DRAFT

Next Generation Internet Program Update--Department of Energy

Accelerating Accomplishment of DOE Missions through Advanced Networks

Mary Anne Scott
Advanced Scientific Computing Research
Office of Science
October 6, 1999

Presentation Outline

- Process
- General Program Overview
- Non-NGI Applications and Measurements
- NGI Program Specifics
  - Applications
  - Measurements
  - Testbeds
  - PI meeting results and issues

What is DOE doing in NGI?

Theme

Wide Area Data Intensive and Collaborative Computing

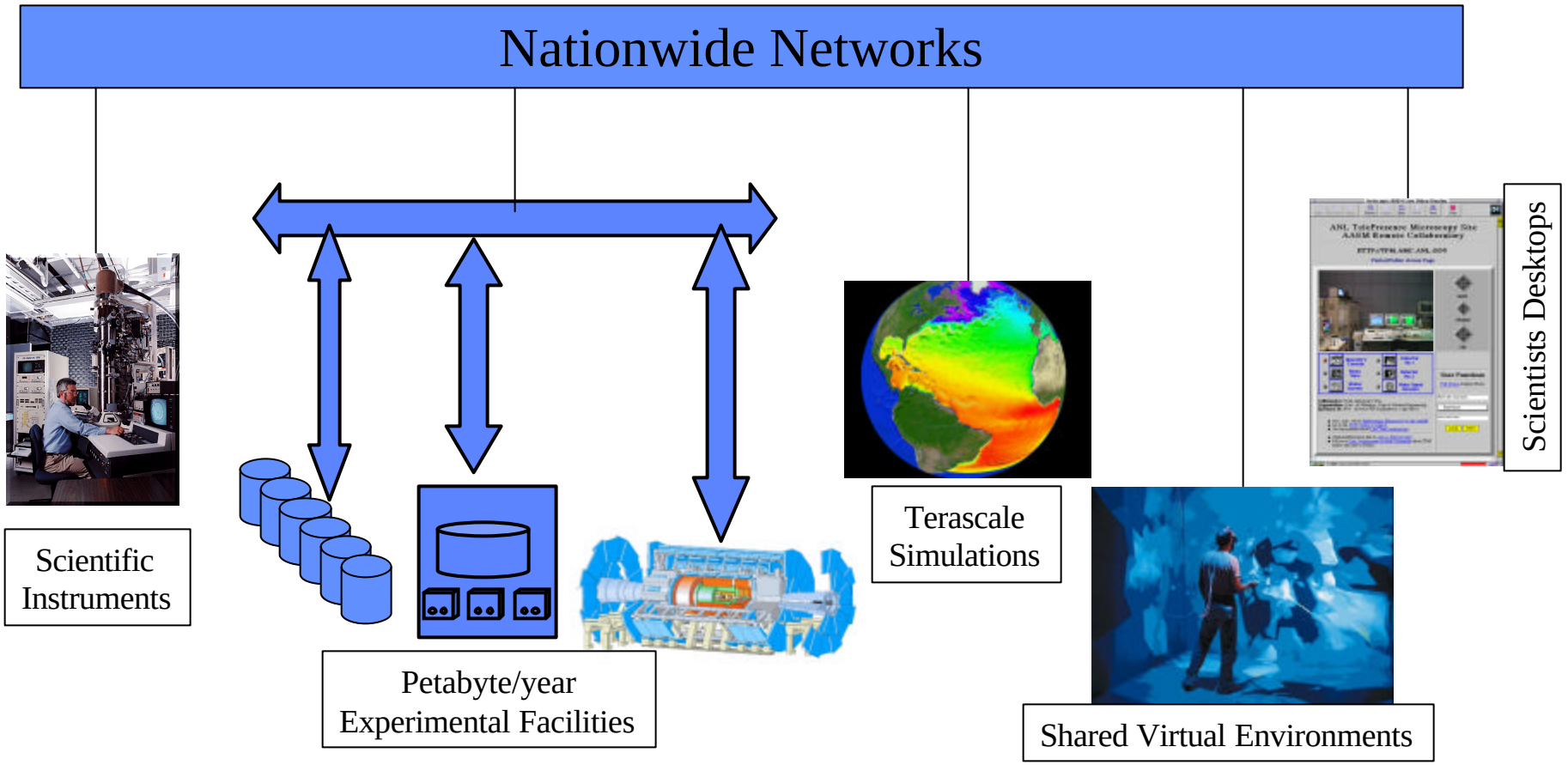
Why?

- Key enabling technology for next generation collaboratories to link users with DOE experimental and computational facilities
- DOE requirements will not be met by commercial R&D;
- Fills a critical gap in NGI research portfolio between DARPA and NSF basic research and applications.

DOE NGI Program

Integrates Basic Research, Applications Research and Testbeds

This approach is critical to research progress in Wide Area Data Intensive and Collaborative Computing



DOE NGI Program Execution

Time Line

- *Jan 8, 1999* Notices of Program Interest published
- *Feb 12, 1999* Preapplications received and reviewed
- *Mar 31, 1999* Applications received
- *Apr 28, 1999* Panel review of applications complete
- *May 6, 1999* Program recommendations complete
- *May 14, 1999* Awards announced
- *Jun 10, 1999* Applications Partnerships and Testbed Providers meet to develop joint strategy
- *Jul 21-22, 1999* Testbed Taskforce Meeting
- *Oct 4-5, 1999* PI meeting to ensure appropriate coupling of basic research with applications partnerships and testbeds.

DOE NGI Program Summary

- Fully Competitive, Peer Reviewed program put in place less than 6 months.
- Reviewers chosen to represent academia, industry, and other Federal agencies ensure DOE program complements other investments.
- Integrated program with basic research, DOE mission critical applications, and testbeds addresses critical research issues in wide area data intensive and collaborative computing.
- PI meeting reinforced integration and identified further areas for closer collaboration and interaction.

The DOE Office of Science's Next Generation Internet Program

- Develop sophisticated, high-performance NGI applications aligned with DOE mission
- Advance middleware & networking technologies to provide the performance guarantees required by applications
- Concept of an integrated grid architecture serves to unify applications, middleware, and networking efforts

DOE NGI Program Awards

Research in Basic Technologies

Twelve projects, for developing advanced middleware services, advanced network architectures and components, and advanced network monitoring tools and services.

- **Architecture** - *Univ. of Tenn; USC*
- **Hardware** - *GATech; LANL*
- **Integration and Analysis** - *USC/ISI, LBLN; UCSD*
- **Measurements** - *UIUC, ANL; Ukan, LBNL*
- **Middleware** - *USC/ISI, ANL, Univ of Wisc; NIU, ANL*
- **Visualization** - *Ohio State; LBNL*

DOE NGI Program Awards

Research in Basic Technologies - Architecture

- ***Optimizing Distributed Application Performance Using Logistical Networking*** *- Univ. of Tenn. / middleware layer to pre-fetch and cache data in network storage servers. Need to deploy cots HW & new SW to make overlay testbed*
- ***Application and Network Cognizant Proxies*** *- USC / Server selection algorithms which allow proxies to select the 'best' server. Need to deploy proxy HW and SW. Work will focus on video streaming schemes to reduce client network BW requirements.*

DOE NGI Program Awards

Research in Basic Technologies - Hardware

- ***Active System Area Networks for Data Intensive Applications*** *- GaTech / Build FPGA HW with drivers to have intelligent NIC. Also provide SW to program FPGAs with specific services (I.e., stream switching).*
- ***Network Interface Cards as First-Class Citizens: Alleviating the Application-to-Network Bottleneck*** *- LANL / Build FPGA HW with drivers to have NIC that attaches to PC system bus instead of I/O bus. Run simulation studies to determine the best approach to making the FPGA act like a "cpu" (cache coherency issues)*

DOE NGI Program Awards

Research in Basic Technologies - Integration and Analysis

- ***Secure and Reliable Group Communication*** *- USC/ISI, LBNL / Integration effort to link Akenti, CLIQUES, and Totem systems together to form a reliable/secure multicast environment. Upgrades and enhancements for group communications.*
- ***Scalable Service Models for Quantifying Quality of Service*** *- UCSD / mathematical analysis of traffic service models.*

DOE NGI Program Awards

Research in Basic Technologies - Measurements

- ***A Uniform Instrumentation, Event, and Adaptation Framework for Network-Aware Middleware and Advanced Network Applications*** *- UIUC, ANL / Middleware service layer to gather application performance data from various sensors located throughout the network. This effort will define the semantics of the sensor data, including what data is captured, how it is stored, and how applications find it.*
- ***Network Monitoring for Performance Analysis and for Enabling Network-Aware Applications*** *- UKans, LBNL / Middleware service layer to gather application performance data from hosts and routers in the network. This effort will focus on building tools to analyze data collected by existing netlogger system.*

DOE NGI Program Awards

Research in Basic Technologies - Middleware

- ***Diplomat: Policy-Based Resource Management for Next-Generation Internet Applications*** *- USC/ISI, ANL, Univ. of Wisc. / Middleware effort to integrate Condor's Match-Maker allocation service into the Globus scheduling environment.*
- ***Technologies and Tools for High-Performance Distributed Computing*** *- NIU, ANL / Middleware effort to incorporate MPI-2 functions and features into MPICH-G.*

DOE NGI Program Awards

Research in Basic Technologies - Visualization

- ***Stackable Middleware Service for Advanced Video Networking Applications*** *- Ohio State / middleware toolkit to provide video applications with inline compression and dithering service layer.*
- ***Advanced Visualization Communication Toolkit*** *- LBNL / middleware toolkit dedicated to mapping video application requests to supported network services.*

DOE NGI Program Awards

Applications-Network Technology-Network Testbed Partnerships

Five applications, all collaborations with multiple sites that include universities and national laboratories

- ***A Grid-based Collaboratory for Real-time Data Acquisition, Reduction and Visualization for Macromolecular X-Ray Crystallography Using the LBL Advanced Light Source*** *- Indiana Univ and ANL*
- ***CorridorOne: An Integrated Distance Visualization Environments for SSI and ASCI Applications*** *- ANL, LBNL, LANL, UIC, Univ of Utah, and Princeton Univ*
- ***Prototyping an Earth System Grid*** *- UCAR, USC, Univ of Wisc, ANL, LANL, LBNL, and LLNL*
- ***Prototyping a Combustion Corridor*** *-LBNL, ANL, LANL, Univ of Wisc*
- ***The Particle Physics Data Grid*** *- Caltech, SDSC, USC, ANL, BNL, FNAL, Jlab, LBNL, and SLAC*

DOE NGI Program Awards

DOE-University Technology Testbeds

Two testbeds, for demonstrating advanced services to university sites, improving capabilities and access for university researchers involved in applications including combustion, climate, and high-energy physics

- ***EMERGE: ESnet/Metropolitan Research and Education Network (MREN) Regional Grid Experimental NGI Testbed -*** *Univ. of IL-Chicago, Univ. of IL-Urbana-Champaign, Northwestern Univ., Univ. of Wisc., Univ. of Chicago*
- ***QUALIT: QBone University and Lab Interconnect Testbed (a collaborative testbed project with the Internet2 QoS Working Group)-*** *University Corporation for Advanced Internet Development*

Non-NGI Applications and Measurements

- predate NGI
- relevant base

Non-NGI Applications and Measurements

“China Clipper” Project

Computational grids providing middleware that supports applications requiring configurable, distributed, high-performance computing and data resources.

Application Demonstration

Integration into an important DOE Science application is critical to demonstrate the successful use of China Clipper technology to construct an application-specific, data-intensive computational grid.

