TeraGrid Communication and Computation

Tal Lavian
tlavian@cs.berkeley.edu

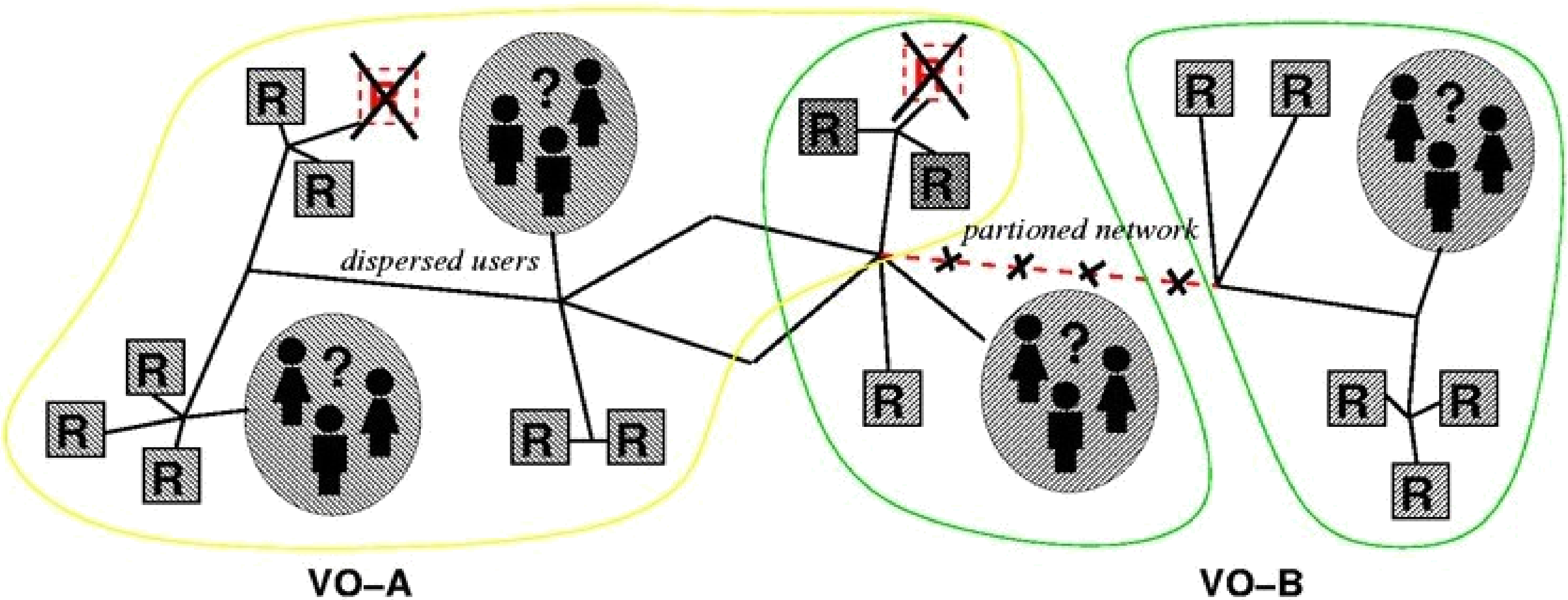
Brainstorm and concepts
Many slides and most of the graphics are taken from other slides

Agenda

- Introduction
- Some applications
- TeraGrid Architecture
- Globus toolkit
- Future comm direction
- Summary

The Grid Problem

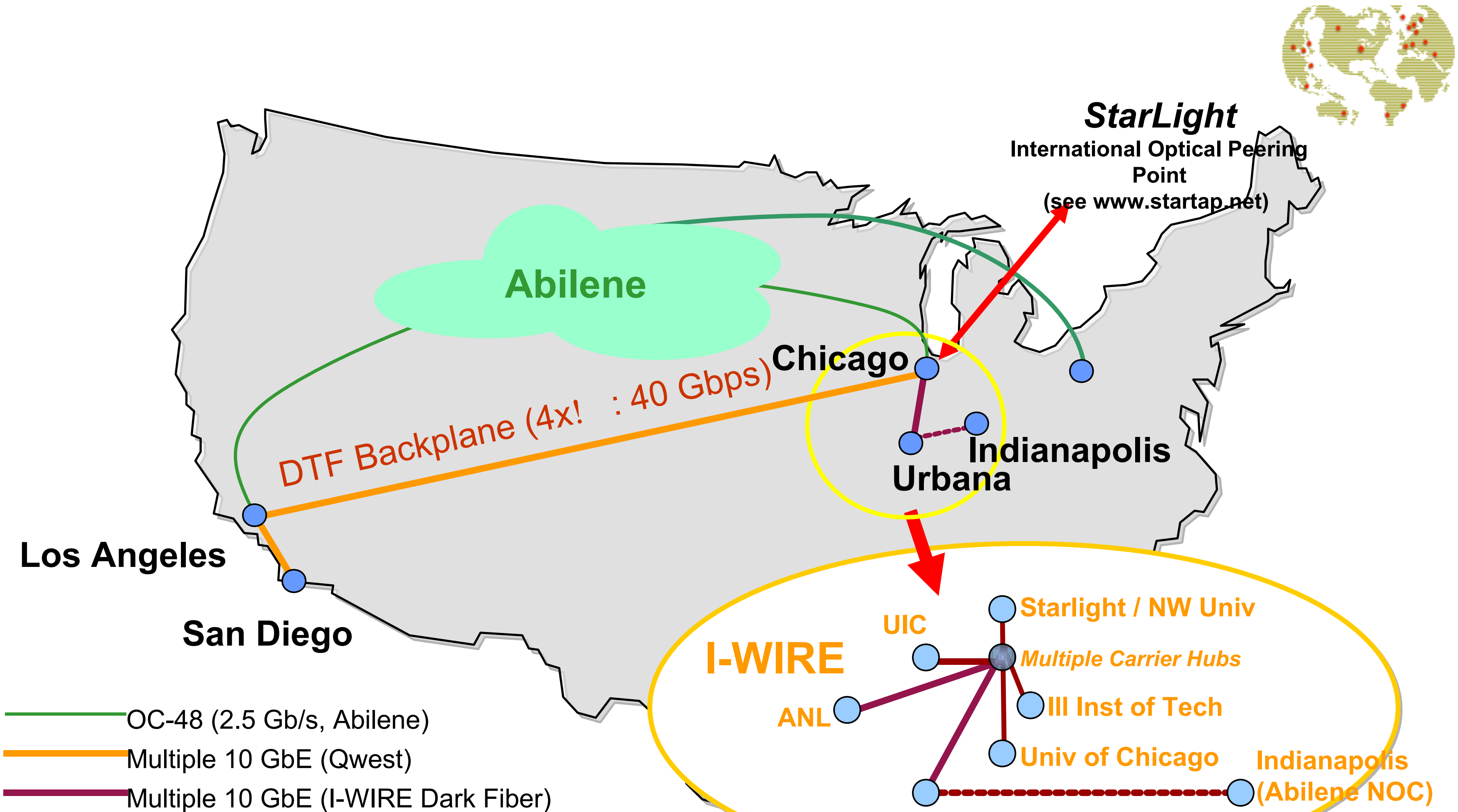
Resource sharing & coordinated problem solving in dynamic, multi-institutional virtual organizations



Some relation to Sahara

- **Service composition: computation, servers, storage, disk, network…**
- **Sharing, cooperating, peering, brokering…**

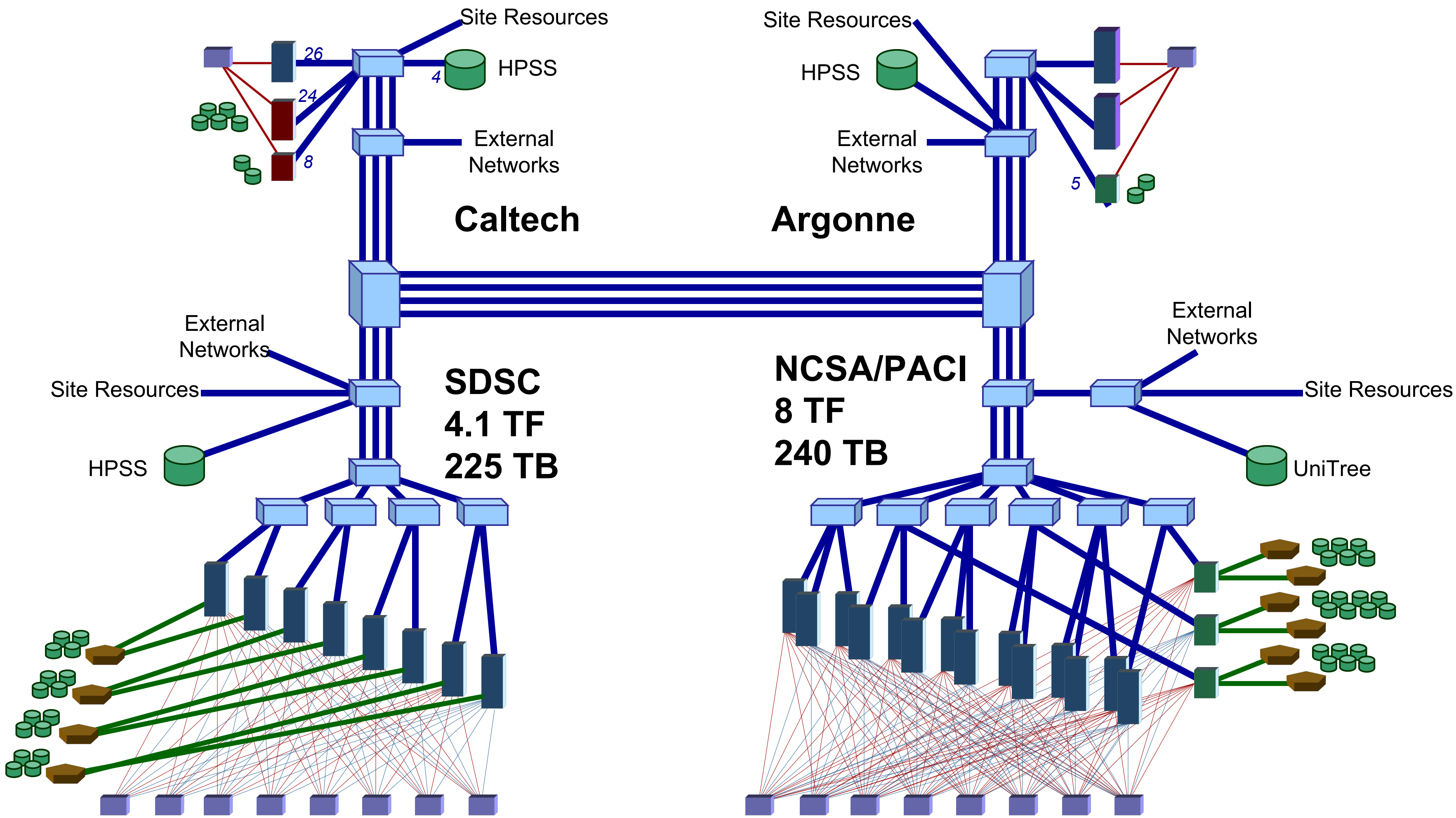
TeraGrid Wide Area Network - NCSA, ANL, SDSC, Caltech



- Solid lines in place and/or available by October 2001
- Dashed I-WIRE lines planned for summer 2002

