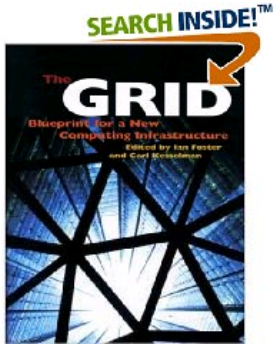
Living the Grid Dream – A Report from the Trenches

Miron Livny
Computer Sciences Department
University of Wisconsin-Madison
miron@cs.wisc.edu

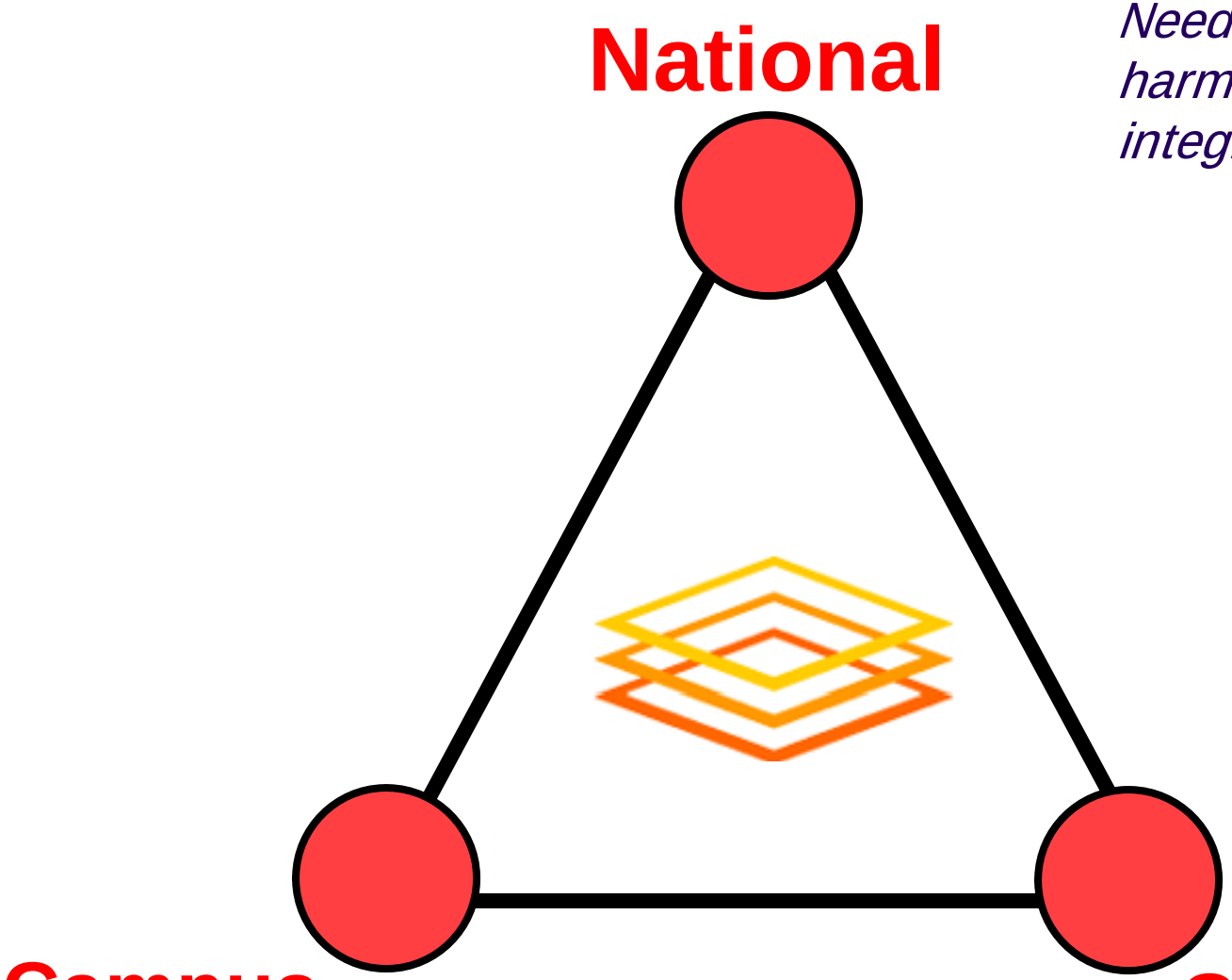
The Grid: Blueprint for a New Computing Infrastructure

Edited by Ian Foster and Carl Kesselman
July 1998, 701 pages.



