the globus alliance
www.globus.org

USC
SCHOOL OF ENGINEERING

Current and Future Grid Infrastructure

Ewa Deelman
Center for Grid Technologies
USC Information Sciences Institute
Globus Alliance

Some material provided by Y. Gil, C. Kesselman and S. Tuecke
Some material obtained from the iVDGL NSF review

The Globus Alliance

- A group of people with a common mission:
  - *"Make Grid computing an everyday reality."*
- Argonne/U.Chicago, USC-ISI, EPCC, & KTH-PDC
  - Led by Ian Foster (Argonne), Carl Kesselman (ISI), Malcolm Atkinson (EPCC), Lennart Johnsson (PDC)
  - Includes researchers, software developers, software architects & designers, systems engineers, & others
  - Collaborations (or at least acquaintances) with most Grid activities in the world
- All activities contribute to our common mission
  - Research
  - Software development (prototypes, reference implns)
    - Globus Toolkit
  - Application and Infrastructure consulting

the globus alliance
www.globus.org

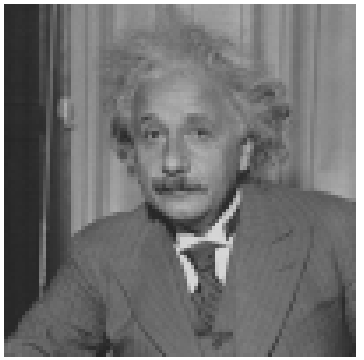
Outline

- Grid technologies in Science and Industry
- Developments in Grid Services and Globus Toolkit
  - Convergence with web services
  - The Grid as a resource management problem
- Research activities within the Globus Alliance, a flavor of things to come
  - Pegasus (Planning for Execution in Grids)
    - Large-scale Workflow management on the Grid
  - The Metadata Catalog Service
    - Metadata management on the Grid
- Conclusions

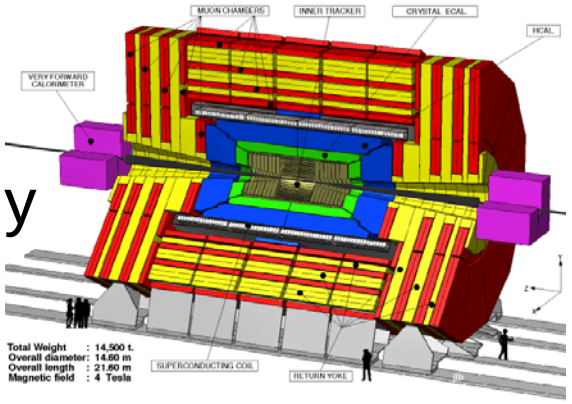
the globus alliance
www.globus.org

Revolution in Science

- Pre-Internet
  - Theorize &/or experiment, alone or in small teams; publish paper



- Post-Internet
  - Construct and mine large databases of observational or simulation data
  - Develop simulations & analyses
  - Access specialized devices remotely
  - Exchange information within distributed multidisciplinary teams



the globus alliance
www.globus.org

Problem Solving in the 21st Century

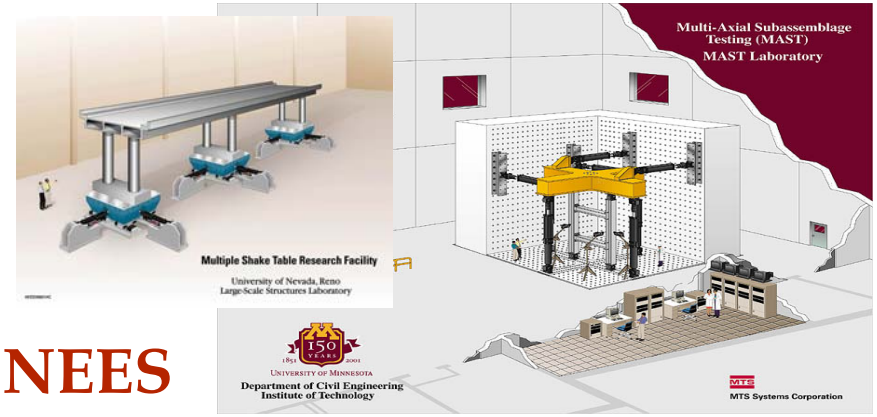
- Teams organized around common goals
  - Communities: “Virtual organizations”
- With diverse membership & capabilities
  - Heterogeneity is a strength not a weakness
- And geographic and political distribution
  - No location/organization possesses all required skills and resources
- Must adapt as a function of the situation
  - Adjust membership, reallocate responsibilities, renegotiate resources

The Grid: enabling technology

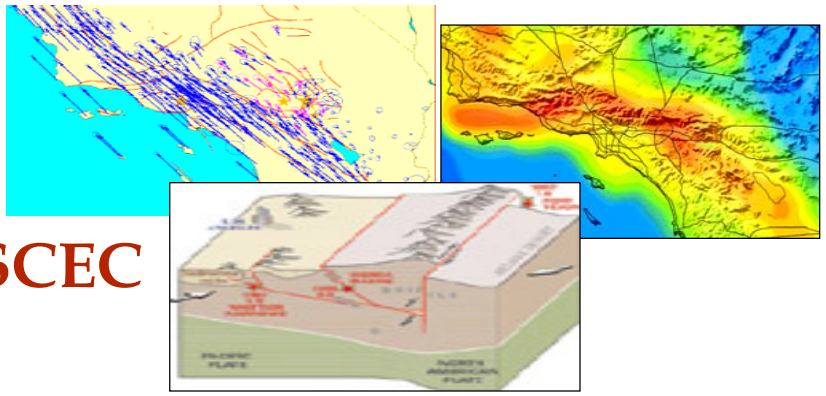
- Infrastructure ("middleware" & "services") for establishing, managing, and evolving multi-organizational federations
  - Dynamic, autonomous, domain independent
  - On-demand, ubiquitous access to computing, data, and services
- Mechanisms for creating and managing workflows within such federations
  - New capabilities constructed dynamically and transparently from distributed services
  - Service-oriented, virtualization

