4

~10s of Petabytes/yr by 2007-8
~1000 Petabytes in < 10 yrs?



Most IT Resources Outside CERN



Driver for Transatlantic Networks

	2001	2002	2003	2004	2005	2006
CMS	100	200	300	600	800	2500
ATLAS	50	100	300	600	800	2500
BaBar	300	600	1100	1600	2300	3000
CDF	100	300	400	2000	3000	6000
D0	400	1600	2400	3200	6400	8000
BTeV	20	40	100	200	300	500
DESY	100	180	210	240	270	300
CERN BW	155- 310	622	2500	5000	10000	20000

- BW in Mbps (2001 estimates)
- Now seen as conservative!



HEP & Network Land Speed Records

- 9/01 102 Mbps CIT-CERN
- 5/02 450-600 Mbps SLAC-Manchester
- 9/02 1900 Mbps Chicago-CERN
- 11/02 [LSR] 930 Mbps California-CERN
- 11/02 [LSR] 9.4 Gbps in 10 Flows California-Chicago
- 2/03 [LSR] 2.38 Gbps in 1 Stream California-Geneva
- 10/03 [LSR] 5 Gbps in 1 Stream

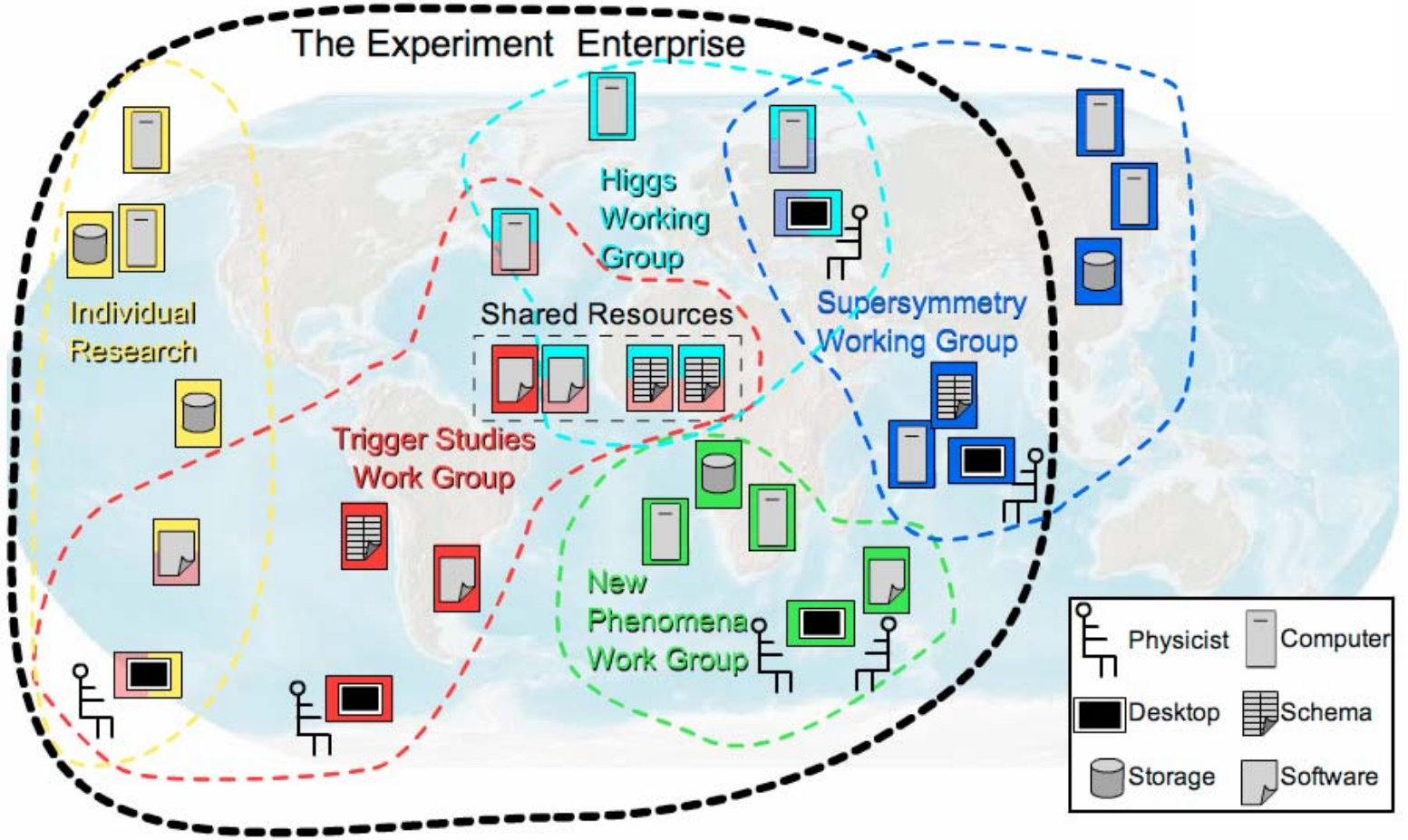
HEP/LHC driving network developments

- New network protocols
- Land speed records
- ICFA networking chair
- HENP working group in Internet2



Analysis by Globally Distributed Teams

Non-hierarchical: Chaotic analyses + productions





Grid Analysis Environment

- **GAE is crucial for LHC experiments**
 - ◆ Large, diverse, distributed community of users
 - ◆ Support for 100s-1000s of analysis tasks, over dozens of sites
 - ◆ Dependent on high-speed networks
- **GAE is where the physics gets done ⇒ analysis teams**
 - ◆ Team structure: Local, national, global
 - ◆ Teams share computing, storage & network resources
- **But the global system has finite resources**
 - ◆ Widely varying task requirements and priorities
 - ◆ Need for robust authentication and security
 - ◆ Need to define and implement collaboration policies & strategies



Data Grid Projects



Global Context: Data Grid Projects

➤ U.S. Infrastructure Projects

- ◆ GriPhyN (NSF)
- ◆ iVDGL (NSF)
- ◆ Particle Physics Data Grid (DOE)
- ◆ PACIs and TeraGrid (NSF)
- ◆ DOE Science Grid (DOE)
- ◆ NSF Middleware Infrastructure (NSF)

➤ EU, Asia major projects

- ◆ European Data Grid (EU)
- ◆ EDG-related national Projects
- ◆ LHC Computing Grid (CERN)
- ◆ EGEE (EU)
- ◆ CrossGrid (EU)
- ◆ DataTAG (EU)
- ◆ GridLab (EU)
- ◆ Japanese Grid Projects
- ◆ Korea Grid project

- **Not exclusively HEP (LIGO, SDSS, ESA, Biology, ...)**
- **But most driven/led by HEP**
- **Many \$M brought into the field**

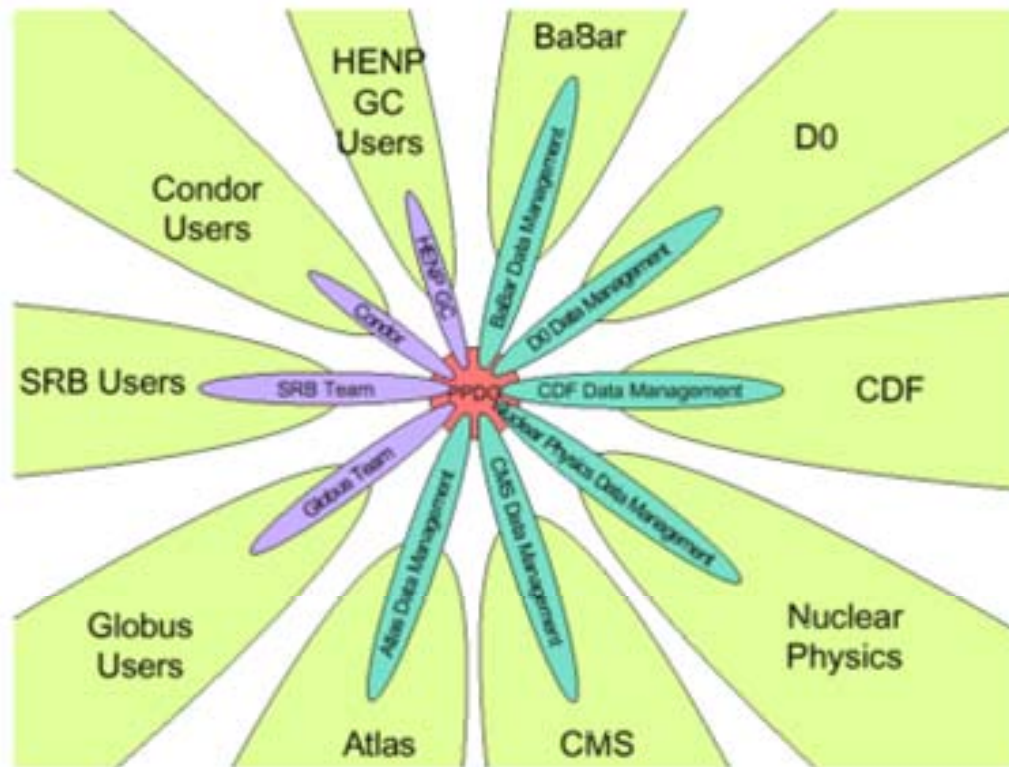


EU DataGrid Project

-
-
-
-
-
-
-
-

Work Package	Work Package title	Lead contractor
WP1	Grid Workload Management	INFN
WP2	Grid Data Management	CERN
WP3	Grid Monitoring Services	PPARC
WP4	Fabric Management	CERN
WP5	Mass Storage Management	PPARC
WP6	Integration Testbed	CNRS
WP7	Network Services	CNRS
WP8	High Energy Physics Applications	CERN
WP9	Earth Observation Science Applications	ESA
WP10	Biology Science Applications	INFN
WP11	Dissemination and Exploitation	INFN
WP12	Project Management	CERN

U.S. Particle Physics Data Grid



DOE funded

- ◆ Funded 1999 – 2004 @ US\$9.5M (DOE)
- ◆ Driven by HENP experiments: D0, BaBar, STAR, CMS, ATLAS
- ◆ Maintains practical orientation: Grid tools for experiments