The grid promises to fundamentally change the way we think about and use computing. This infrastructure will connect multiple regional and national computational **grids**, creating a universal source of **pervasive and dependable** computing power that supports dramatically new classes of applications.

The Three Cornerstones

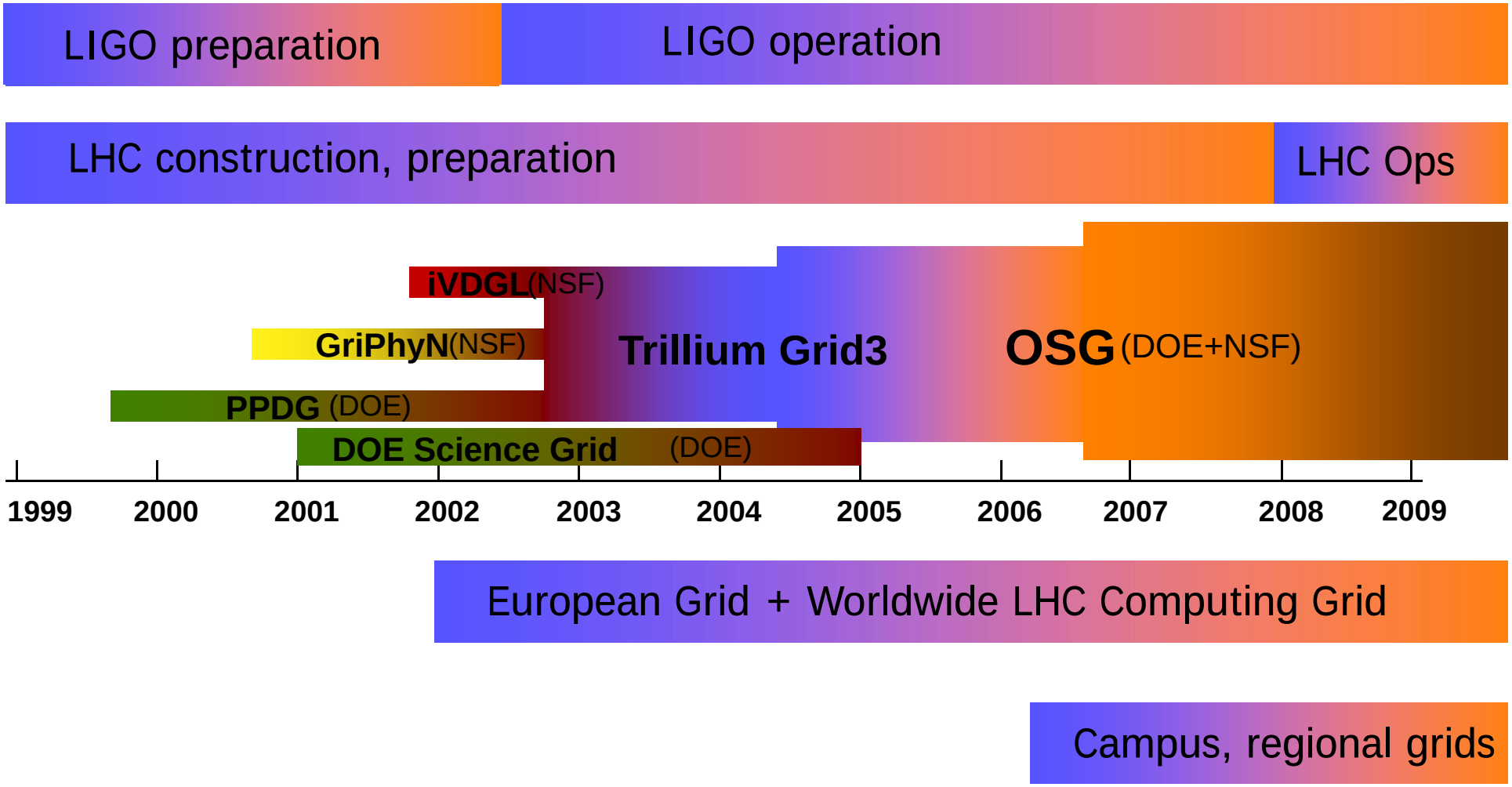


Need to be harmonized into a well integrated whole.

