Major
Grid Computing Initatives

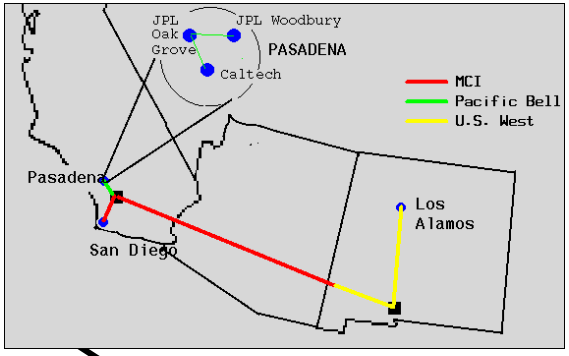
Ian Foster
Mathematics and Computer Science Division
Argonne National Laboratory
and
Department of Computer Science
The University of Chicago

Overview

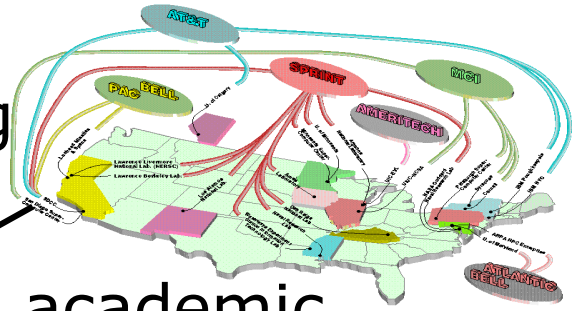
- The Grid concept
- Historical background
- Grid computing initiatives
- Grid technology roadmap
- Data Grid projects
- Summary

The Grid Concept

- Enable communities ("virtual organizations") to share geographically distributed resources as they pursue common goals—in the absence of central control, omniscience, trust relationships
- Via investigations of
  - New applications that become possible when resources can be shared in a coordinated way
  - Protocols, algorithms, persistent infrastructure to facilitate sharing



A Little History



- Early 90s
  - Gigabit testbeds, metacomputing
- Mid to late 90s
  - Early experiments (e.g., I-WAY), academic software projects (e.g., Globus), application experiments
- 2000
  - Major application communities emerging
  - Major infrastructure deployments
  - Clear architecture picture, rich technology base
  - Grid Forum: >300 people, >90 orgs, 11 countries

Layered Grid Architecture (By Analogy to Internet Architecture)

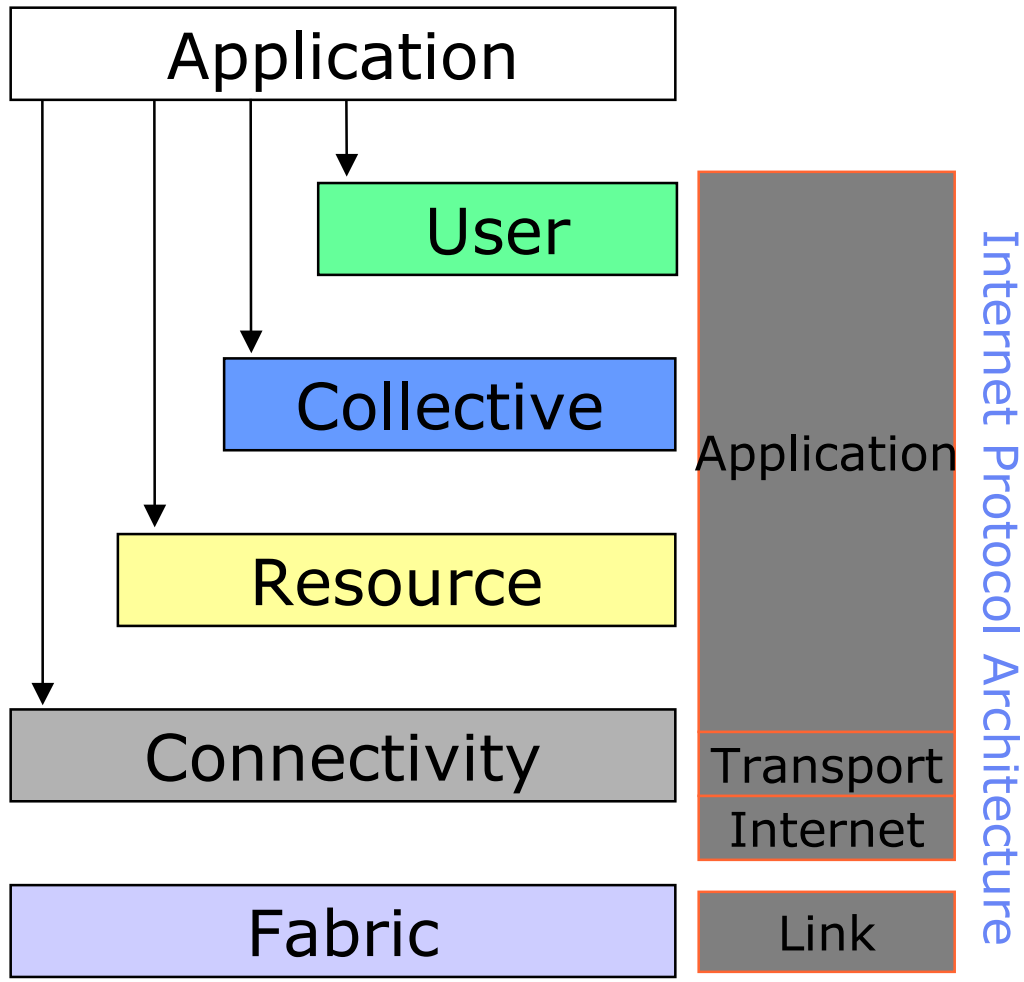
"Specialized services": user- or appln-specific distributed services

