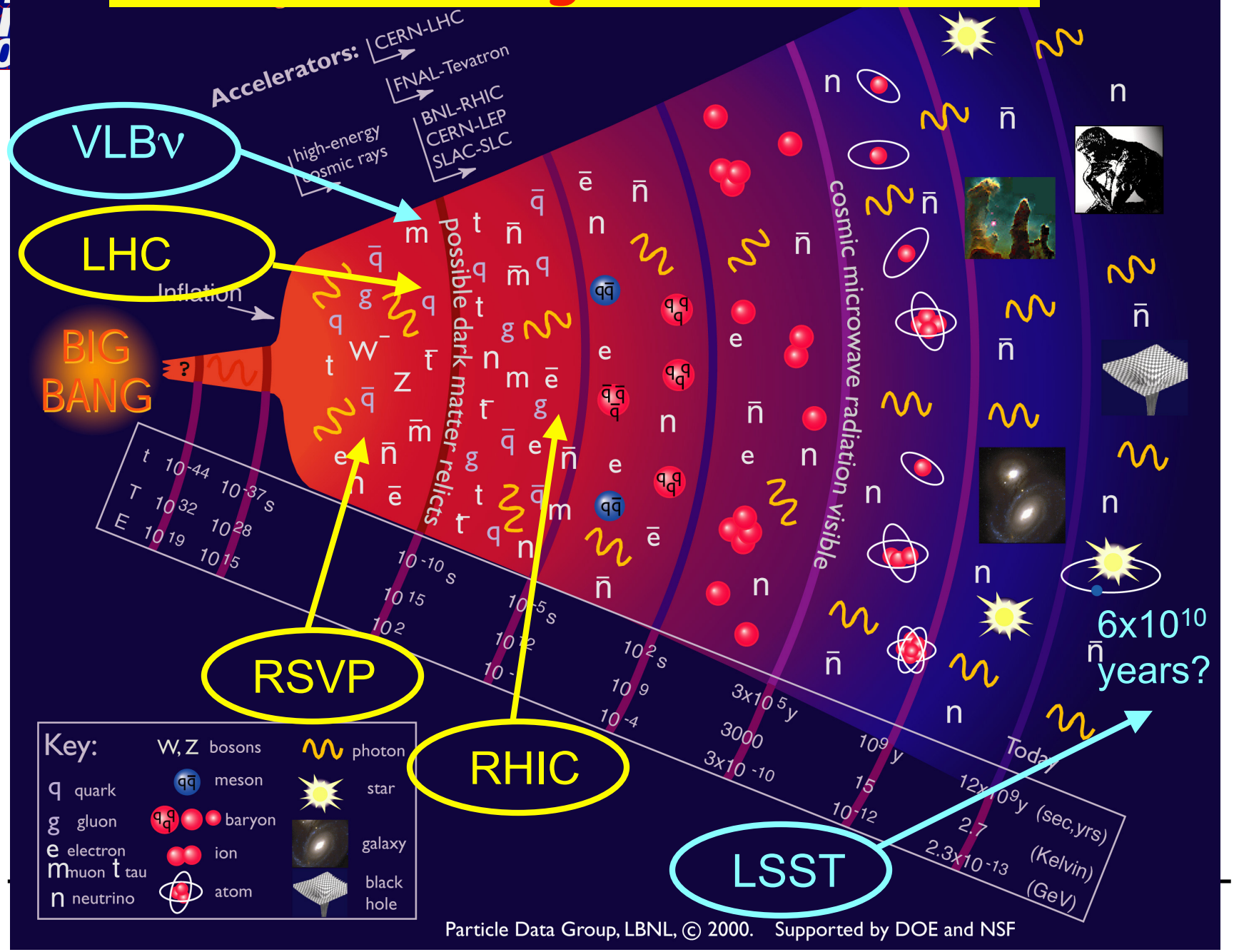
Grids: How Universities Participate in Global Science

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

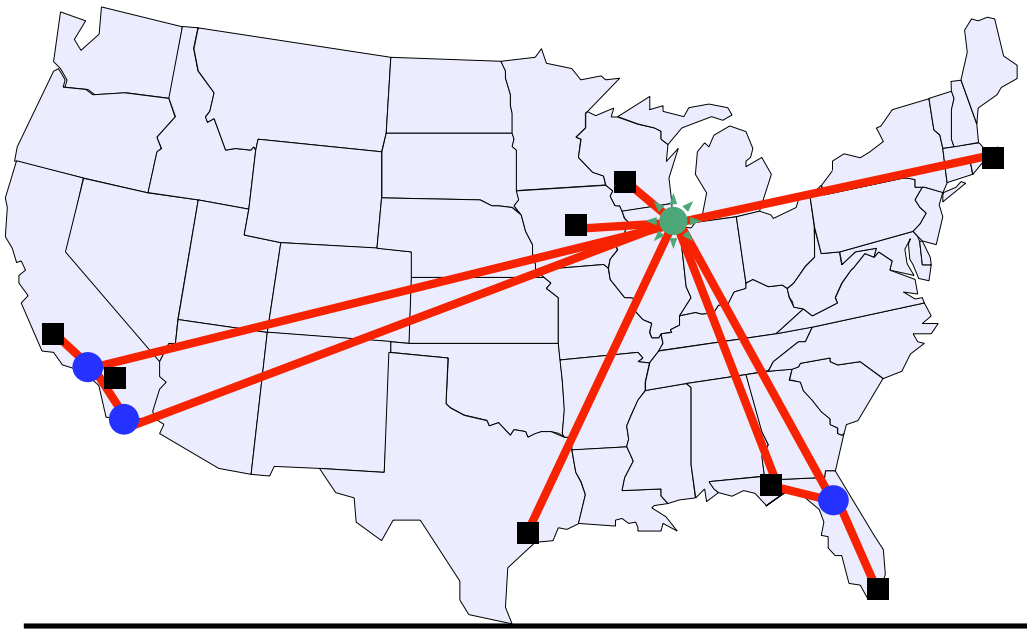
Understanding the Universe
Education / Outreach Discussion Meeting
Arlington, Virginia
April 8, 2004

Understanding the Universe



The Grid Concept

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



US-CMS
"Virtual Organization"