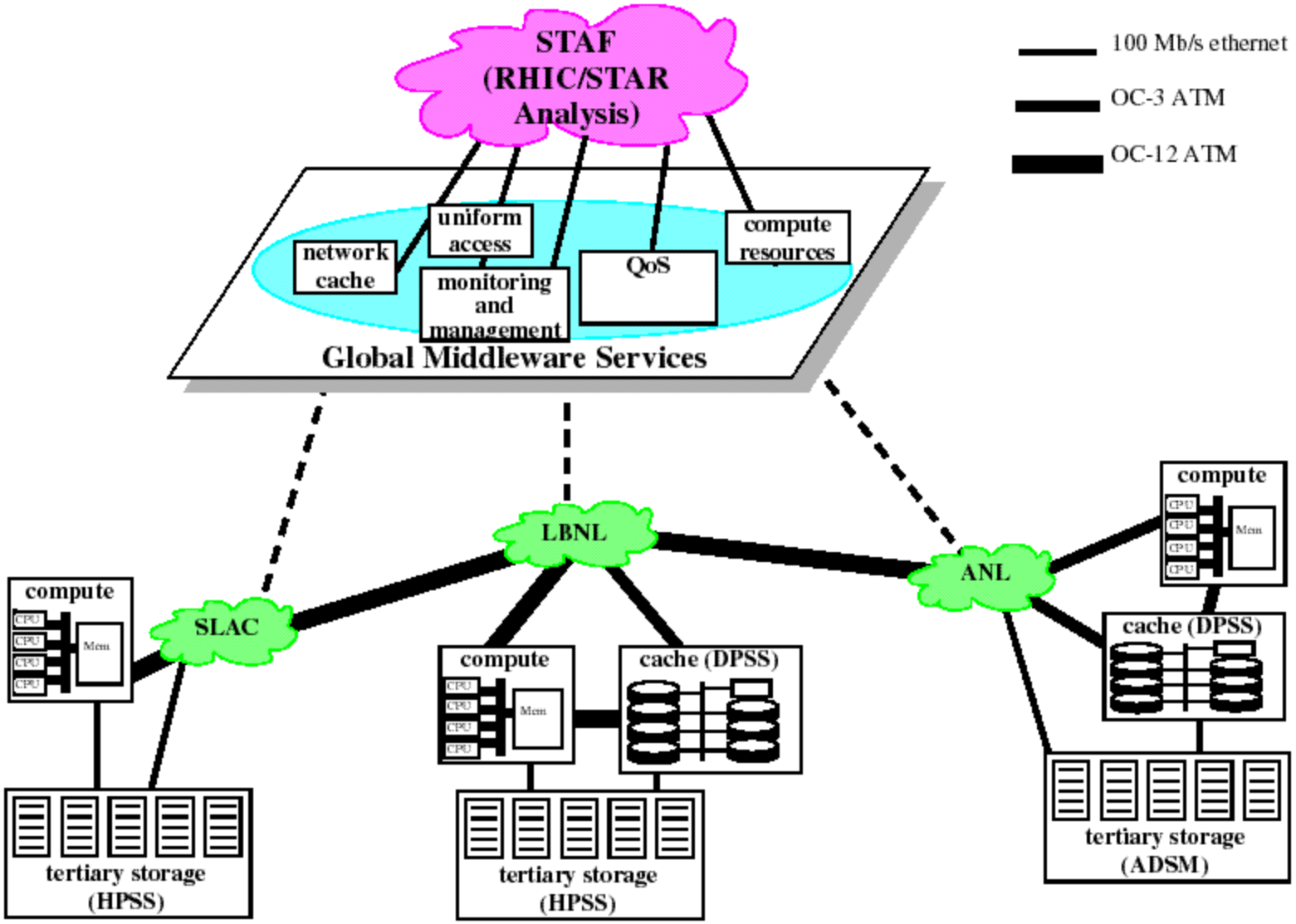
China Clipper goals:

- Demonstrate STAF accessing data from HPSS (using DPSS cache) with sustained data transfer rate of 50MB/sec
- Demonstrate STAF choosing to access data from SLAC, ANL, or LBNL depending on network congestion and quality of respective data sets

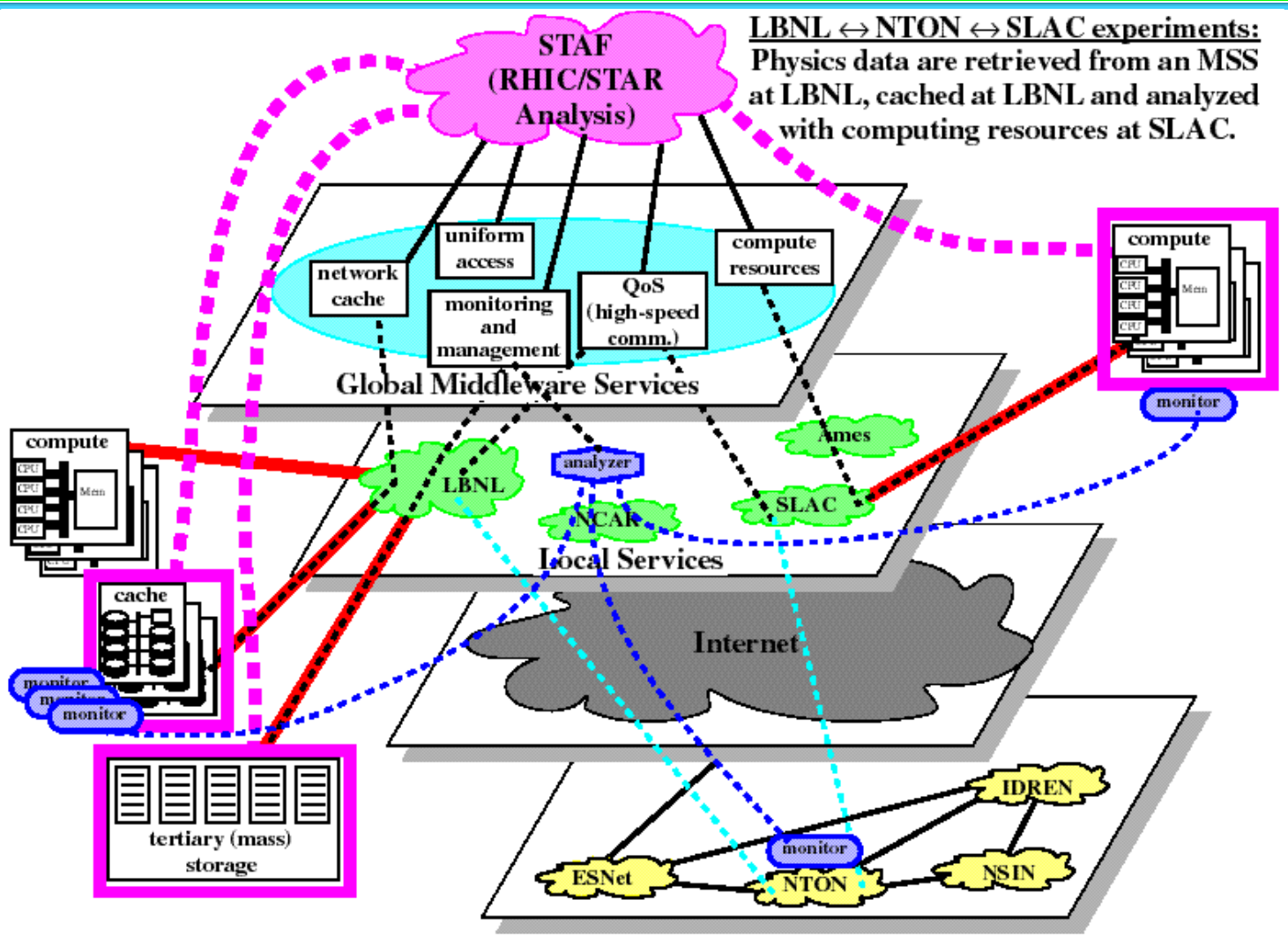
China Clipper Elements

- High-Speed Testbed
  - Computing and networking infrastructure
- Differentiated Network Services
  - Traffic shaping on ESnet
- Monitoring Architecture
  - Traffic analysis to support traffic shaping and CPU scheduling
- Data Architecture
  - Transparent management of data
- Application Demonstration
  - Standard Analysis Framework (STAF)

Clipper Architecture



Clipper Architecture



Bandwidth Broker

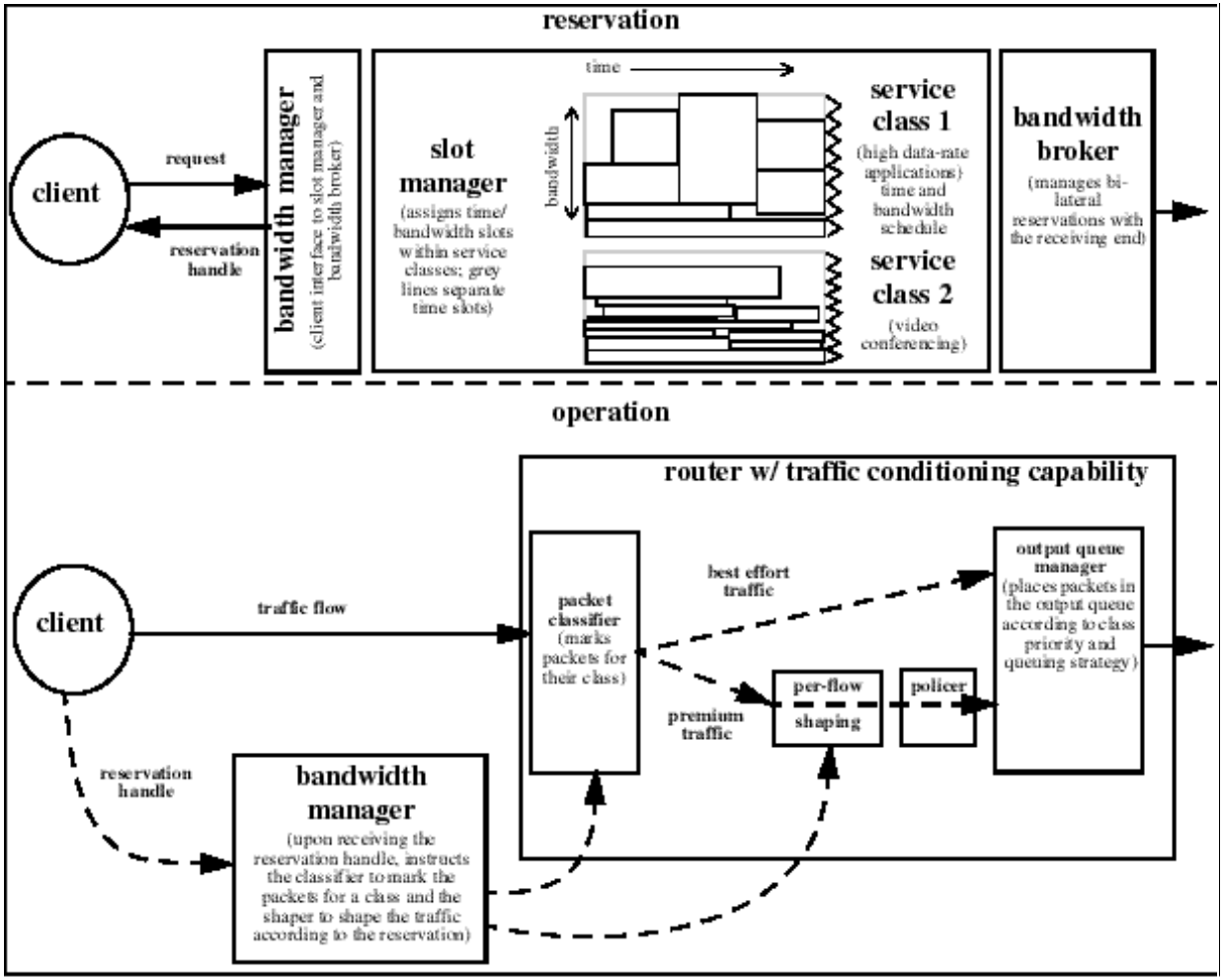
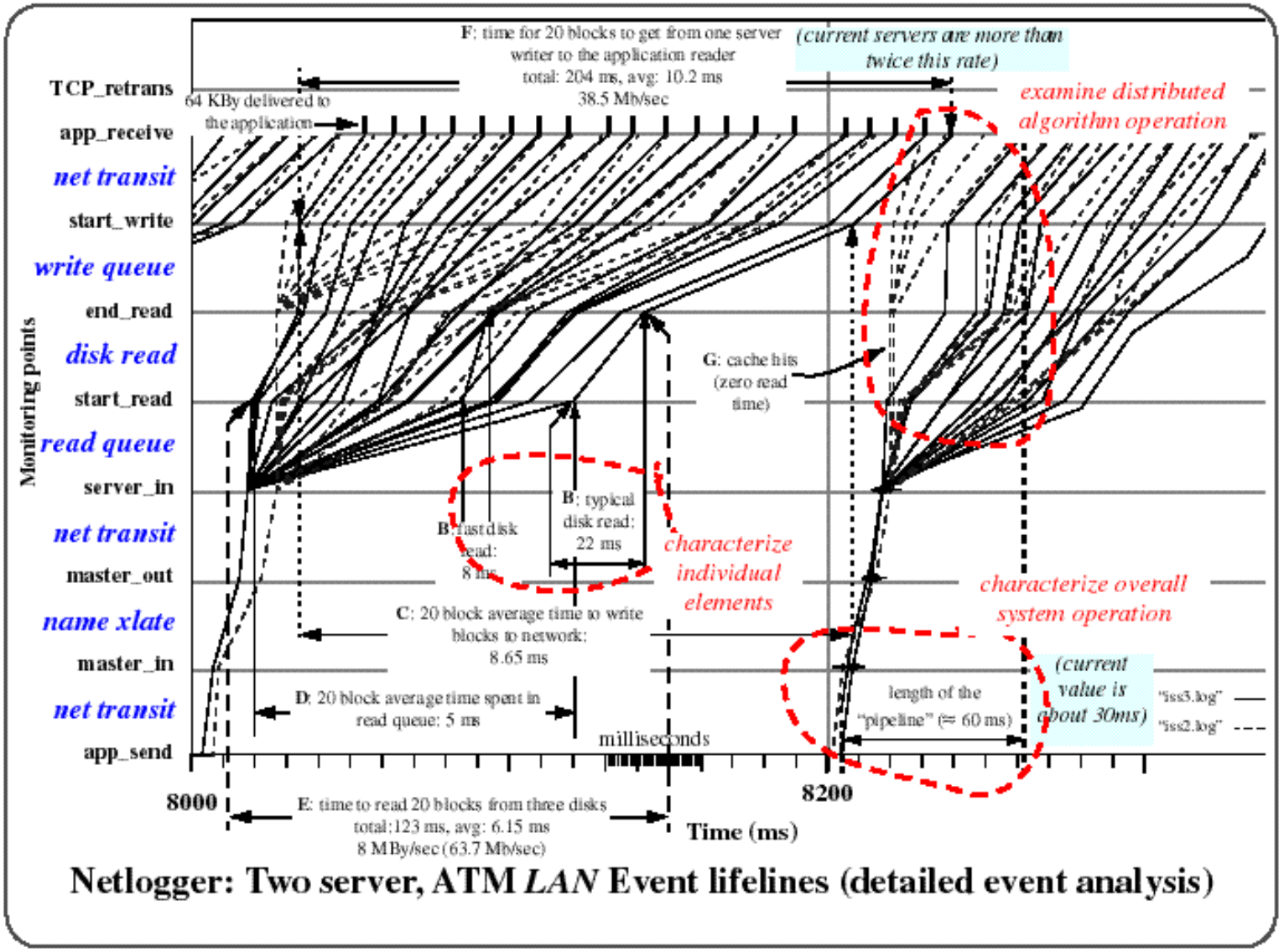