"Managing multiple resources": ubiquitous infrastructure services

"Sharing single resources": negotiating access, controlling use

"Talking to things": communication (Internet protocols) & security

"Controlling things locally": Access to, & control of, resources



Grid Technology Base

- Development of Grid protocols & services
  - Protocol-mediated access to remote resources
  - New services: e.g., resource brokering
  - "On the Grid" = speak Intergrid protocols
  - Mostly (extensions to) existing protocols
- Development of Grid APIs & SDKs
  - Facilitate application development by supplying higher-level abstractions
- The (hugely successful) model is the Internet
- The Grid is not a distributed OS!

U.S. Grid Computing Activities (Excluding Data Grid Projects)

- NSF: Fundamental IT research, plus
  - NSF PACI program (~$3M/yr)
  - NEESgrid ($10M over 3 years)
- DOE SC: Fundamental IT research, plus
  - NGI program (done), SciDAC (perhaps)
- DOE DP: DISCOM ($3M/yr?)
- DARPA: Parts of Quorum ($2M/yr?)
- NASA: Information Power Grid (~$5M/yr)

Funds [inadequate] support for research, development, deployment, operations

Data Intensive Computing and Grids

- The term "Data Grid" is often used
  - Unfortunate as it implies a distinct infrastructure, which it isn't; but easy to say
- Data-intensive computing shares numerous requirements with collaboration, instrumentation, computation, …
- Important to exploit commonalities as very unlikely that multiple infrastructures can be maintained
- Fortunately this seems easy to do!