Global Context of Data Grid Projects

Collaborating Grid infrastructure projects

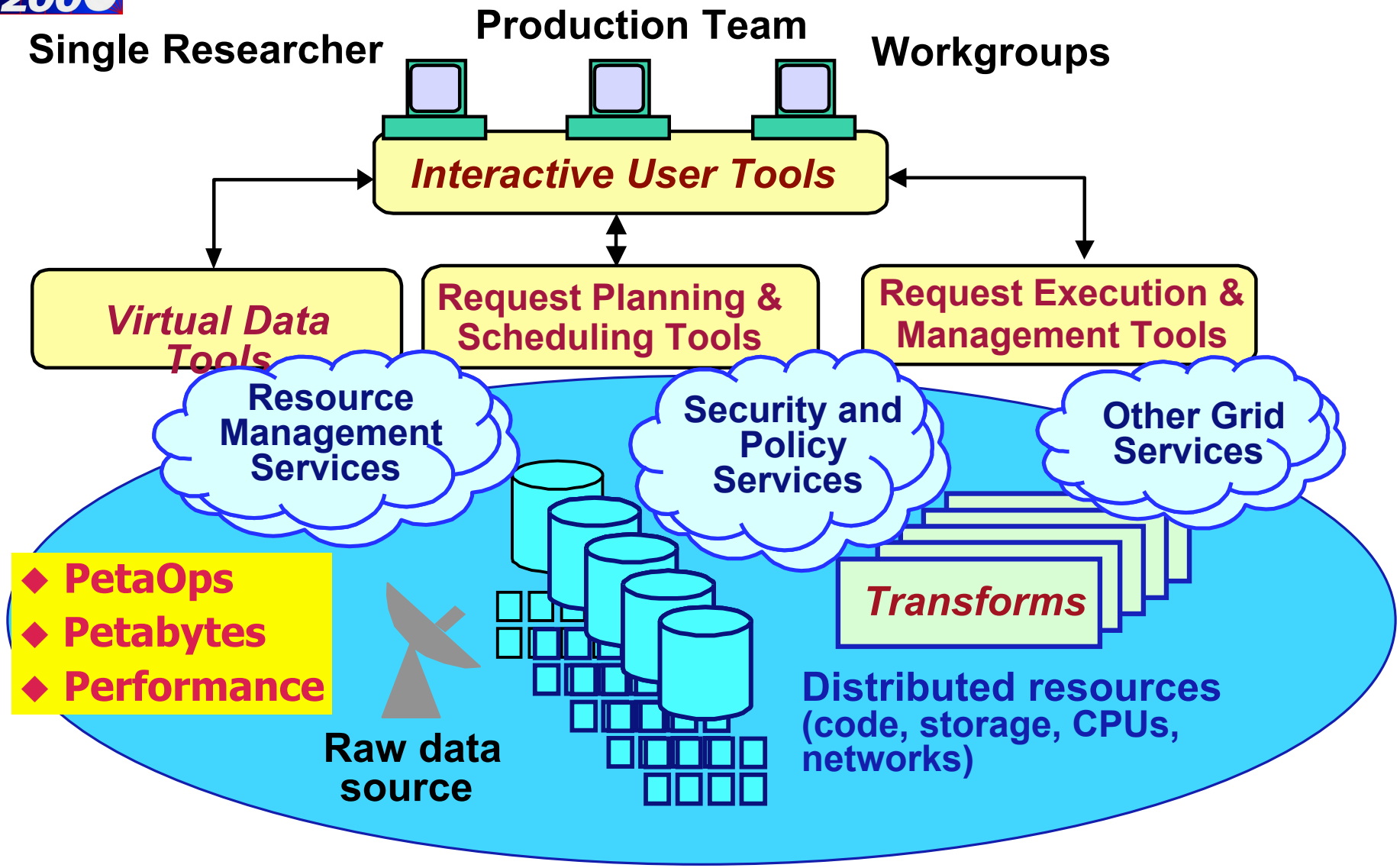
- U.S. Projects
  - GriPhyN (NSF)
  - iVDGL (NSF)
  - Particle Physics Data Grid (DOE)
  - PACIs and TeraGrid (NSF)
  - DOE Science Grid (DOE)
  - NEESgrid (NSF)
  - NSF Middleware Initiative (NSF)
- EU, Asia projects
  - European Data Grid (EU)
  - EDG-related national Projects
  - DataTAG (EU)
  - LHC Computing Grid (CERN)
  - EGEE (EU)
  - CrossGrid (EU)
  - GridLab (EU)
  - Japanese, Korea Projects

- **Two primary project clusters: US + EU**
- **Not exclusively HEP, but driven/led by HEP (with CS)**

U.S. "Trillium" Grid Projects

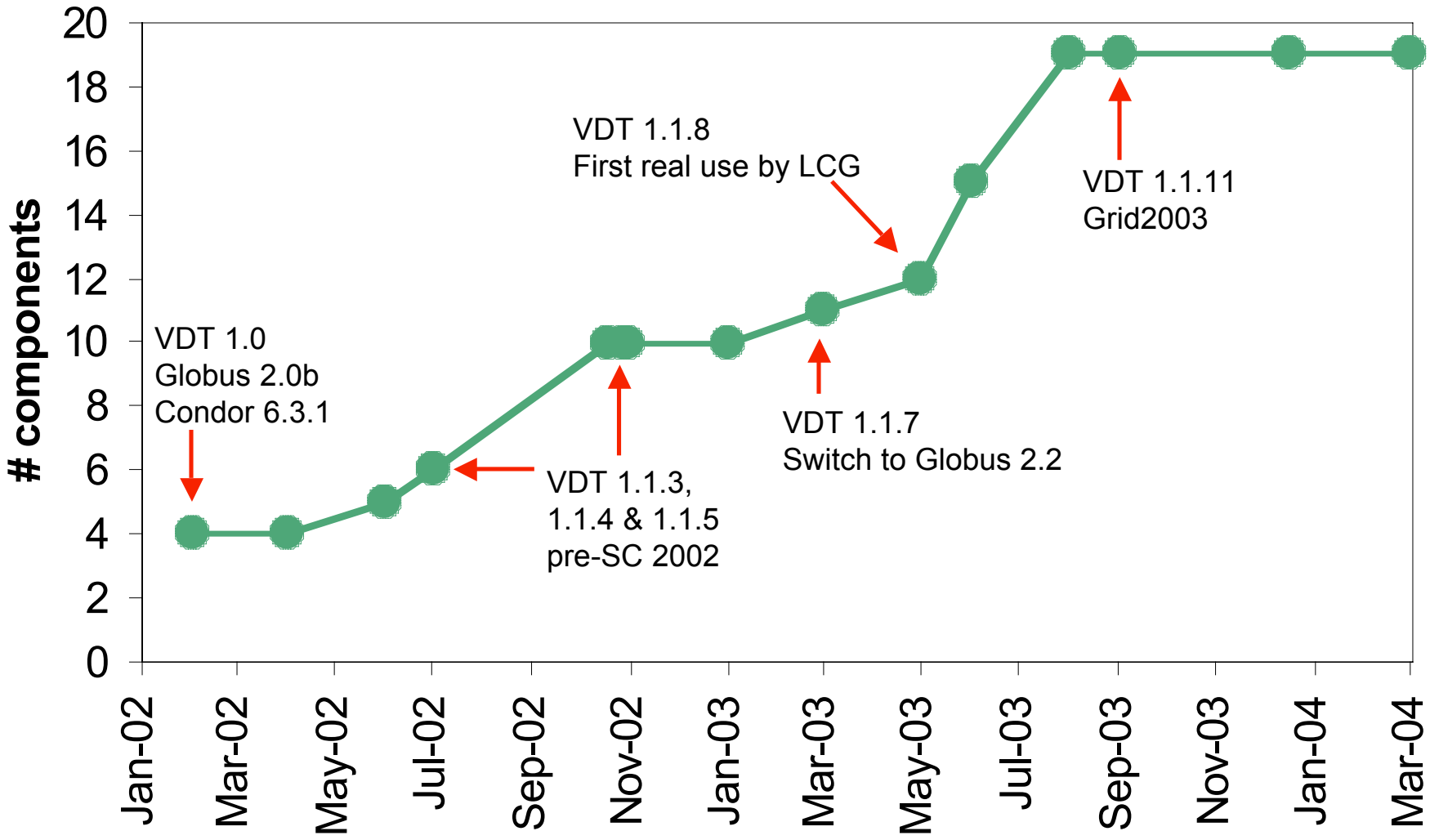
- Trillium = PPDG + GriPhyN + iVDGL
  - PPDG: $10M (DOE) (1999 – 2004+)
  - GriPhyN: $12M (NSF) (2000 – 2005)
  - iVDGL: $14M (NSF) (2001 – 2006)
- Basic composition (~150 people)
  - PPDG: 4 universities, 6 labs
  - GriPhyN: 12 universities, SDSC, 3 labs
  - iVDGL: 18 universities, SDSC, 4 labs, foreign partners
  - Expts: BaBar, D0, STAR, Jlab, CMS, ATLAS, LIGO, SDSS/NVO
- Complementarity of projects
  - GriPhyN: CS research, Virtual Data Toolkit (VDT) development
  - PPDG: "End to end" Grid, services, monitoring, analysis
  - iVDGL: Grid laboratory deployment using VDT
  - Experiments provide frontier challenges