Source: Charlie Catlett, Argonne

The 13.6 TF TeraGrid: Computing at 40 Gb/s



TeraGrid/DTF: NCSA, SDSC, Caltech, Argonne www.teragrid.org

4 TeraGrid Sites Have Focal Points

SDSC – The Data Place

- Large-scale and high-performance data analysis/handling
- Every Cluster Node is Directly Attached to SAN

NCSA – The Compute Place

- Large-scale, Large Flops computation

Argonne – The Viz place

- Scalable Viz walls

Caltech – The Applications place

- Data and flops for applications – Especially some of the GriPhyN Apps

Specific machine configurations reflect this

TeraGrid building blocks

Distributed, multisite facility

- single site and "Grid enabled" capabilities
  - **uniform compute node selection and interconnect networks at 4 sites**
  - **central "Grid Operations Center"**
- at least one 5+ teraflop site and newer generation processors
  - **SDSC at 4+ TF, NCSA at 6.1-8 TF with McKinley processors**
- at least one additional site coupled with the first
  - **four core sites: SDSC, NCSA, ANL, and Caltech**

Ultra high-speed networks (Static configured)

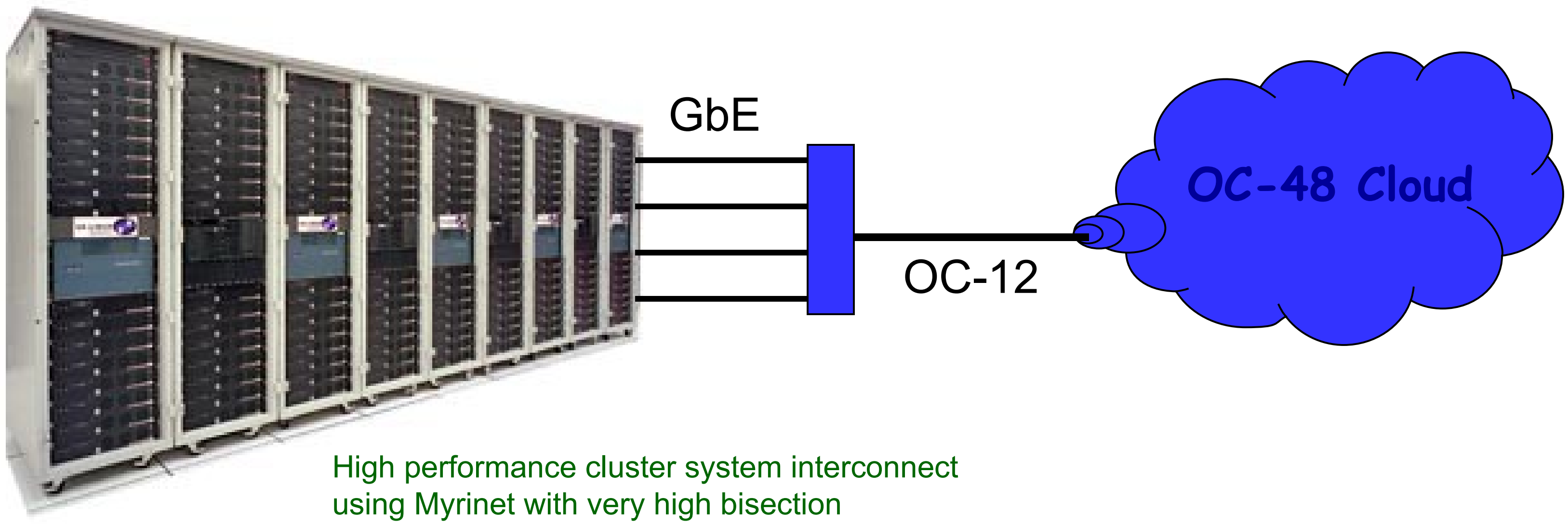
- multiple gigabits/second
  - **modular 40 Gb/s backbone (4 x 10 GbE)**

Remote visualization

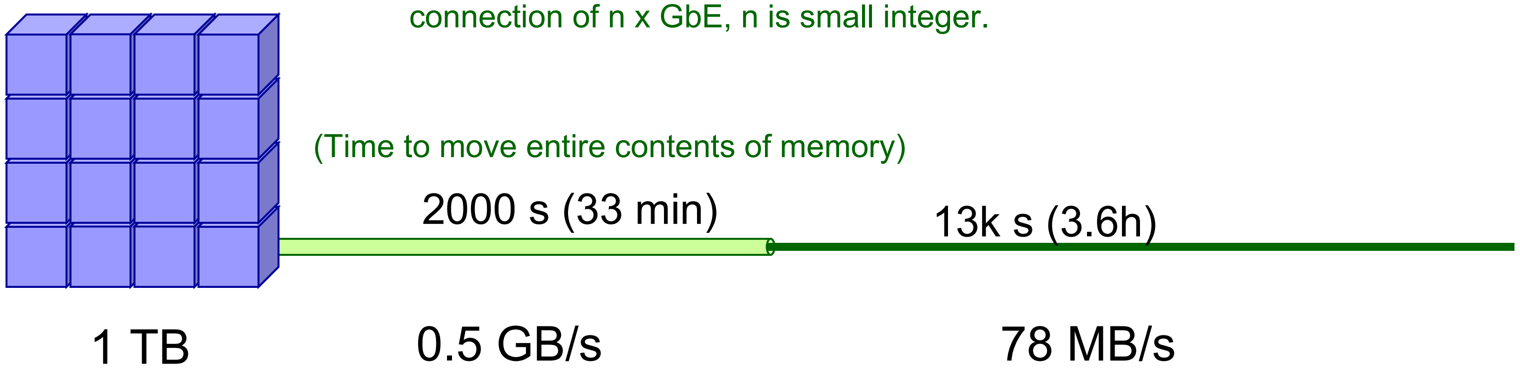
- data from one site visualized at another
  - **high-performance commodity rendering and visualization system**
  - **Argonne hardware visualization support**
  - **data serving facilities and visualization displays**

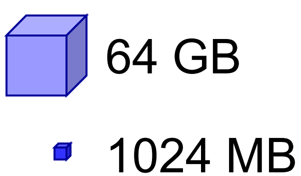
NSF - $53M award in August 2001

Traditional Cluster Network Access



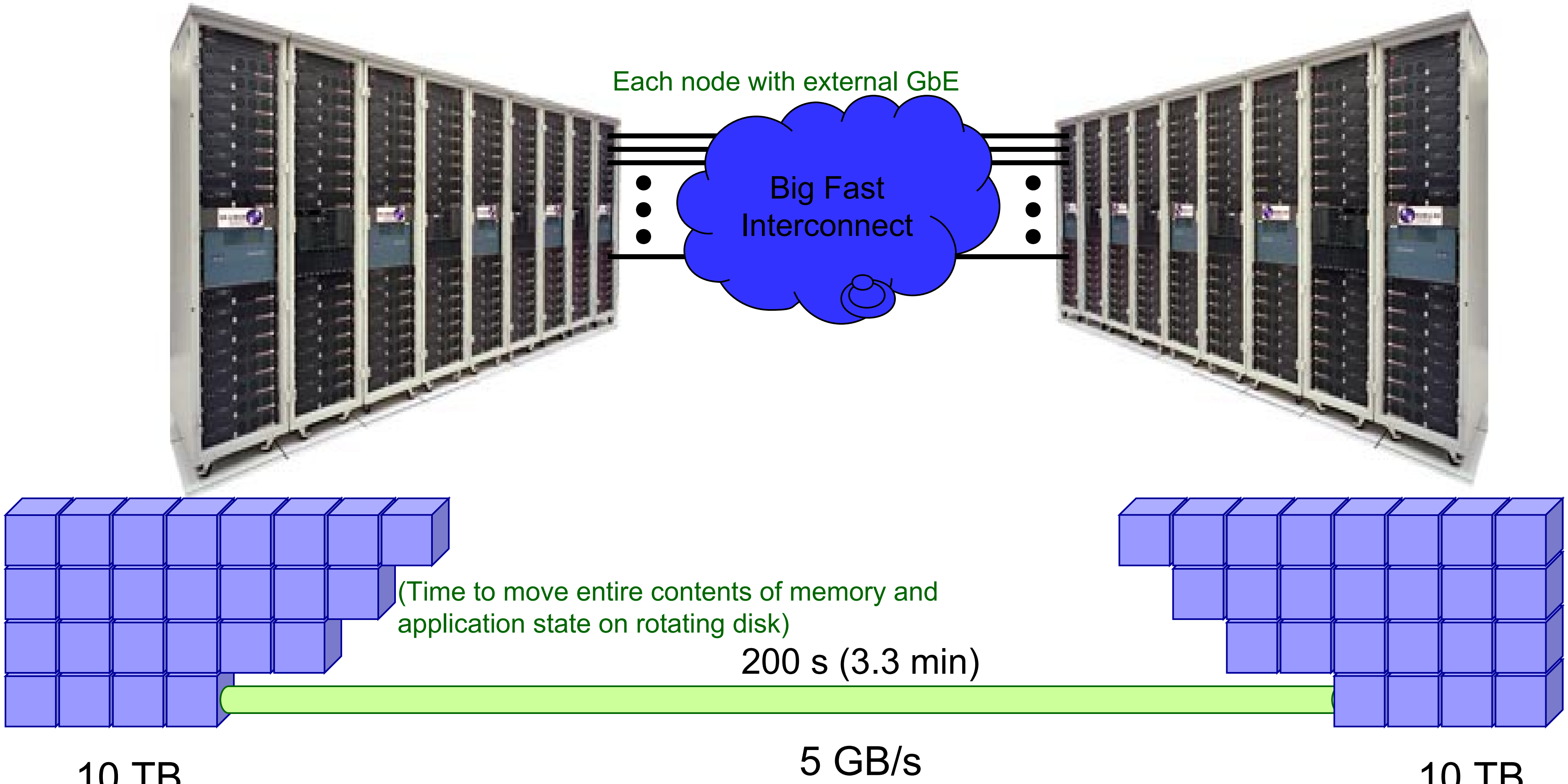
High performance cluster system interconnect using Myrinet with very high bisection bandwidth (hundreds of GB/s) with external connection of n x GbE, n is small integer.



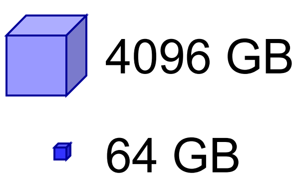


Traditionally, high-performance computers have been islands of capability separated by wide area networks that provide a fraction of a percent of the internal cluster network bandwidth.

To Build a Distributed Terascale Cluster…



5 GB/s = 200 nodes x 25 MB/s (=20% of GbE per node)



TeraGrid is building a "machine room" network across the country while increasing external cluster bandwidth to many GbE. Requires edge systems that handle n x 10 GbE and hubs that handle minimum 10 x 10 GbE.

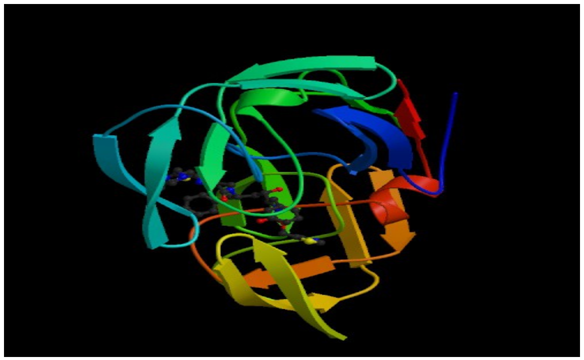
Agenda

- Introduction
- Some applications
- TeraGrid Architecture
- Globus toolkit
- Future comm direction
- Summary

What applications are being targeted for Grid-enabled computing? Traditional

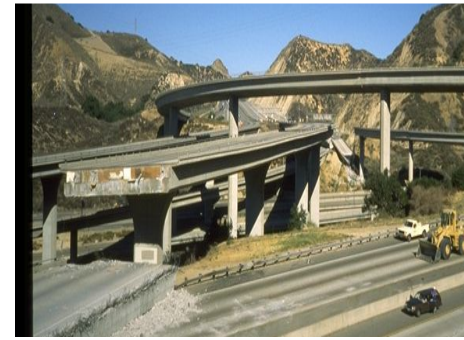
- Quantum Chromodynamics
- Biomolecular Dynamics
- Weather Forecasting
- Cosmological Dark Matter
- Biomolecular Electrostatics
- Electric and Magnetic Molecular Properties

Beginning of the Digital Millennium: The Data Decade!