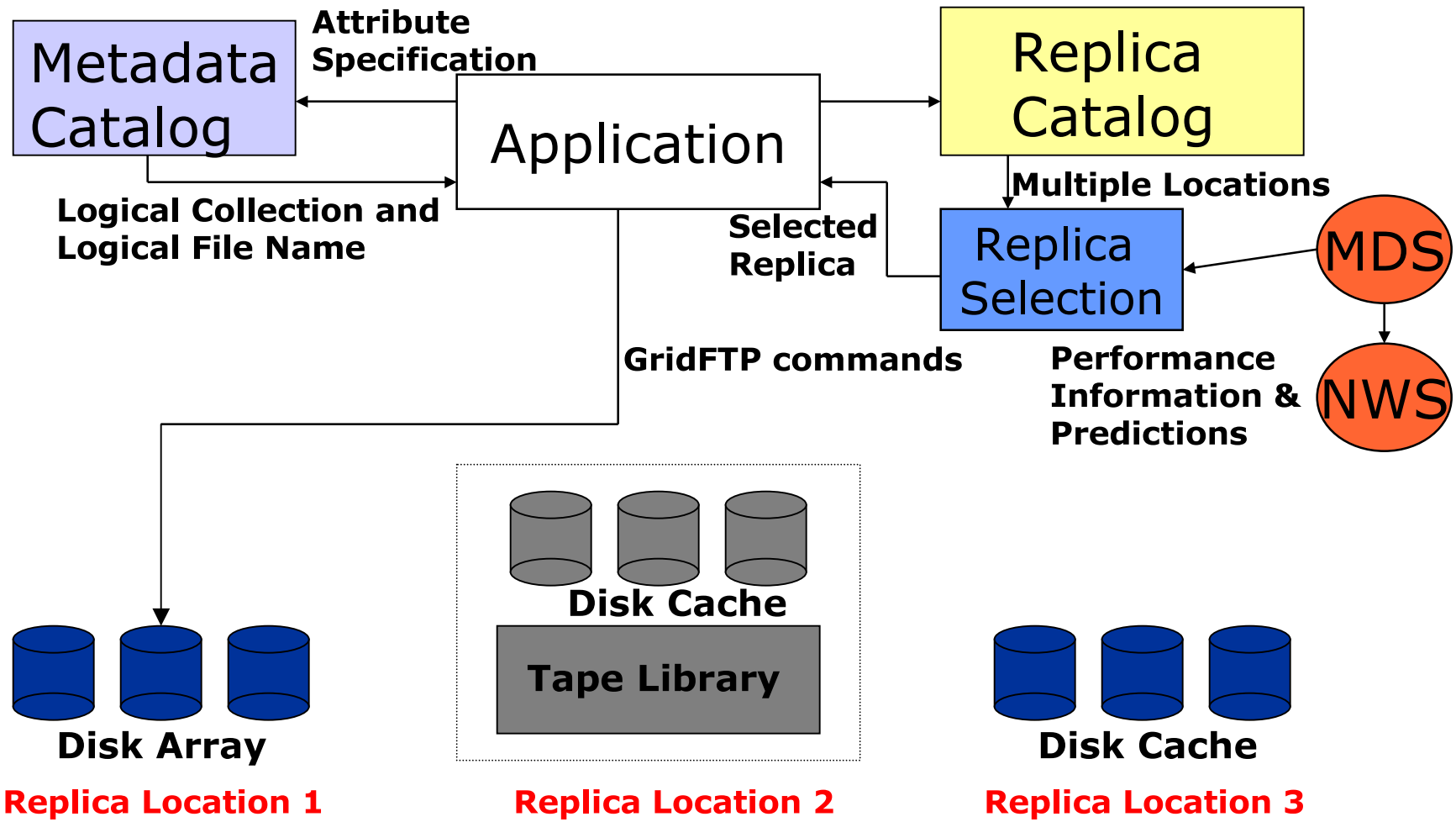
Emerging Data Grid Architecture

Appln	Discipline-Specific Data Grid Application
User	Coherency control, replica selection, task management, virtual data catalog, virtual data code catalog, …
Collective	Replica catalog, replica management, co-allocation, certificate authorities, metadata catalogs,
Resource	Access to data, access to computers, access to network performance data, …
Connect	Communication, service discovery (DNS), authentication, authorization, delegation
Fabric	Storage systems, clusters, networks, network caches, …

Major Data Grid Projects

- Clipper (DOE Science)
  - Technologies for reliable high-speed transfer
- Earth System Grid (DOE Office of Science)
  - DG technologies, climate applications
- European Data Grid (EU)
  - DG technologies & deployment in EU
- GriPhyN (NSF ITR)
  - Investigation of “Virtual Data” concept
- Particle Physics Data Grid (DOE Science)
  - DG applications for HENP experiments

High-Level View of Earth System Grid: A Model Architecture for Data Grids



GriPhyN Overview (www.griphyn.org)