Trillium: "PetaScale" Virtual-Data Grids

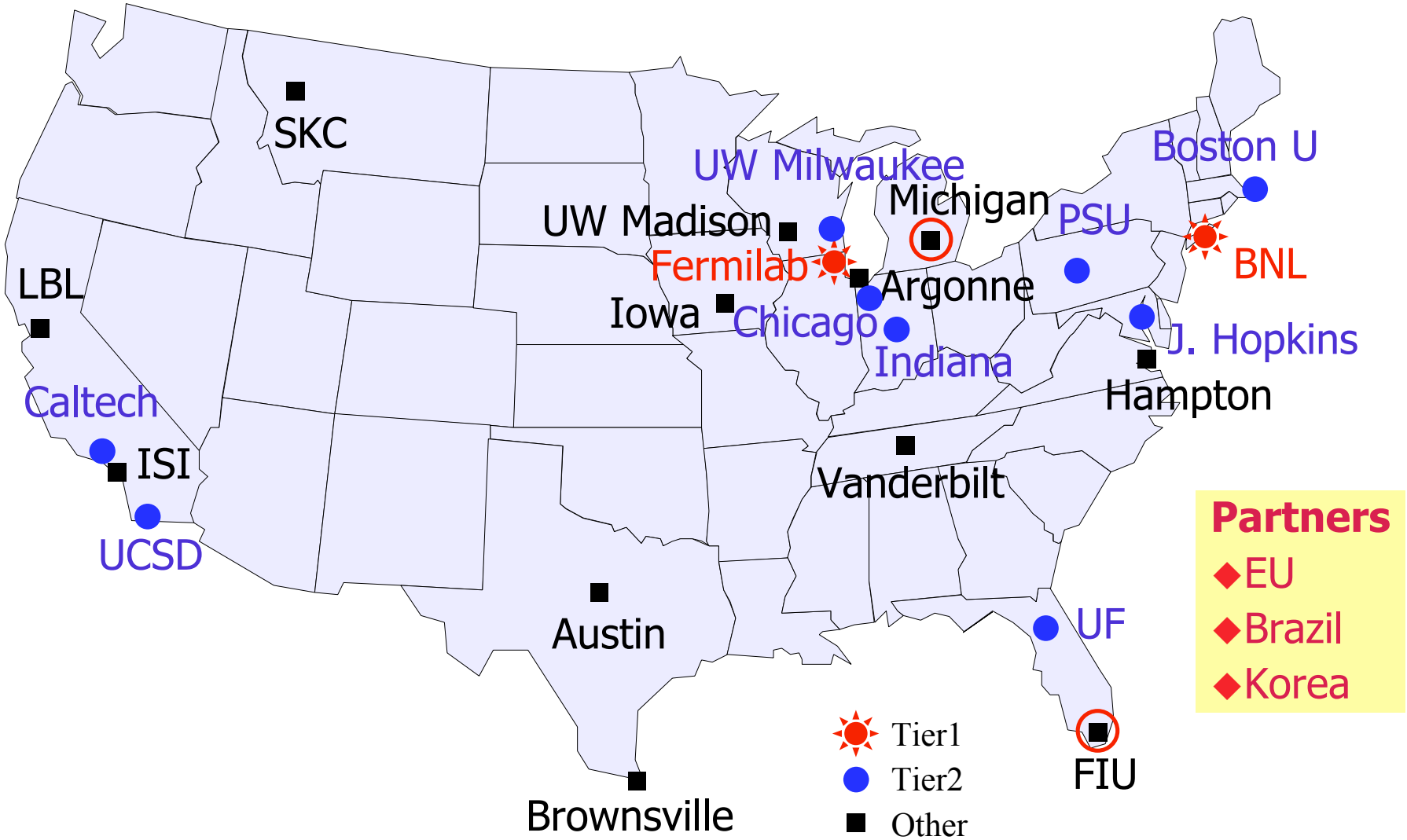


- PetaOps
- Petabytes
- Performance

GriPhyN: Virtual Data Toolkit

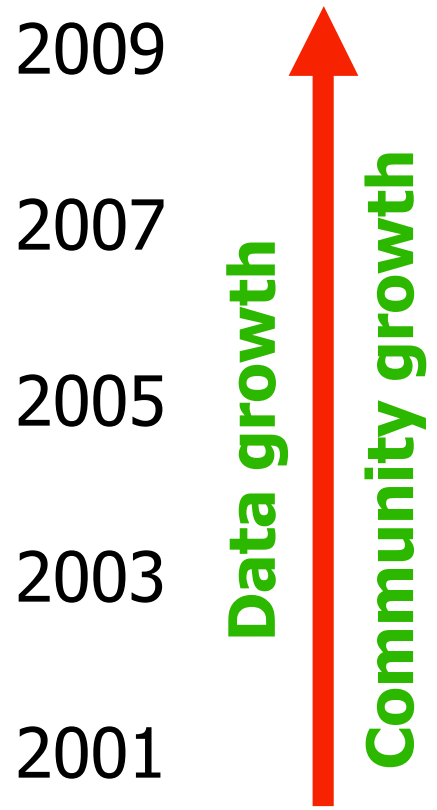


International Virtual Data Grid Laboratory



Trillium Science Drivers

- LHC experiments
  - 100s of Petabytes
- High Energy Physics experiments
  - ~1 Petabyte (1000 TB)
- LIGO (gravity wave search)
  - 100s of Terabytes
- Sloan Digital Sky Survey
  - 10s of Terabytes