U.S. GriPhyN and iVDGL Projects

- Both funded by NSF (ITR/CISE + Physics)
 - ◆ GriPhyN: \$11.9M (NSF) (2000 – 2005)
 - ◆ iVDGL: \$14.0M (NSF) (2001 – 2006)
- Basic composition (~120 people)
 - ◆ GriPhyN: 12 universities, SDSC, 3 labs
 - ◆ iVDGL: 20 universities, SDSC, 3 labs, foreign partners
 - ◆ Expts: CMS, ATLAS, LIGO, SDSS/NVO
- Grid research/infrastructure vs Grid deployment
 - ◆ GriPhyN: CS research, Virtual Data Toolkit (VDT) development
 - ◆ iVDGL: Grid laboratory deployment using VDT
 - ◆ 4 physics experiments provide frontier challenges
- Extensive student involvement
 - ◆ Undergrads, grads, postdocs participate at all levels
 - ◆ Strong outreach component



GriPhyN/iVDGL Science Drivers

- **LHC experiments**

- ◆ High energy physics
- ◆ 100s of Petabytes

2007

- **LIGO**

- ◆ Gravity wave experiment
- ◆ 100s of Terabytes

2002

- **Sloan Digital Sky Survey**

- ◆ Digital astronomy (1/4 sky)
- ◆ 10s of Terabytes

2001



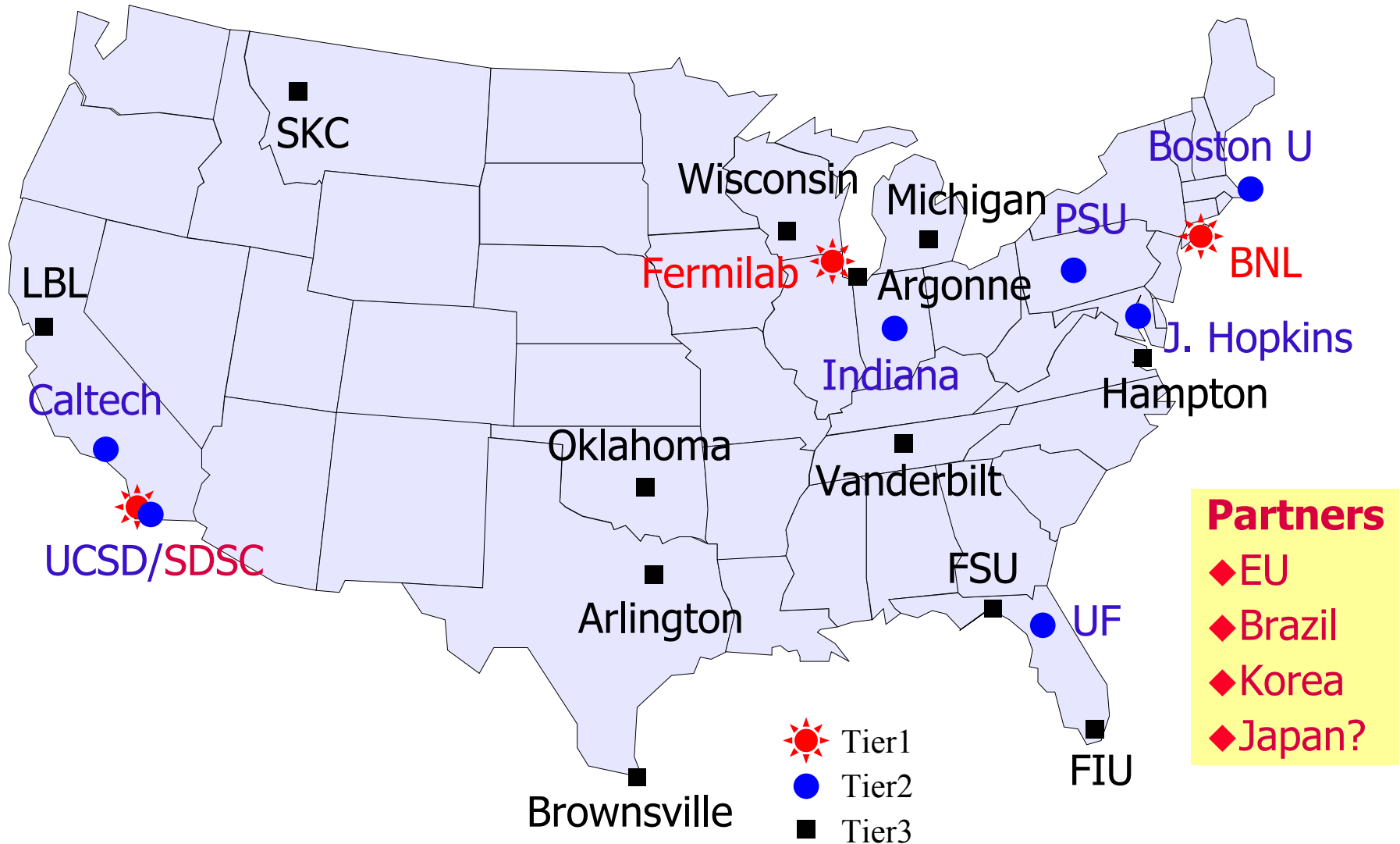
- **Massive CPU (PetaOps)**

- **Large distributed datasets (>100PB)**

- **International (global) communities (1000s)**



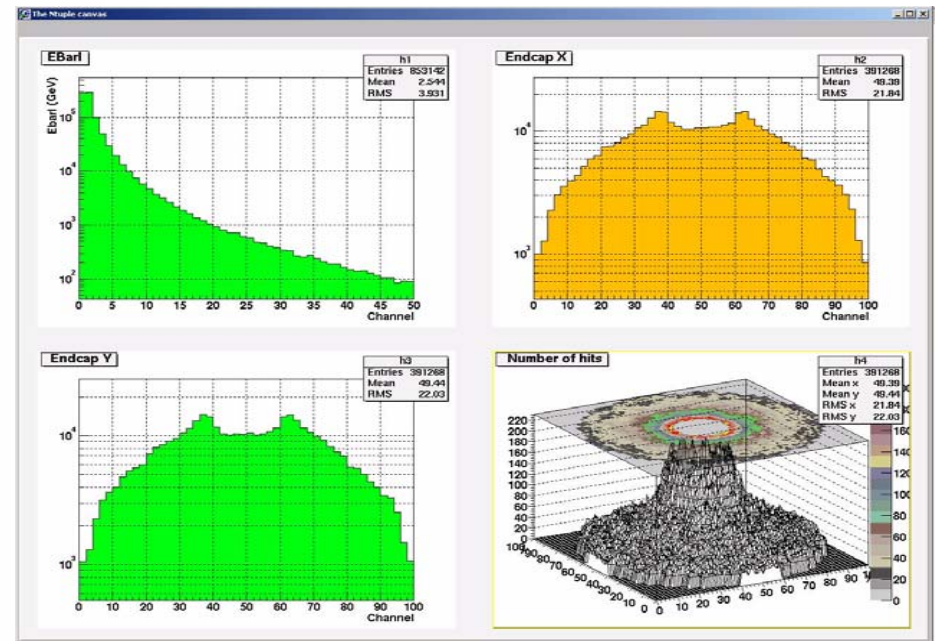
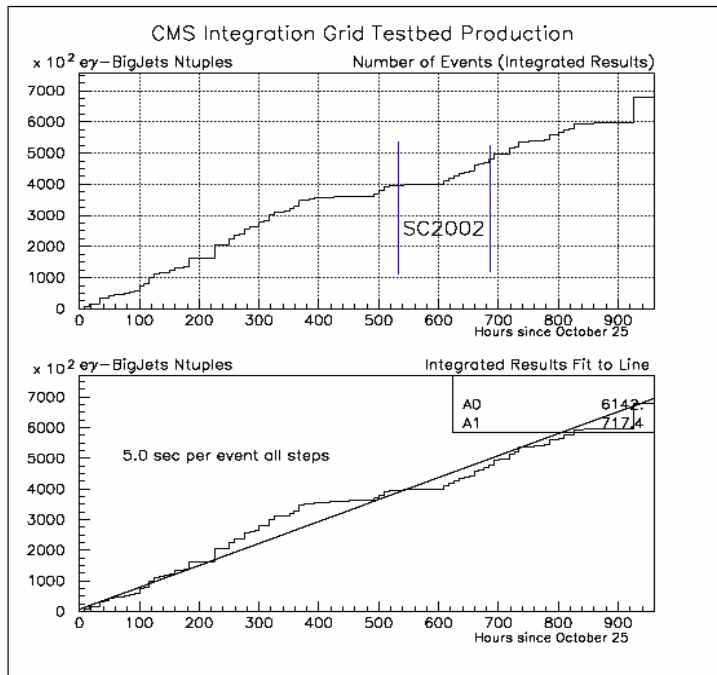
International Virtual Data Grid Laboratory (Fall 2003)





Testbed Successes: US-CMS Example

- Extended runs for Monte Carlo data production
 - ◆ 200K event test run identified many bugs in core Grid middleware
 - ◆ 2 months continuous running across 5 testbed sites (1.5M events)
 - ◆ Demonstrated at Supercomputing 2002





LCG: LHC Computing Grid Project

- Prepare & deploy computing environment for LHC expts
 - ◆ Common applications, tools, frameworks and environments
 - ◆ Emphasis on collaboration, coherence of LHC computing centers
 - ◆ Deployment only: no middleware development
- Move from testbed systems to real production services
 - ◆ Operated and supported 24x7 globally
 - ◆ Computing fabrics run as production physics services
 - ◆ A robust, stable, predictable, supportable infrastructure
- Need to resolve some issues
 - ◆ Federation vs integration
 - ◆ Grid tools and technologies



Sep. 29, 2003 announcement

PR13.0
29.09.2003



Renovation of the Computer Centre at CERN at this moment which "looks like a grid" ...

LHC Computing Grid Goes Online



The world's particle physics community today announced the launch of the first phase of the [LHC computing Grid \(LCG\)](#). The LCG is designed to handle the unprecedented quantities of data that will be produced by experiments at CERN's [Large Hadron Collider \(LHC\)](#) from 2007 onwards.

"The LCG will provide a vital test-bed for the new Grid computing technologies that are set to revolutionise the way scientists use the world's computing resources in areas ranging from fundamental research to medical diagnosis," said [Les Robertson](#), CERN's LCG project manager.