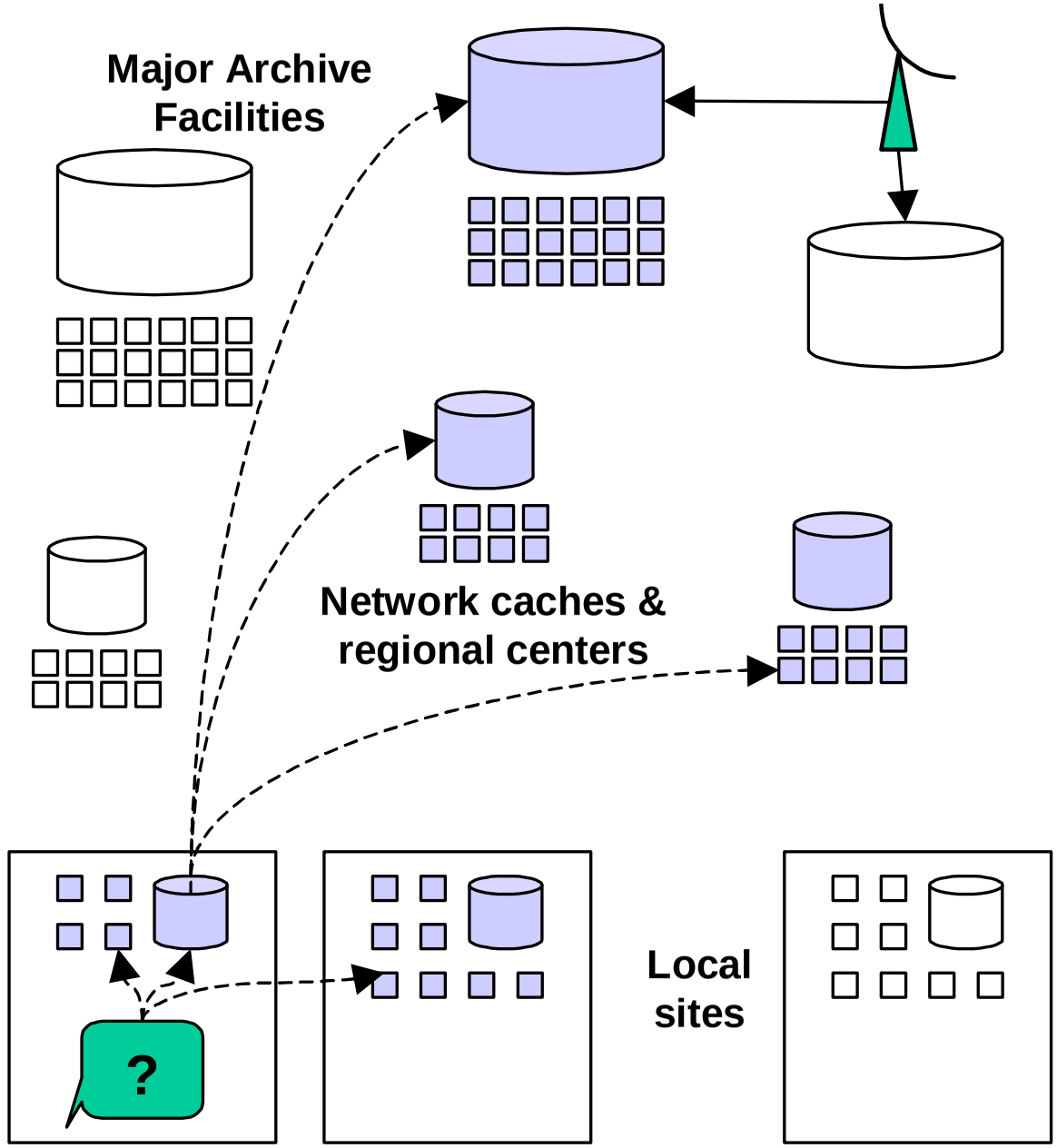
- 5-year, $12M NSF ITR proposal to realize the concept of virtual data, via:
  1) CS research on
     - Virtual data technologies (info models, management of virtual data software, etc.)
     - Request planning and scheduling (including policy representation and enforcement)
     - Task execution (including agent computing, fault management, etc.)
  2) Development of Virtual Data Toolkit (VDT)
  3) Applications: ATLAS, CMS, LIGO, SDSS
- PIs=Avery (Florida), Foster (Chicago)