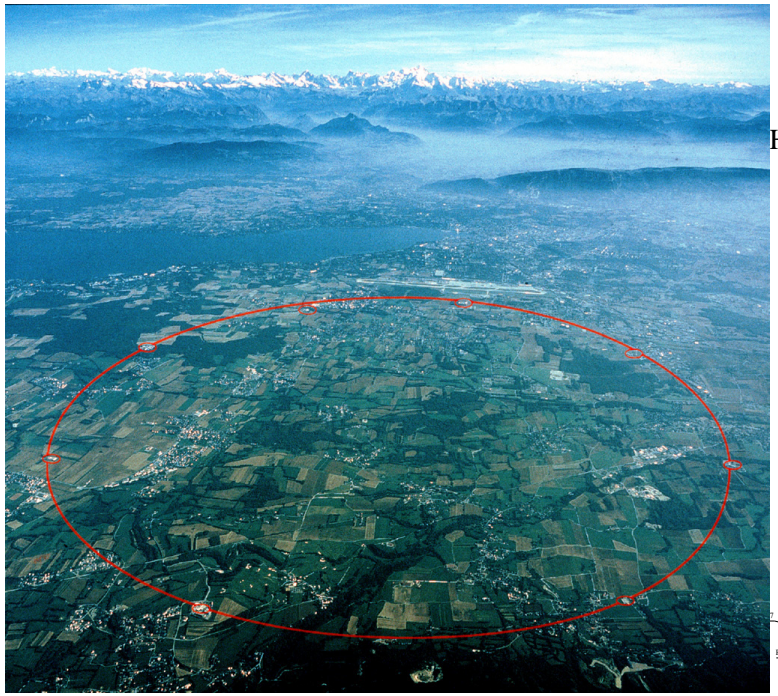


Future Grid resources

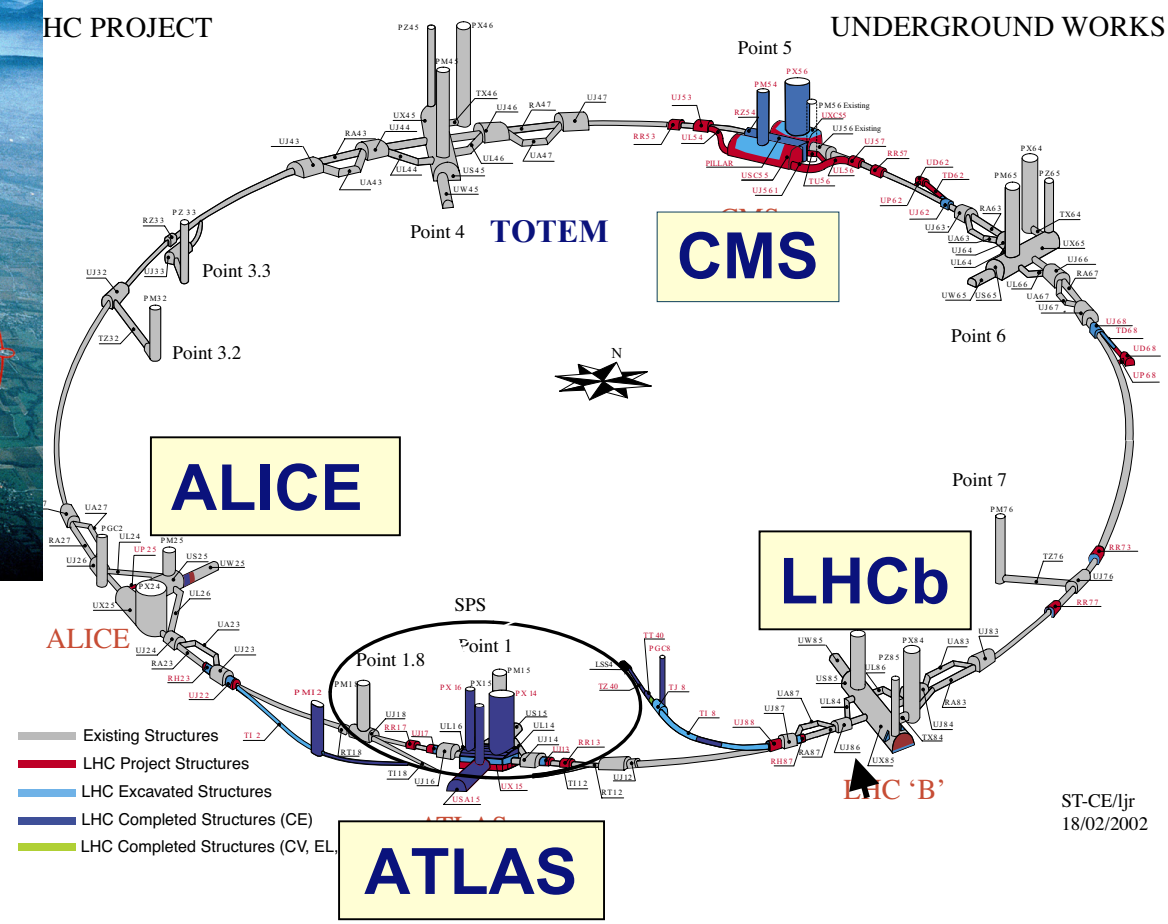
- **Massive CPU (PetaOps)**
- **Large distributed datasets (>100PB)**
- **Global communities (1000s)**

Data Grids and The Large Hadron Collider

Large Hadron Collider (LHC) @ CERN



✸ **27 km Tunnel in Switzerland & France**

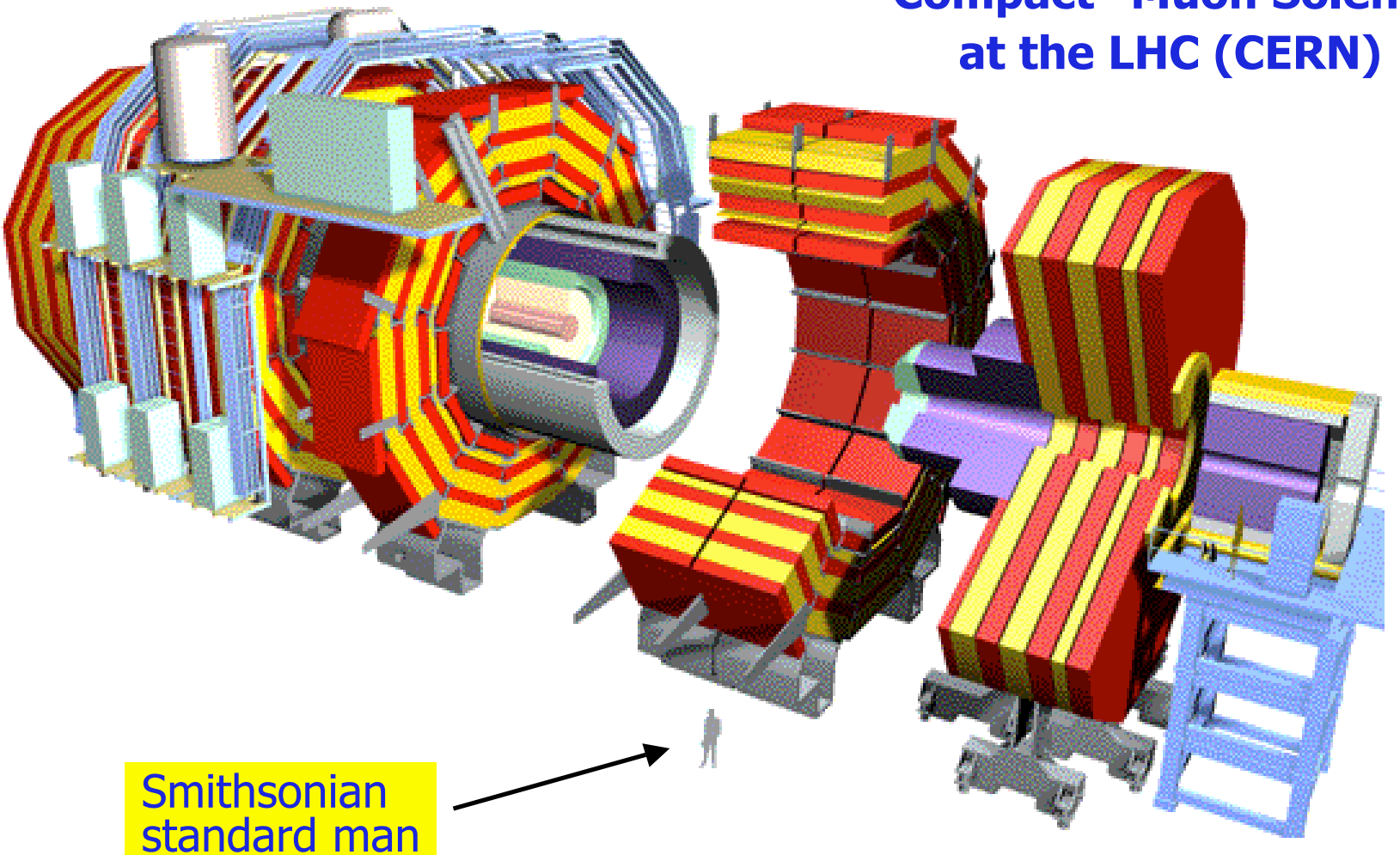


2007 – ?