The computational requirements of the experiments that will operate at the LHC are enormous. Some 12-14 petabytes of data will be generated each year, the equivalent of more than 20 million CDs. Analysing this data will require the equivalent of 70,000 of today's fastest PC computers. The LCG will meet these needs by deploying a worldwide computational Grid, integrating the resources of scientific computing centres spread across Europe, America and Asia into a global virtual computing service.

The first phase of the project, LCG-1, will operate a series of prototype services, gradually increasing in scale and complexity as its builders develop an understanding of the functional and operational complexities involved in building a Grid of such unprecedented scale. LCG-1 uses so-called 'middleware' developed mainly by the [European Data Grid](#) project in Europe and the Globus, Condor and related projects contributing to the Virtual Data Toolkit in the US. It allows physicists to access worldwide distributed computing resources from their desktops as if they were local.



Current LCG Sites



Promising New Directions



U.S. Project Coordination: Trillium

- **Trillium = GriPhyN + iVDGL + PPDG**
 - ◆ Large overlap in leadership, people, experiments
 - ◆ Driven primarily by HEP, particularly LHC experiments
 - ◆ But includes other disciplines

- **Benefit of coordination**
 - ◆ Common software base + packaging: **VDT + PACMAN**
 - ◆ Collaborative / joint projects: **monitoring, demos, security, ...**
 - ◆ Wide deployment of new technologies, e.g. Virtual Data
 - ◆ Stronger, broader outreach effort

- **Forum for US Grid projects**
 - ◆ Joint view, strategies, meetings and work
 - ◆ Unified U.S. entity to interact with international Grid projects





Open Science Grid

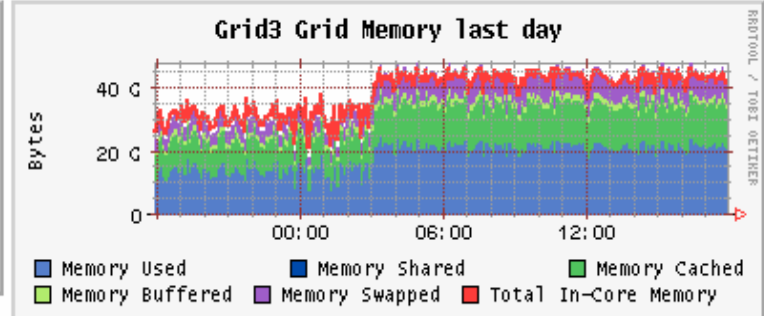
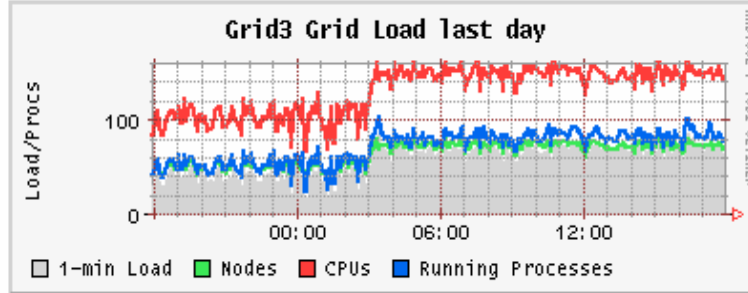
- <http://www.opensciencegrid.org/>
 - ◆ Specific goal: Support US-LHC research program
 - ◆ General goal: U.S. Grid supporting other disciplines
- **Funding mechanism: DOE/NSF**
 - ◆ Laboratories (DOE) and universities (NSF)
 - ◆ Strong interest from Educators: NSF/EHR, QuarkNet, ...
- **Getting there: "Functional Demonstration Grids"**
 - ◆ New release every 6-12 months, increasing functionality & scale
 - ◆ Constant participation in LHC computing exercises
- **Grid2003 ~15-20 sites**
 - ◆ Labs (Fermilab, Brookhaven, Argonne) + Universities
 - ◆ Korea
 - ◆ ~1000 CPUs

Grid3 Grid (8 sources) (tree view)

CPU's Total: **156**
Hosts up: **76**
Hosts down: **1**

Avg Load (15, 5, 1m):
49%, 48%, 49%

Localtime:
2003-09-14 17:54

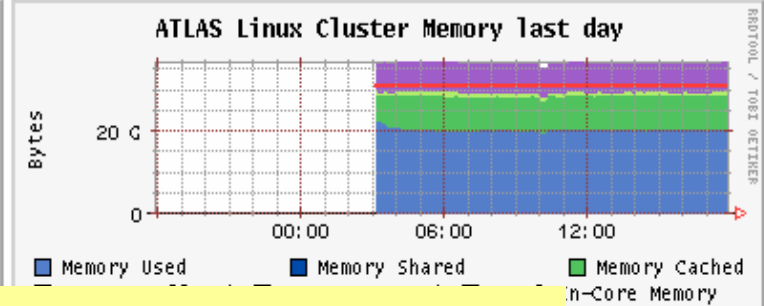
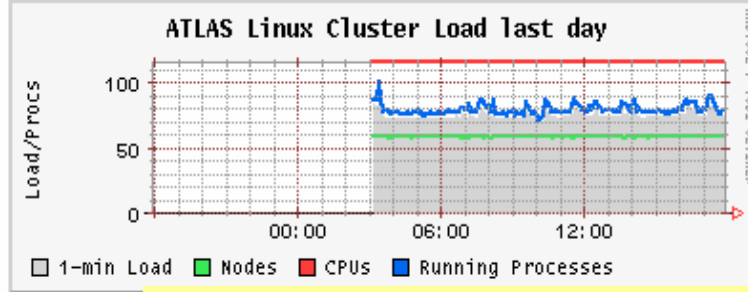


ATLAS Linux Cluster (physical view)

CPU's Total: **116**
Hosts up: **58**
Hosts down: **0**

Avg Load (15, 5, 1m):
64%, 63%, 64%

Localtime:
2003-09-14 17:51



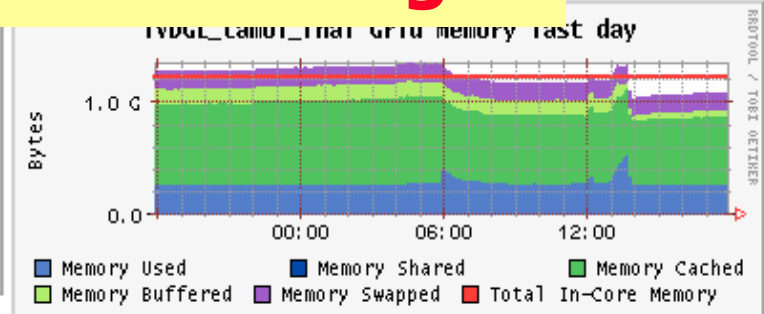
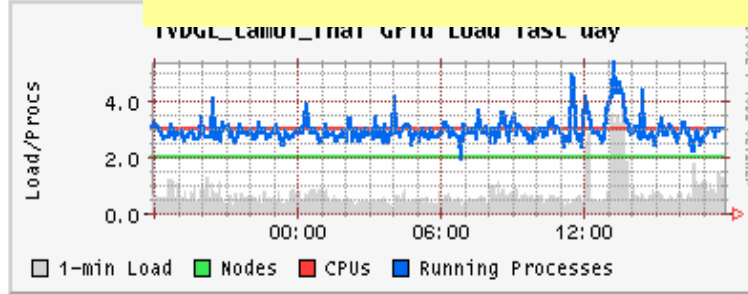
Grid2003 Site Monitoring

iVDGL_tam01_fnal Grid (tree view)

CPU's Total: **3**
Hosts up: **2**
Hosts down: **1**

Avg Load (15, 5, 1m):
36%, 33%, 29%

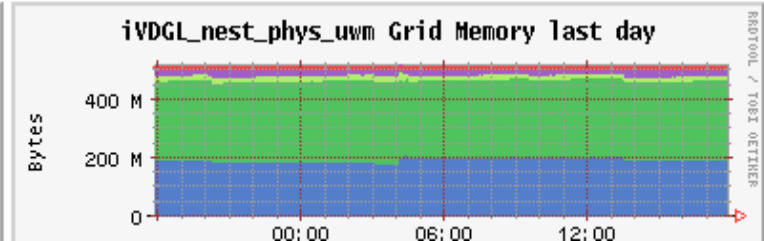
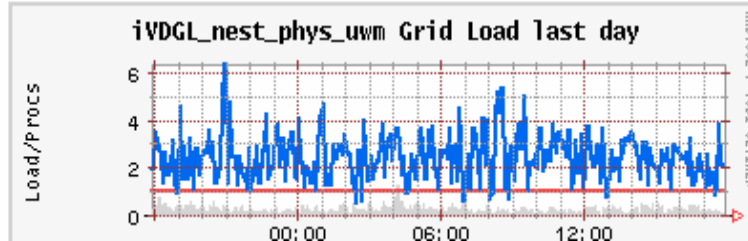
Localtime:
2003-09-14 17:52

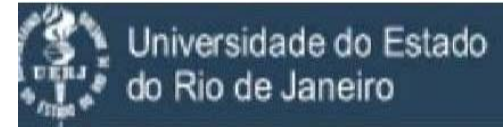


iVDGL_nest_phys_uwm Grid (tree view)

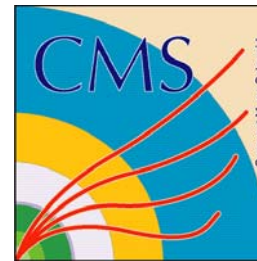
CPU's Total: **1**
Hosts up: **1**
Hosts down: **0**

Avg Load (15, 5, 1m):
13%, 16%, 9%





An Inter-Regional **C**enter for **H**igh **E**nergy **P**hysics **R**esearch and **E**ducational **O**utreach (**CHEPREO**) at Florida International University