Figure ZZ An architecture for bandwidth reservation in IP networks based on temporal reservation based admission control for high priority IP differentiated services classes.

Monitoring Architecture



Netlogger: Two server, ATM *LAN* Event lifelines (detailed event analysis)

China Clipper Status

- LBNL <-> SLAC
  - NTON OC12
  - ATM connection
  - 58MB sustained
- LBNL <-> ANL
  - ESnet OC12
  - IP through SPRINT cloud
  - 38 MB sustained

Networking Issues

- Underlying network and routers owned by service providers.
  - Multiple administrative domains.
  - We don’t have control!
- Who gets access to Testbed PVC?
  - Finite resource.
  - Can this be brokered?
  - Security issues.

Slide or slides on PingER measurements with emphasis on correlation with Surveyor

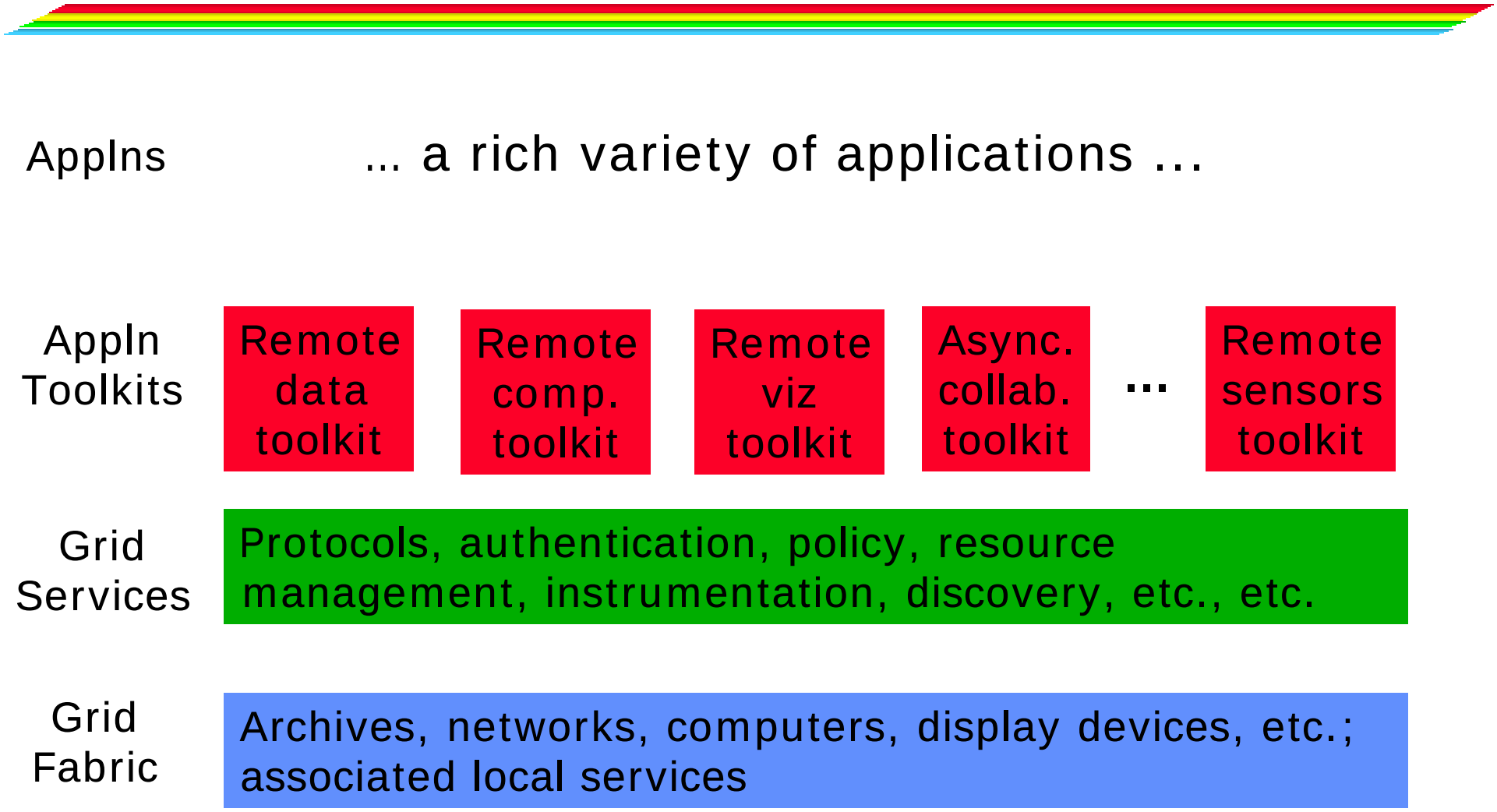
NGI Program Specifics

DOE FY 1999 NGI Program

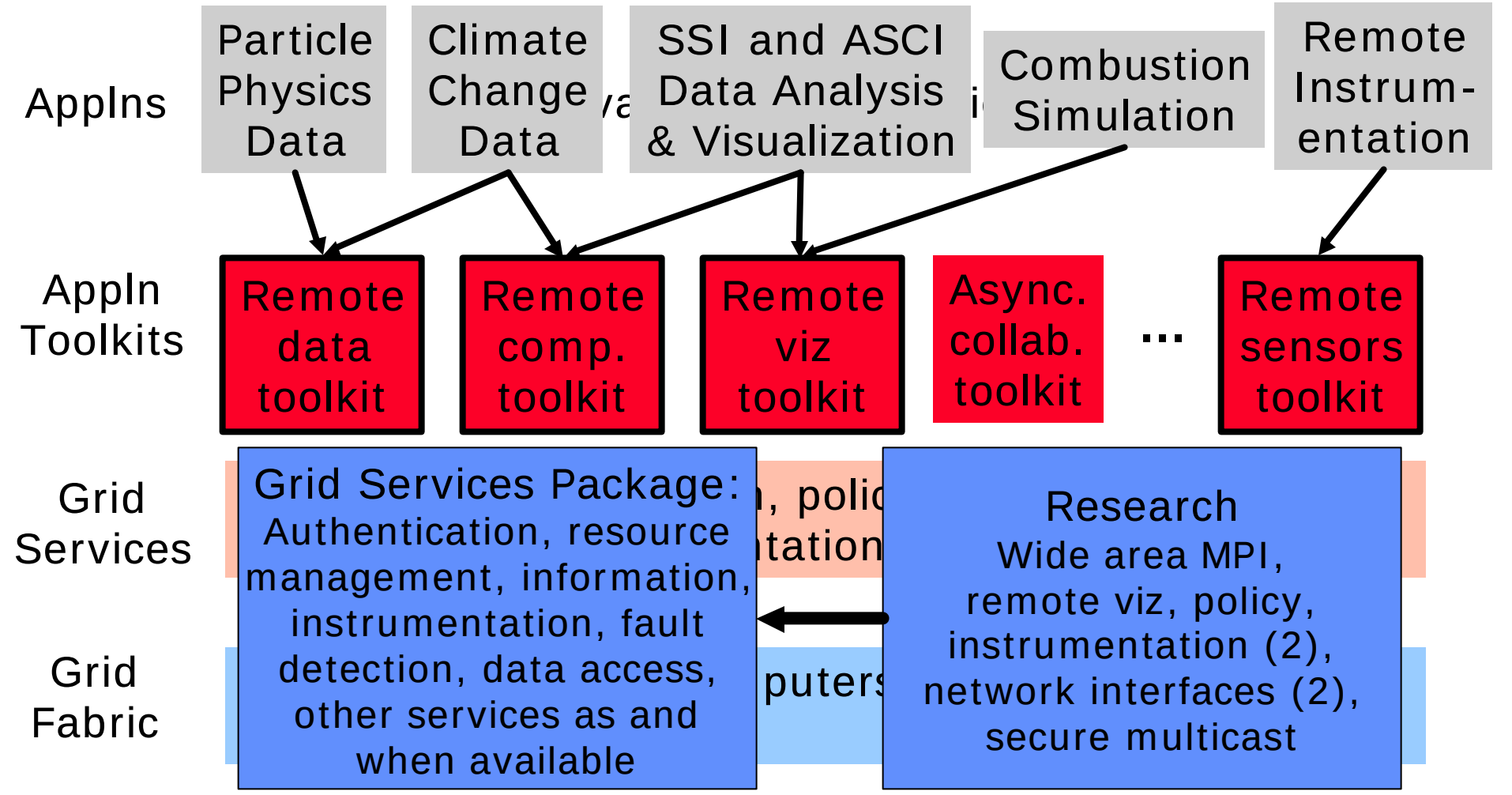
- **Wide Area Data Intensive and Collaborative Computing**
  - **Ensure that the underlying technologies are developed**
  - **Integrate and test the technologies on DOE mission applications**
  - **Ensure that the tools developed can be used by researchers at universities**

Concept of an integrated grid architecture serves to unify applications, middleware, and networking efforts