Genomics

Sensors



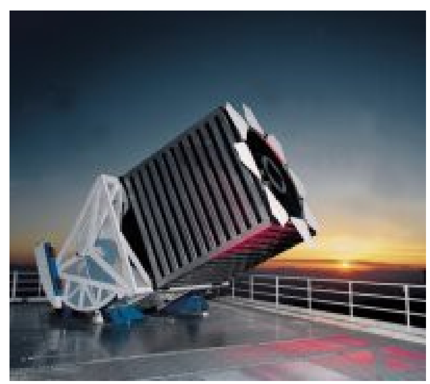
Disaster response

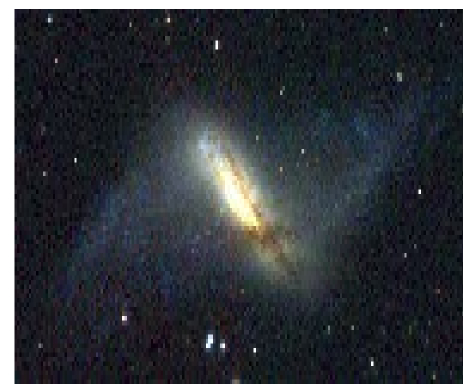
GriPhyN

Data Intensive Science

Physics

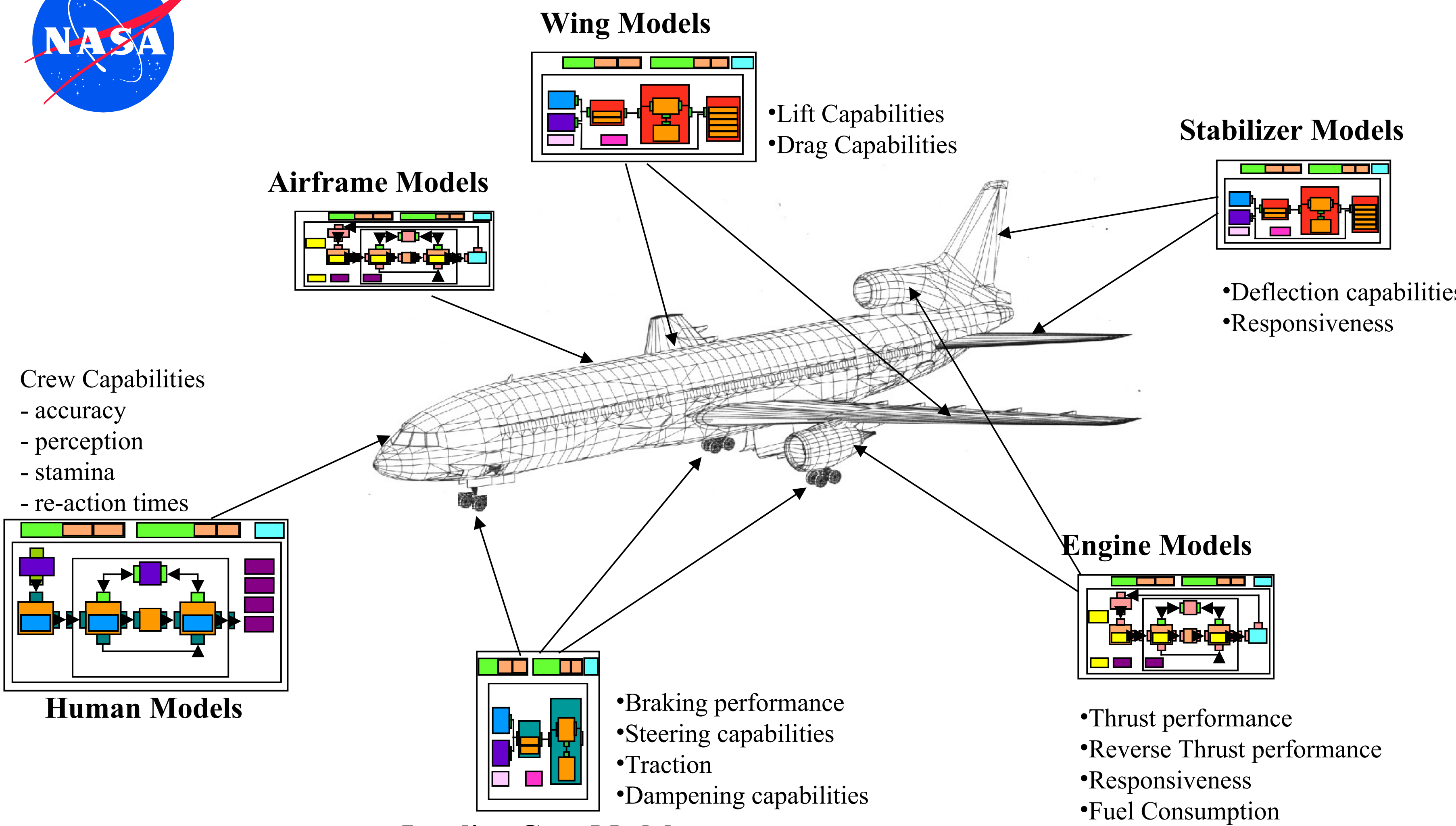
Digital Libraries





Astronomy

Multi-disciplinary Simulations: Aviation Safety



Source NASA

Whole system simulations are produced by coupling all of the sub-system simulations

New Results Possible on TeraGrid

Biomedical Informatics Research Network (National Inst. Of Health):

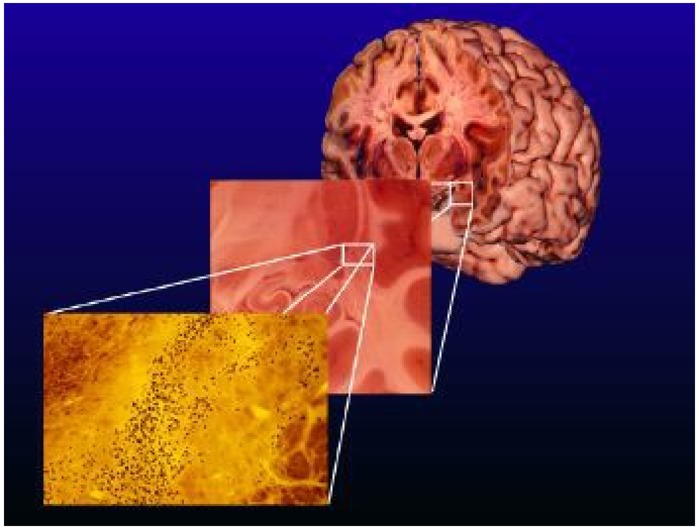
Evolving reference set of brains provides essential data for developing therapies for neurological disorders (Multiple Sclerosis, Alzheimer's, etc.).

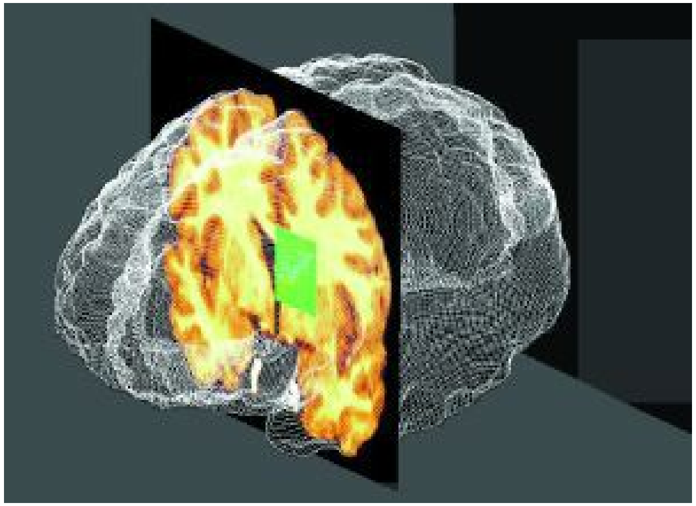
Pre-TeraGrid:

- One lab
- Small patient base
- 4 TB collection

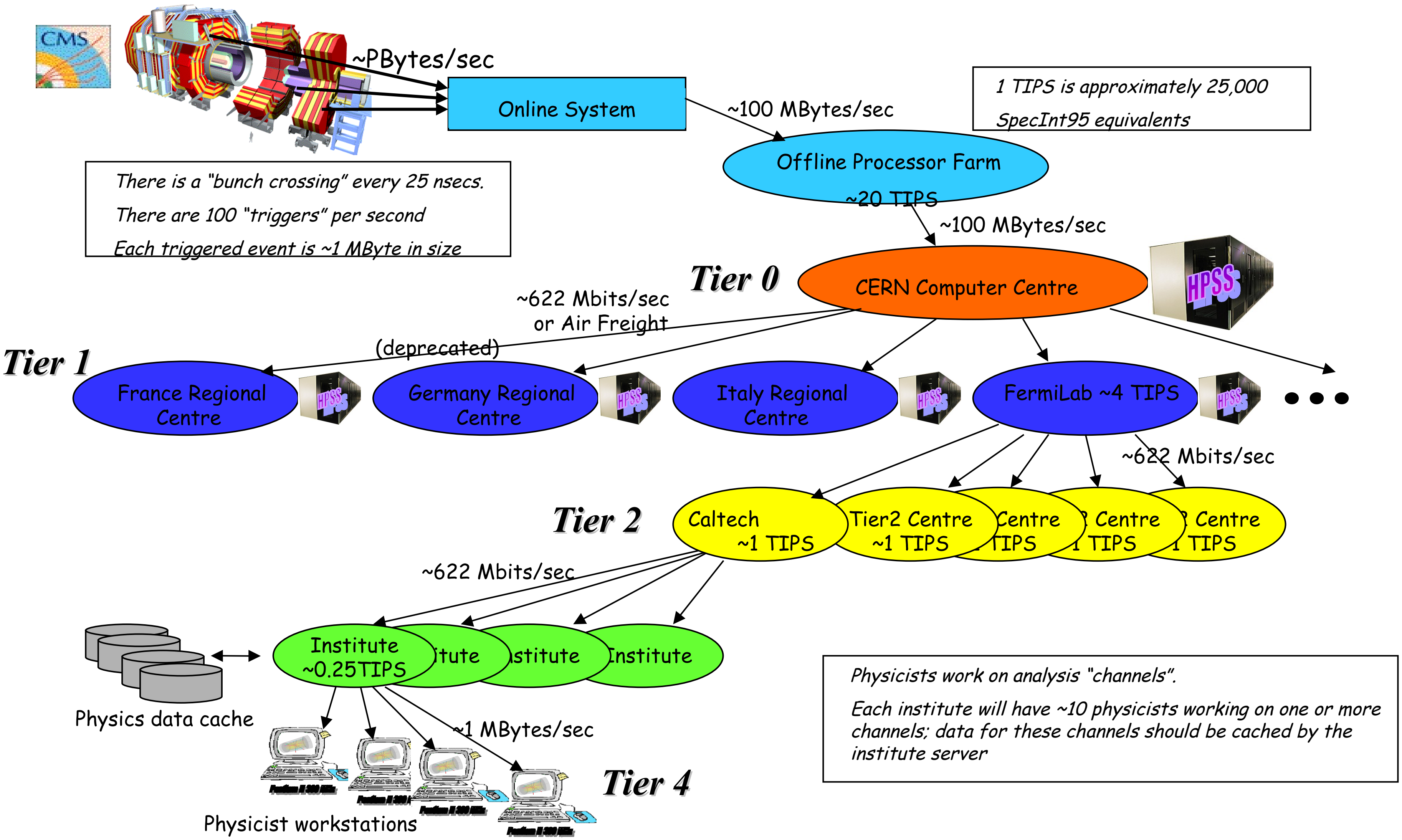
Post-TeraGrid:

- Tens of collaborating labs
- Larger population sample
- 400 TB data collection: more brains, higher resolution
- Multiple scale data integration and analysis





Grid Communities & Applications: Data Grids for High Energy Physics



Source Harvey Newman, Caltech

Agenda

- Introduction
- Some applications
- TeraGrid Architecture
- Globus toolkit
- Future comm direction
- Summary

Grid Computing Concept

- **New applications enabled by the coordinated use of geographically distributed resources**
  - E.g., distributed collaboration, data access and analysis, distributed computing
- **Persistent infrastructure for Grid computing**
  - E.g., certificate authorities and policies, protocols for resource discovery/access
- **Original motivation, and support, from high-end science and engineering; but has wide-ranging applicability**

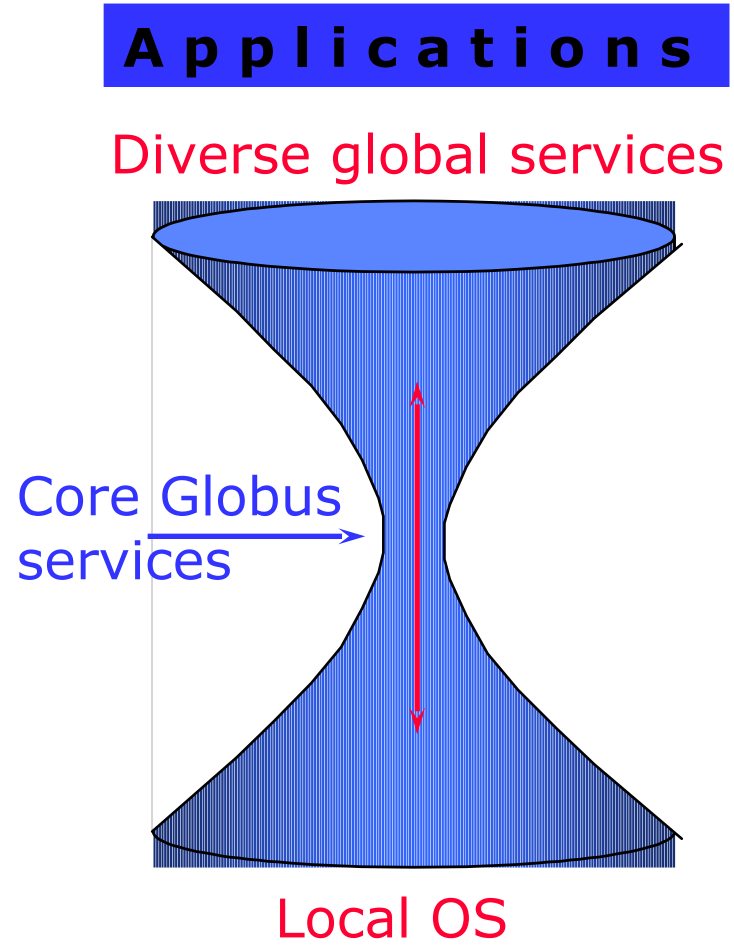
Globus Hourglass

Focus on architecture issues

- Propose set of core services as basic infrastructure
- Use to construct high-level, domain-specific solutions

Design principles

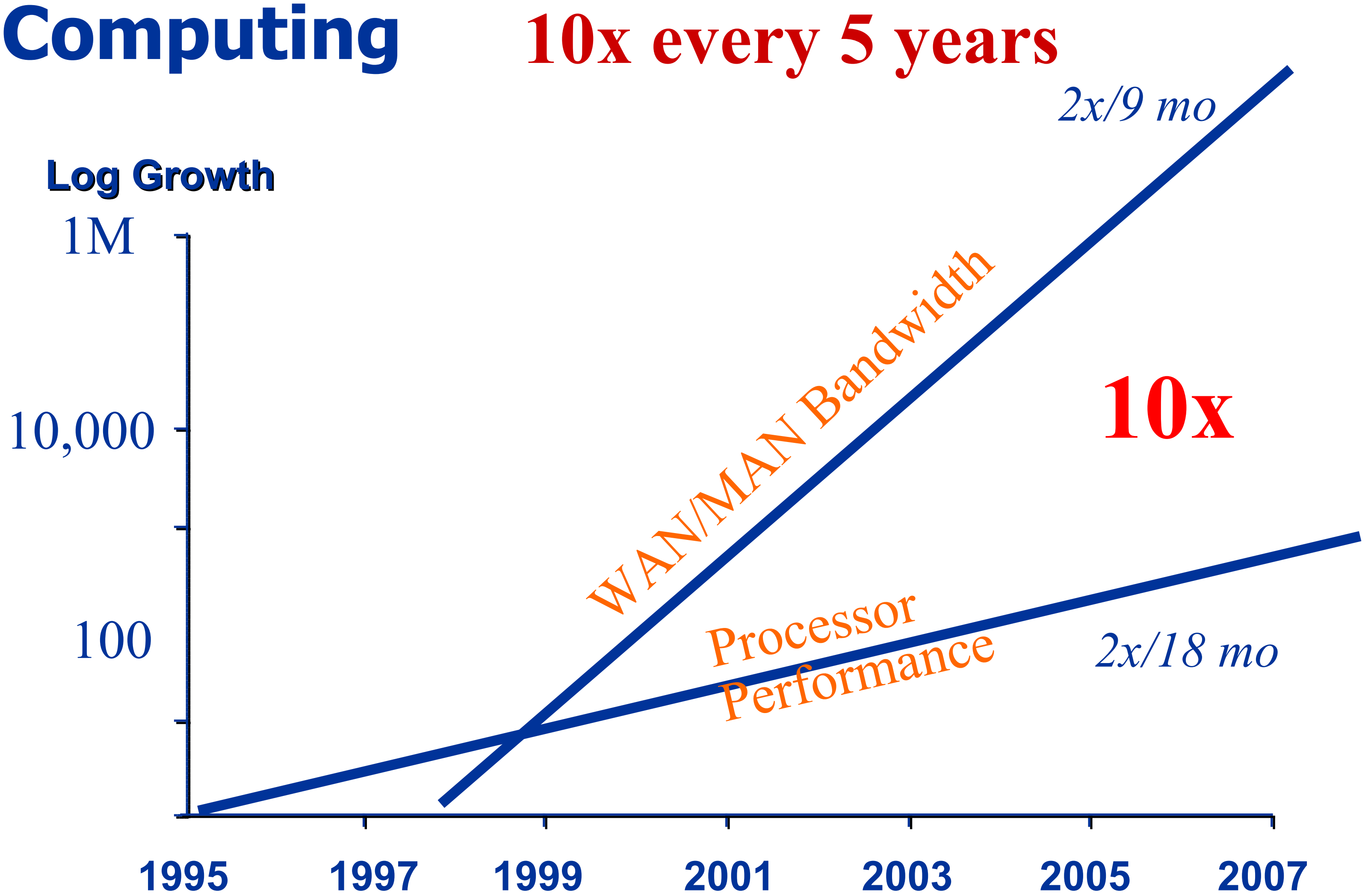
- Keep participation cost low
- Enable local control
- Support for adaptation
- “IP hourglass” model



Elements of the Problem

- **Resource sharing**
  - Computers, storage, sensors, networks, …
  - Sharing always conditional: issues of trust, policy, negotiation, payment, …
- **Coordinated problem solving**
  - Beyond client-server: distributed data analysis, computation, collaboration, …
- **Dynamic, multi-institutional virtual orgs**
  - Community overlays on classic org structures
  - Large or small, static or dynamic

Gilder vs. Moore – Impact on the Future of Computing



Improvements in Large-Area Networks

Network vs. computer performance

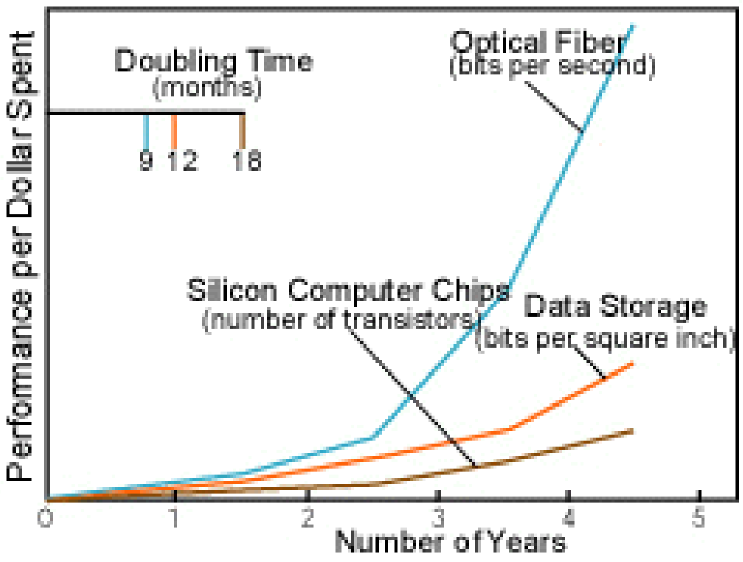
- Computer speed doubles every 18 months
- Network speed doubles every 9 months
- Difference = order of magnitude per 5 years