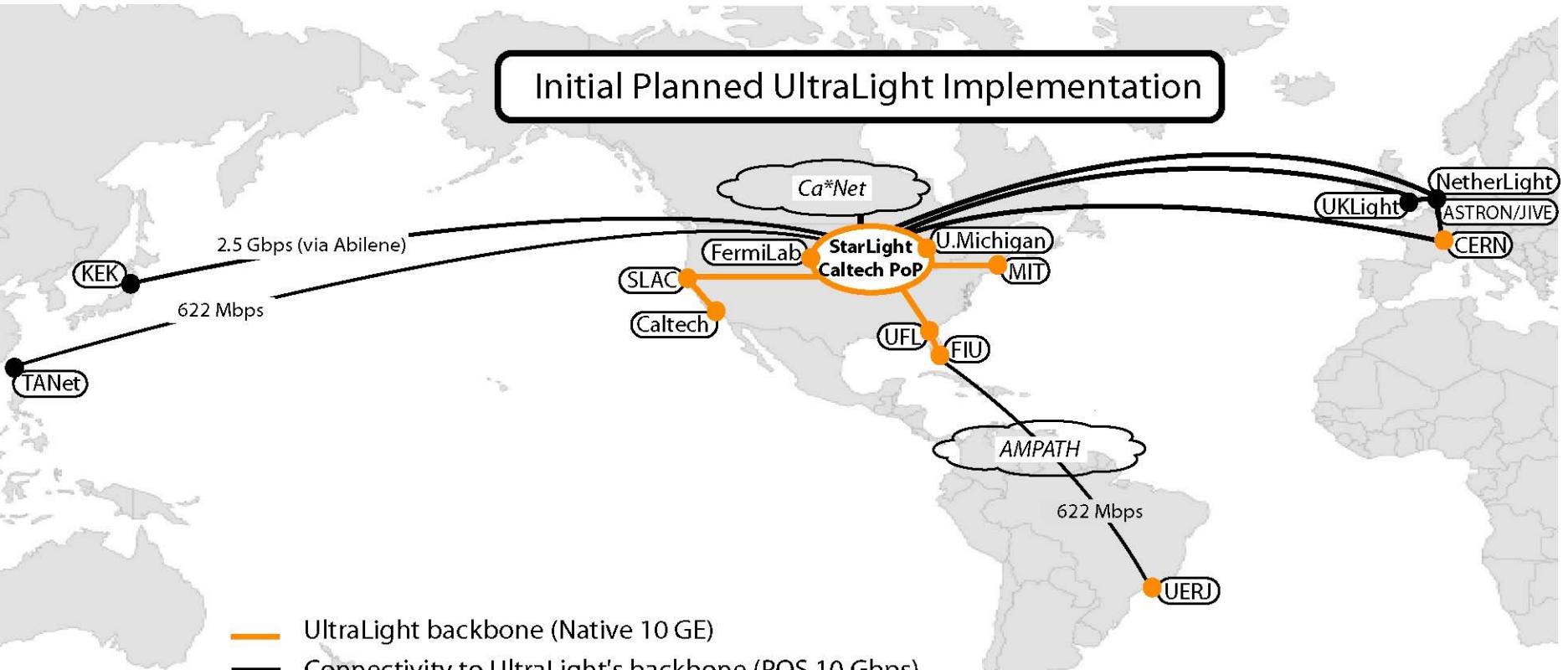


- E/O Center in Miami area
- iVDGL Grid Activities
- CMS Research
- AMPATH network (S. America)

Funded September 2003



UltraLight: 10 Gb/s Network



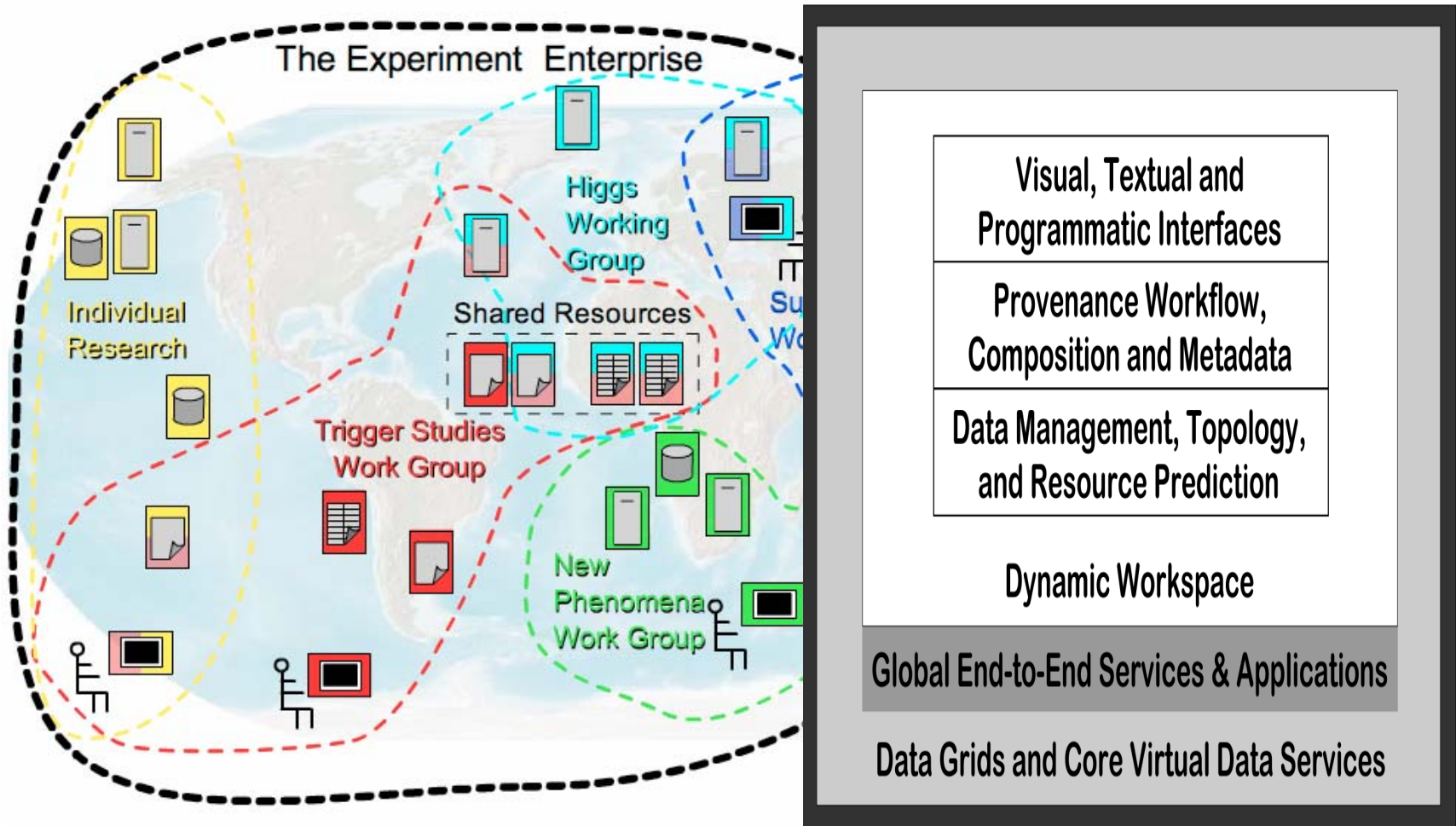
- UltraLight backbone (Native 10 GE)
- Connectivity to UltraLight's backbone (POS 10 Gbps)
- Partners sites
- Peer sites

10 Gb/s+ network

- Caltech, UF, FIU, UM, MIT
- SLAC, FNAL
- Int'l partners
- Level(3), Cisco, NLR

Dynamic Workspaces

Enabling Global Analysis Communities






GLORIAD: US-Russia-China Network

- New 10 Gb/s network linking US-Russia-China
 - ◆ Plus Grid component linking science projects
- Meeting at NSF April 14 with US-Russia-China reps.
 - ◆ HEP people (Hesheng, et al.)
- Broad agreement that HEP can drive Grid portion
 - ◆ Other applications will be solicited
- More meetings planned



Grids: Enhancing Research & Learning

➤ **Fundamentally alters conduct of scientific research**

- 
- ◆ “Lab-centric”: Activities center around large facility
 - ◆ “Team-centric”: Resources shared by distributed teams
 - ◆ “Knowledge-centric”: Knowledge generated/used by a community

➤ **Strengthens role of universities in research**

- ◆ Couples universities to data intensive science
- ◆ Couples universities to national & international labs
- ◆ Brings front-line research and resources to students
- ◆ Exploits intellectual resources of formerly isolated schools
- ◆ Opens new opportunities for minority and women researchers

➤ **Builds partnerships to drive advances in IT/science/eng**

- ◆ HEP ⇔ Physics, astronomy, biology, CS, etc.
- ◆ “Application” sciences ⇔ Computer Science
- ◆ Universities ⇔ Laboratories
- ◆ Scientists ⇔ Students
- ◆ Research Community ⇔ IT industry

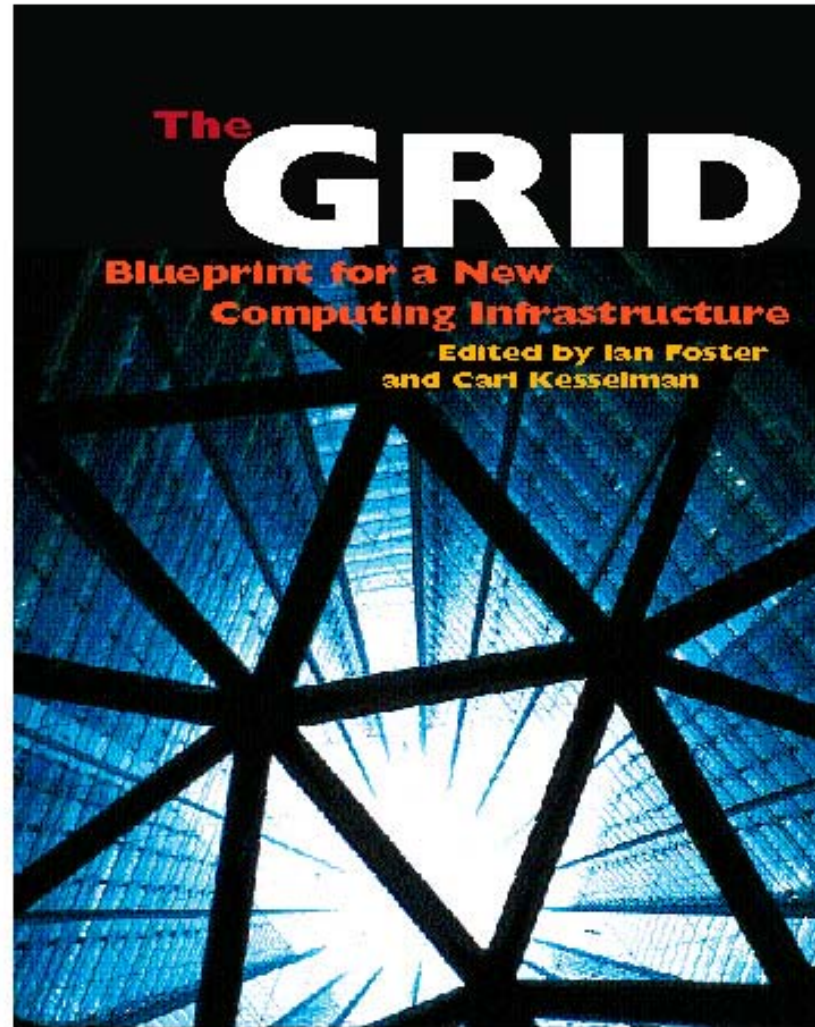


HEP's Broad Impact and Relevance

- HEP is recognized as the strongest science driver for Grids
 - ◆ (In collaboration with computer scientists)
 - ◆ LHC a particularly strong driving function
- Grid projects are driving important network developments
 - ◆ "Land speed records" attract much attention
 - ◆ ICFA-SCIC, I-HEPCCC, US-CERN link, ESNET, Internet2
- We are increasing our impact on education and outreach
 - ◆ Providing technologies, resources for training, education, outreach
- HEP involvement in Grid projects has helped us!
 - ◆ Many \$M brought into the field
 - ◆ Many visible national and international initiatives
 - ◆ Partnerships with other disciplines ⇒ increasing our visibility
 - ◆ Recognition at high levels (NSF, DOE, EU, Asia)