- **Origin of Mass**
- **Supersymmetry**
- **New particles**

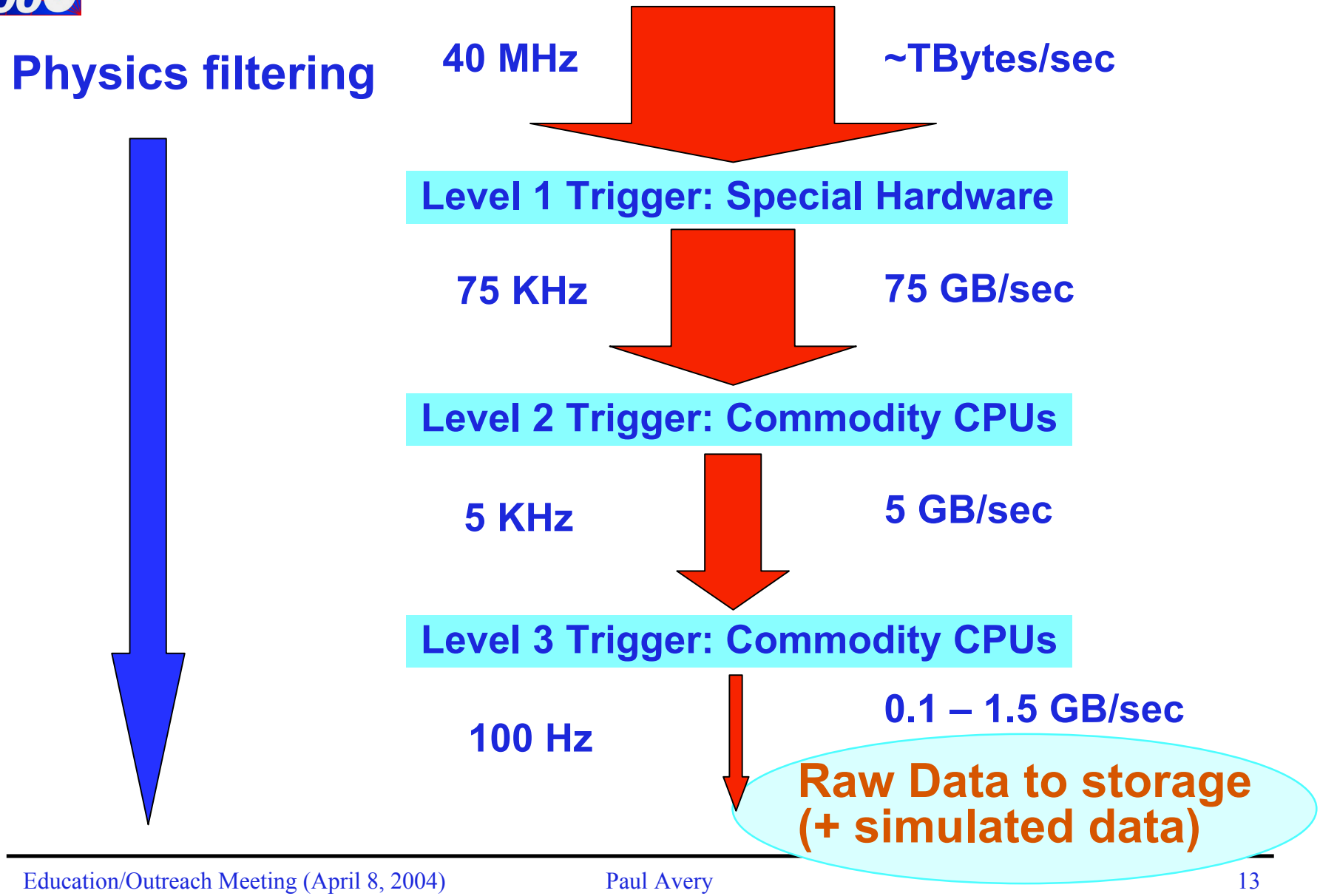
Example: CMS Experiment at LHC

"Compact" Muon Solenoid at the LHC (CERN)