National:

The Open Science Grid (OSG)

The Evolution of the OSG



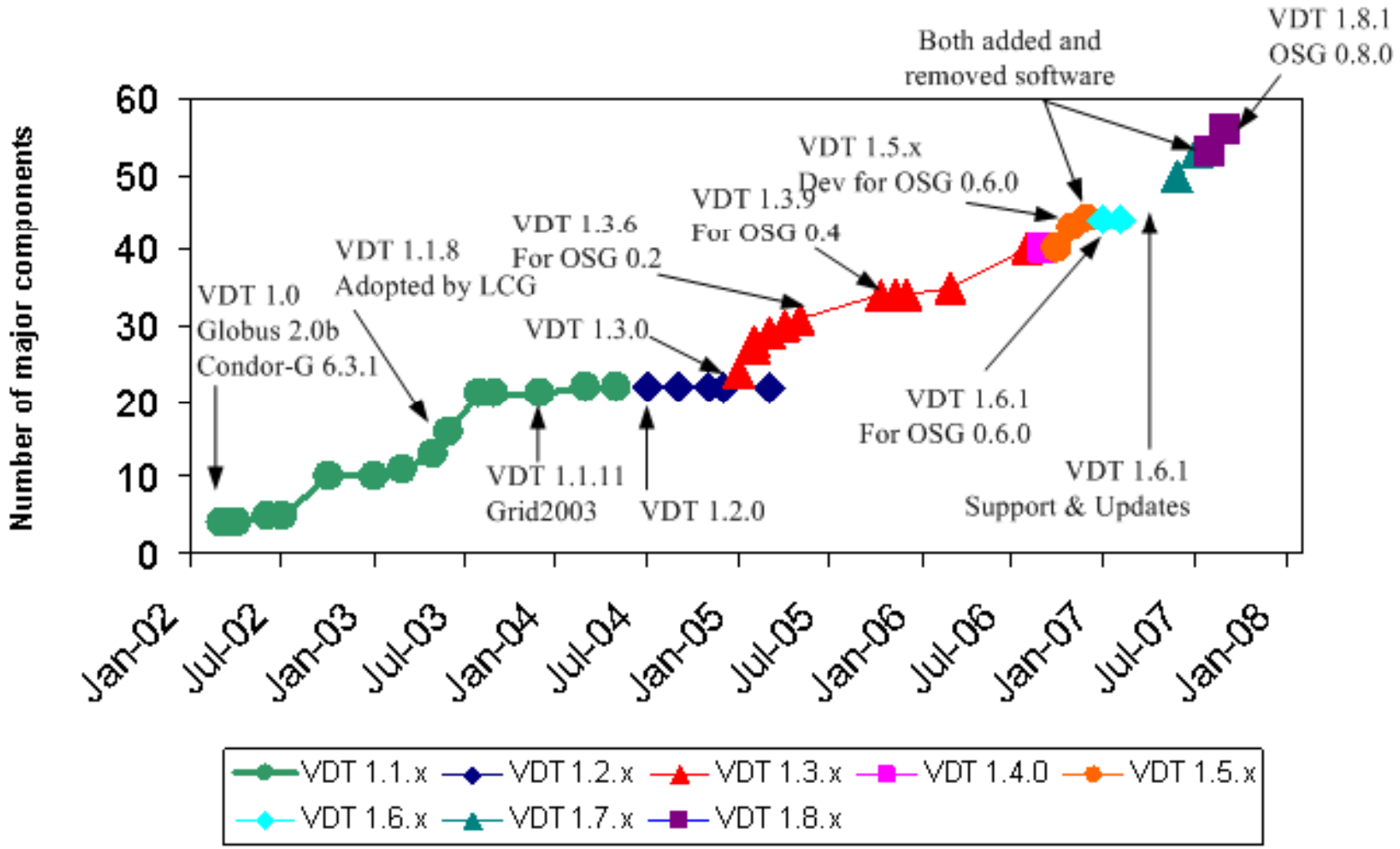
The OSG Consortium

- **> 20 Scientific Virtual Organizations:** US-LHC, LIGO, SBGrid etc.
- **> 25 Resource Providers:** DOE National Labs, University Facilities etc.
  - 10 Storage focused resources
- **>10 Software Providers (including External Projects):** Condor, Globus, Storage Resource Manager, Internet2, ESNET, CEDPS, etc.
- **>4 Partners - Ex-Officio:** EGEE, TeraGrid, NWICG, etc.

Some (weekly) OSG Numbers

- ~500 unique Distinguished Names (DNs)
- ~1M jobs
- ~1.5M CPU hours
- ~15 Virtual Organizations
- ~60 contributing sites

How much software?