Grid References

- Grid Book
 - ◆ www.mkp.com/grids
- Globus
 - ◆ www.globus.org
- Global Grid Forum
 - ◆ www.gridforum.org
- PPDG
 - ◆ www.ppdg.net
- GriPhyN
 - ◆ www.griphyn.org
- iVDGL
 - ◆ www.ivdgl.org
- TeraGrid
 - ◆ www.teragrid.org
- EU DataGrid
 - ◆ www.eu-datagrid.org



Extra Slides



Some (Realistic) Grid Examples

- **High energy physics**
 - ◆ 3,000 physicists worldwide pool Petaflops of CPU resources to analyze Petabytes of data
- **Fusion power (ITER, etc.)**
 - ◆ Physicists quickly generate 100 CPU-years of simulations of a new magnet configuration to compare with data
- **Astronomy**
 - ◆ An international team remotely operates a telescope in real time
- **Climate modeling**
 - ◆ Climate scientists visualize, annotate, & analyze Terabytes of simulation data
- **Biology**
 - ◆ A biochemist exploits 10,000 computers to screen 100,000 compounds in an hour



GriPhyN Goals

- Conduct CS research to achieve vision
 - ◆ “Virtual Data” as unifying principle
- Disseminate through Virtual Data Toolkit (VDT)
 - ◆ Primary deliverable of GriPhyN
- Integrate into GriPhyN science experiments
 - ◆ Common Grid tools, services
- Impact other disciplines
 - ◆ HEP, biology, medicine, virtual astronomy, eng.
- Educate, involve, train students in IT research
 - ◆ Undergrads, grads, postdocs, underrepresented groups



iVDGL Goals and Context

- **International Virtual-Data Grid Laboratory**
 - ◆ A global Grid laboratory (US, EU, E. Europe, Asia, S. America, ...)
 - ◆ A place to conduct Data Grid tests “at scale”
 - ◆ A mechanism to create common Grid infrastructure
 - ◆ A laboratory for other disciplines to perform Data Grid tests
 - ◆ A focus of outreach efforts to small institutions
- **Context of iVDGL in LHC computing program**
 - ◆ Develop and operate proto-Tier2 centers
 - ◆ Learn how to do Grid operations (GOC)
- **International participation**
 - ◆ DataTag partner project in EU
 - ◆ New international partners: Korea and Brazil
 - ◆ UK e-Science programme: support 6 CS Fellows per year in U.S.