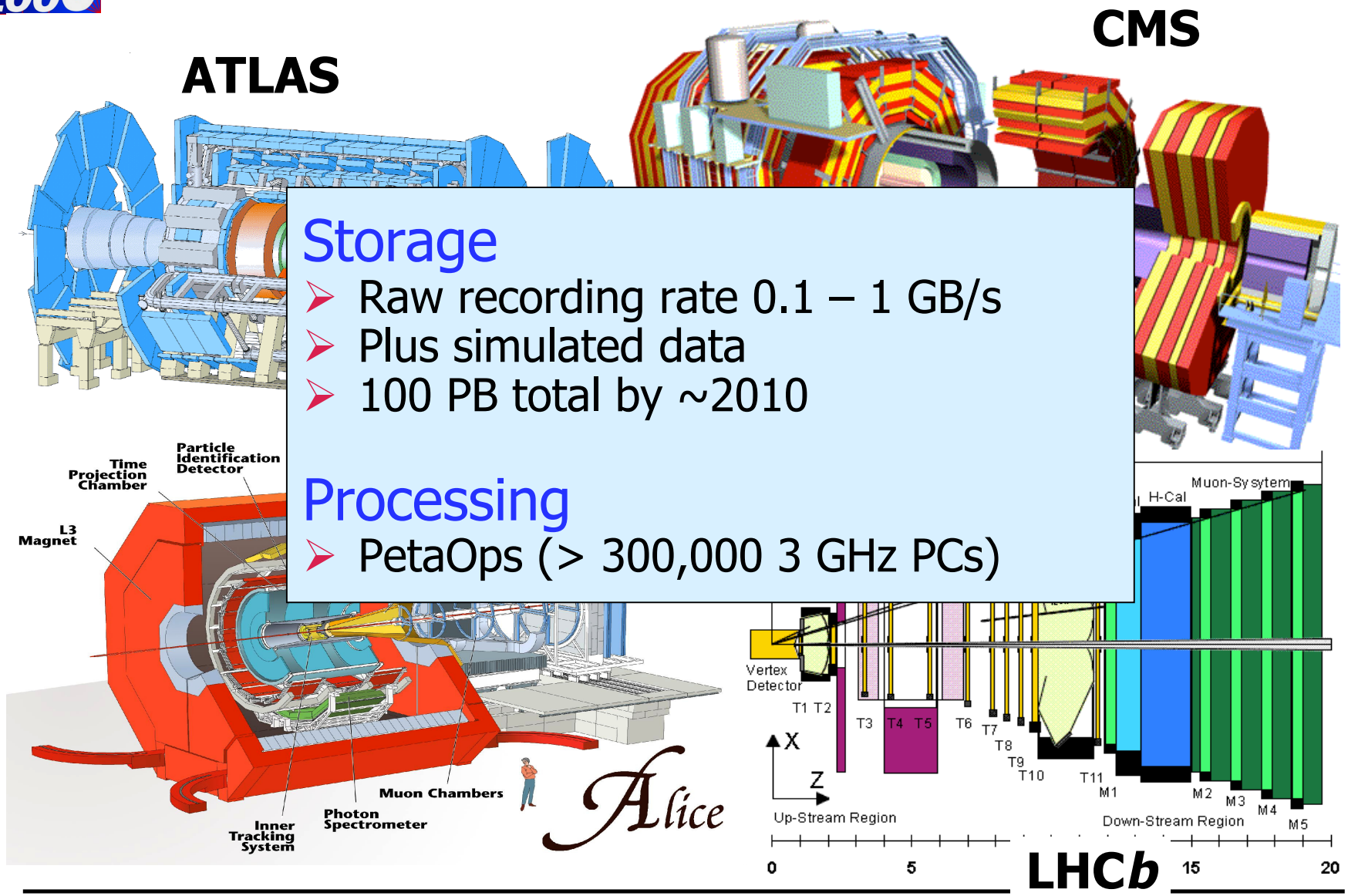


Smithsonian standard man

LHC Data Rates: Detector to Storage



LHC Data and CPU Requirements



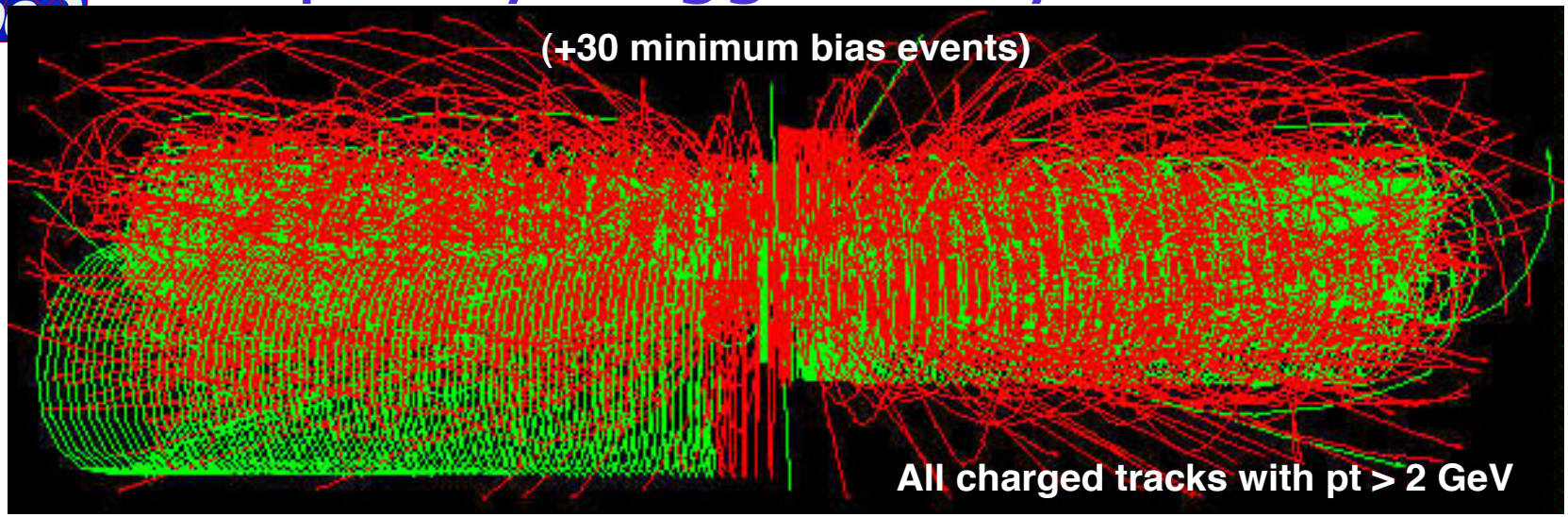
Storage

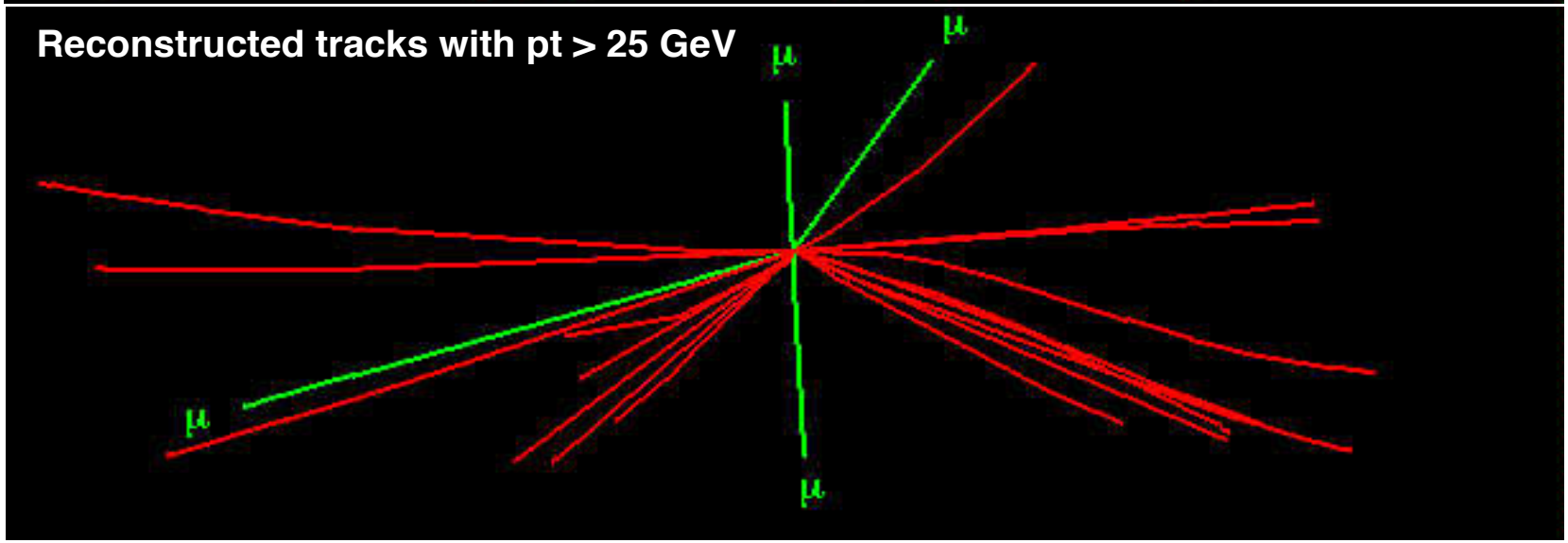
- Raw recording rate 0.1 – 1 GB/s
- Plus simulated data
- 100 PB total by ~2010

Processing

- PetaOps (> 300,000 3 GHz PCs)

Complexity: Higgs Decay into 4 muons

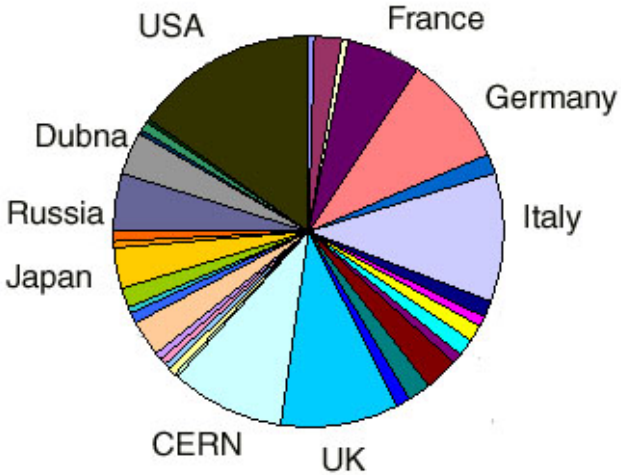




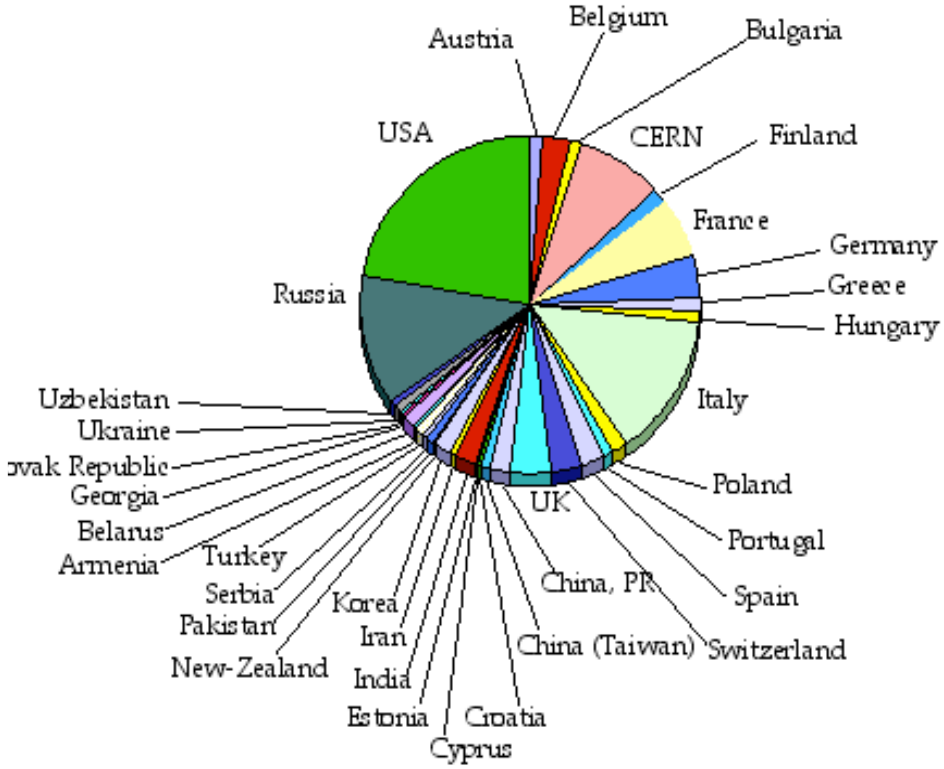
10^9 collisions/sec, selectivity: 1 in 10^{13}