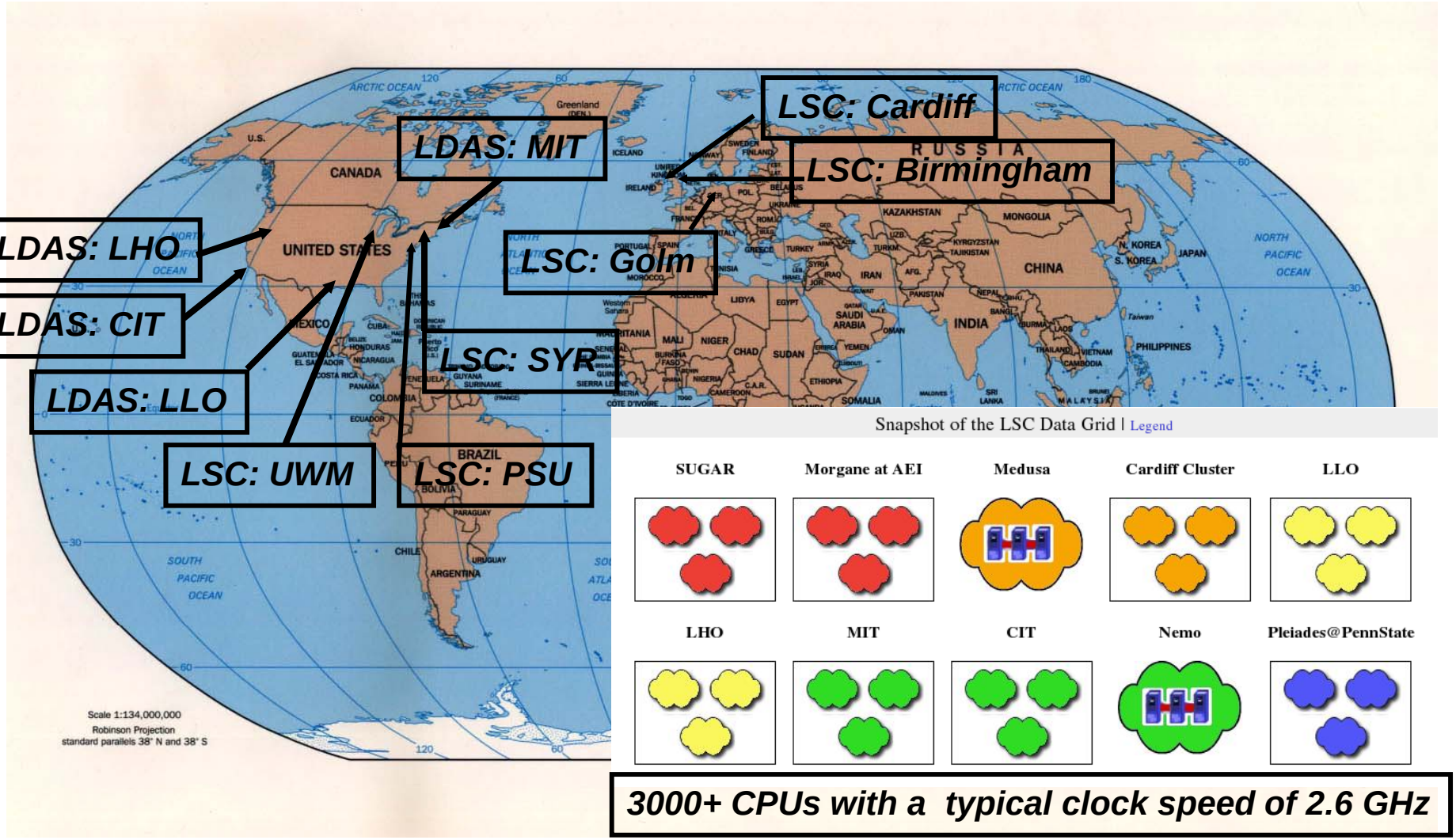


Community:

The LIGO Data Grid (LDG)

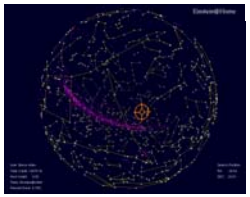
Worldwide Interferometers

Worldwide Data Analysis using The LIGO Data Grid (LDG)



The LIGO Data Grid

- Users
  - 500+ scientist in the LIGO Scientific Collaboration
  - 200+ doing data analysis on the LIGO Data Grid
- Grid middleware
  - LDG Client/Server
    - Virtual Data Toolkit (VDT)
      - Globus Toolkit
      - GSI and X.509 certificates
      - pyGlobus, tclGlobus, Pegasus, etc…
    - In house packages
  - Glue: LSC Data Location & Pipeline Tools
  - LDR: LSC Lightweight Data Replication
    - GridFTP for moving data and files
    - Replica Location Service (RLS)
  - Onaysis: LSC Online Analysis System
- High throughput computing
  - **Condor** for most analyses
  - BOINC for Einstein@Home
- LSC Analysis Software
  - LAL, Matapps,
  - DMT, etc..