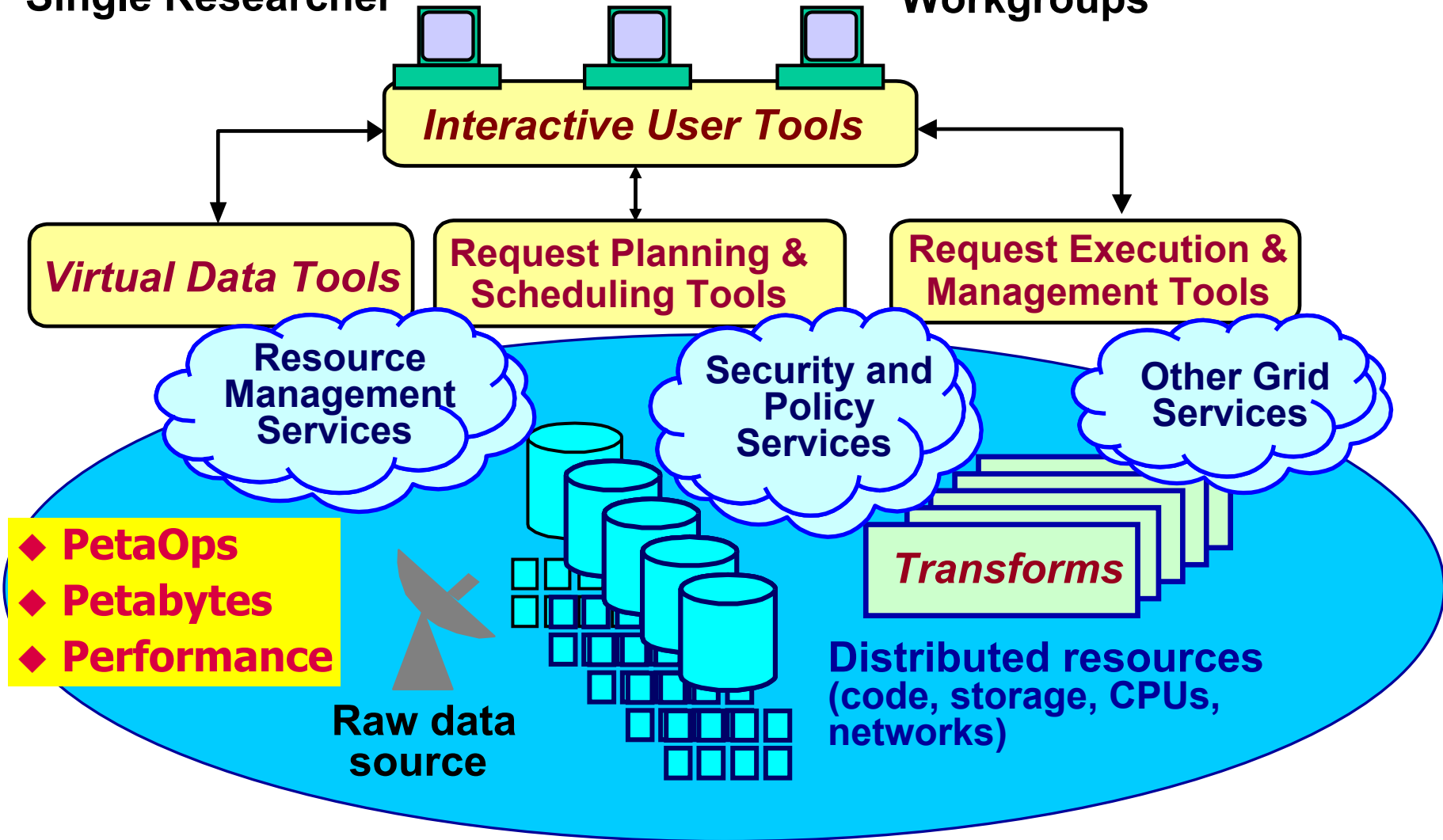


Goal: PetaScale Virtual-Data Grids

Single Researcher

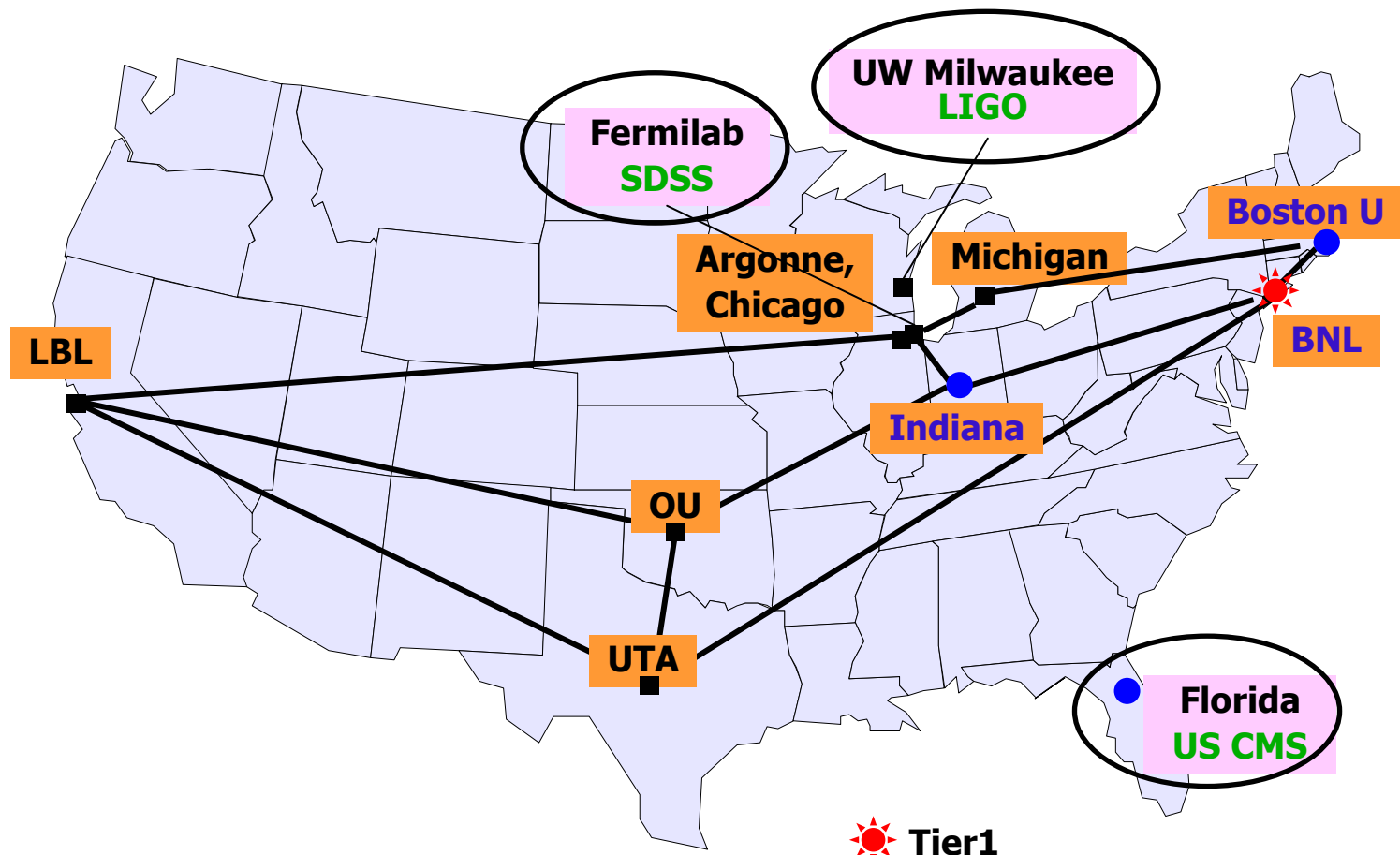
Production Team

Workgroups





ATLAS Simulations on iVDGL Resources

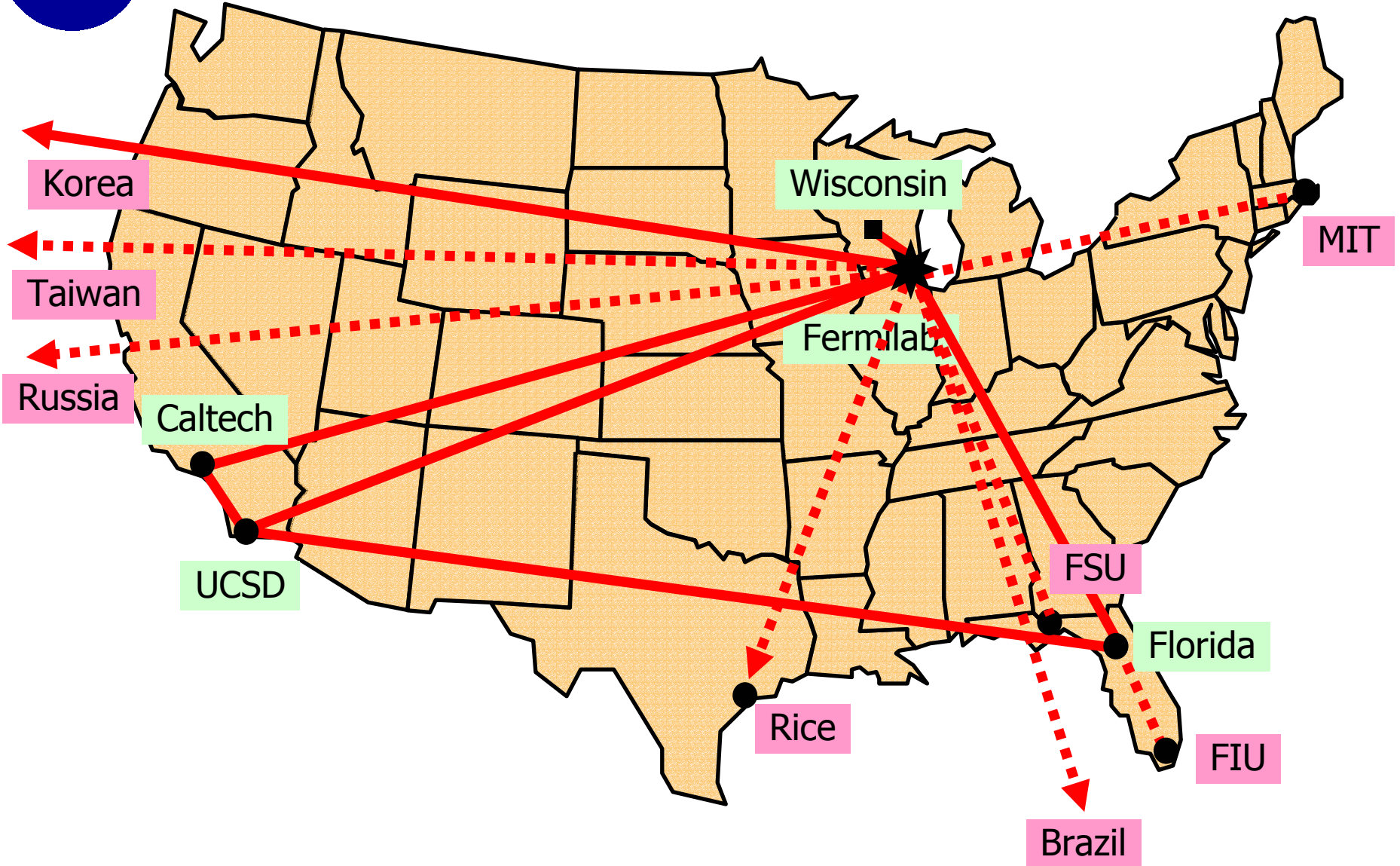


-  Tier1
-  Prototype Tier2
-  Testbed sites

Joint project with iVDGL



US-CMS Testbed





WorldGrid Demonstration (Nov. 2002)

➤ Joint iVDGL + EU effort

- ◆ Resources from both sides (15 sites)
- ◆ Monitoring tools (Ganglia, MDS, NetSaint, ...)
- ◆ Visualization tools (Nagios, MapCenter, Ganglia)

➤ Applications

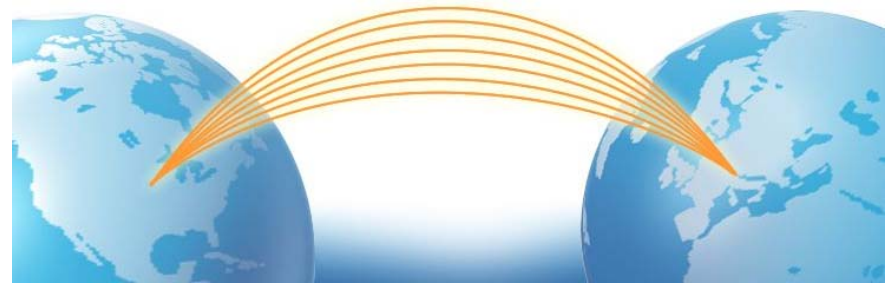
- ◆ CMS: CMKIN, CMSIM
- ◆ ATLAS: ATLSIM

➤ Submit jobs from US or EU

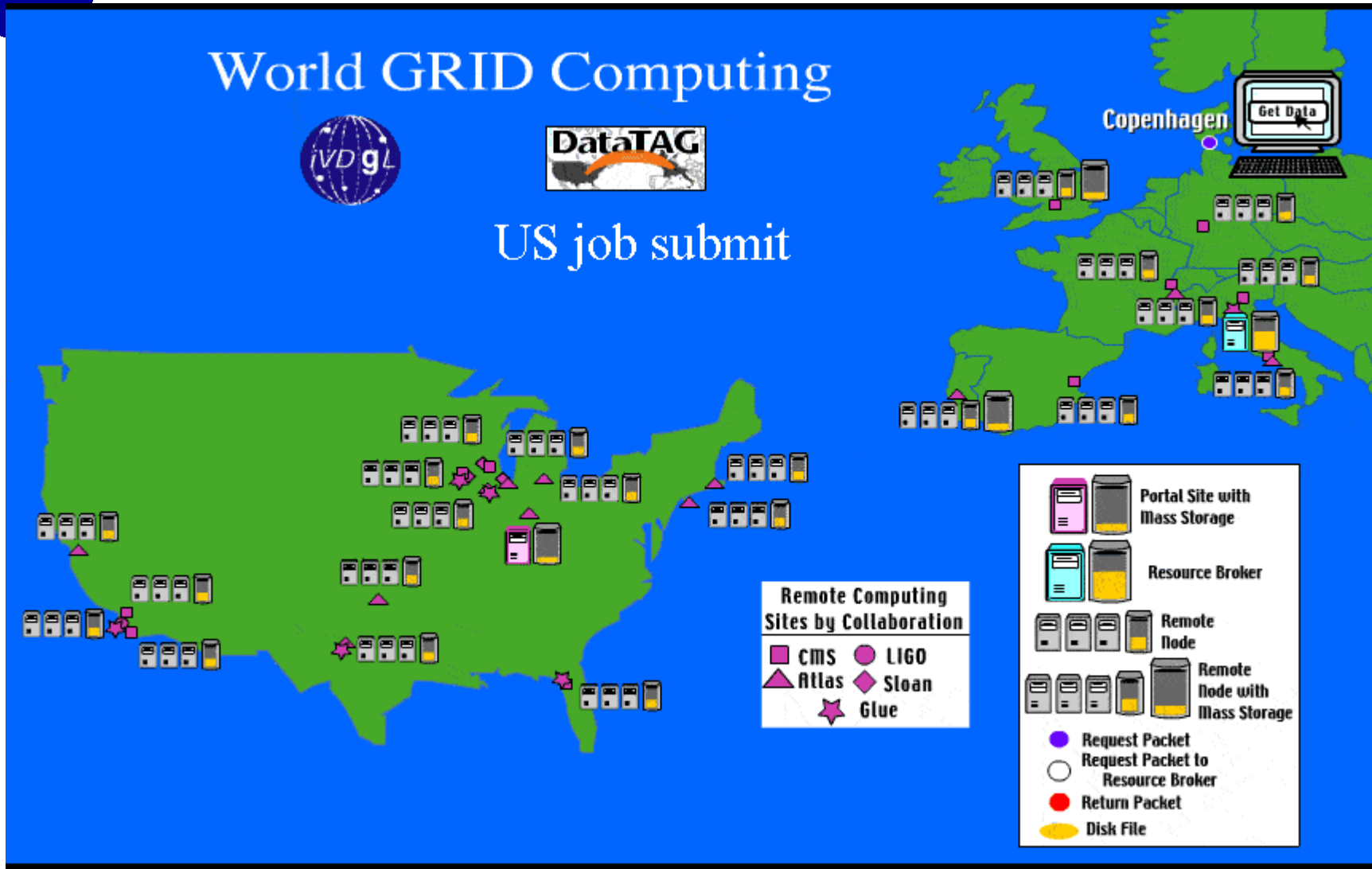
- ◆ Jobs can run on any cluster

➤ Major demonstrations

- ◆ IST2002 (Copenhagen)
- ◆ SC2002 (Baltimore)