1986 to 2000

- Computers: x 500
- Networks: x 340,000

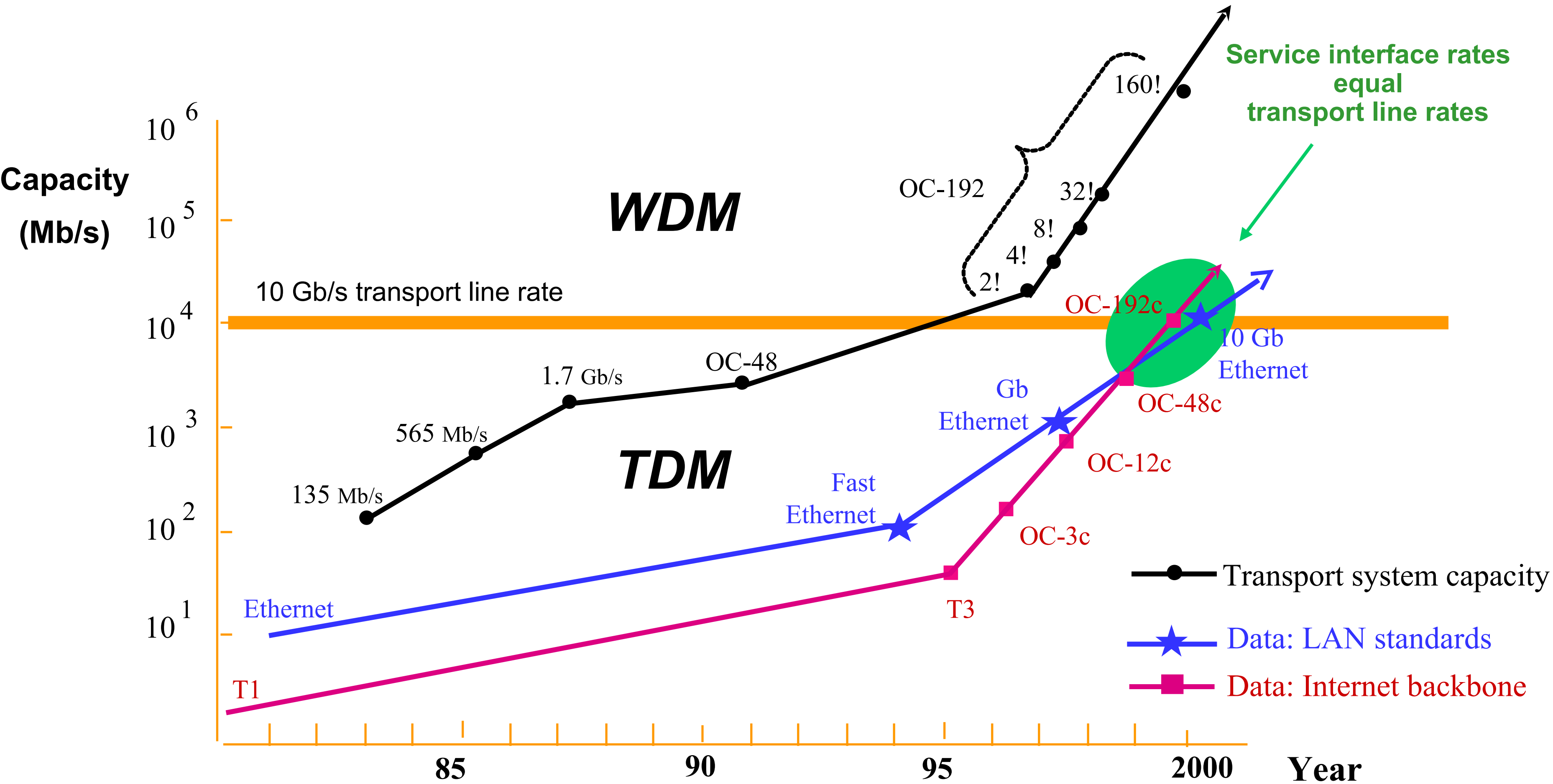
2001 to 2010

- Computers: x 60
- Networks: x 4000



Moore's Law vs. storage improvements vs. optical improvements. Graph from **Scientific American** (Jan-2001) by Cleo Vilett, source Vined Khoslan, Kleiner, Caufield and Perkins.

Evolving Role of Optical Layer



Source: IBM WDM research

Scientific Software Infrastructure

One of the Major Software Challenges

Peak Performance is skyrocketing (more than Moore's Law)

but ...

- **Efficiency has declined from 40-50% on the vector supercomputers of 1990s to as little as 5-10% on parallel supercomputers of today and may decrease further on future machines**

Research challenge is software

- **Scientific codes to model and simulate physical processes and systems**
- **Computing and mathematics software to enable use of advanced computers for scientific applications**
- **Continuing challenge as computer architectures undergo fundamental changes: *Algorithms that scale to thousands-millions processors***

Agenda

- Introduction
- Some applications
- TeraGrid Architecture
- Globus toolkit
- Future comm direction
- Summary

Globus Approach

A toolkit and collection of services addressing key technical problems

- Modular "bag of services" model
- Not a vertically integrated solution
- General infrastructure tools (aka middleware) that can be applied to many application domains

Inter-domain issues, rather than clustering

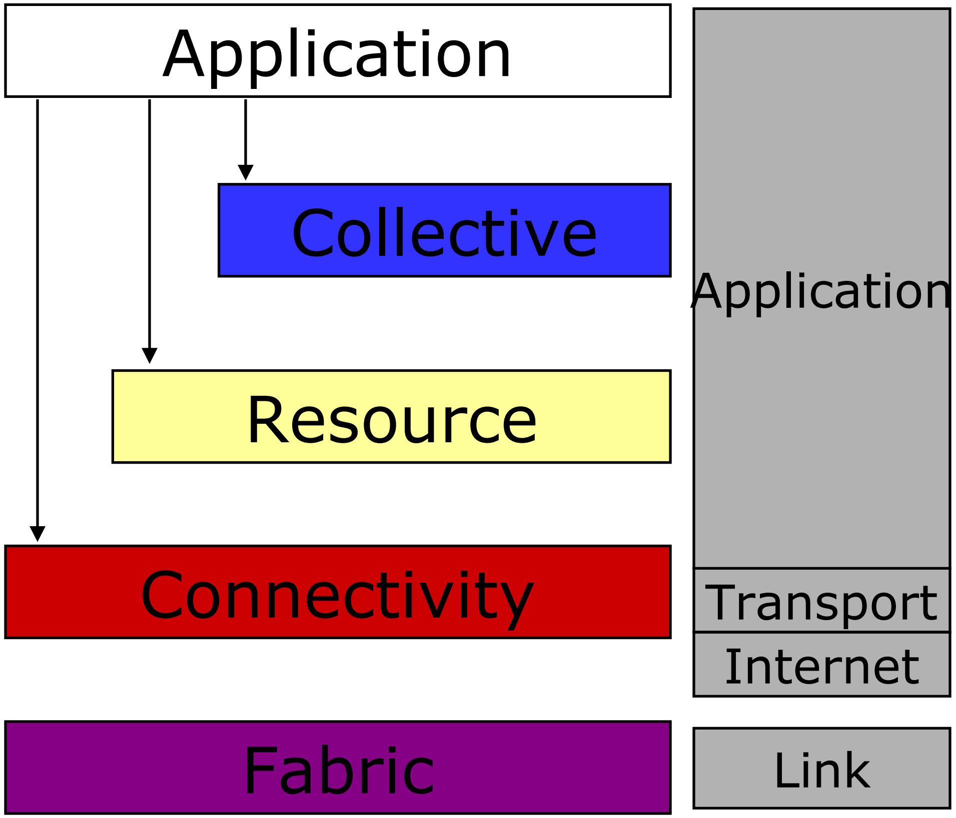
- Integration of intra-domain solutions

Distinguish between local and global services

Globus Technical Focus & Approach

- **Enable incremental development of grid-enabled tools and applications**
  - *Model neutral*: Support many programming models, languages, tools, and applications
  - Evolve in response to user requirements
- **Deploy toolkit on international-scale production grids and testbeds**
  - Large-scale application development & testing
- **Information-rich environment**
  - Basis for configuration and adaptation

Layered Grid Architecture (By Analogy to Internet Architecture)



"Coordinating multiple resources": ubiquitous infrastructure services, app-specific distributed services

"Sharing single resources": negotiating access, controlling use

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

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

For more info: www.globus.org/research/papers/anatomy.pdf

Globus Architecture?

No "official" standards exist

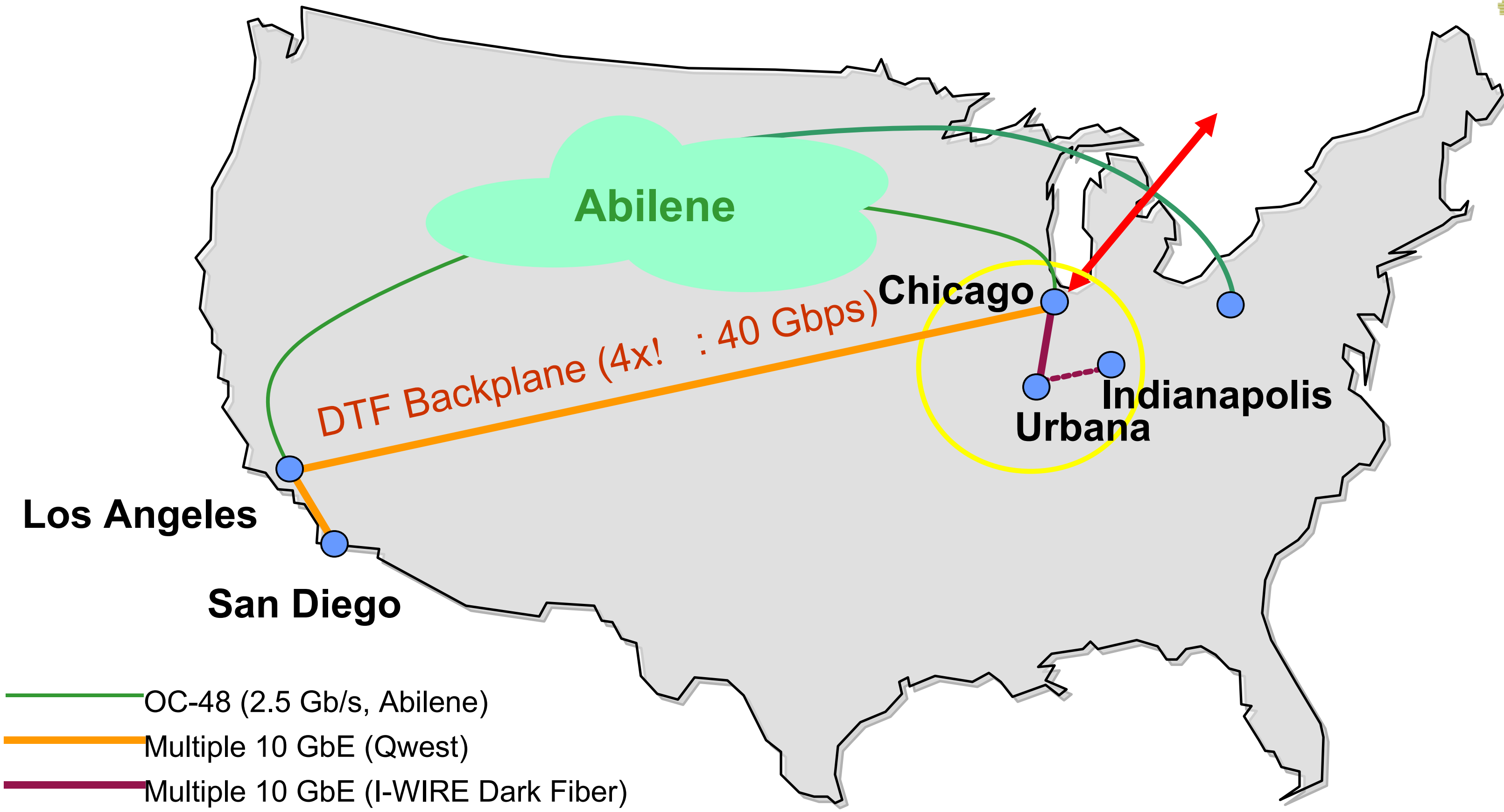
But:

- Globus Toolkit has emerged as the de facto standard for several important Connectivity, Resource, and Collective protocols
- Technical specifications are being developed for architecture elements: e.g., security, data, resource management, information

Agenda

- Introduction
- Some applications
- TeraGrid Architecture
- Globus toolkit
- Future comm direction
- Summary