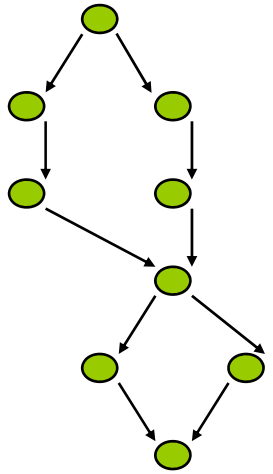


Use of Condor by the LIGO Scientific Collaboration

Condor
High Throughput Computing

• Condor handles 10's of millions of jobs per year running on the LDG, and up to 500k jobs per DAG.

• Condor standard universe check pointing widely used, saving us from having to manage this.



• At Caltech, 30 million jobs processed using 22.8 million CPU hrs. on 1324 CPUs in last 30 months.

• For example, to search 1 yr. of data for GWs from the inspiral of binary neutron star and black hole systems takes ~2 million jobs, and months to run on several thousand ~2.6 GHz nodes.

Campus:

The Boiler Grid

Rosen Center for Advanced Computing

BoilerGrid

- Purdue Condor Grid (BoilerGrid)
  - Comprised of Linux HPC clusters, student labs, machines from academic department, and Purdue regional campuses
- **8900** batch slots today..
- **14,000** batch slots in a few weeks