WorldGrid Sites (Nov. 2002)





International Grid Coordination

- **Global Grid Forum (GGF)**
 - ◆ International forum for general Grid efforts
 - ◆ Many working groups, standards definitions
- **Close collaboration with EU DataGrid (EDG)**
 - ◆ Many connections with EDG activities
- **HICB: HEP Inter-Grid Coordination Board**
 - ◆ Non-competitive forum, strategic issues, consensus
 - ◆ Cross-project policies, procedures and technology, joint projects
- **HICB-JTB Joint Technical Board**
 - ◆ Definition, oversight and tracking of joint projects
 - ◆ GLUE interoperability group
- **Participation in LHC Computing Grid (LCG)**
 - ◆ Software Computing Committee (SC2)
 - ◆ Project Execution Board (PEB)
 - ◆ Grid Deployment Board (GDB)

- **Most scientific data are not simple “measurements”**
 - ◆ They are computationally corrected/reconstructed
 - ◆ They can be produced by numerical simulation
- **Science & eng. projects are more CPU and data intensive**
 - ◆ Programs are significant community resources (transformations)
 - ◆ So are the executions of those programs (derivations)
- **Management of dataset transformations important!**
 - ◆ Derivation: Instantiation of a potential data product
 - ◆ Provenance: Exact history of any existing data product

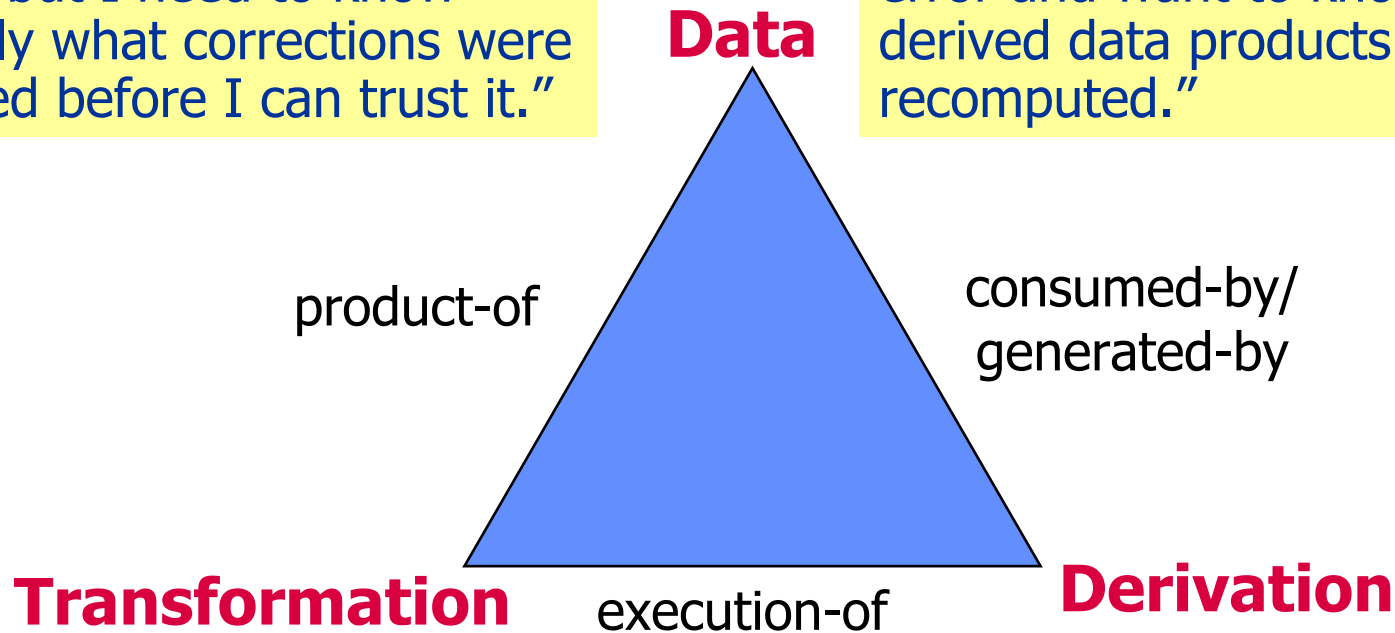
We already do this, but manually!



Virtual Data Motivations (1)

"I've found some interesting data, but I need to know exactly what corrections were applied before I can trust it."

"I've detected a muon calibration error and want to know which derived data products need to be recomputed."

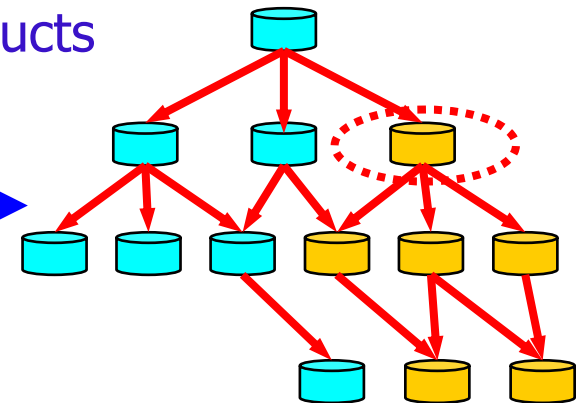


"I want to search a database for 3 muon SUSY events. If a program that does this analysis exists, I won't have to write one from scratch."

"I want to apply a forward jet analysis to 100M events. If the results already exist, I'll save weeks of computation."

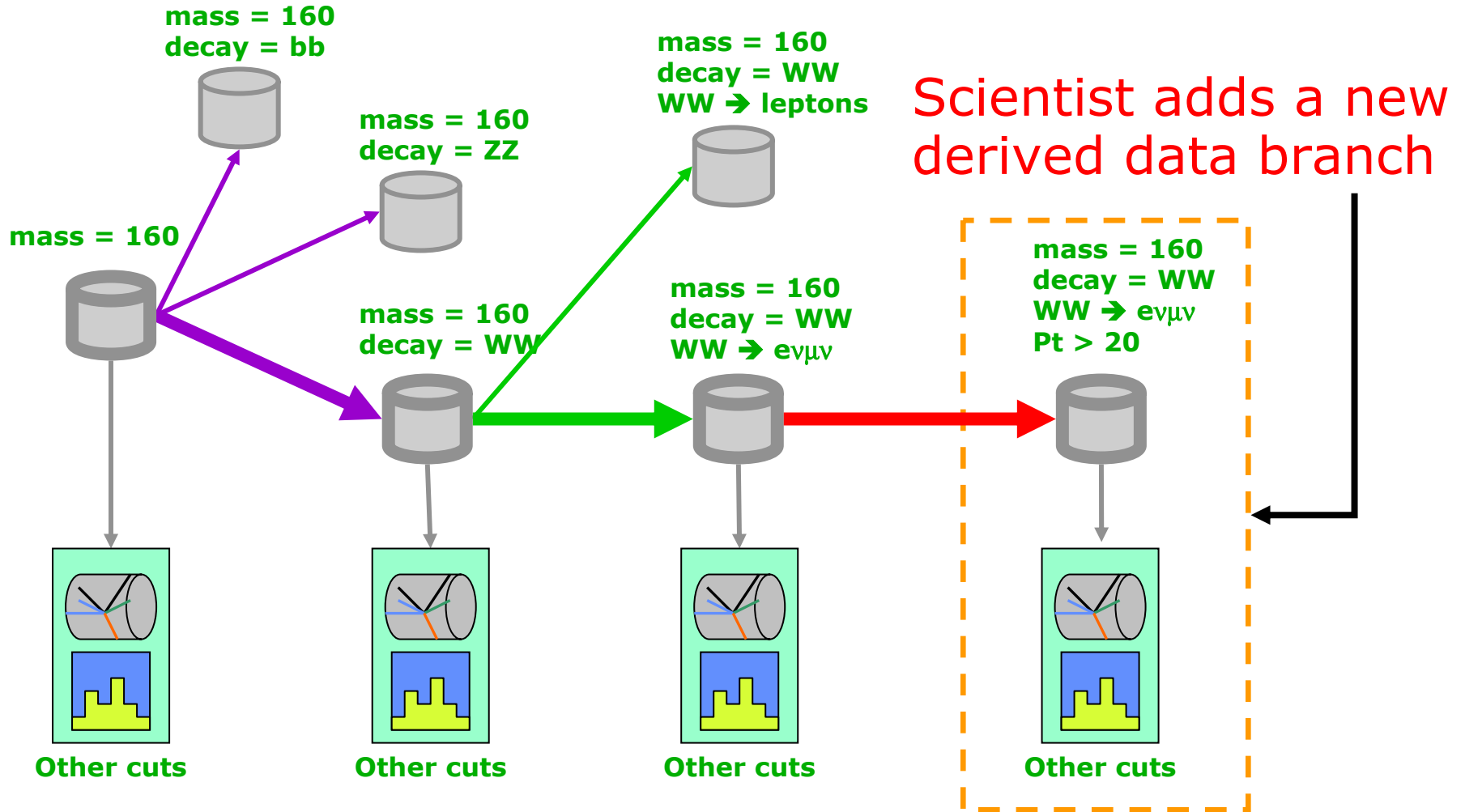
Virtual Data Motivations (2)

- Data track-ability and result audit-ability
 - ◆ Universally sought by scientific applications
- Facilitate resource sharing and collaboration
 - ◆ Data is sent along with its recipe
 - ◆ A new approach to saving old data: economic consequences?
- Manage workflow
 - ◆ Organize, locate, specify, request data products
- Repair and correct data automatically
 - ◆ Identify dependencies, apply x-tions
- Optimize performance
 - ◆ Re-create data or copy it (caches)