Static lightpath setting NCSA, ANL, SDSC, Caltech



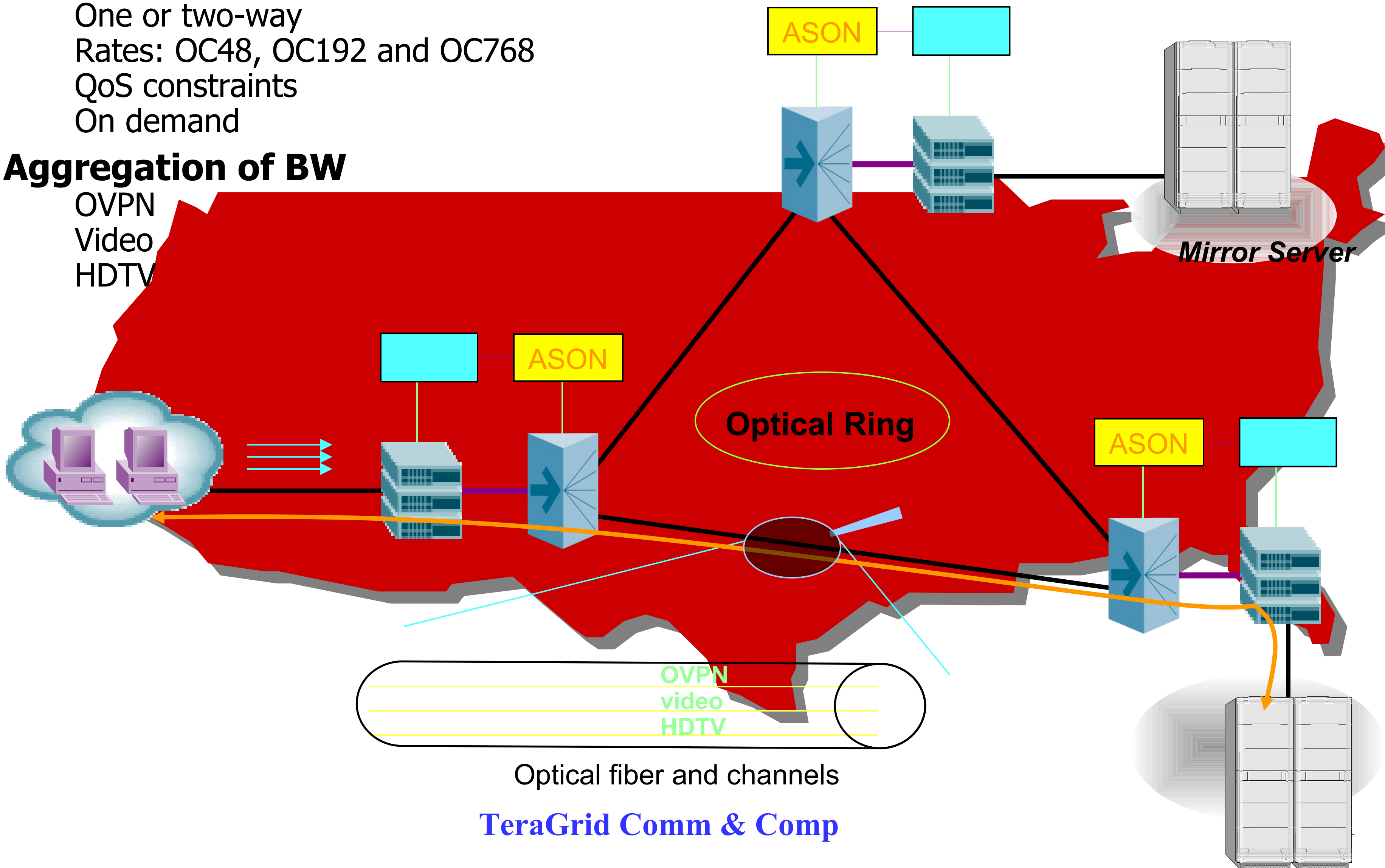
- Solid lines in place and/or available by October 2001
- Dashed I-WIRE lines planned for summer 2002

Source: Charlie Catlett, Argonne

Lightpath for OVPN

Lightpath setup
- One or two-way
- Rates: OC48, OC192 and OC768
- QoS constraints
- On demand

Aggregation of BW
- OVPN
- Video
- HDTV

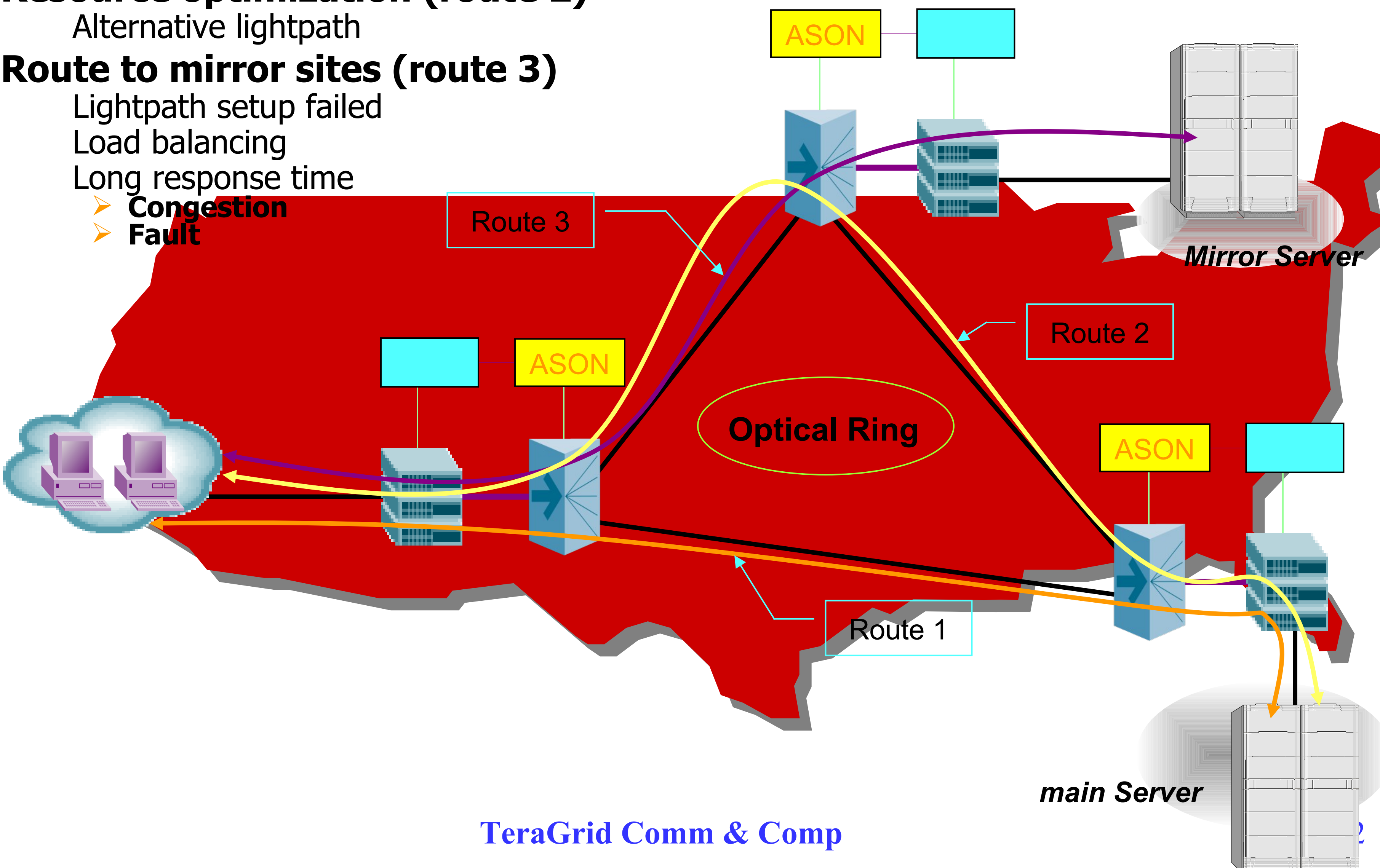


TeraGrid Comm & Comp

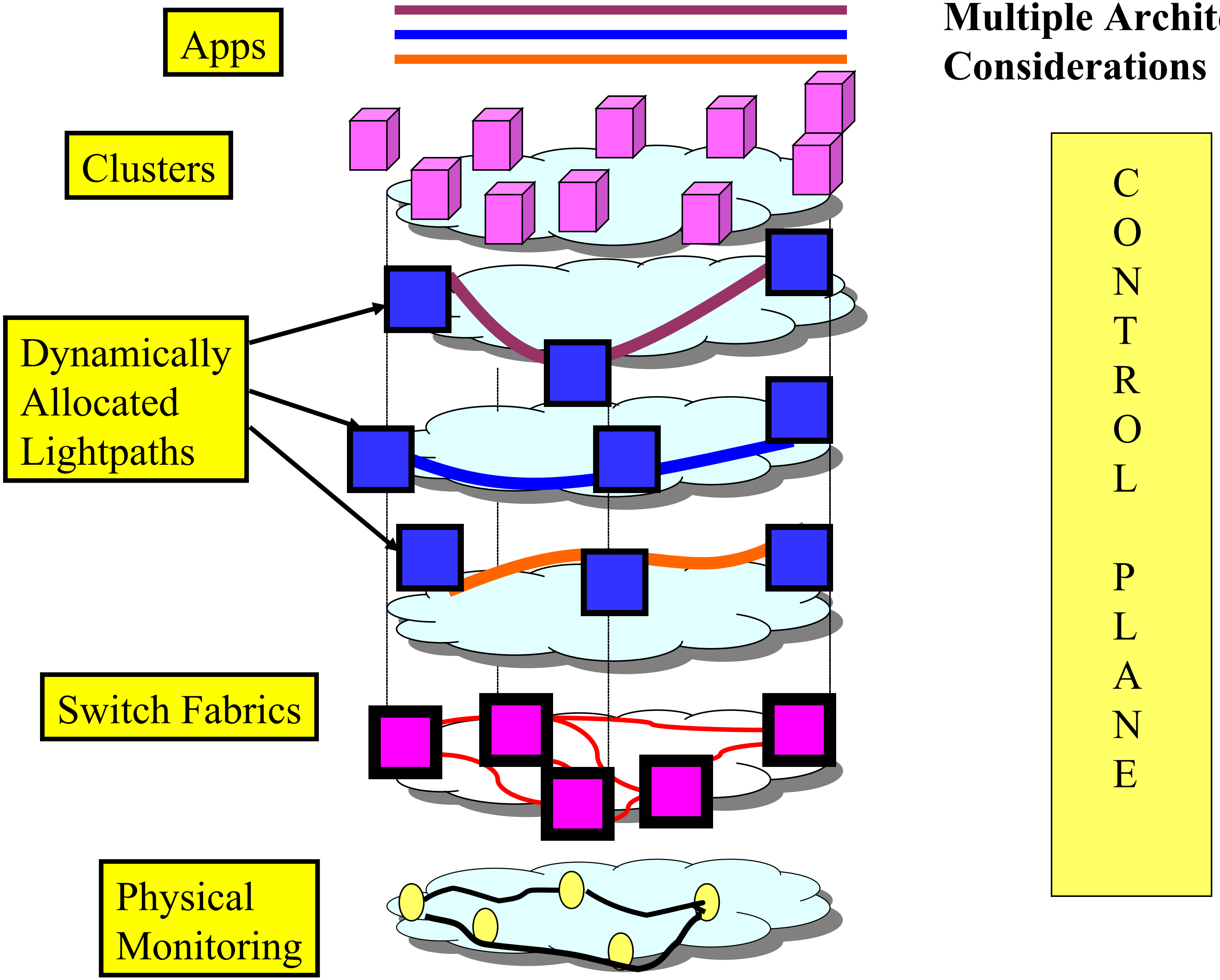
Dynamic Lightpath setting

Resource optimization (route 2)
- Alternative lightpath

Route to mirror sites (route 3)
- Lightpath setup failed
- Load balancing
- Long response time
  - **Congestion**
  - **Fault**