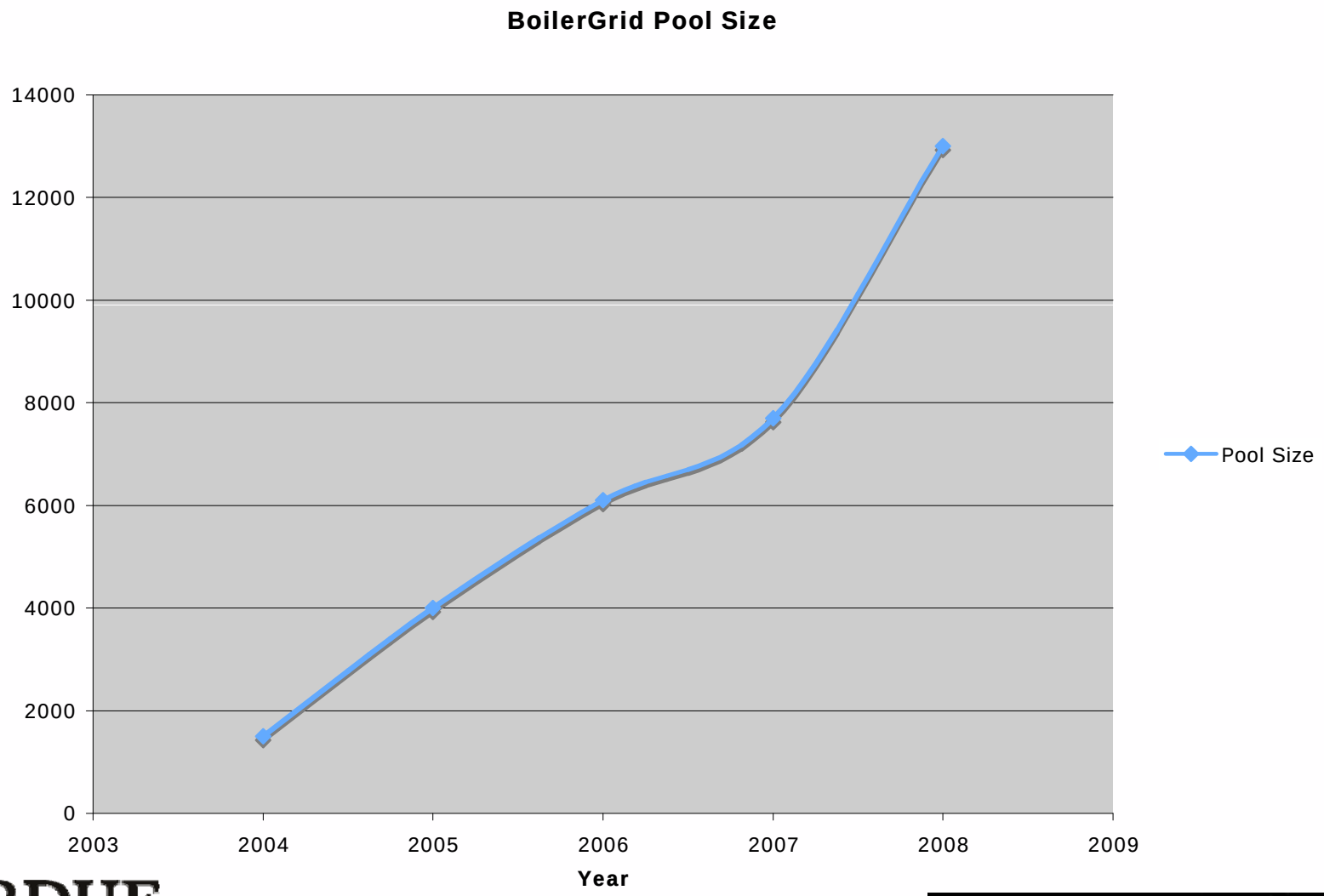
- 2007 - Delivered over 10 million CPU-hours to high-throughput science to Purdue and national community through Open Science Grid and TeraGrid