the globus alliance
www.globus.org

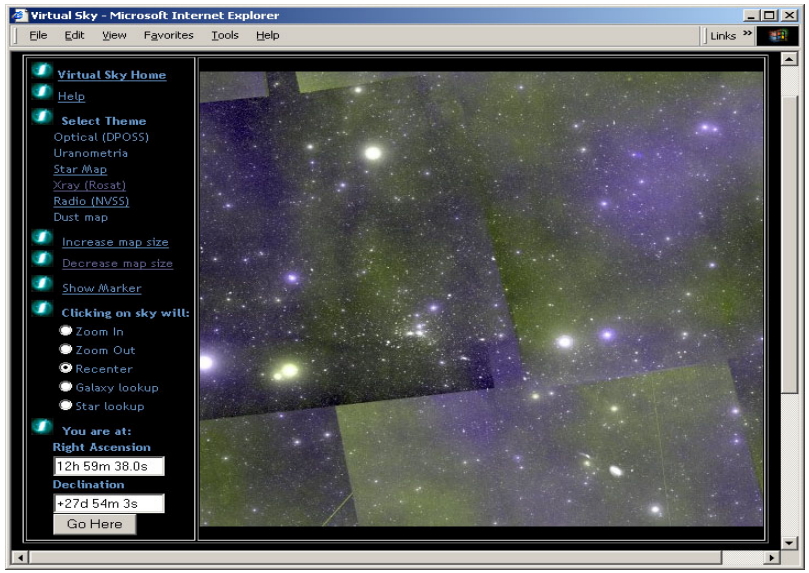
Science is Adopting Grid Technology:



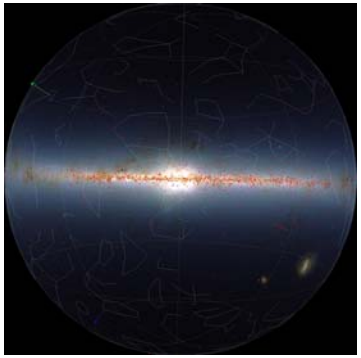
NEES



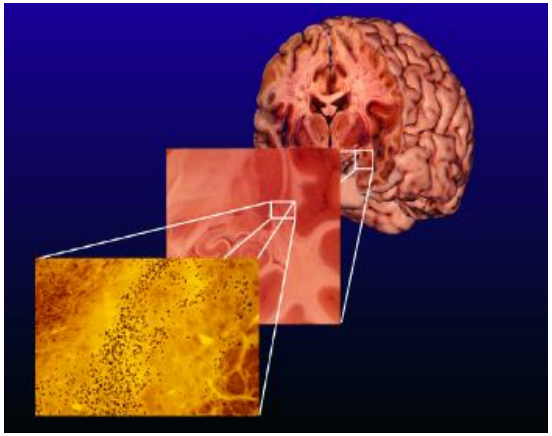
SCEC



NVO



GriPhyN



BIRN

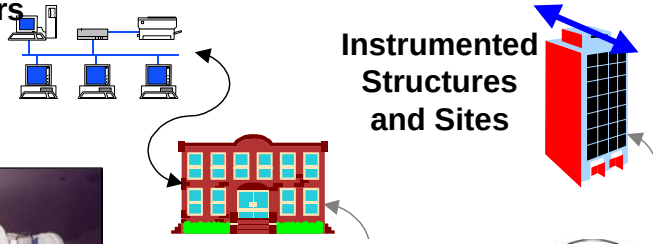
the globus alliance
www.globus.org

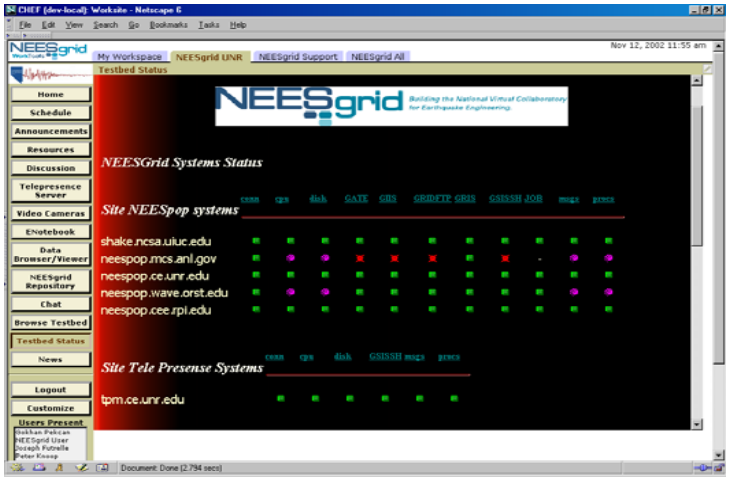
NEESgrid Earthquake Engineering Collaboratory