LHC Global Collaborations

ATLAS



CMS



- **1000 – 4000 per experiment**
- **USA is 20 – 25% of total**

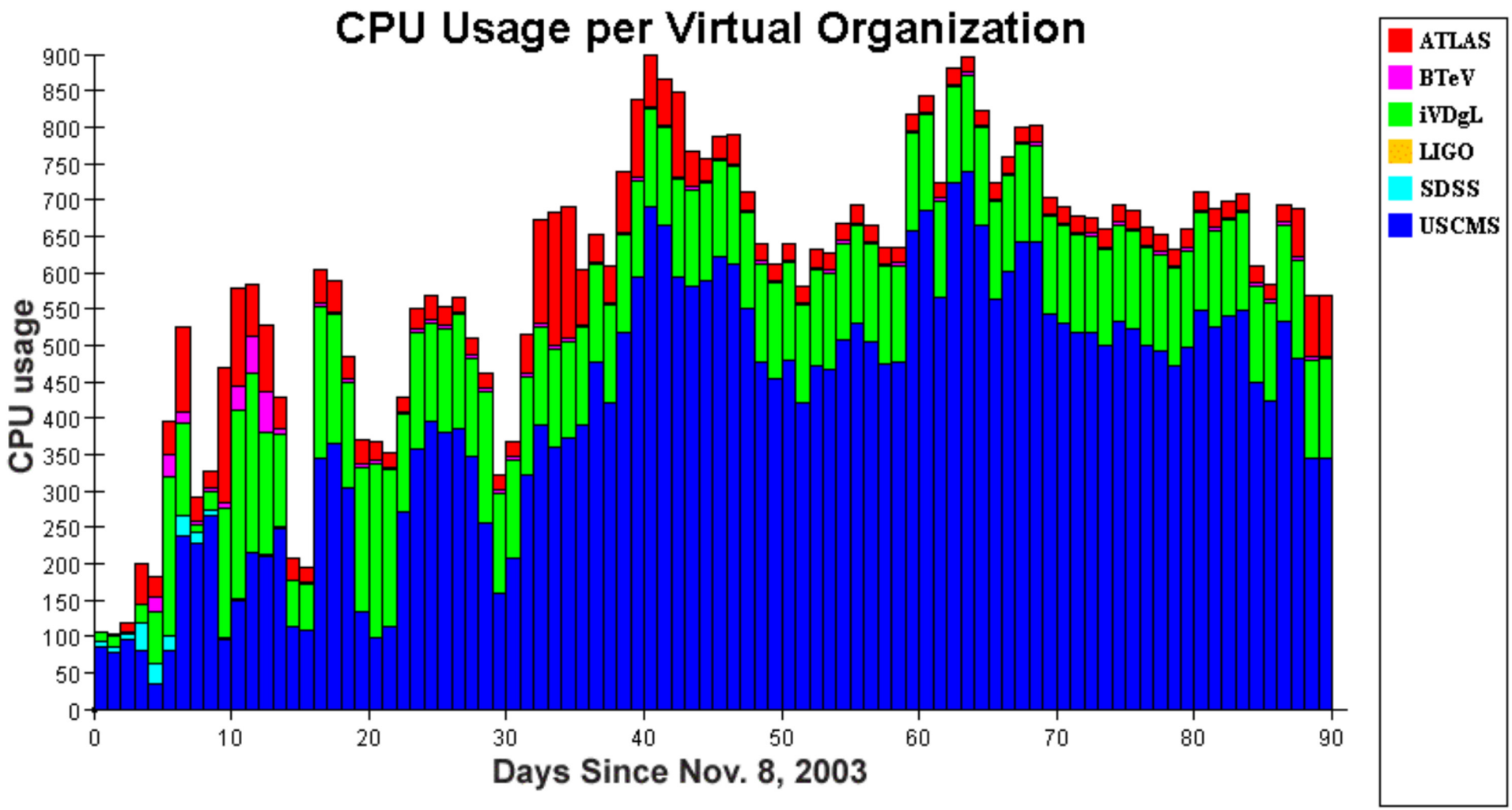
Universities and the Global LHC Data Grid Hierarchy

Grid2003: An Operational National Grid

- 28 sites: Universities + national labs (Fermilab, Brookhaven, Argonne, Berkeley)
- 2800 CPUs, 400–1300 jobs
- Running since October 2003
- HEP, LIGO, SDSS, Genomics

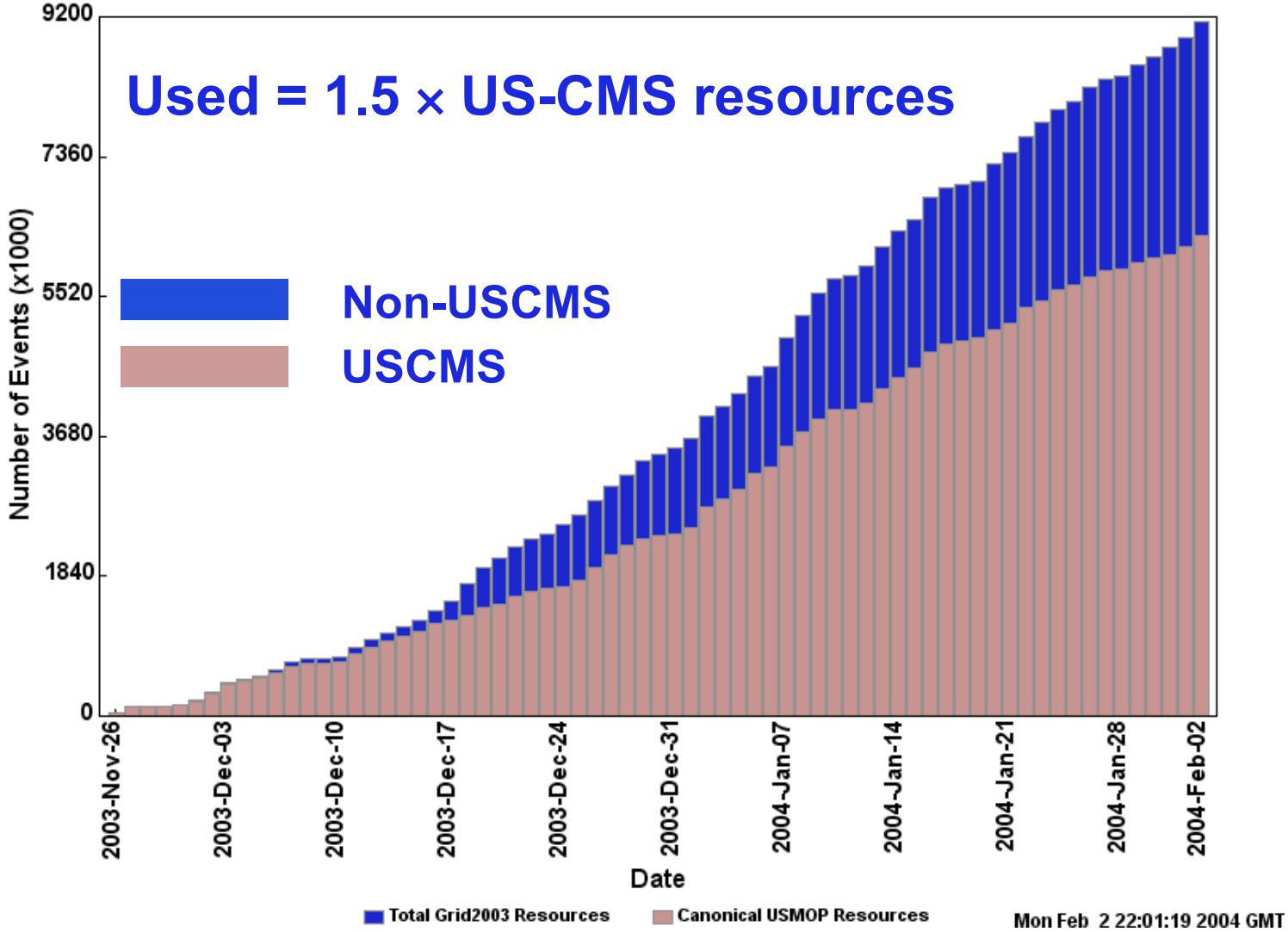
http://www.ivdgl.org/grid2003

Grid2003: Three Months Usage



Production Simulations on Grid2003

US-CMS Monte Carlo Simulation



Grid2003: A Necessary Step

- Learning how to operate a Grid
  - Add sites, recover from errors, provide info, update, test, etc.
  - Tools, services, procedures, documentation, organization
  - Security, account management, multiple organizations
- Coping with "large" scale
  - "Interesting" failure modes as scale increases
  - Dedicated operations team
- Developing strategies for future Grids
  - Delegation of responsibilities to conserve human resources
    - Project, Virtual Org., Grid service, site, application
  - Better services, documentation, packaging