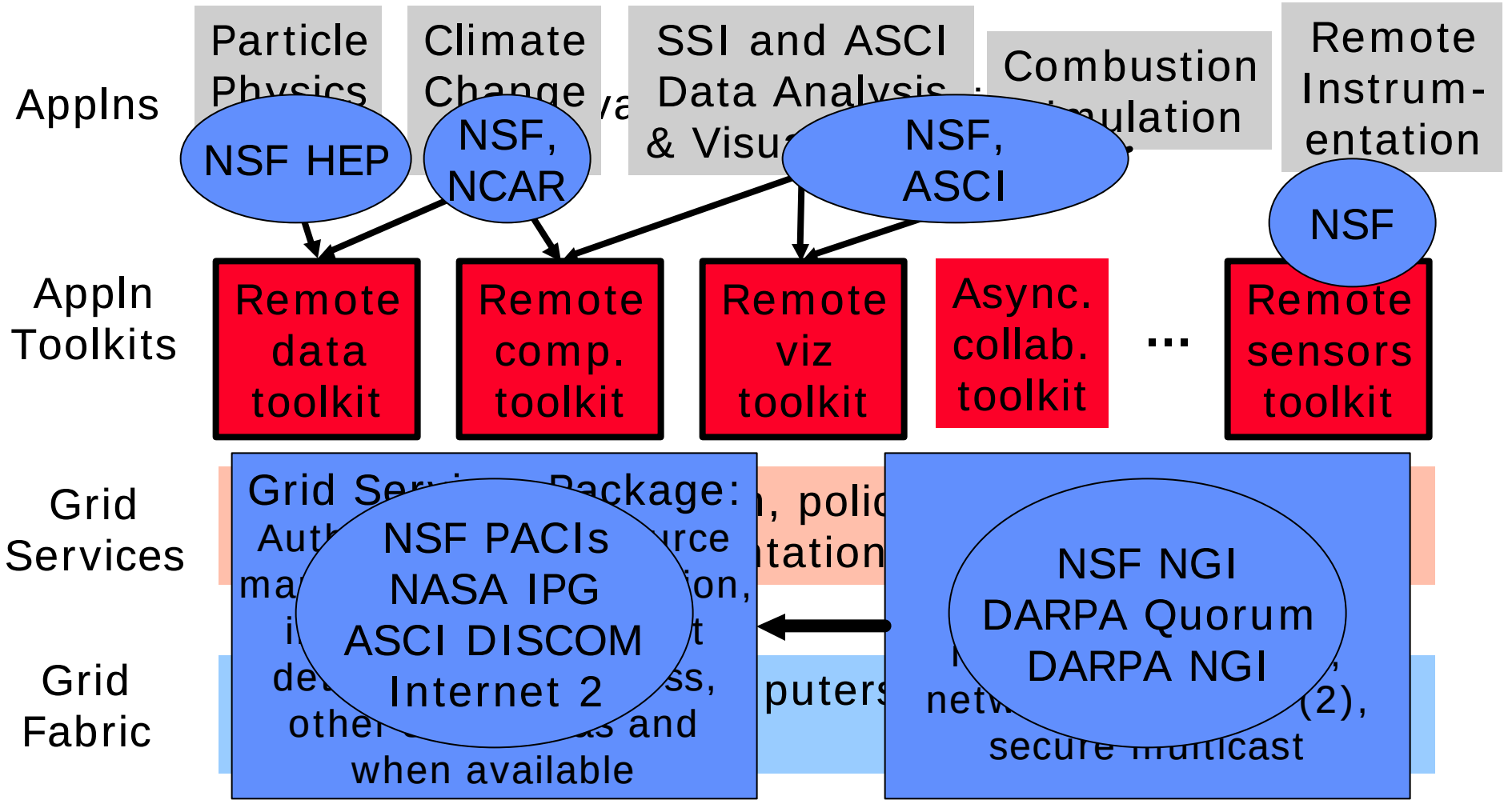
An Integrated Grid Architecture



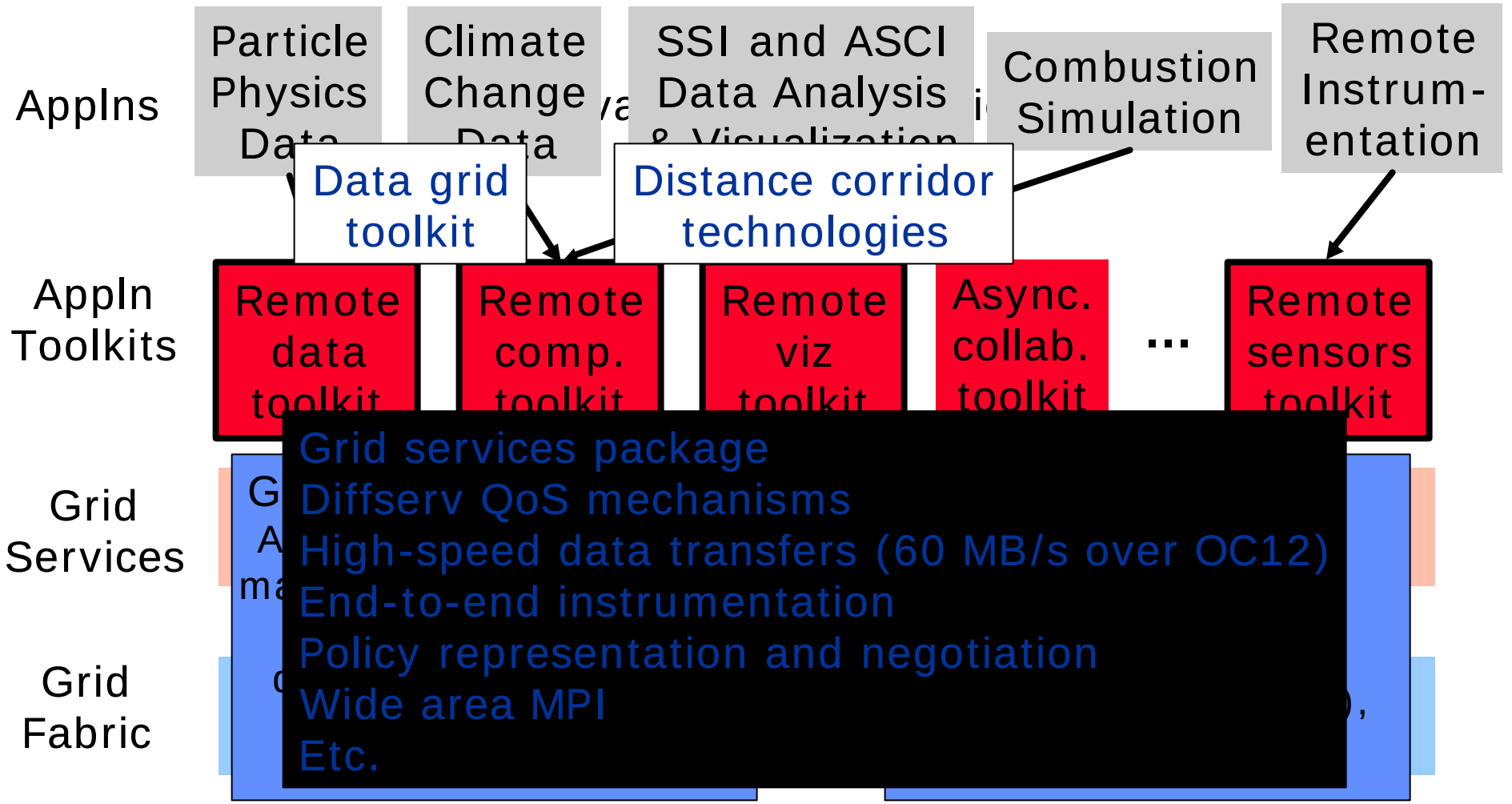
DOE's NGI Program and the Integrated Grid Architecture



DOE's NGI Program Builds on and Complements other Agency Programs



DOE's NGI Program Makes Unique and Important Contributions



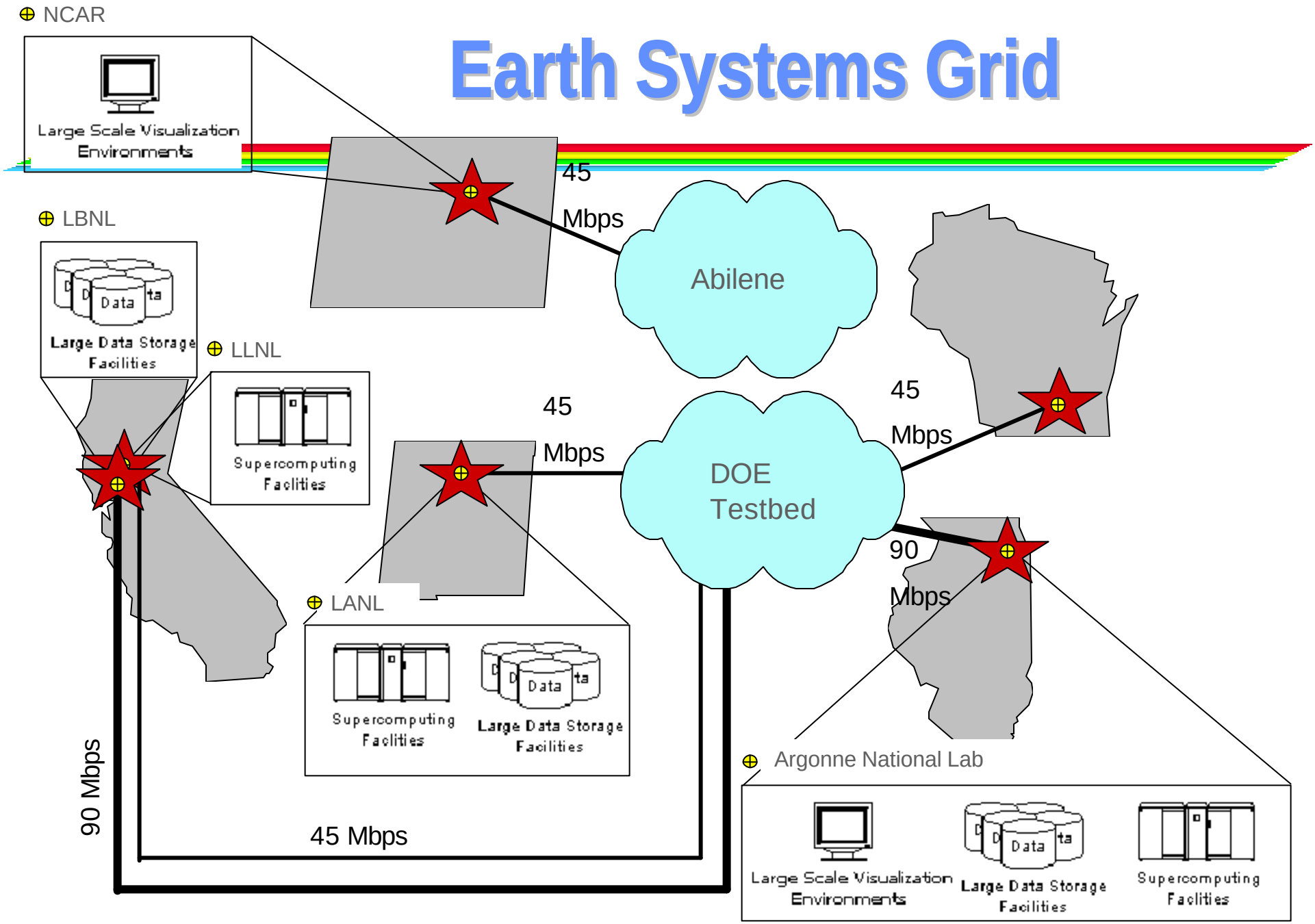
Prototyping an Earth System Grid

Context:: The Accelerated Climate Prediction Initiative

- Model development and evaluation
  - Atmosphere/ocean/land surface/sea ice components, coupled models, process parameterizations, simulation design, algorithm development
- Climate projections (PBs of data)
  - Optimization of computer performance, multi-decadal ensemble simulations, archives of petabyte datasets, generation of probability distributions
- Impact assessment, analysis (1000s users)
  - Policy evaluation, impacts research, regional/national/international assessments, specialized tools and datasets, information dissemination

Why Climate Change Presents a Networking Problem

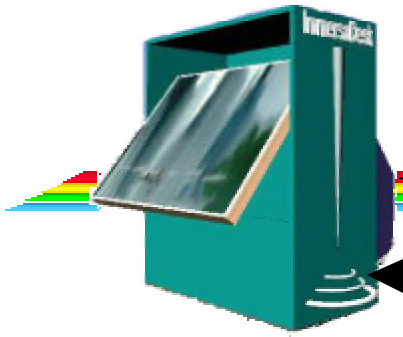
- Data sources: high data rates
  - Multiple, high-resolution, multi-component runs: 100 MB/s (@5 TF/s) => 3 PB/yr
- Data analyses: sophisticated, intensive
  - Regional impacts studies => downscaling
  - Intercomparisons with data & other models
- User community: large and diverse
  - O(100+) laboratories, centers, universities
  - Highly multidisciplinary
  - International connections


NCAR
Earth Systems Grid
Large Scale Visualization Environments
45 Mbps
LBNL
Data
Large Data Storage Facilities
LLNL
Supercomputing Facilities
Abilene
45 Mbps
45 Mbps
DOE Testbed
90 Mbps
LANL
Supercomputing Facilities
Data
Large Data Storage Facilities
90 Mbps
45 Mbps
Argonne National Lab
Large Scale Visualization Environments
Data
Large Data Storage Facilities
Supercomputing Facilities

Approach: Earth System Grid

- Integrate data archives into a distributed data management and analysis “Grid”
- More than storage & network, also e.g.
  - Caching and mirroring to exploit locality
  - Intelligent scheduling to determine appropriate replica, site for (re)computation, etc.
  - Coordinated resource management for performance guarantees
  - Embedded security, policy, agent technologies for effective distributed analysis

Illustrative Scenario



"How do midwest flood frequencies under 2xCO_2 scenario compare with historical data?"