Focus on remote experimentation

U.Nevada Reno

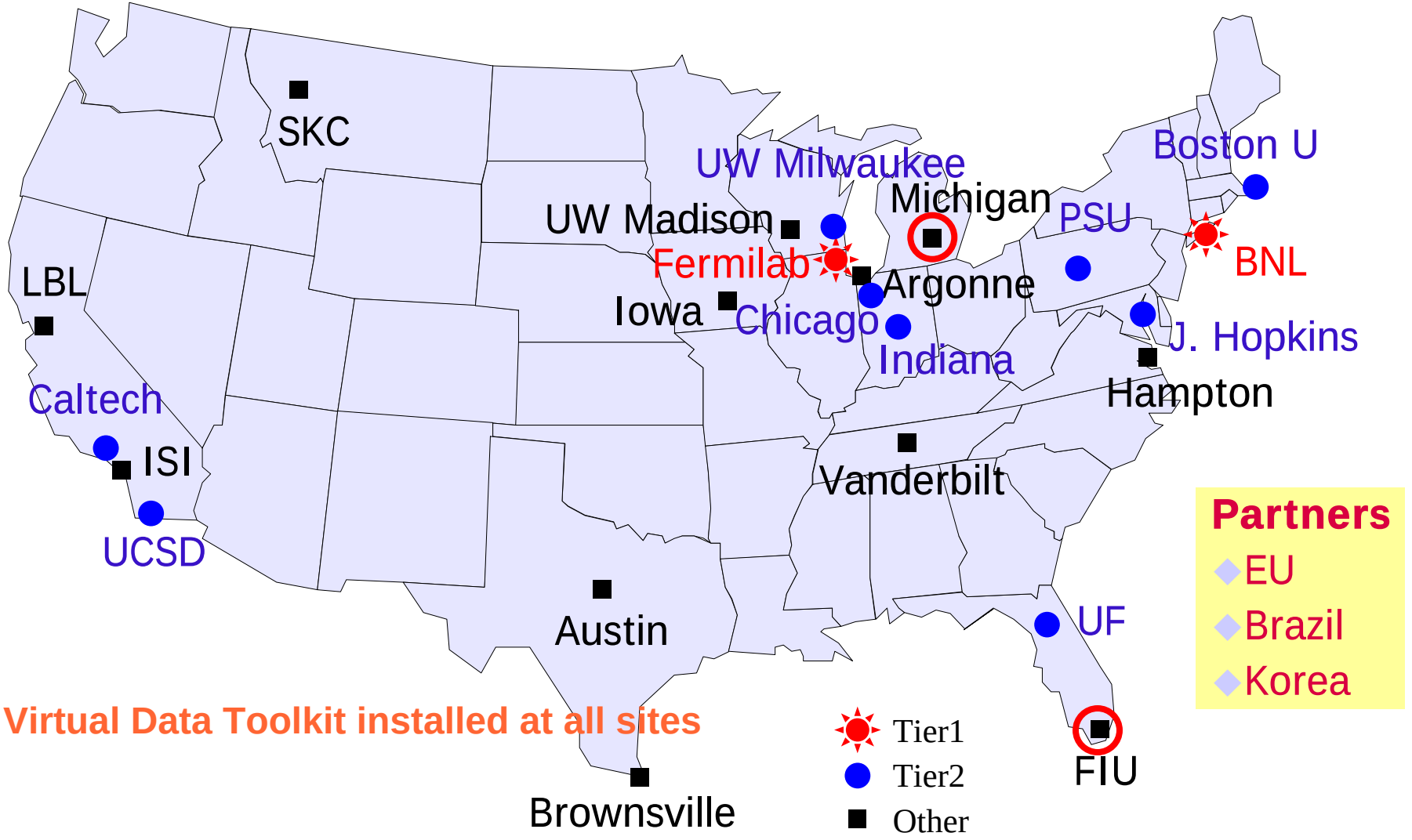
www.neesgrid.org





the globus alliance
www.globus.org

iVDGL Sites (February 2004)
iVDGL supports GriPhyN and PPDG



Partners
- EU
- Brazil
- Korea

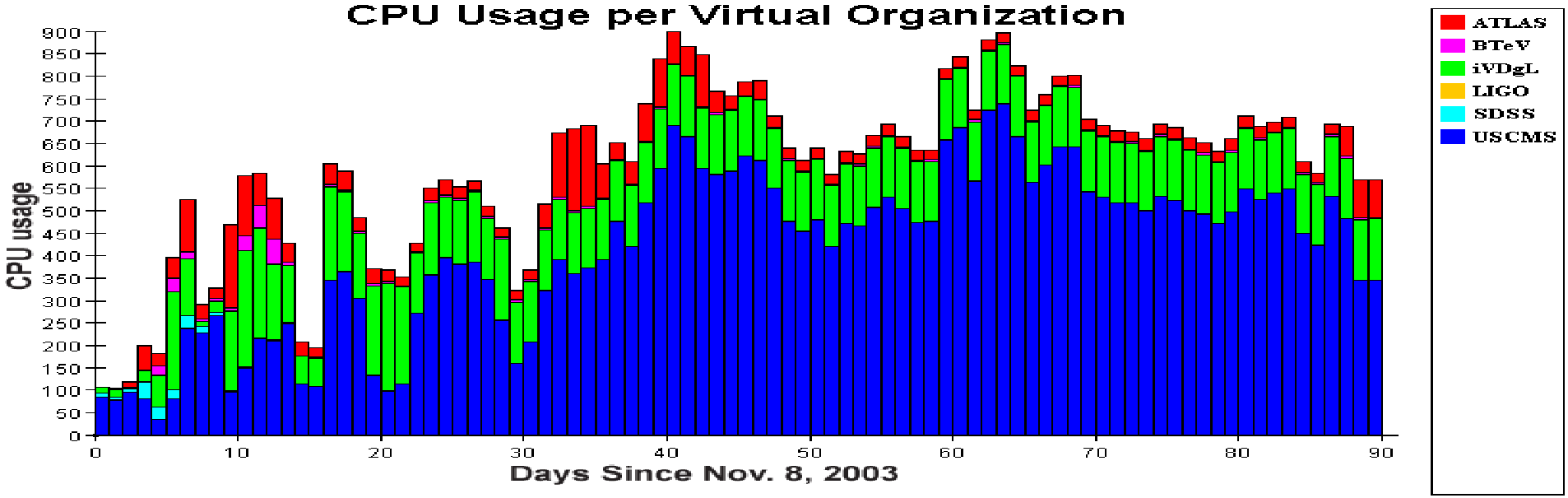
Virtual Data Toolkit installed at all sites