Multiple Architectural Considerations



Agenda

- Introduction
- Some applications
- TeraGrid Architecture
- Globus toolkit
- Future comm direction
- Summary

Summary

- The Grid problem: Resource sharing & coordinated problem solving in dynamic, multi-institutional virtual organizations
- Grid architecture: Emphasize protocol and service definition to enable interoperability and resource sharing
- Globus Toolkit a source of protocol and API definitions, reference implementations
- Current static communication. Next wave dynamic optical VPN
- Some relation to Sahara
  - Service composition: computation, servers, storage, disk, network…
  - Sharing, cooperating, peering, brokering…

References

globus.org

griphyn.org

gridforum.org

grids-center.org

nsf-middleware.org

Backup

Wavelengths and the Future

Wavelength services are causing a network revolution:

- Core long distance SONET Rings will be replaced by meshed networks using wavelength cross-connects
- Re-invention of pre-SONET network architecture

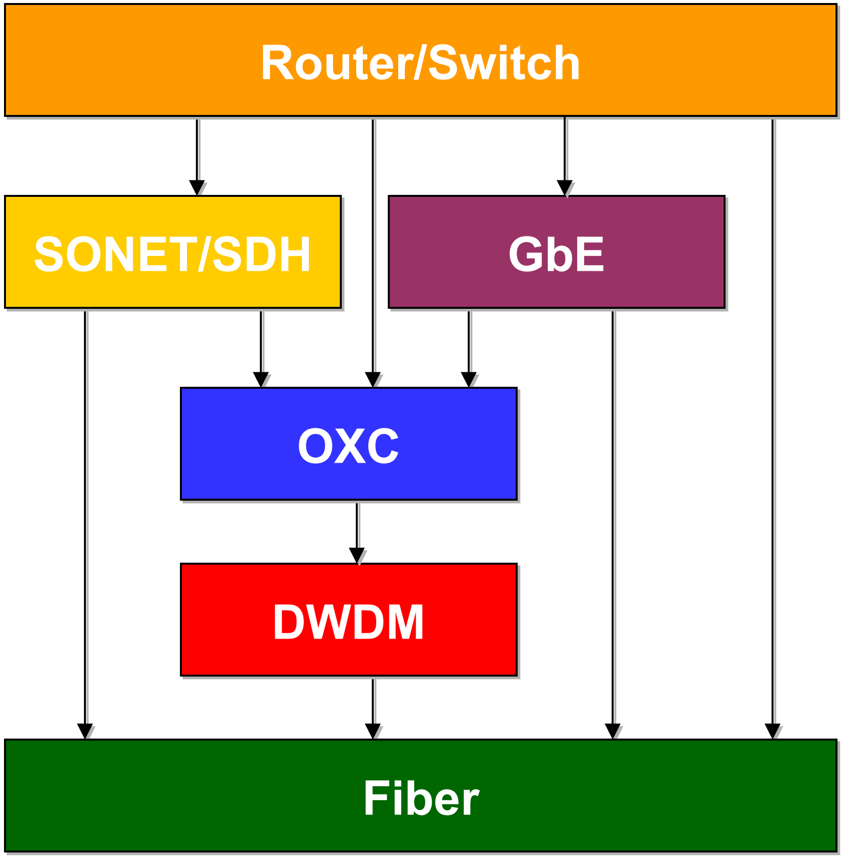
Improved transport infrastructure will exist for IP/packet services

Electrical/Optical grooming switches will emerge at edges

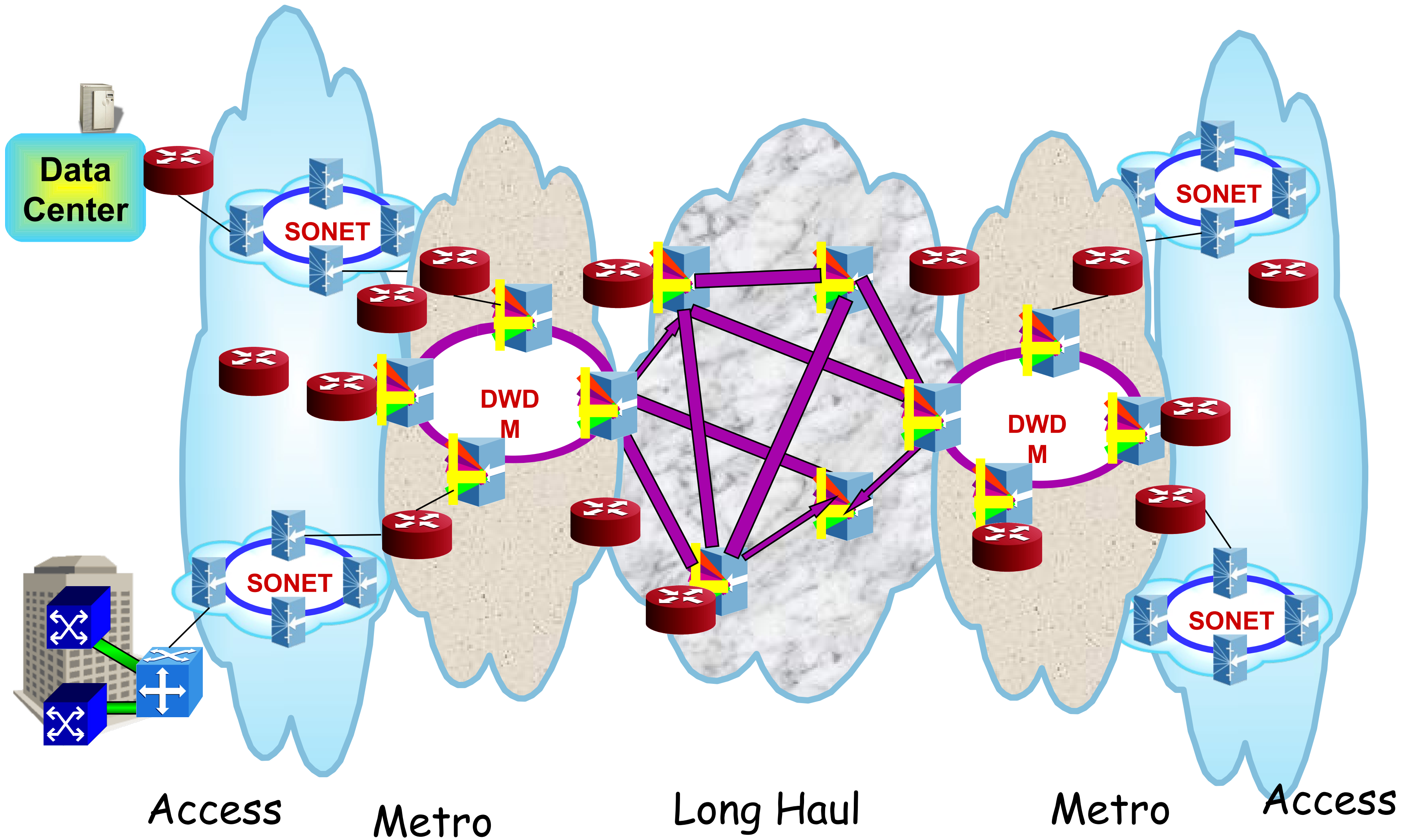
Automated Restoration (algorithm/GMPLS driven) becomes technically feasible.