The Petascale Virtual Data Grid (PVDG) Model

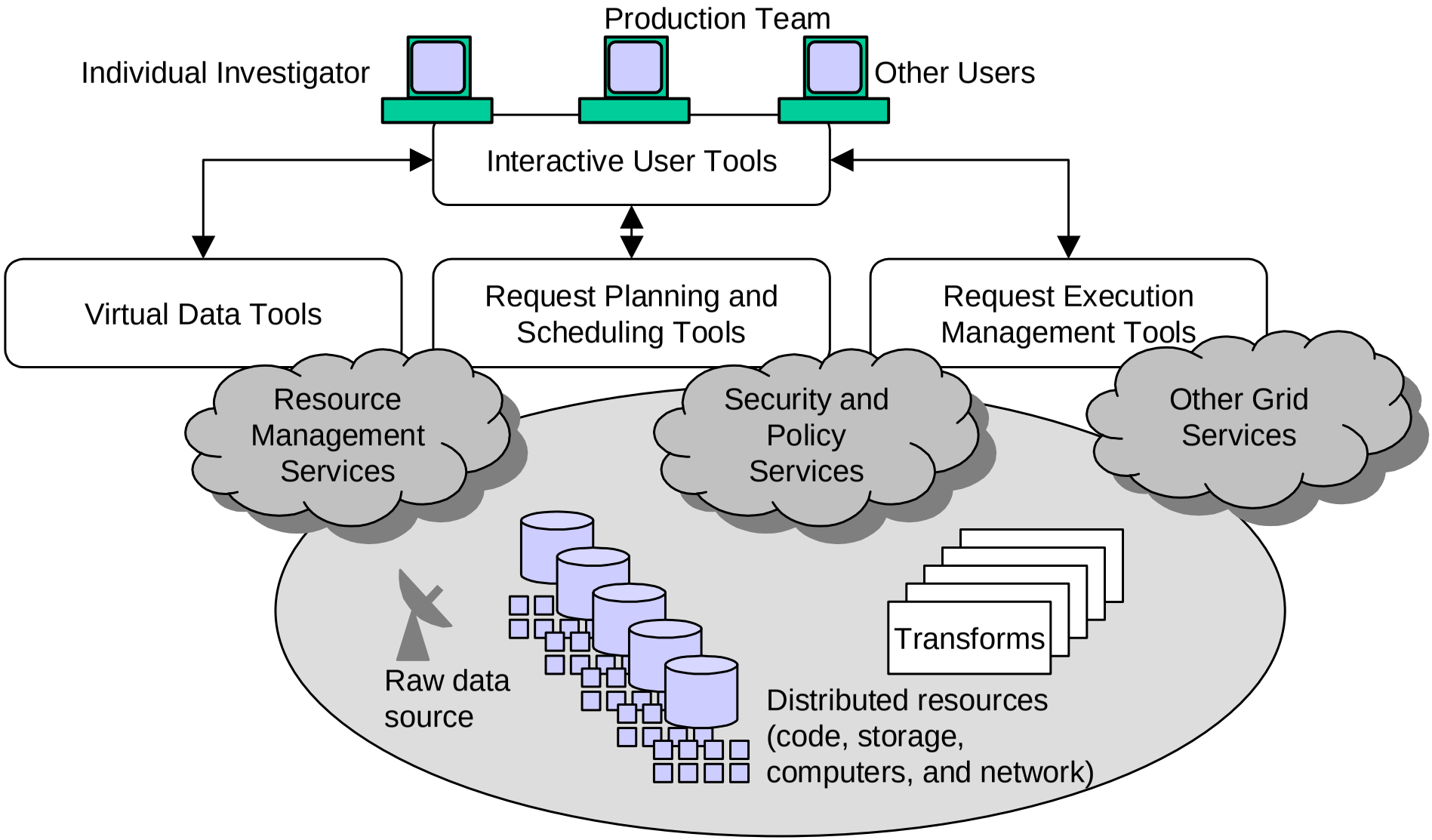
- Data suppliers publish data to the Grid
- Users request raw or derived data from Grid, without needing to know
  - Where data is located
  - Whether data is stored or computed
- User can easily determine
  - What it will cost to obtain data
  - Quality of derived data
- PVDG serves requests efficiently, subject to global and local policy constraints

PVDG Scenario

User requests may be satisfied via a combination of data access and computation at local, regional, and central sites