Query manager

"Access datasets A, B; run A->meso->hydro; compare result with B"

Cache manager

Historical data archive

Resource manager

Cache

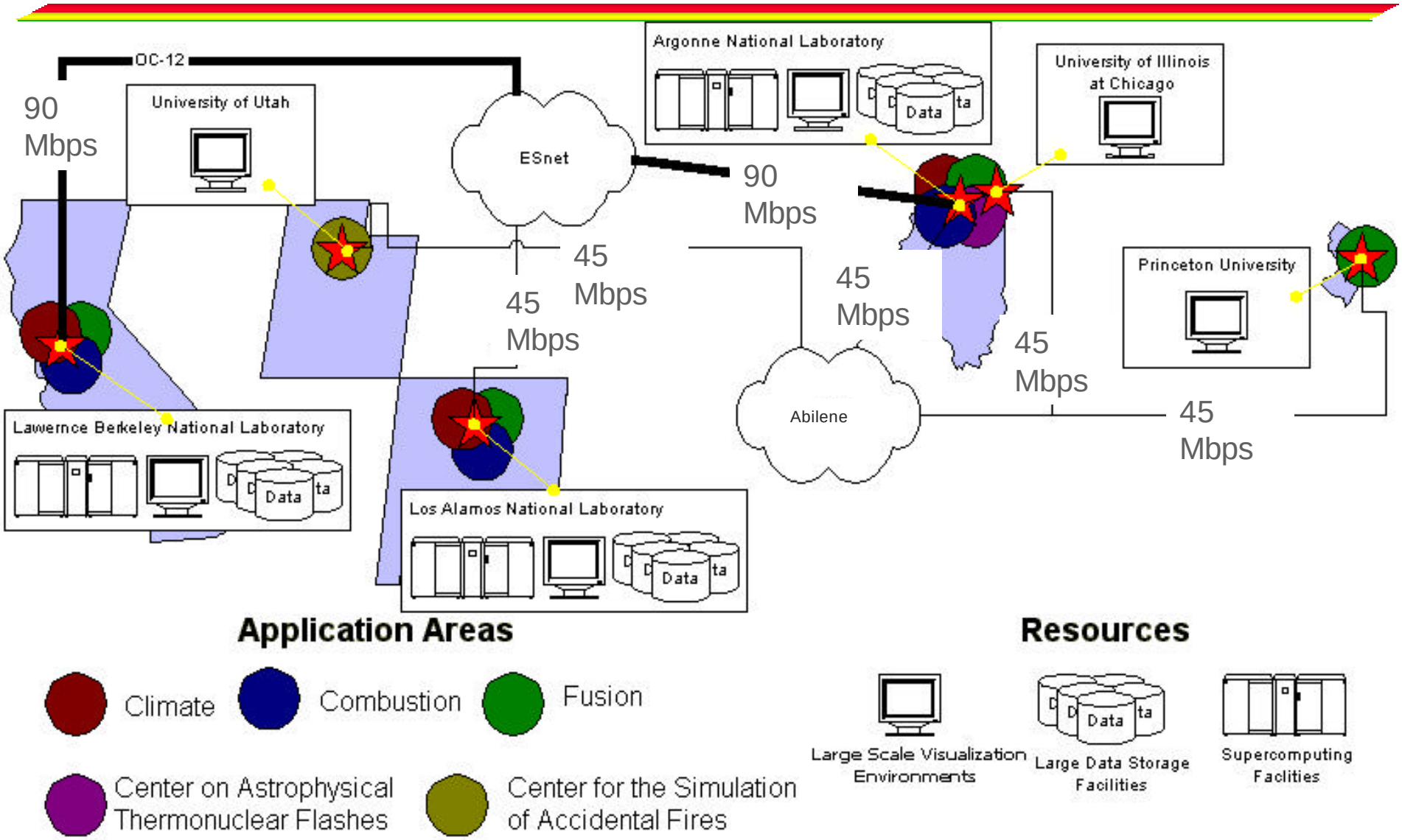
Simulation data archive

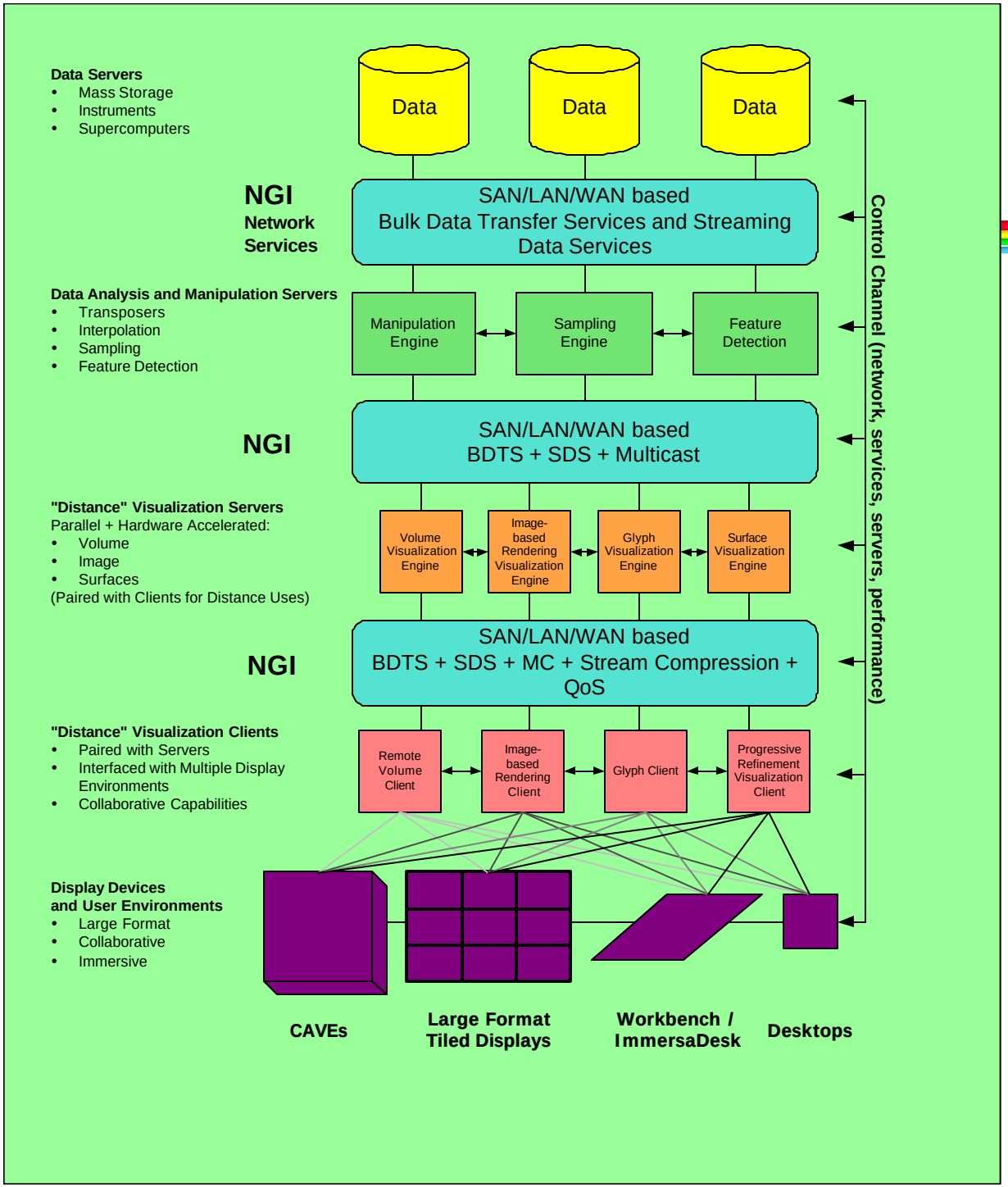
Cache

Cache

meso → hydro → compare

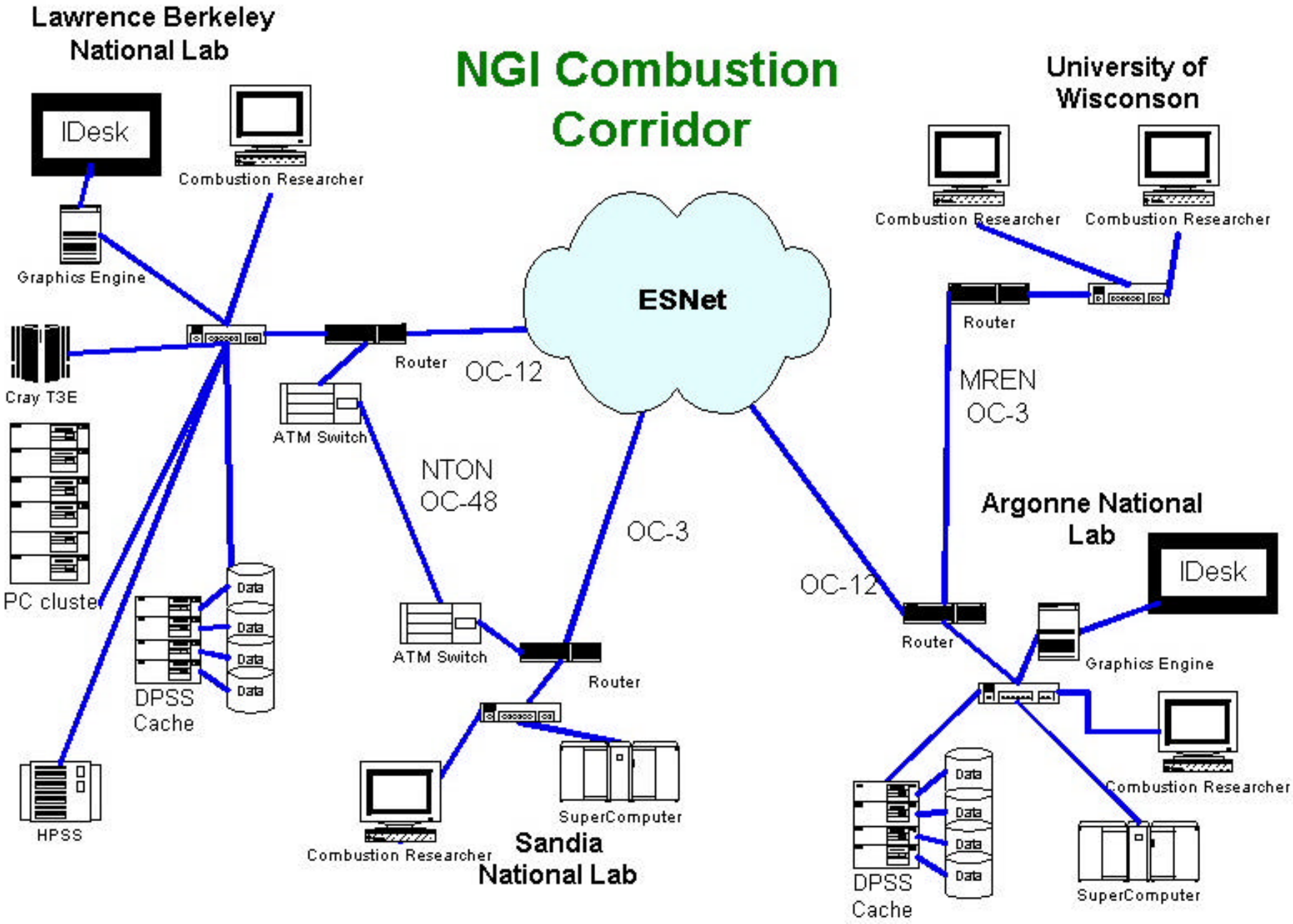
Corridor One Þ Attacking the Distance Visualization Problem