Operational implementation will take some time

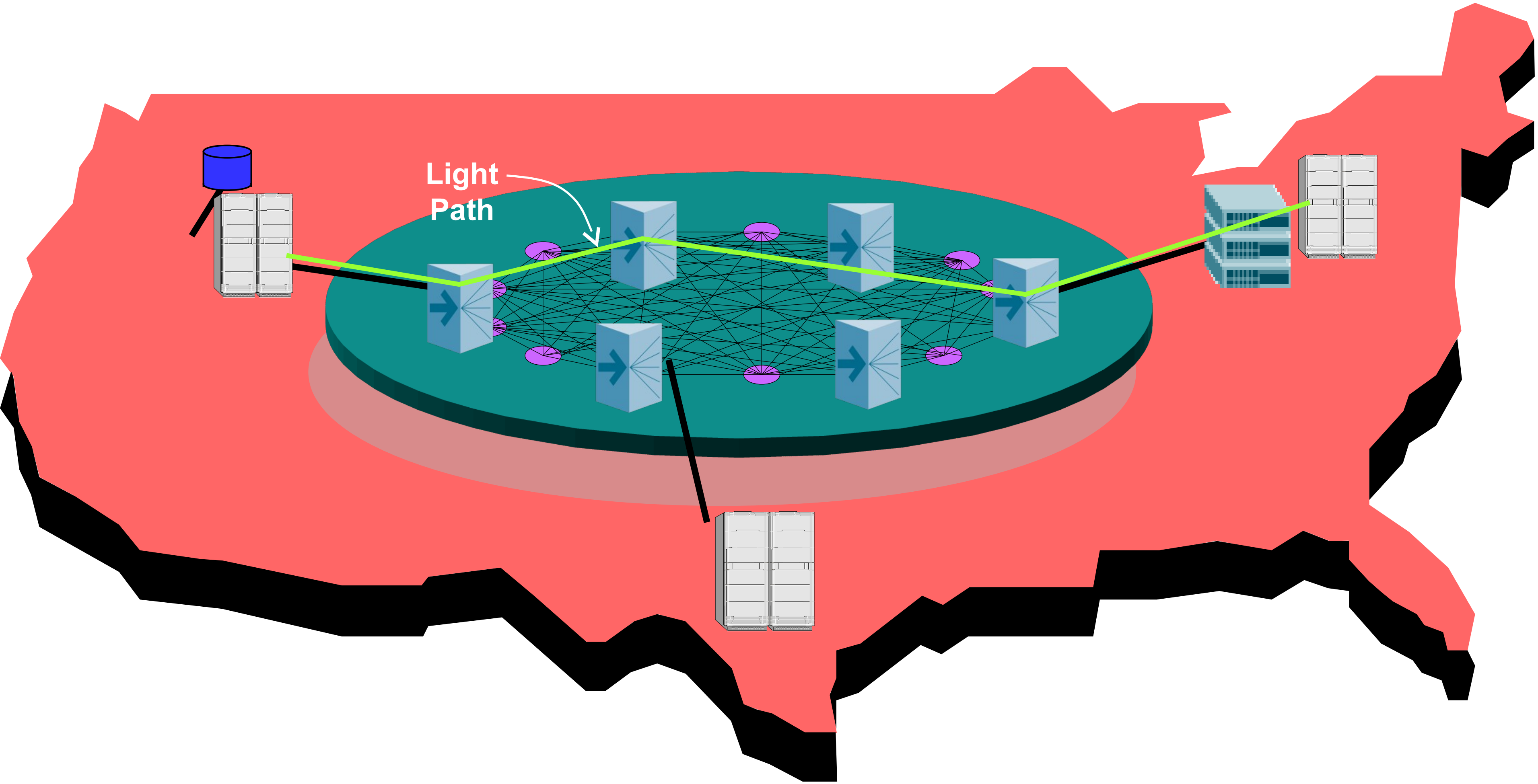
Optical components



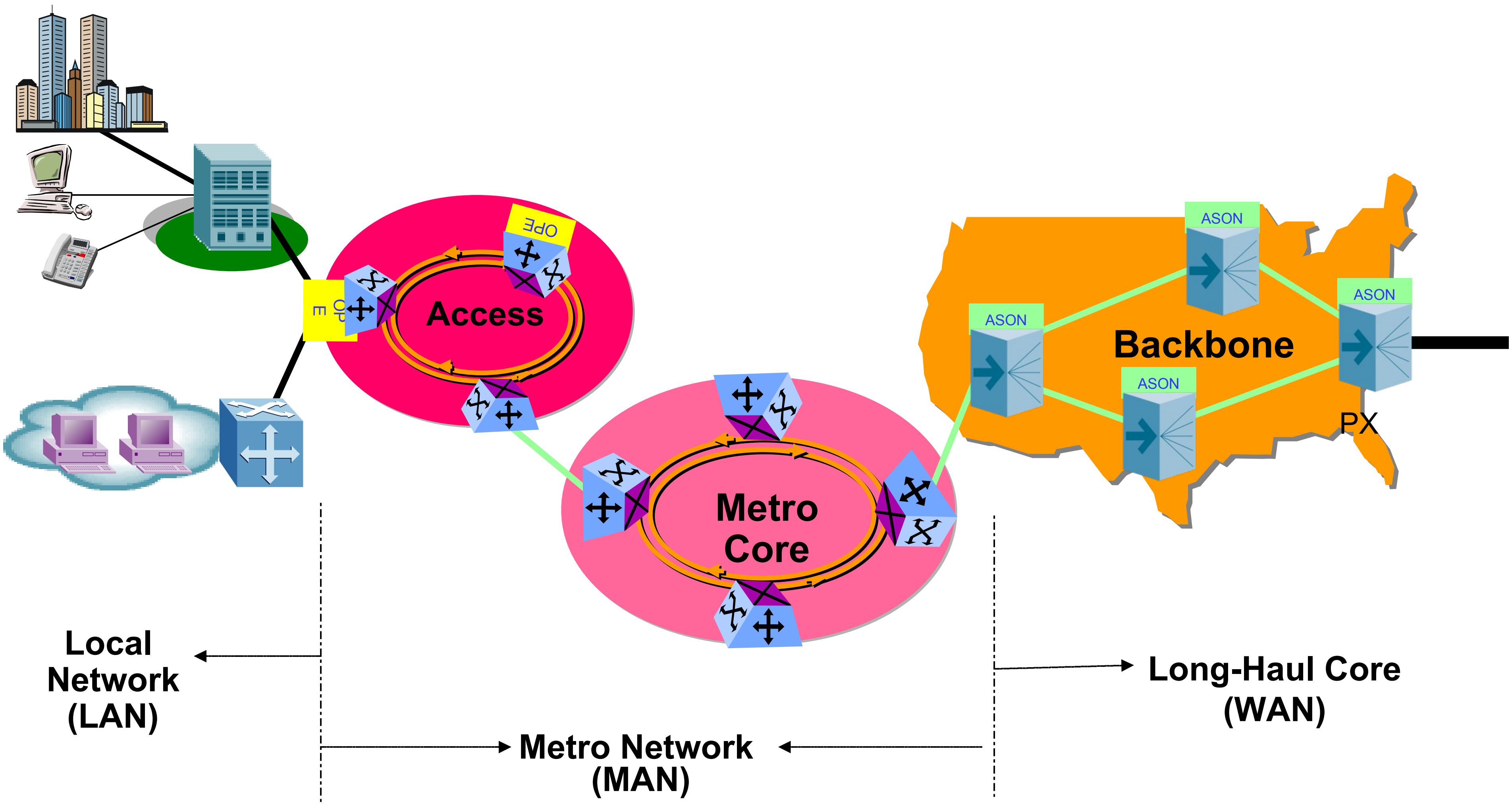
Internet Reality



OVPN on Optical Network



Three networks in The Internet



Data Transport Connectivity

Packet Switch

data-optimized
- Ethernet
- TCP/IP

Network use
- LAN

Advantages
- Efficient
- Simple
- Low cost

Disadvantages
- Unreliable

Circuit Switch

Voice-oriented
- SONET
- ATM

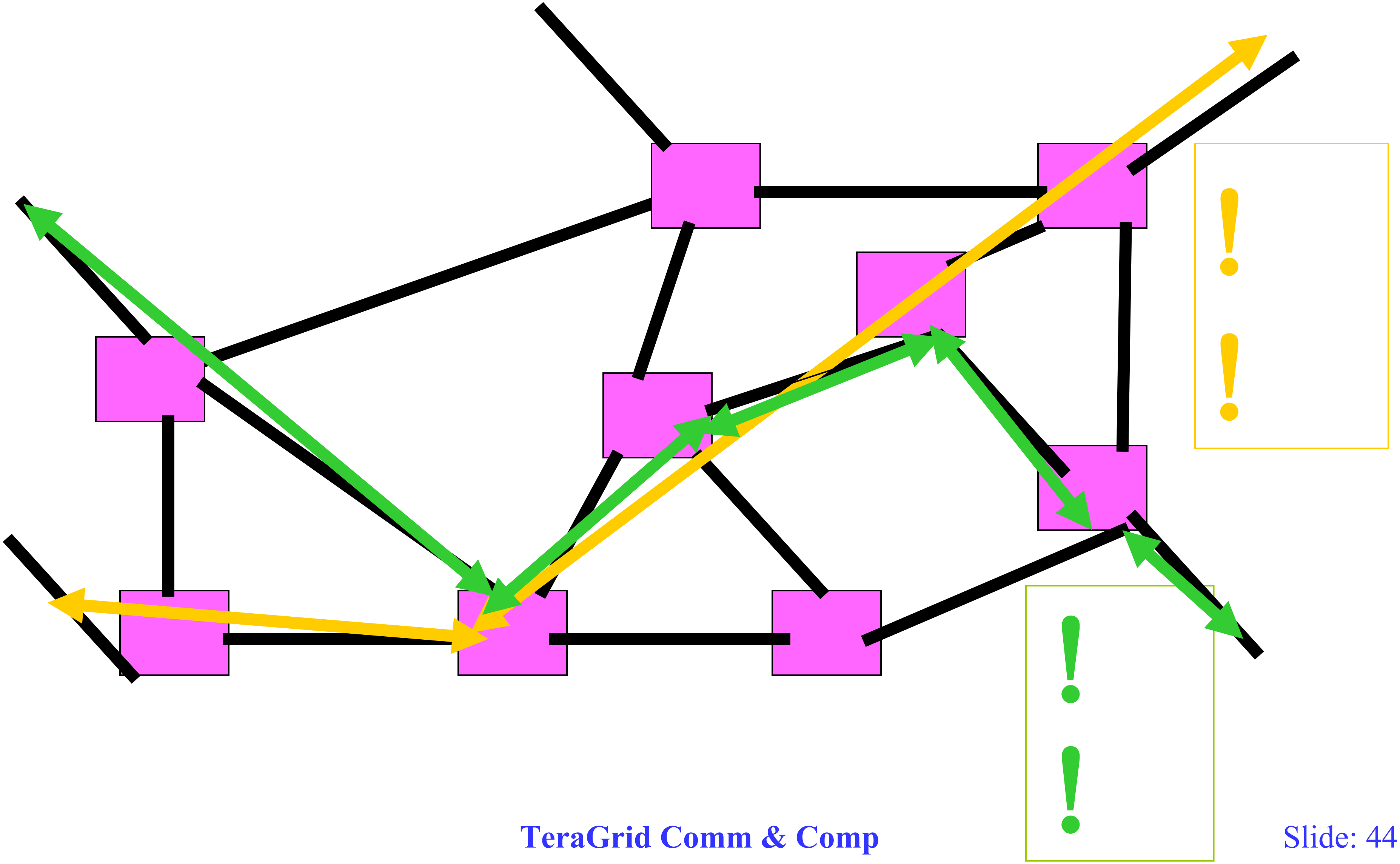
Network uses
- Metro and Core

Advantages
- Reliable

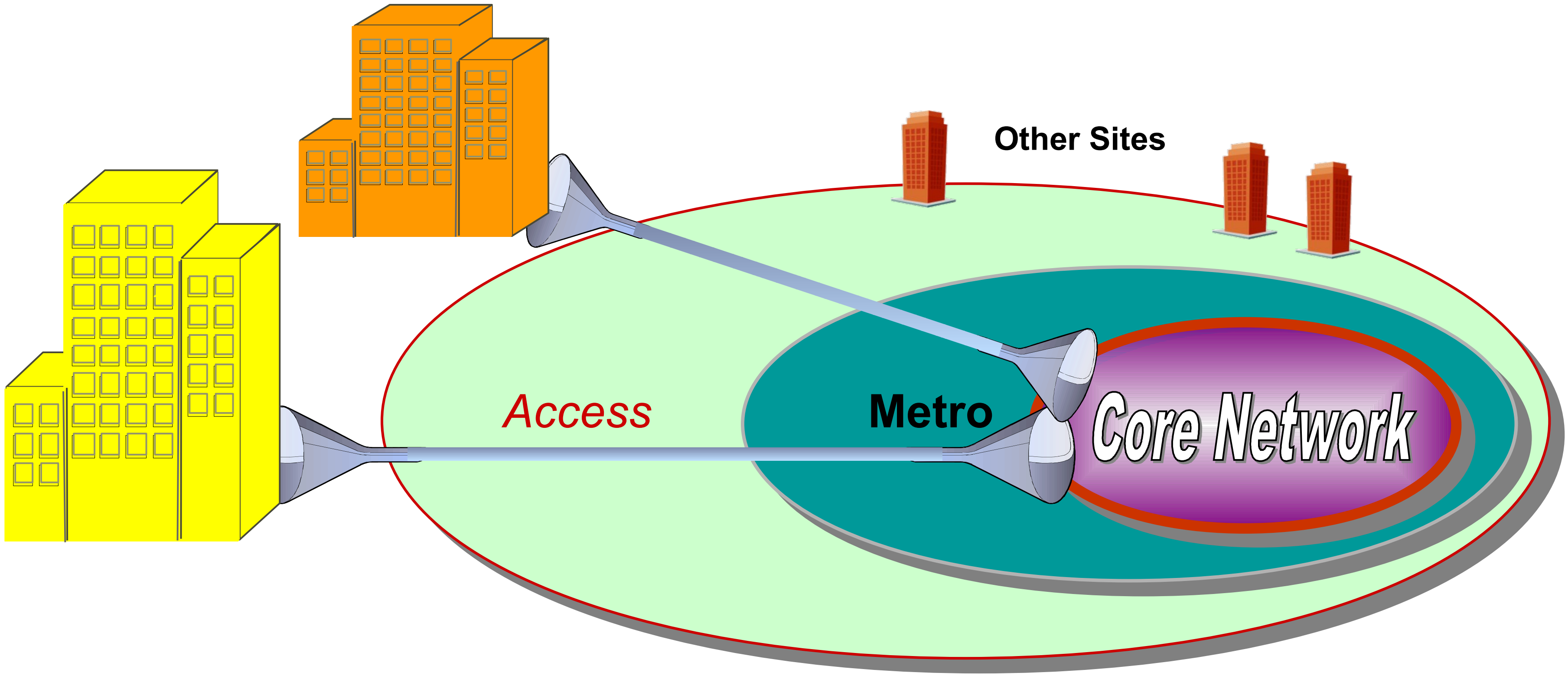
Disadvantages
- Complicate
- High cost

Efficiency ? Reliability

Global Lambda Grid - Photonic Switched Network



The Metro Bottleneck



End User	Access	Metro	Core
Ethernet LAN	DS1 DS3	OC-12 OC-48 OC-192	OC-192 DWDM n x !
IP/DATA 1GigE	LL/FR/ATM 1-40Meg	10G	40G+