Slide by Paul Avery, iVDGL NSF review 2004

the globus alliance
www.globus.org

Grid2003: Three Months Usage

- Hardware resources
  - Total of 2762 CPUs
  - Total of 27 sites
    - 27 administrative domains with local policies in effect
    - All across US and Korea
- Running jobs
  - Peak number of jobs 1100



Slide by Paul Avery, iVDGL NSF review 2004

Grid the basic scientific tool

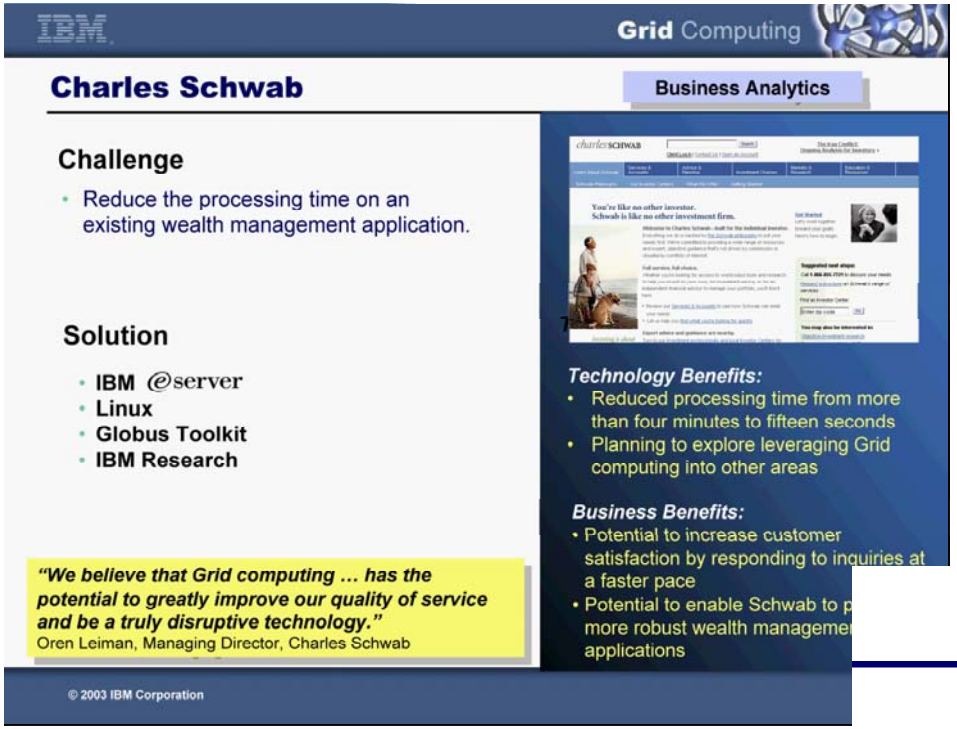
- Proposed infrastructure for University of Texas (provided by IBM)
  - Largest university grid-computing project in the nation
  - Will support more than 50,000 students and 20,000 faculty and staff members
- Grid technology will be used for
  - Simulations
  - data sharing
  - data-intensive calculations
  - genomics investigations; climate modeling; petroleum exploration; and environmental remediation.

eScience is the Initial Motivator …

- New research approaches based on
  - Complex analysis of huge quantities of data
  - Interdisciplinary collaboration
  - Large-scale simulation
  - Smart instrumentation
- Enabled by an infrastructure based on the Globus Toolkit that enables access to, and integration of, resources & services without regard for location

… but eBusiness is catching up rapidly

Industry Adopts Grid Technology



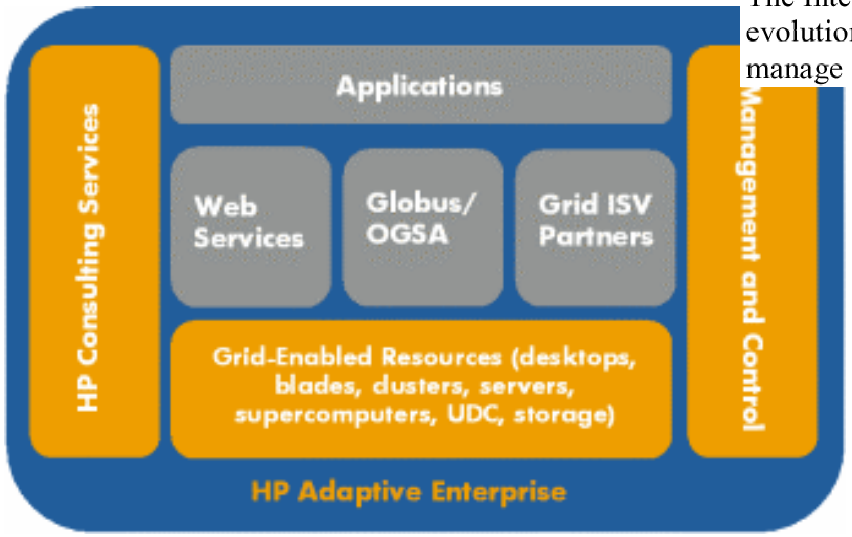
I.B.M. Making a Commitment to Next Phase of the Internet

By STEVE LOHR — The New York Times

I.B.M. is announcing today a new initiative to support and exploit a technology known as grid computing, which the company and much of the computer research community say is the

Merrill Lynch

DataSynapse



Globus Grid Computing—the Next Internet

by John Roy/Steve Milunovich

The Internet was first a network and is now a communications platform. The next evolutionary step could be to a platform for distributed computing. This ability to manage applications and share data over the network is called "grid computing."