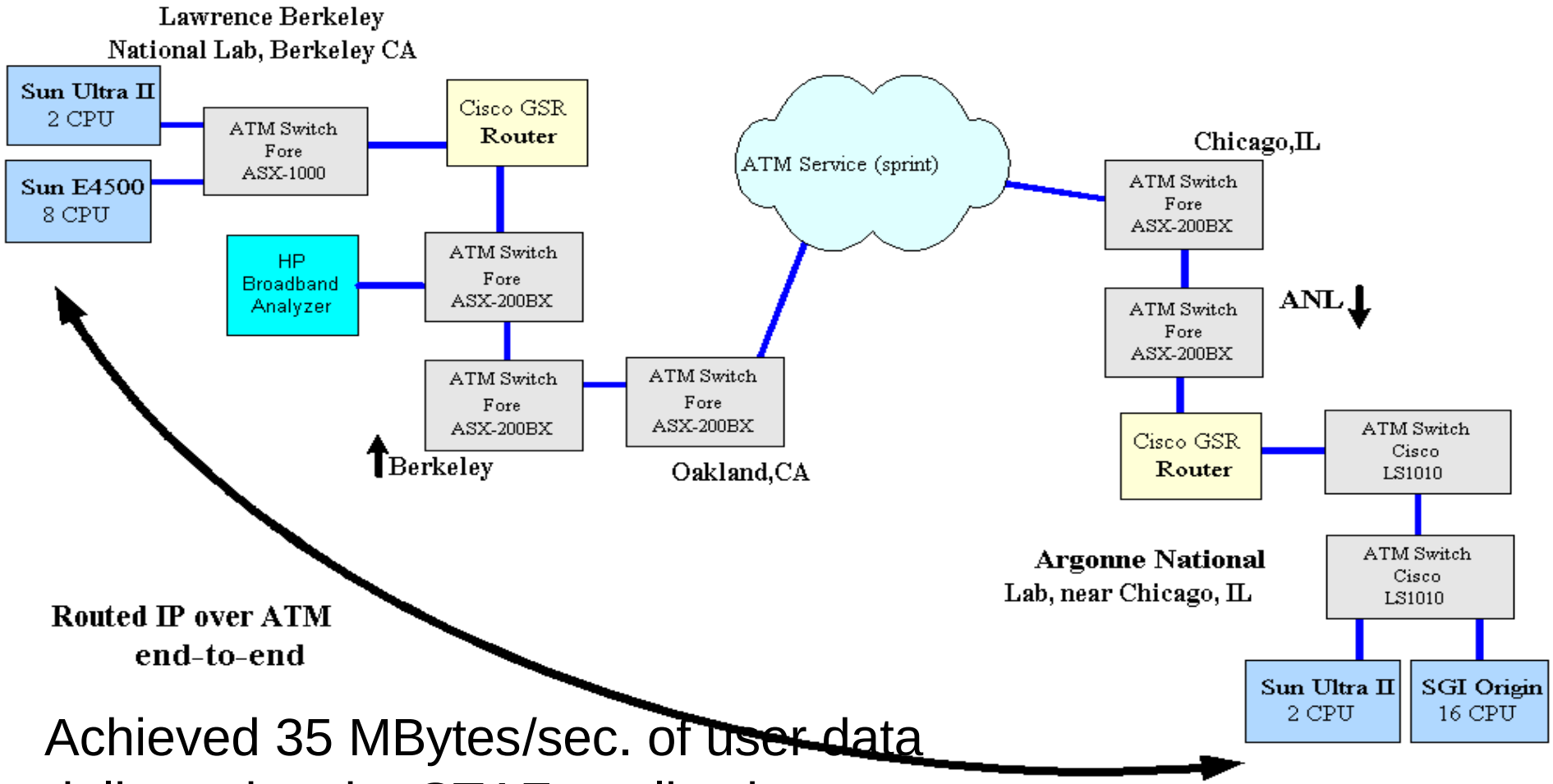


Corridor One

- Data Servers
- Analysis and Manipulation Engines
- Visualization Backend Servers
- Visualization Clients
- Display Device Interfaces
- Advanced Networking Services


NGI Combustion Corridor
Lawrence Berkeley National Lab
IDesk
Combustion Researcher
Graphics Engine
Cray T3E
PC cluster
HPSS
DPSS Cache
Data
Data
Data
Data
Router
ATM Switch
OC-12
ESNet
NTON OC-48
OC-3
ATM Switch
Router
SuperComputer
Combustion Researcher
Sandia National Lab
University of Wisconson
Combustion Researcher
Combustion Researcher
Router
MREN OC-3
OC-12
Argonne National Lab
Router
IDesk
Graphics Engine
Combustion Researcher
SuperComputer
DPSS Cache
Data
Data
Data
Data

High Bandwidth Data Distribution



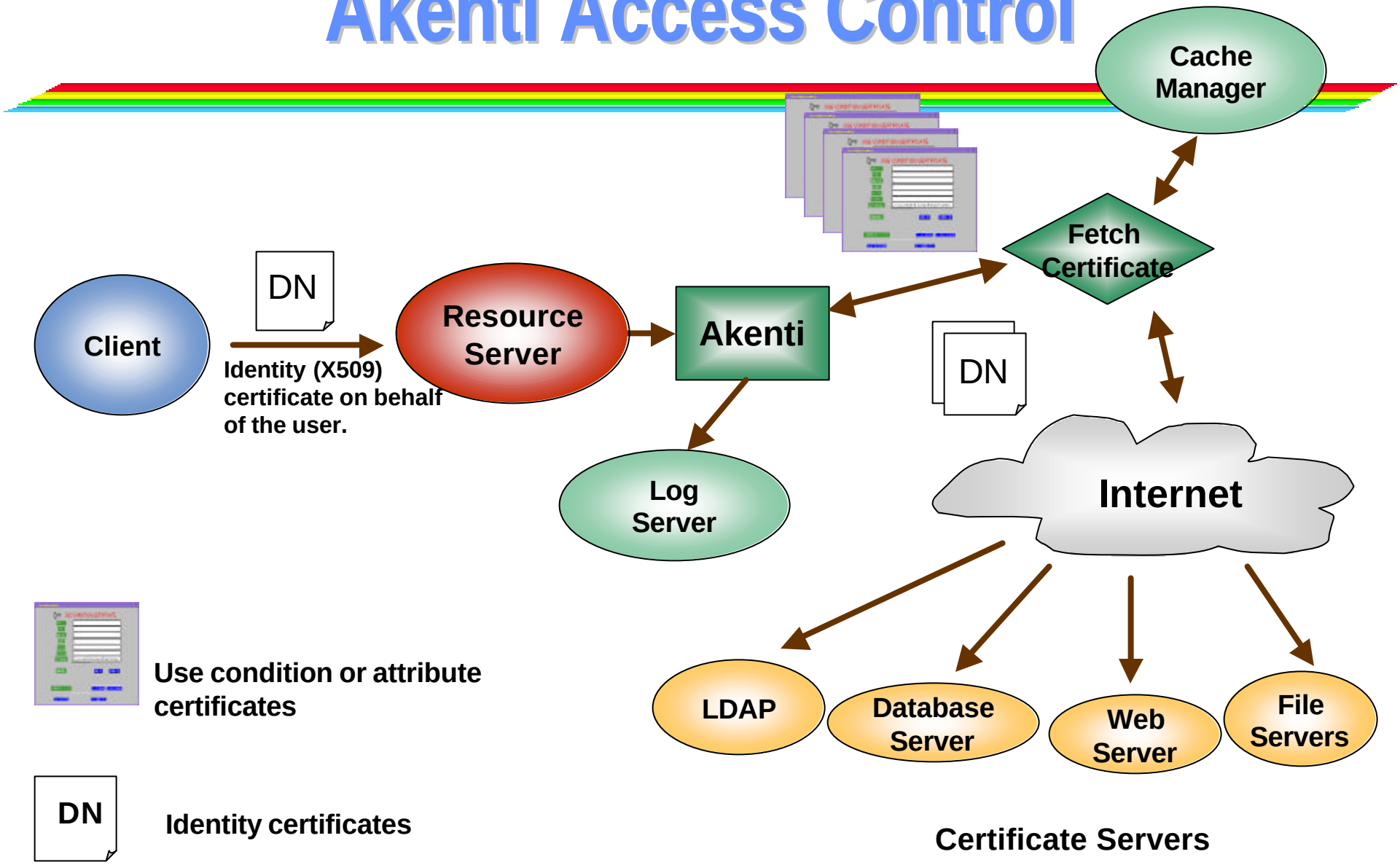
Achieved 35 MBytes/sec. of user data delivered to the STAF application

2/9/99 - RLN

Resource Management

- Grid Storage API
  - New Globus component
  - Joint work involving ANL, LBNL, USC/ISI and SDSC
  - Single API for file and block level access to NFS, HPSS, DPSS, FTP, and HTTP
  - Design for high-speed transfers
- Global Naming Service
  - Derived from SRB and LDAP
- Resource Reservation System
  - Globus GARA
  - Reserve networks, disk caches, PC clusters, etc.

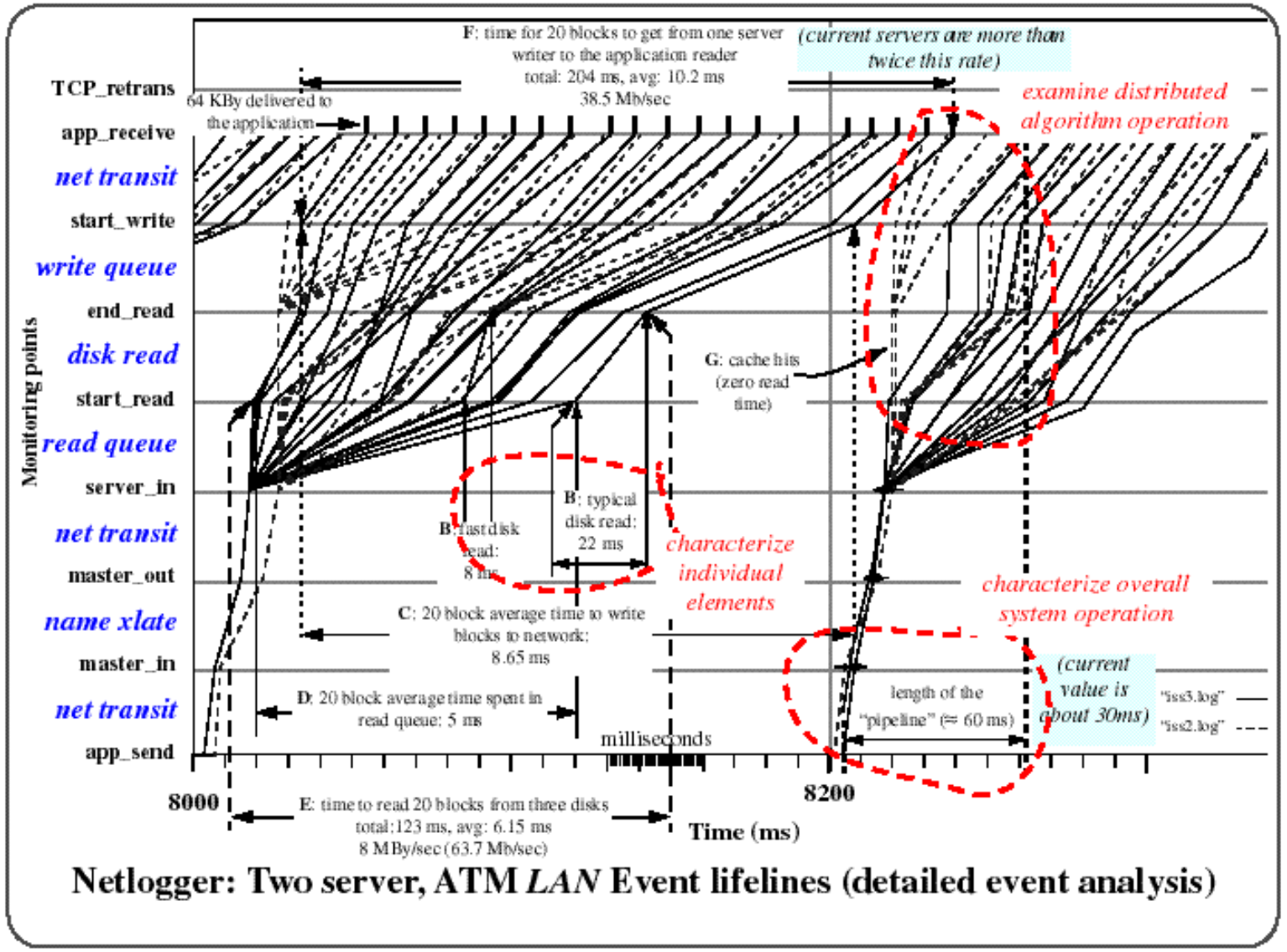
Akenti Access Control



Stream Management

- Image-Based Rendering Assisted Volume Rendering
  - High B/W critical
- Control channel to manipulate images
  - Low latency and reliability are critical
- Video of collaborators
- Audio from collaborators
- Synchronization of the above