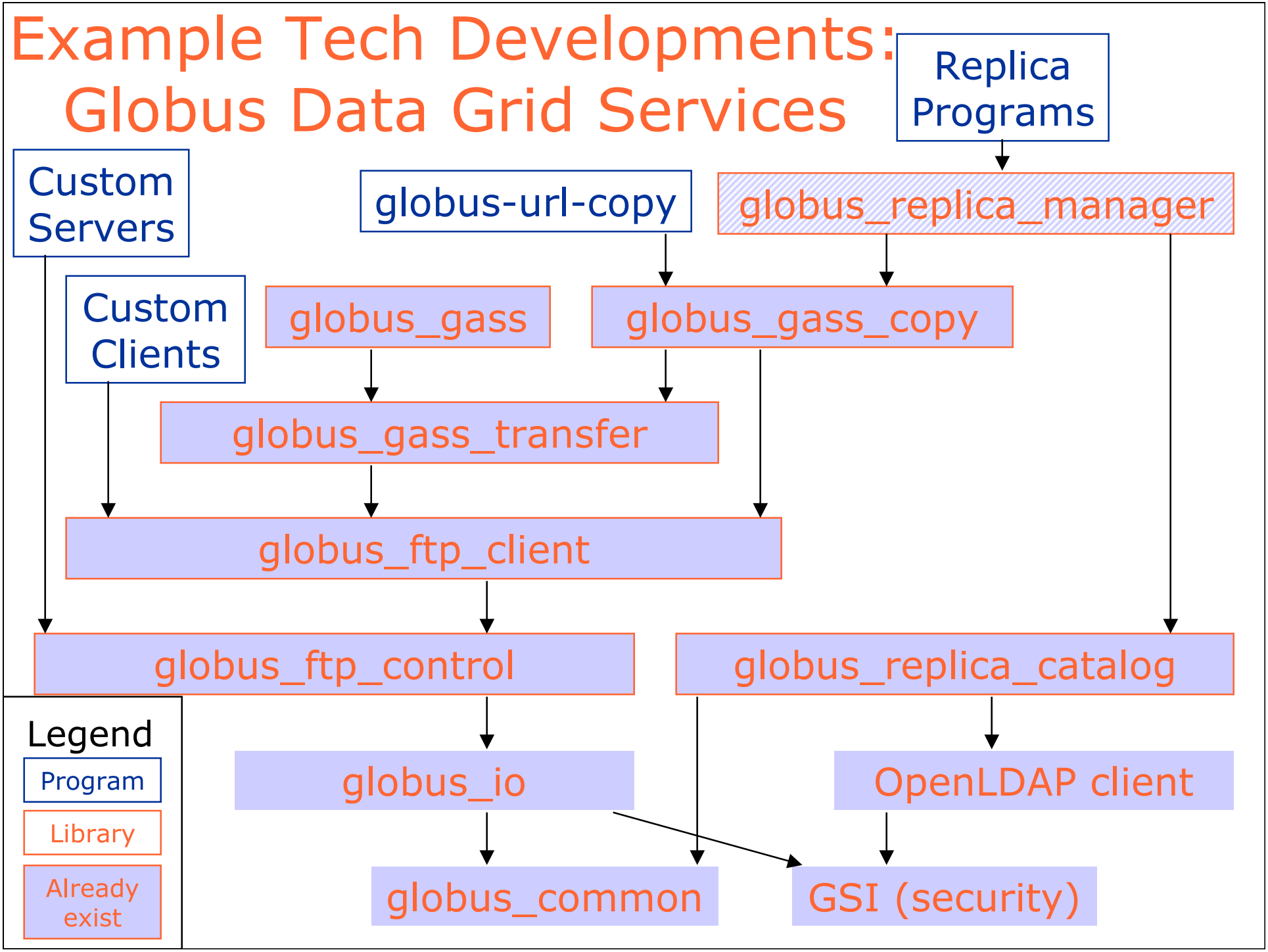
User View of PVDG Architecture



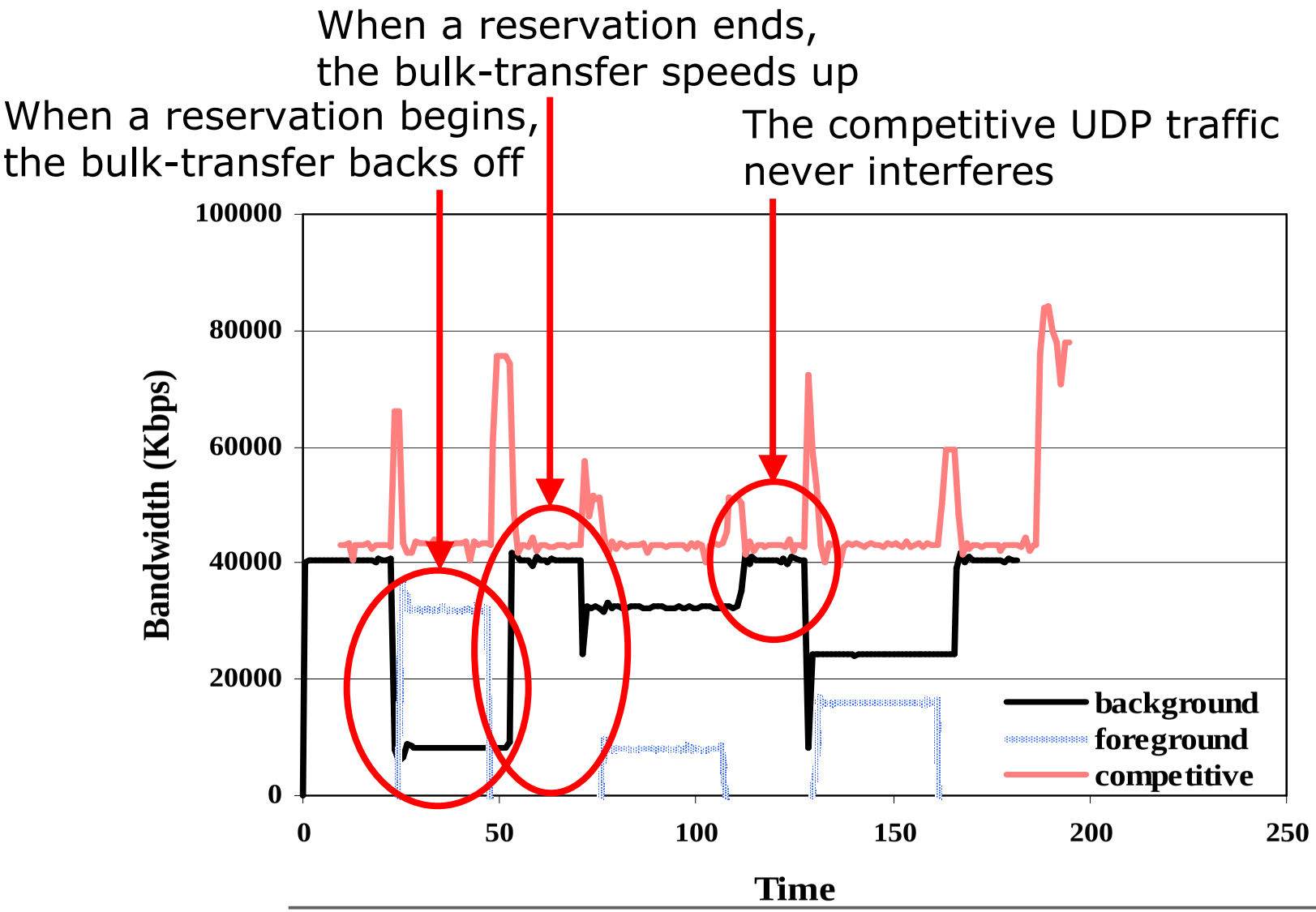
Other Activities Relevant to Data Grids

- Simulation activities
  - MONARC, MicroGrid
- Globus Data Grid/replica mgmt services
  - GridFTP: secure high-performance FTP
  - Replica catalog/replica management
- Grid Data Management Pilot (GDMP)
  - Being used to move data CERN->Caltech
  - Uses GridFTP
  - http://cmsdoc.cern.ch/cms/grid/

Example Tech Developments: Globus Data Grid Services



Example Technology Developments: Quality of Service for Bulk Transfer



GARA: www.mcs.anl.gov/qos

Summary

- New data-intensive applications require a new type of infrastructure: “Data Grids”
- Concerns and infrastructure requirements have much in common with other “Grids”
- Development requires substantial R&D in caching, security, policy, QoS, etc., etc.
- Existing technology base enables contruction of Data Grids to start now

www.globus.org www.griphyn.org

www.gridforum.org www.ppdg.net

grid.web.cern.ch