COVER STORY

ORACLE DATABASE 10g

THE WORLD'S FIRST

SELF-MANAGING, GRID-READY DATABASE ARRIVES

Oracle's new self-managing database increases performance and availability while enabling commercial grid computing.

BY KELLI WISETH

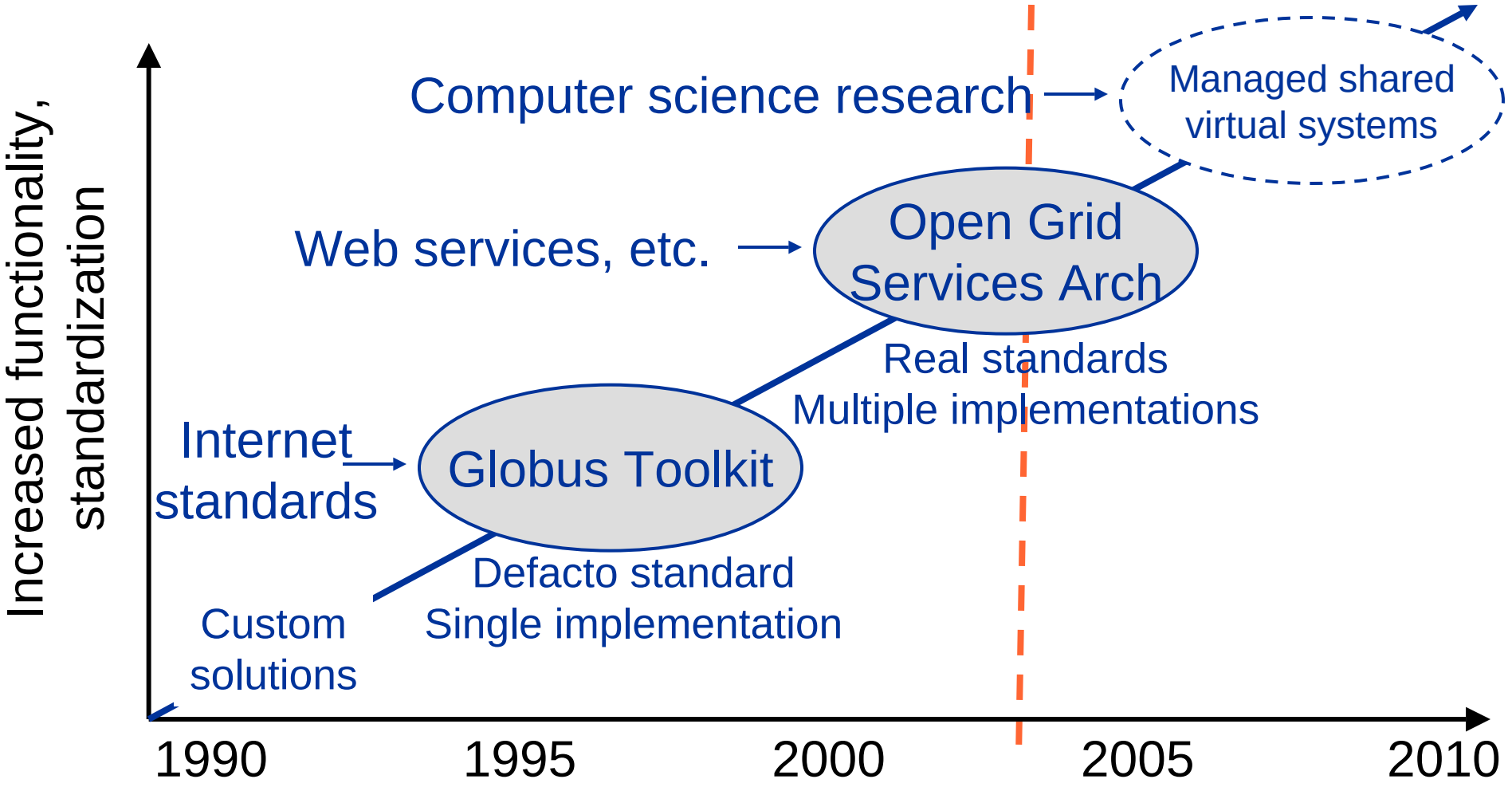
Ewa Deelman, USC Information Scienc

Grid Solutions in Business Environments

- Financial markets
  - Better market forecasts, risk analysis of stock portfolios
- Life sciences
  - Better drugs faster
- Automotive and aerospace
  - Better collaborative designs in shorter amount of time
- Government
  - Facilitate information sharing and analysis
- Benefits of Grid Solutions:
  - Increase the computation power delivered to businesses
  - Exploit available, underutilized compute and storage resources
  - Enable data sharing and distributed workflow across partners
  - Help with collaborative design