Monitoring Architecture



Netlogger: Two server, ATM *LAN* Event lifelines (detailed event analysis)

Part of a Broader Collaboration

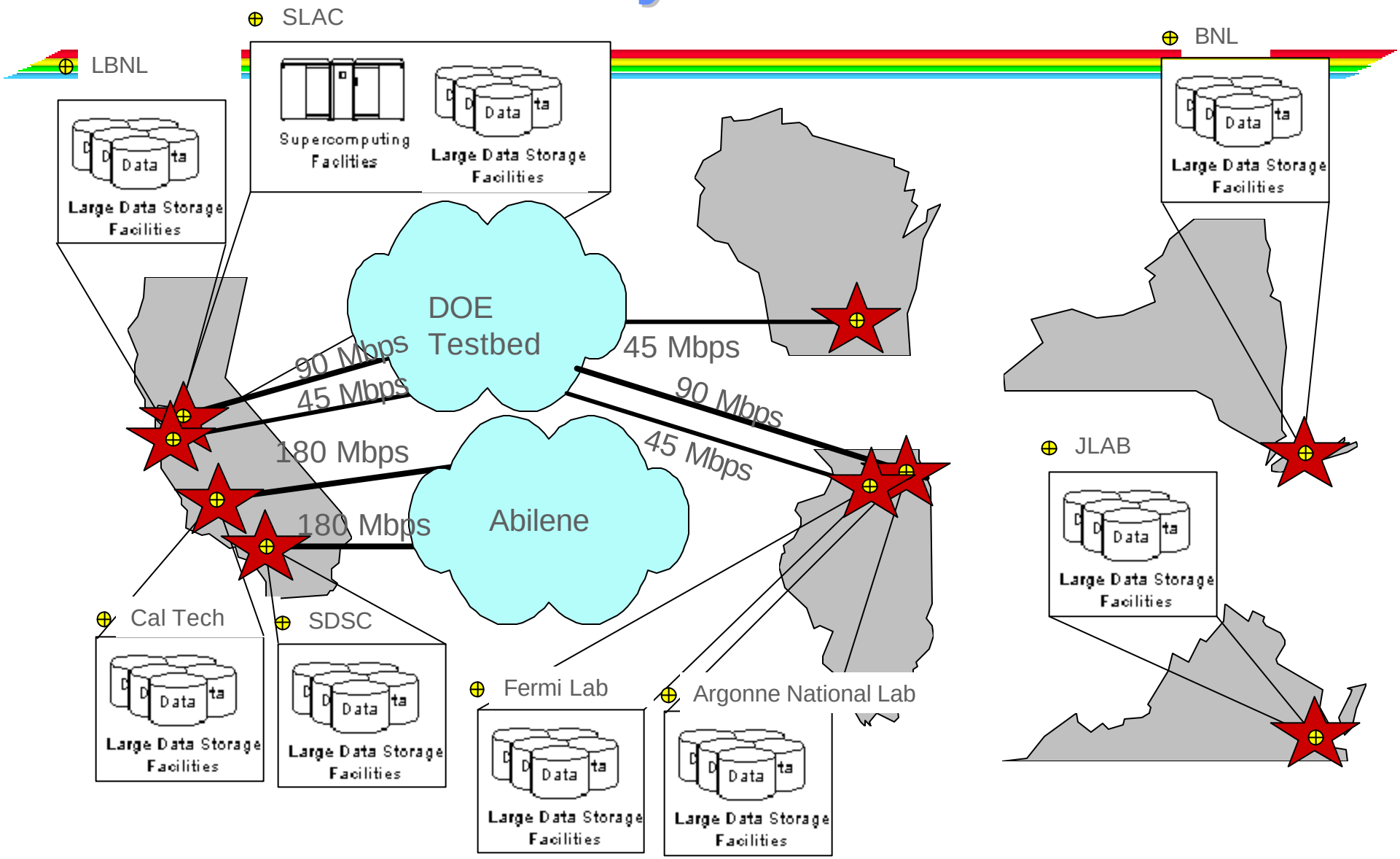
- DARPA Supernet
- Related DOE NGI Applications Efforts
  - Corridor One
  - Earth Systems Grid
  - Partical Physics Data Grid
- Related DOE NGI Network Research Efforts
  - Advanced Visualization Toolkit
  - ENABLE
  - Secure and Reliable Group Communications
- Related Network Testbeds
  - ESnet (China Clipper)
  - MREN (EMERGE)
  - NTON

Particle Physics Data Grid

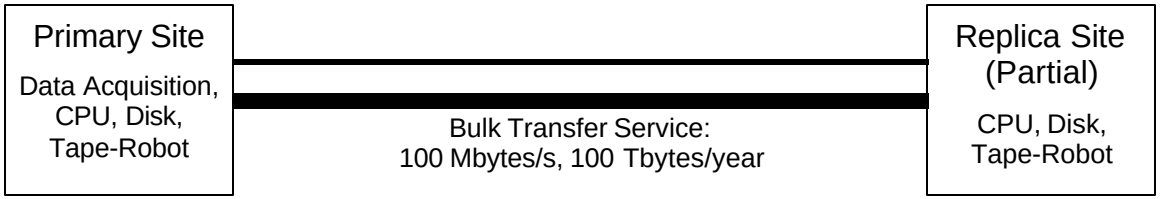
Universities, DOE Accelerator Labs, DOE Computer Science

- Particle Physics: a Network-Hungry Collaborative Application
  - Petabytes of compressed experimental data;
  - Nationwide and worldwide university-dominated collaborations analyze the data;
  - Close DOE-NSF collaboration on construction and operation of most experiments;
  - The PPDG lays the foundation for lifting the network constraint from particle-physics research.
- Short-Term Targets:
  - High-speed site-to-site replication of newly acquired particle-physics data (> 100 Mbytes/s);
  - Multi-site cached file-access to thousands of ~10 Gbyte files.