- **Grid2003 experience critical for building "useful" Grids**
- **Frank discussion in "Grid2003 Project Lessons" doc**

Grid2003 Lessons (1)

- Need for financial investment
  - Grid projects: PPDG + GriPhyN + iVDGL: $35M
  - National labs: Fermilab, Brookhaven, Argonne, LBL
  - Other expts: LIGO, SDSS
- Critical role of collaboration
  - CS + 5 HEP + LIGO + SDSS + CS + biology
  - Makes possible large common effort
  - Collaboration sociology ~50% of issues
- Building "stuff": Prototypes ⇒ testbeds ⇒ production Grids
  - Build expertise, debug Grid software, develop Grid tools & services
- Need for "operations team"
  - Point of contact for problems, questions, outside inquiries
  - Resolution of problems by direct action, consultation w/ experts

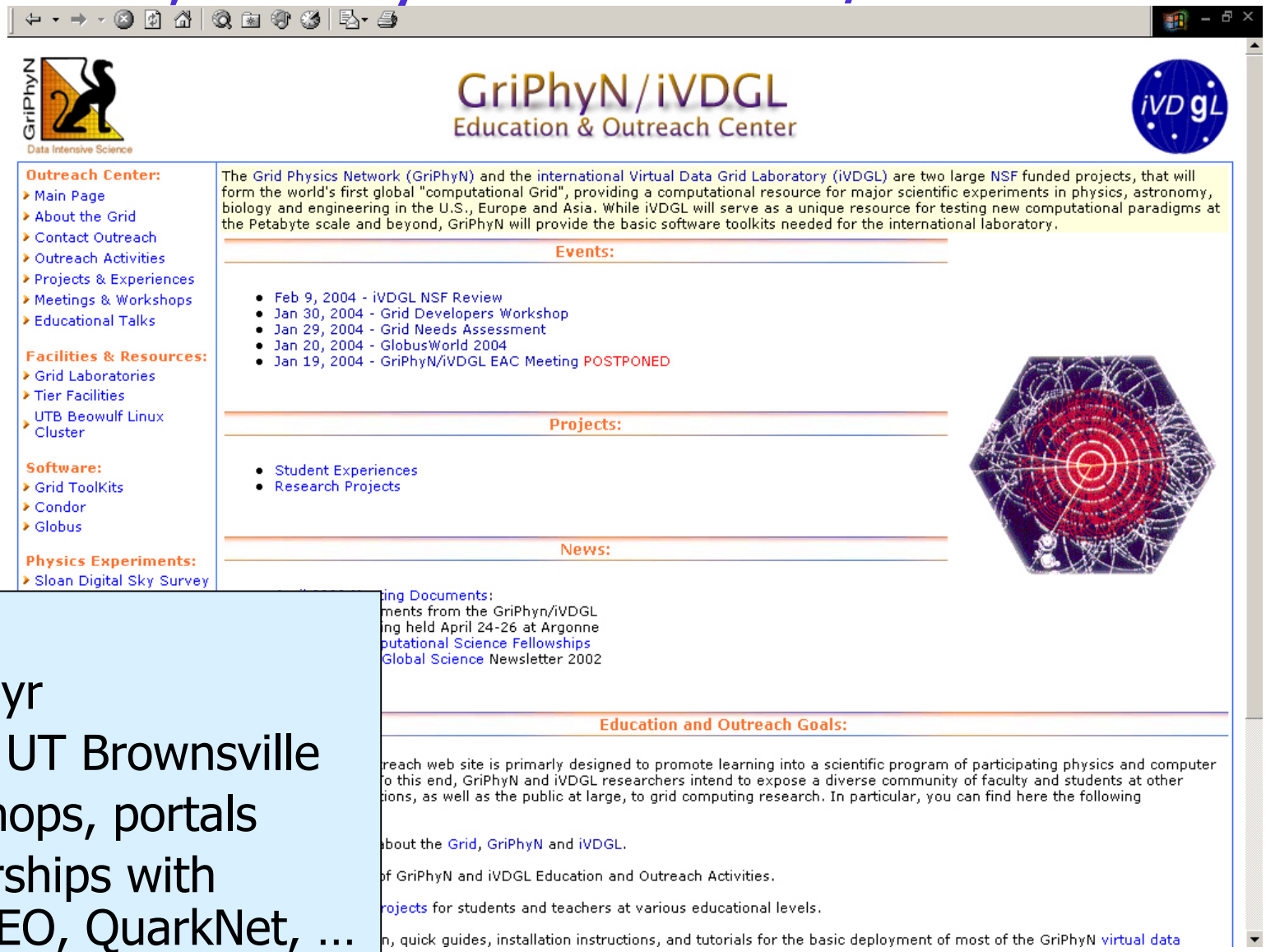
Grid2003 Lessons (2): Packaging

- VDT + Pacman: Installation and configuration
  - Simple installation, configuration of Grid tools + applications
  - Major advances over 14 VDT releases
- Why this is critical for us
  - Uniformity and validation across sites
  - Lowered barriers to participation ⇒ more sites!
  - Reduced FTE overhead & communication traffic
  - The next frontier: remote installation

U.S. Open Science Grid

- Goal: An integrated U.S. Grid infrastructure
  - Grid computing infrastructure to support US scientific efforts
  - CPU & storage resources from laboratories and universities
  - DOE and NSF partnership
- Getting there: OSG-1 (Grid3+), OSG-2, …
  - Series of releases ⇒ increasing functionality & scale
- Initial meetings
  - Sep. 17 @ NSF: Educators, scientists, etc.
  - Jan. 12 @ Fermilab: Public discussion, planning sessions
- Next steps
  - White paper to be expanded into roadmap
  - Presentation to funding agencies (June?)

iVDGL, GriPhyN Education / Outreach



Basics

- $200K/yr
- Led by UT Brownsville
- Workshops, portals
- Partnerships with CHEPREO, QuarkNet, …

CHEPREO: Center for High Energy Physics Research and Educational Outreach

Florida International University

- Physics Learning Center
- CMS Research
- iVDGL Grid Activities
- AMPATH network (S. America)

Funded September 2003
$4M initially (3 years)

CGWA : Center for Gravitational Wave Astronomy - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Search Favorites Media

Links Google LCG Customize Links ECS Audio ESNet Audio2 Geographic Resources Amazon Media PDG UF Directory HEP Logos Reason

Address C:\avery\public_html\griphyn\talks\TMP817afvuxbz.htm Go

GriPhyN

Data Intensive Science

CWGA

Center for Gravitational Wave Astronomy

iVDgL

25-30 students

GRID Summer Workshop
June 21-25 2004

Center for Gravitational Wave Astronomy
The University of Texas at Brownsville

Grid Physics Network (GriPhyn)

International Virtual Data Grid Laboratory (iVDGL)

The Center for Gravitational Wave Astronomy, a NASA University Research Center at the University of Texas at Brownsville, the Grid Physics Network (GriPhyn) and the International Virtual Data Grid Laboratory (iVDGL) will sponsor a one-week intensive summer workshop in Grid Computing and Scientific Data Analysis.

This school would be held on **South Padre Island**, located on the Gulf of Mexico, twenty miles from downtown Brownsville.

- General Information
- Curriculum
- Summer School Location

http://cgwa.phys.utb.edu/ My Computer

Grids and the Digital Divide
Rio de Janeiro, Feb. 16-20, 2004

HEPGRID AND DIGITAL DIVIDE WORKSHOP

ICFA SCIC

NEWS:

Bulletin: ONE TWO
WELCOME BULLETIN
General Information
Registration
Travel Information
Hotel Registration
Participant List
How to Get UERJ/Hotel
Computer Accounts
Useful Phone Numbers
Program
Contact us:
Secretariat
Chairmen

Background

- World Summit on Information Society
- HEP Standing Committee on Inter-regional Connectivity (SCIC)

Themes

- Global collaborations, Grids and addressing the Digital Divide

Next meeting: 2005 (Korea?)

http://www.uerj.br/lishep2004