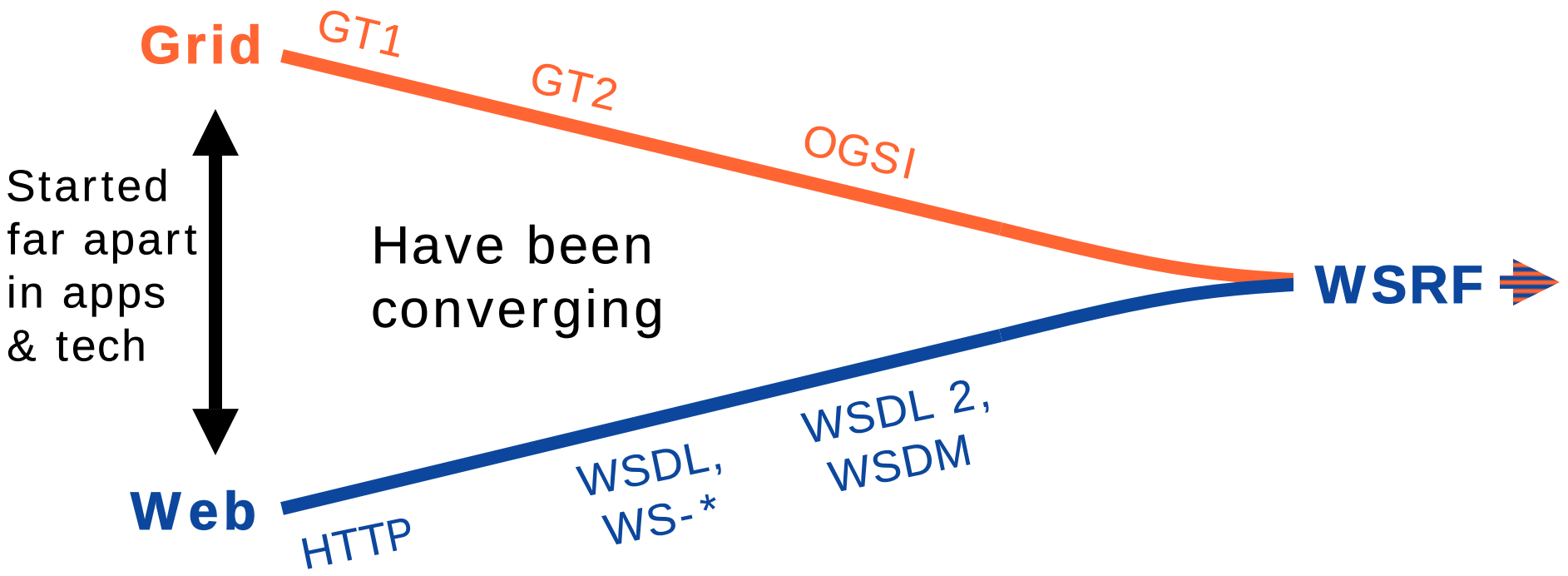
the globus alliance
www.globus.org

Progress through the Emergence of Open Grid Standards



the globus alliance
www.globus.org

Convergence of Grid and Web Services



The definition of WSRF means that Grid and Web communities can move forward on a common base

the globus alliance
www.globus.org

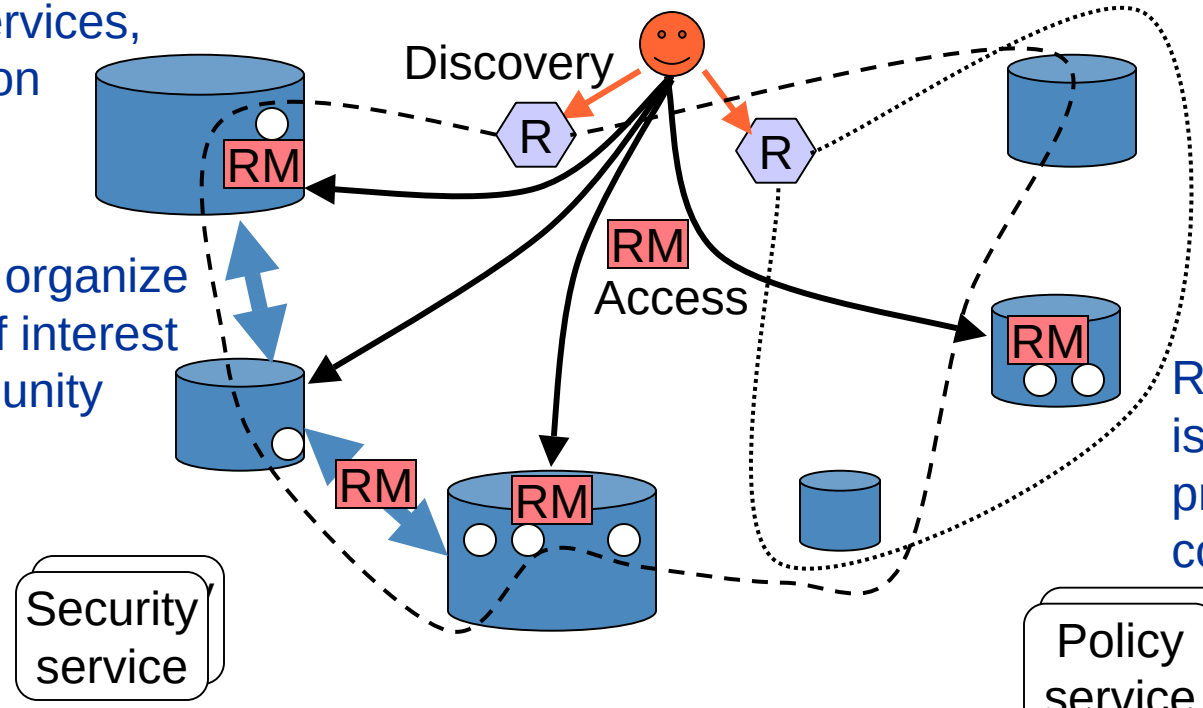
The Grid is about Resource Integration

Many sources of data, services, computation

Registries organize services of interest to a community

Security & policy must underlie access & management decisions

Resource management is needed to ensure progress & arbitrate competing demands



Data integration activities may require access to, & exploration/analysis of, data at many locations