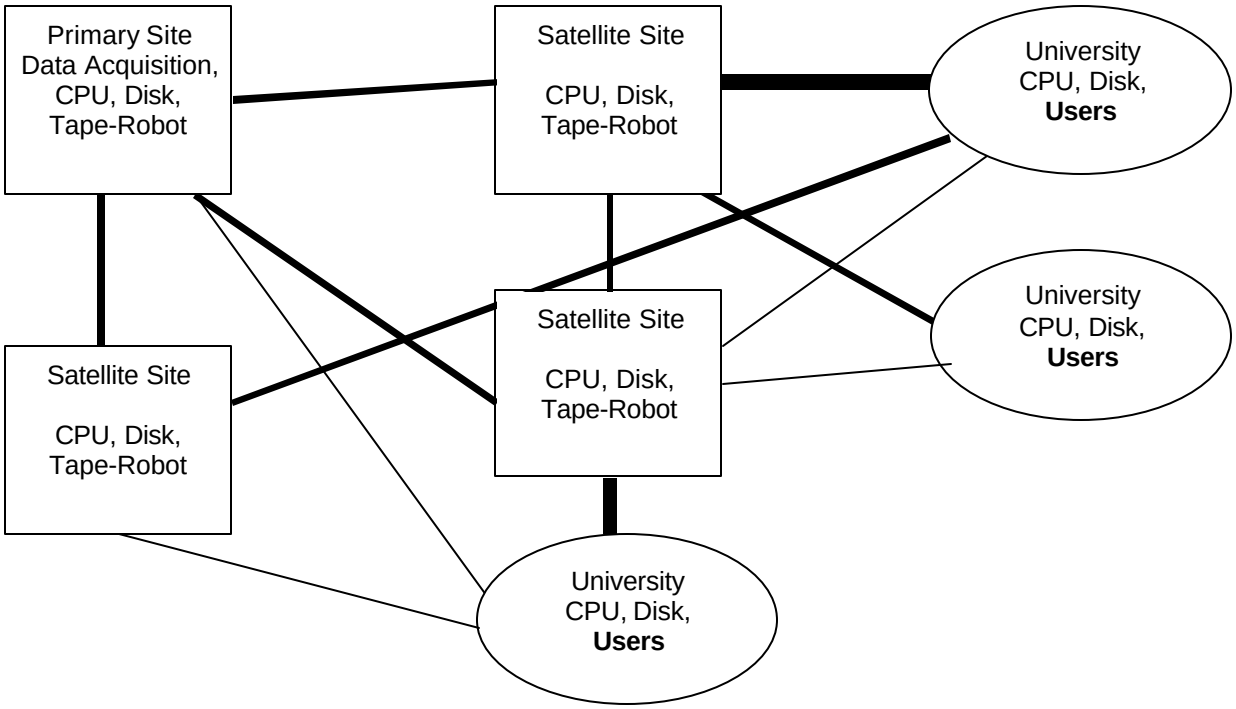
Particle Physics Data Grid



High-Speed Site-to-Site File Replication Service



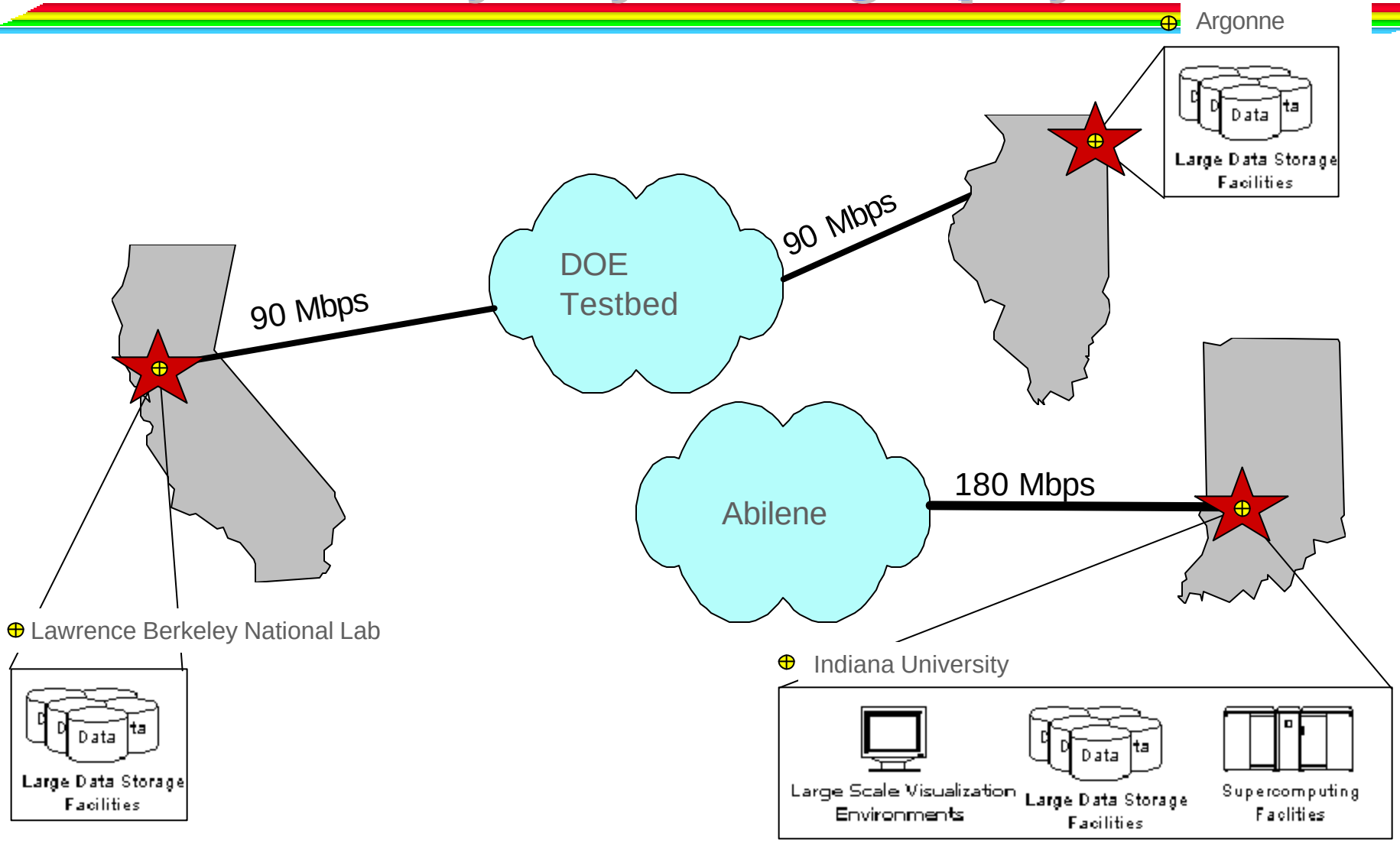
Multi-Site Cached File Access



PPDG Resources

- Network Testbeds:
  - ESNET links at up to 622 Mbits/s (e.g. LBL-ANL)
  - Other testbed links at up to 2.5 Gbits/s (e.g. Caltech-SLAC via NTON)
- Data and Hardware:
  - Tens of terabytes of disk-resident particle physics data (plus hundreds of terabytes of tape-resident data at accelerator labs);
  - Dedicated terabyte university disk cache;
  - Gigabit LANs at most sites.
- Middleware Developed by Collaborators:
  - Many components needed to meet short-term targets (e.g.Globus, SRB, MCAT, Condor,OOFS,Netlogger, Grand-Challenge Cache Manager, Mass Storage Management) already developed by collaborators.
- Existing Achievements of Collaborators:
  - WAN transfer at 57 Mbytes/s;
  - Single site database access at 175 Mbytes/s

Xray Crystallography



Slide/slides to present recent QoS results--ANL internal testbed, then extended to LBNL

DOE NGI Applications & Network Resources

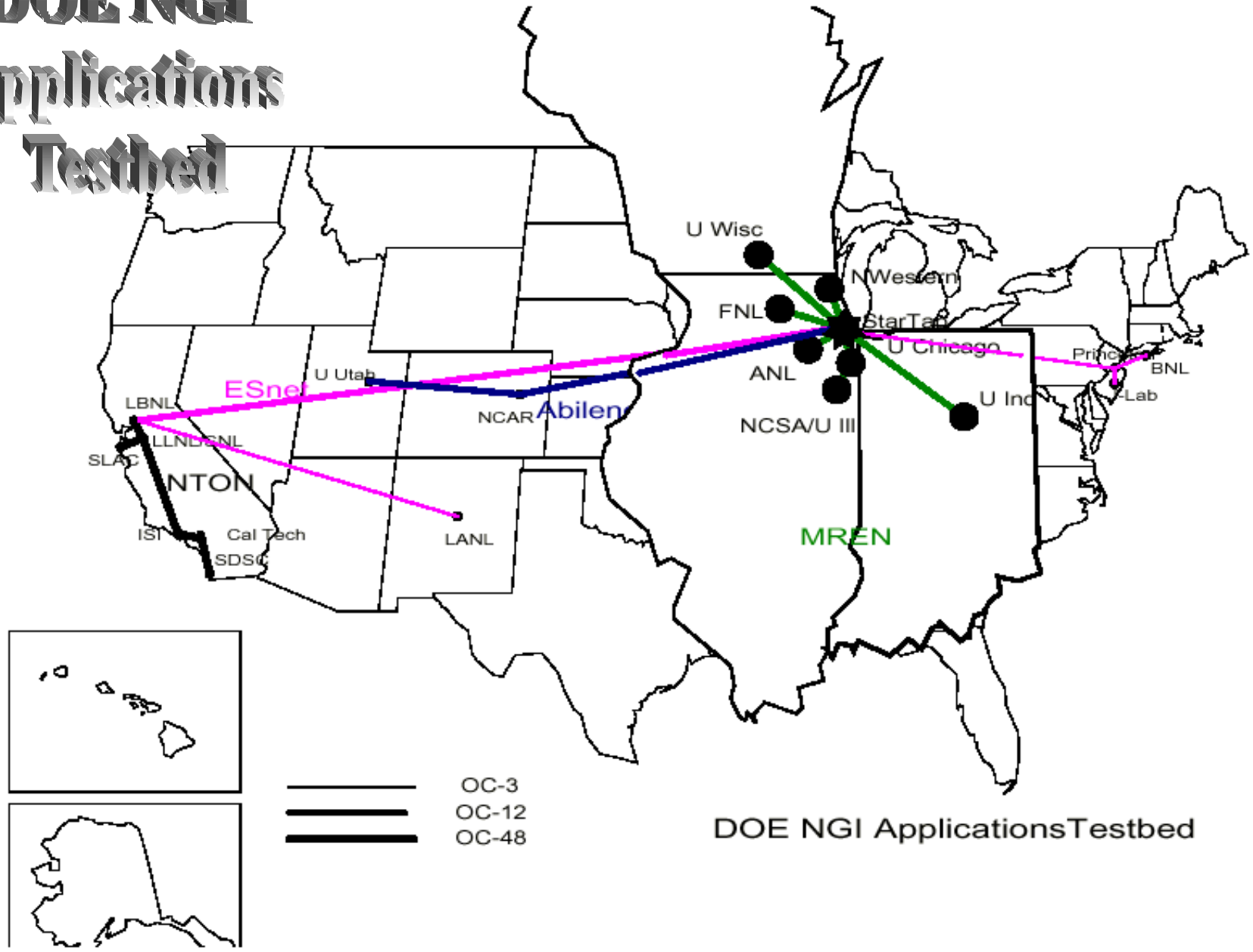
Applications

- **Combustion**
  - remote visualization; 100 Mbps
  - huge bandwidth needs; 1Gbps
  - management and access to large distributed scientific database
- **Corridor One**
  - multipoint distance visualization
  - latency and jitter sensitive
  - 15 Mbps
  - multicast required
  - inter-flow correlation
- **Earth Systems Grid**
  - movement, caching, location of large scientific datasets
  - on-demand discovery and scheduling of caches, networks, and computers
  - 50 Mbps
- **Particle Physics Data Grid**
  - bulk transfer of raw data to and reconstructed data between regional data centers
  - interactive access to analyzed events
- **Xray Crystallography**
  - collect/process/vis detector data
  - 16 Mbps

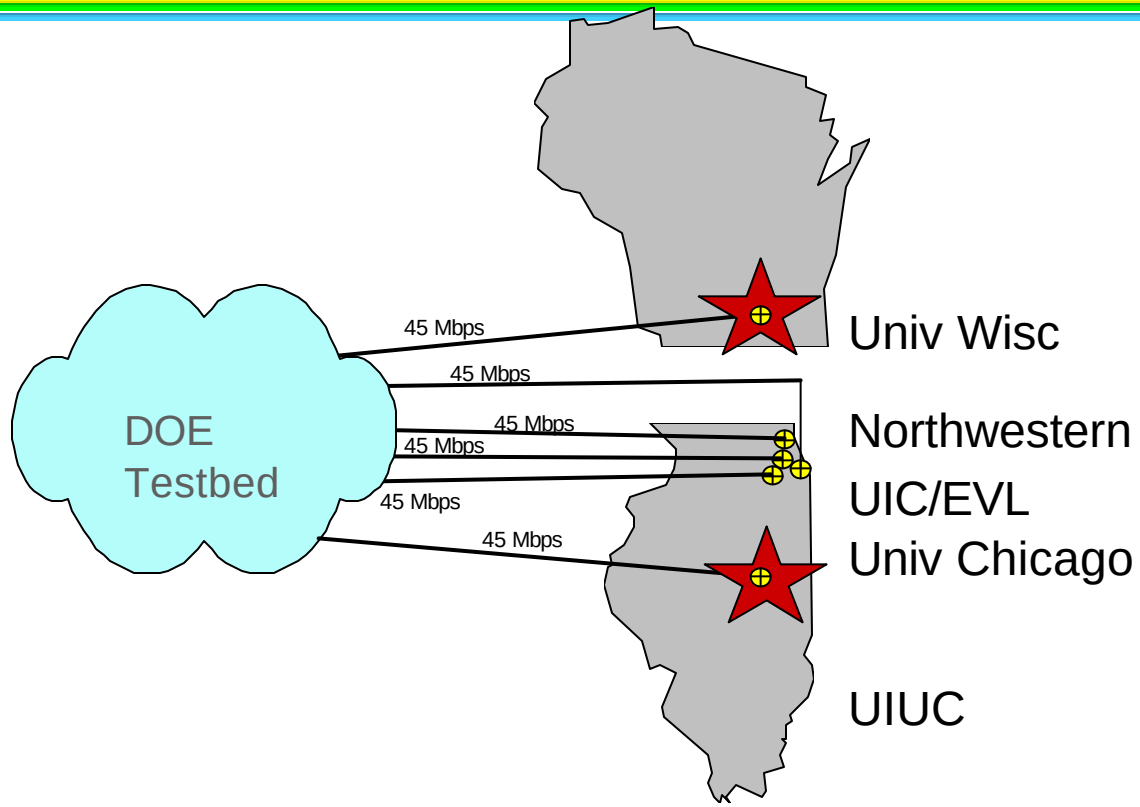
Networks

- **ESnet (ANL, LBNL, LLNL, SNL, SLAC, JLAB, FNAL, BNL)**
- **Abilene (Caltech, SDSC, NCAR, Indiana, Utah, Princeton, ISI)**
- **MREN (Wisconsin, UIC/EVL, Northwestern, UChicago, UIllinois)**

DOE NGI Applications Testbed

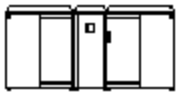


EMERGE



Large Scale Visualization Environments

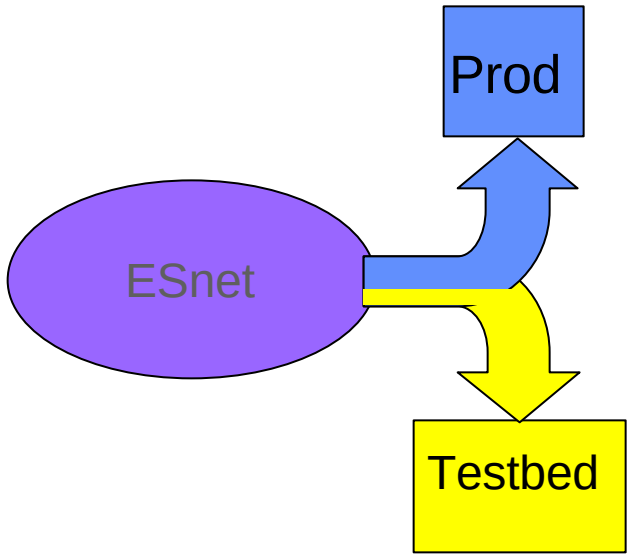
Large Data Storage Facilities



Supercomputing Facilities

Planned DOE/NGI Testbed Capabilities

- Allocate bandwidth out of production ESnet service (30% of access link; 45-90 Mbps per site)
- Testbed will carry Best Effort + Premium traffic
- Resource control/allocation/scheduling
- Provide instrumentation assistance for applications to drive toward adaptive applications



Benefits to Applications

- Identify & address applications requirements bandwidth, latency, jitter, multicast
- Creation of intelligent networks & network aware applications
- Resource Manager (reservations) to handle multiple interdomain flows for a single application
- Identify and implement measurement/performance data for applications

Slide/slides with results and issues from PI meeting (held Oct 4-5)

DOE NGI Program

- For more information:
  - Program information:
    - http://www.er.doe.gov/production/octr/mics/NGI.HTML
  - Testbeds and applications:
    - http://www-itg.lbl.gov/NGI/private/
  - PI meeting information, presentations, etc.
    - *http://www.lbl.gov/CS/NGI/Events/Oct_PImtg/*