Manual /error prone ⇒ Automated /robust

LHC Analysis with Virtual Data

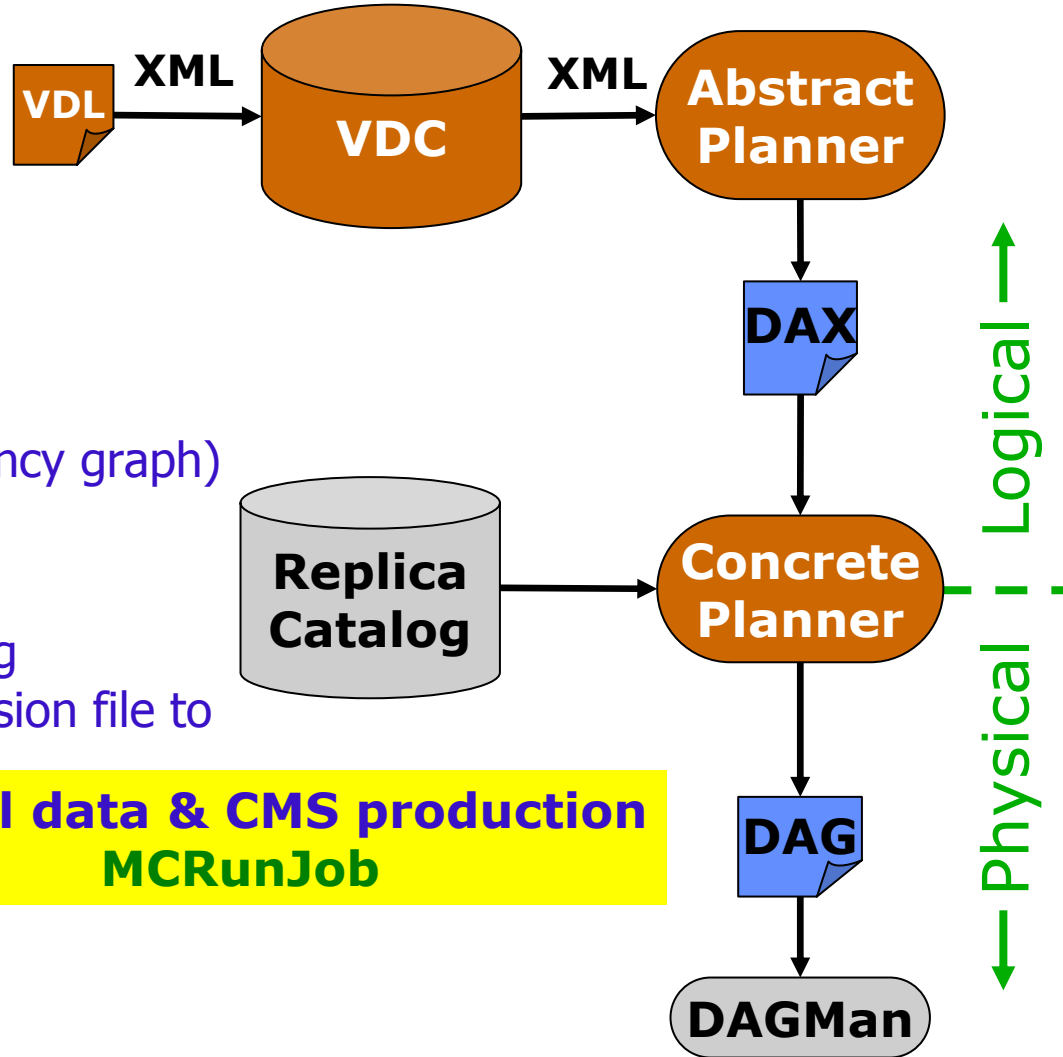


Scientist adds a new derived data branch

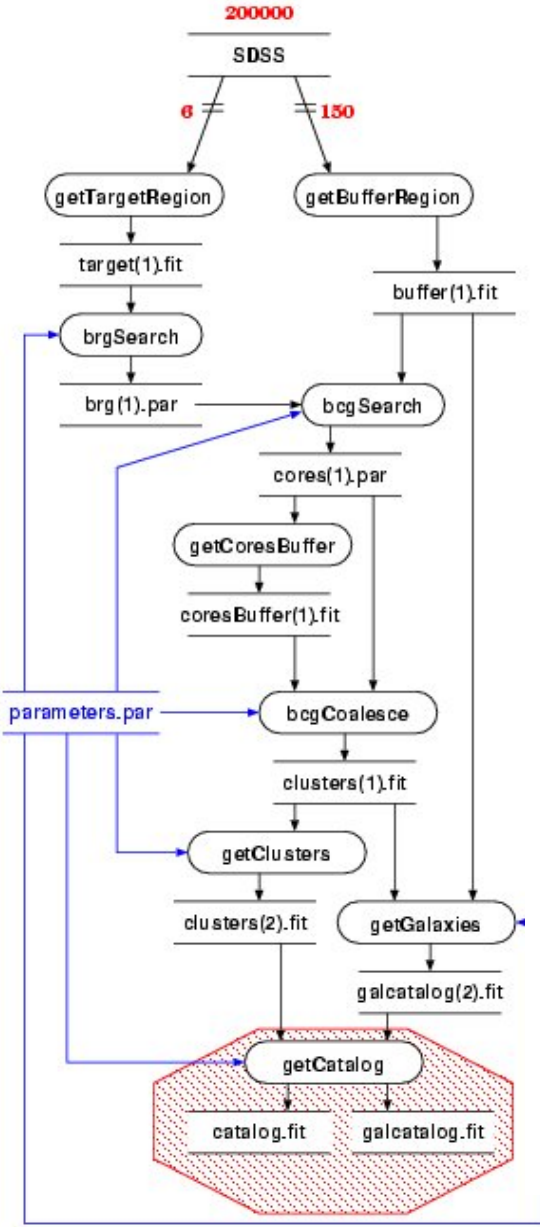


Chimera Virtual Data System

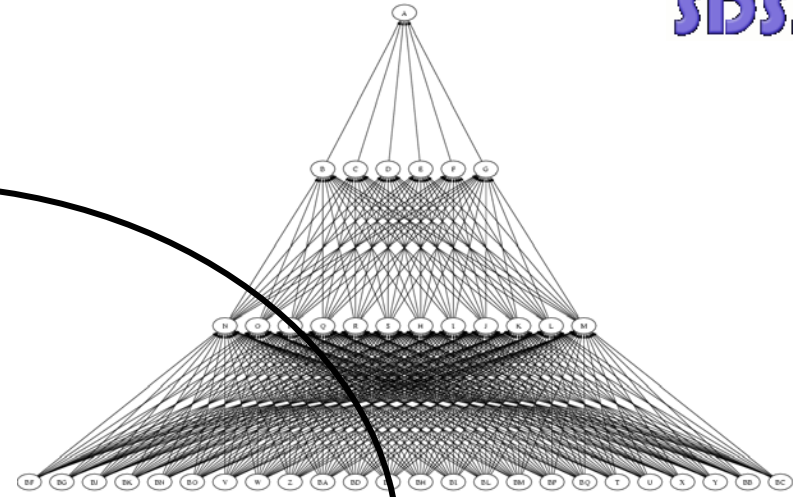
- Virtual Data Language (VDL)
 - ◆ Describes virtual data products
- Virtual Data Catalog (VDC)
 - ◆ Used to store VDL
- Abstract Job Flow Planner
 - ◆ Creates a logical DAG (dependency graph)
- Concrete Job Flow Planner
 - ◆ Interfaces with a Replica Catalog
 - ◆ Provides a physical DAG submission file to Condor-G
- Generic and flexible
 - ◆ As a toolkit and/or a framework
 - ◆ In a Grid environment or locally



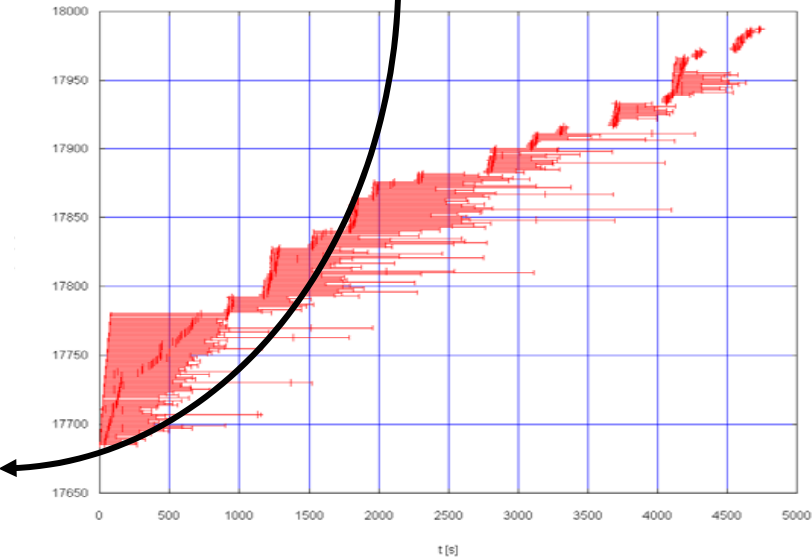
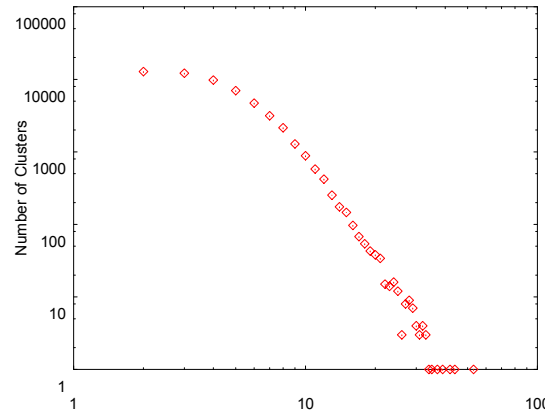
Virtual data & CMS production
MCRunJob



Sloan Data



Galaxy cluster size distribution





A Global Grid Enabled Collaboratory for Scientific Research (GECSR)

- \$4M ITR proposal from
 - ◆ Caltech (HN PI, JB:CoPI)
 - ◆ Michigan (CoPI, CoPI)
 - ◆ Maryland (CoPI)
- Plus senior personnel from
 - ◆ Lawrence Berkeley Lab
 - ◆ Oklahoma
 - ◆ Fermilab
 - ◆ Arlington (U. Texas)
 - ◆ Iowa
 - ◆ Florida State
- First Grid-enabled Collaboratory
- Tight integration between
 - ◆ Science of Collaboratories
 - ◆ Globally scalable work environment
 - ◆ Sophisticated collaborative tools (VRVS, VNC; Next-Gen)
 - ◆ Agent based monitoring & decision support system (MonALISA)

Initial targets are the global HEP collaborations, but applicable to other large scale collaborative scientific endeavors