Exploration & analysis may involve complex, multi-step workflows

the globus alliance
www.globus.org

How to Build a Grid? Guiding Principles

Grid definition implies that we must be able to

- Talk about resources in uniform ways
  - Ubiquitous Service Oriented Architecture base for modeling resources
- Govern the behavior resources that are not under our control &/or oversubscribed
  - Negotiate agreements among entities
- Exploit commonality across resources so as to reduce integration complexity
  - Factor across multiple resource types

Grids Needs Support for Transient Stateful Services

- Web services address discovery & invocation of persistent services
  - Lack dynamic behavior
- In Grids, must also support transient services, created/destroyed dynamically
  - Interfaces to the states of distributed activities
  - E.g. workflow, video conf., dist. data analysis
- Significant implications for how services are managed, named, discovered, and used
  - In fact, much of our work is concerned with the management of services

Key concepts in OGSI and WSRF

Entity: in OGSI---service, in WSRF---resource

- Naming
  - Every entity has a unique name,
- Lifecycle (basis for fault resilient state management)
  - Entities are dynamically created
  - Entities are destroyed explicitly or via soft state
- Information model (basis for introspection, monitoring, discovery)
  - Attributes associated with entities
  - Operations for querying and setting the attributes
  - Asynchronous notification of changes to attribute set
- Service Groups (basis for registries & collective svcs)
  - Group membership rules & membership management

the globus alliance
www.globus.org

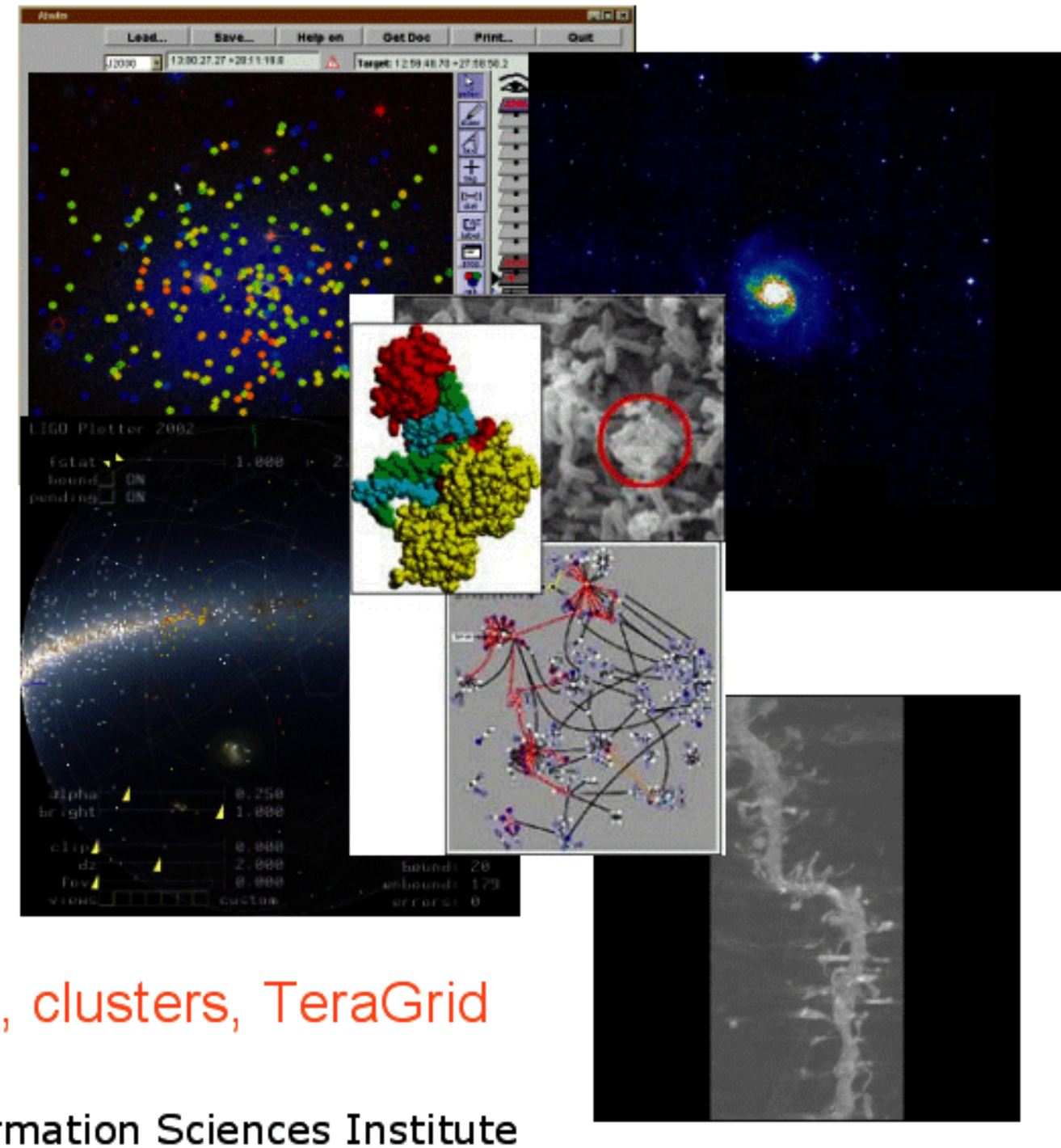
Grid Futures

- Build on standard infrastructure (WSRF)
- Resource management (Service level agreements)
- Large-scale VO management, policy management
- Tele-instrumentation
- Workflow composition, management and planning
  - Pegasus
- High-level data services: metadata, ontologies
  - MCS