PURDUE UNIVERSITY

BOILERGRID

Rosen Center for Advanced Computing

BoilerGrid - Growth

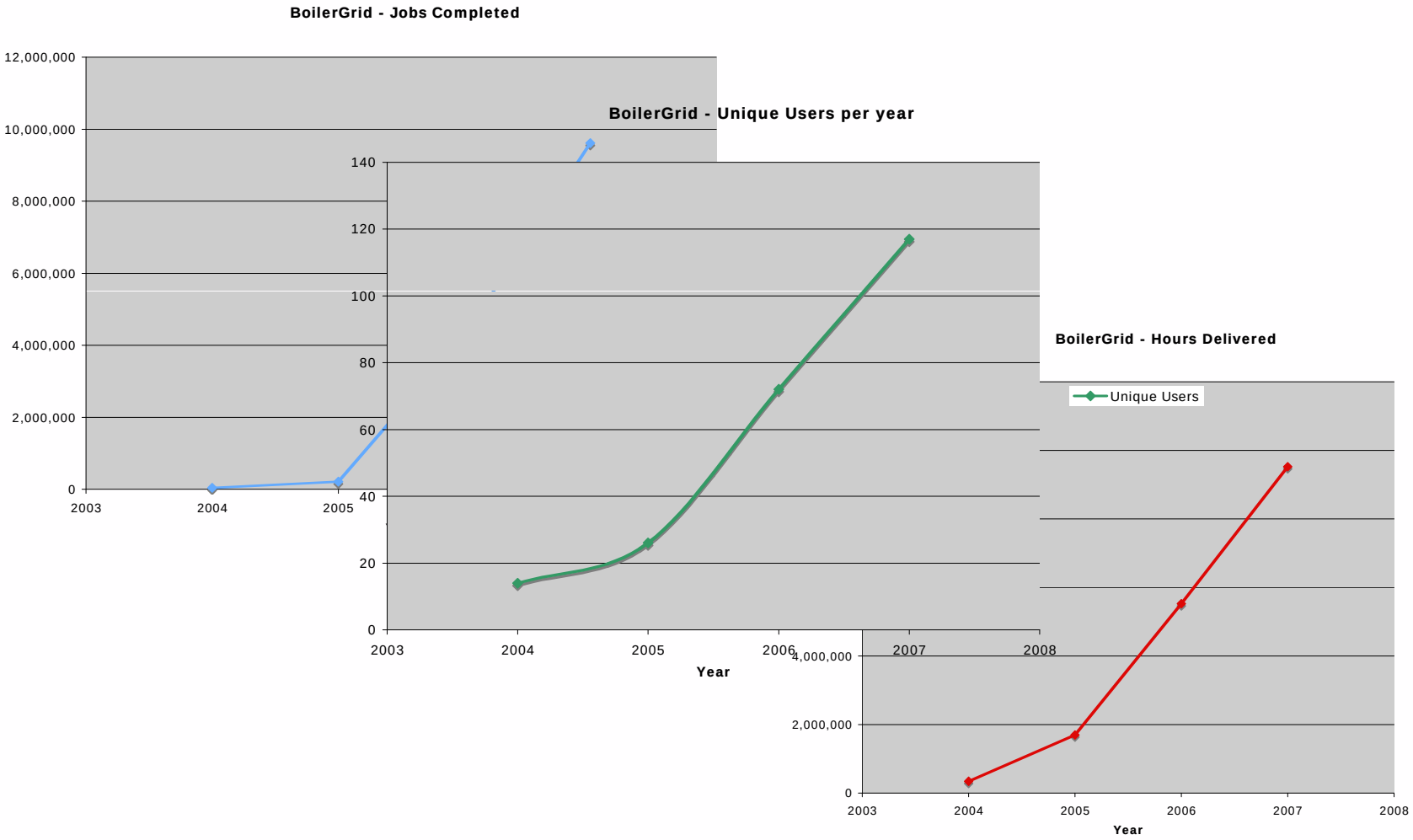


PURDUE UNIVERSITY

BOILERGRID

Rosen Center for Advanced Computing

BoilerGrid - Results



PURDUE UNIVERSITY

BOILERGRID

A partial list of the

many things

we do not know

and the **many** tools

we do not have

(operational) Security

- Do we have the "right" model?
- Do we agree on who is responsible for what?
- Do we understand the threats?
- Do we know how to build secure software?
- Do we follow established processes?
- Do we collect and keep the information we need to deal with security incidences?
- Do we know who is (can be) legally liable for misuse of resources and services?

Operation

- Missing software tools for monitoring and managing (including accounting) the services and resources in the infrastructure
- Limited experience in organizing and managing a distributed infrastructure (not only computers but also people) with services that have high availability requirements
- Un (ill) defined metrics to quantify progress, success and cost

Software Life Cycle

- Stability of the software supply-chain
- Functionality at a large scale and wide area environment (limited access to at-scale testing resources)
- Evolution of the software stack in the context of a distributed production (hot) environment (incremental updates and configuration management)
- Management of software development projects (time lines and effort estimates)

Solutions, Answers and Contributions are welcome!