Pegasus: scientific workflow management

(work with Y. Gil and dozens of scientists)

- Pulsar search for gravitational-wave physics (NSF-funded GriPhyN)
- Galaxy morphology (NSF-funded NVO) and Montage (NASA)
- Tomography for neural structure reconstruction (NIH)
- High-energy physics – Compact Muon Solenoid (GriPhyN)
- Gene sequencing (PACI Data Quest)



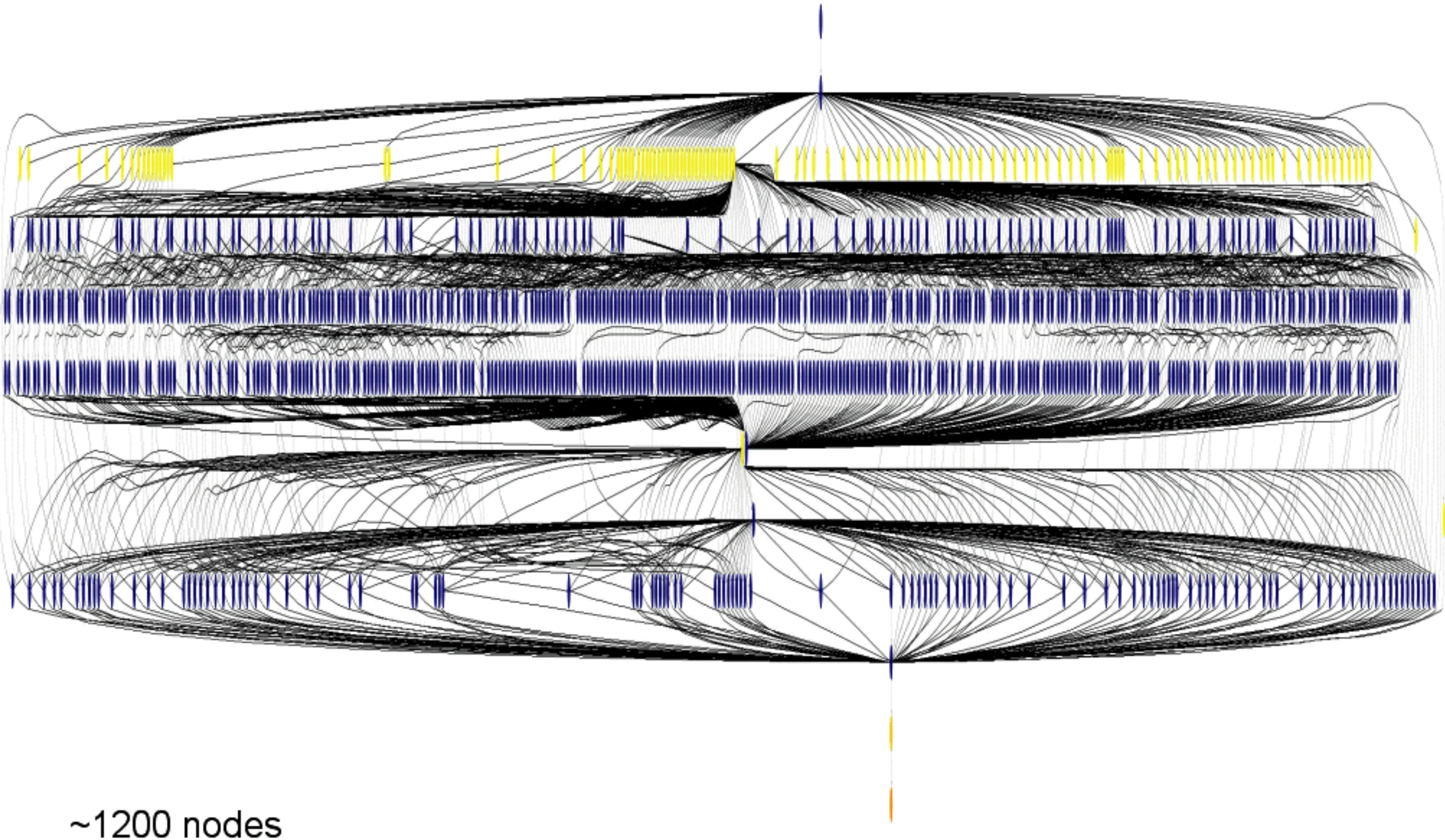
Execution environment: condor pools, clusters, TeraGrid

Pegasus

- Input: a logical workflow that does not indicate locations of the data or the resources used in the computation
- Pegasus
  - Locates the data and executables (possibly replicated)
  - Locates appropriate resources
  - Reduces the workflow where possible (if intermediate data products already exists)
  - Schedules the workflow and hands it off for execution
- Futures: Better performance and reliability

the globus alliance
www.globus.org

Small Montage Workflow



~1200 nodes

Ewa Deelman, USC Information Sciences Institute

Metadata Catalog Service (MCS)

- Separation between management of metadata and data
- Metadata can be organized and published by a collaboration
- Metadata can be organized and personalized by members of a community
- Metadata and data can be easily discovered through attribute-based queries
- Futures: Exploring issues of metadata federation, use of ontologies, query mediators, query planners

Summary

- Grid: resource sharing, integration, virtualization in network systems
- Grid technologies are used in academia and industry
- Grid and Web standards are merging
- The Globus Toolkit provides the basic grid infrastructure
- Ongoing research in many areas

the globus alliance
www.globus.org

For More Information

- Papers and software
  - www.globus.org
  - www.isi.edu/~deelman
  - pegasus.isi.edu
- www.globalgridforum.org