Grid Optical Network Service Architecture for Data Intensive Applications

Control of Optical Systems and Networks OFC/NFOEC 2006

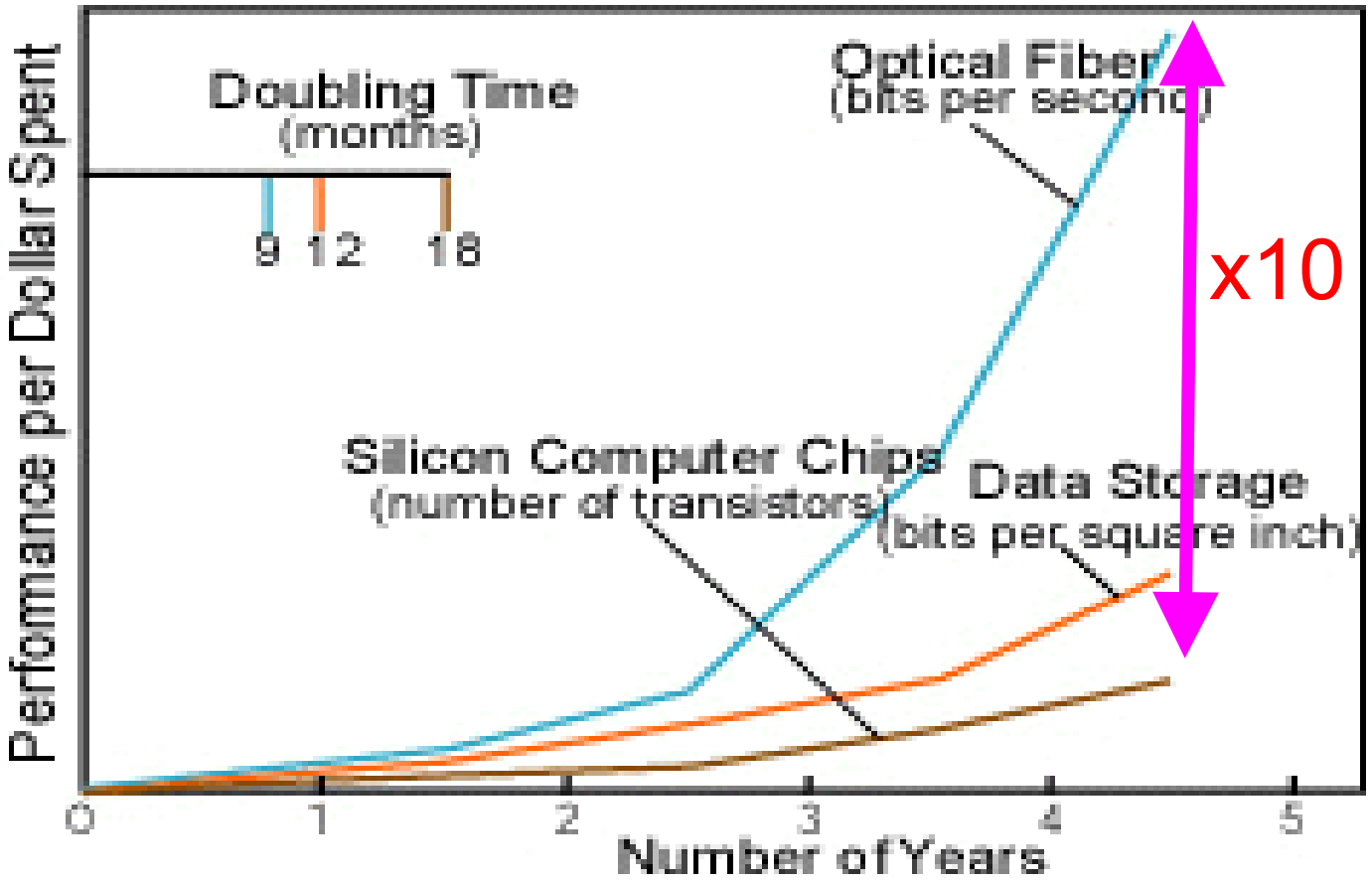
Tal Lavian tlavian@cs.berkeley.edu
UC Berkeley, and Advanced Technology Research , Nortel Networks

•

Randy Katz – UC Berkeley
John Strand – AT&T Research

March 8, 2006

Impedance mismatch:
Optical Transmission vs. Computation



Original chart from Scientific American, 2001
Support – Andrew Odlyzko 2003, and NSF Cyber-Infrastructure Jan 2006

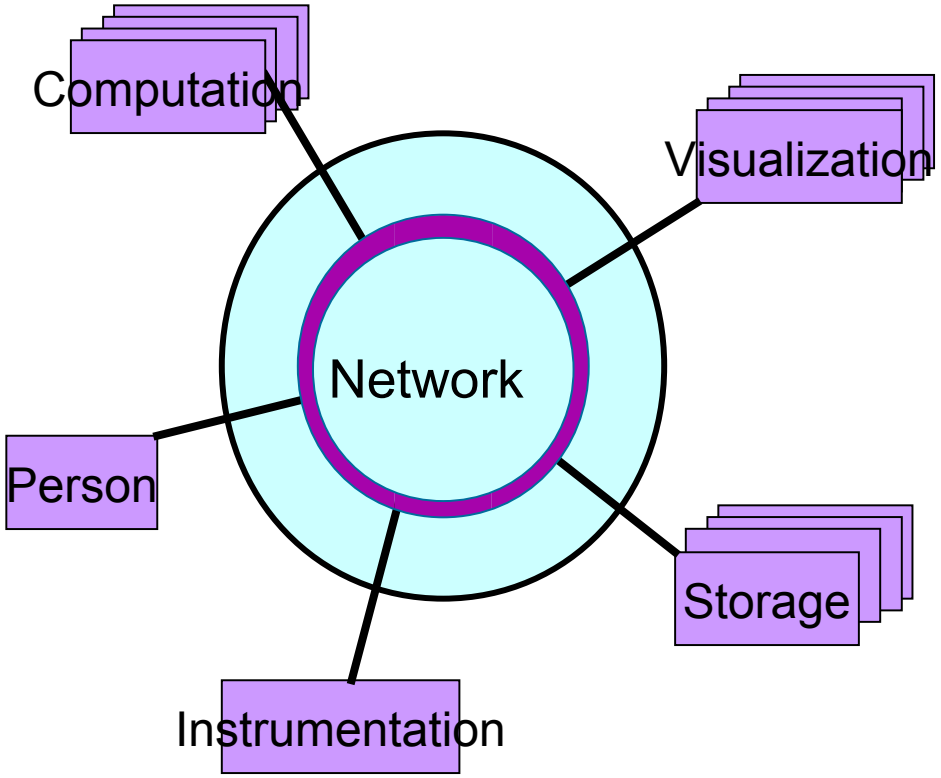
DWDM- fundamental miss-balance between computation and communication
5 Years – x10 gap, 10 years- x100 gap

Waste Bandwidth

"A global economy designed to waste transistors, power, and silicon area -and conserve bandwidth above all- is breaking apart and reorganizing itself to waste bandwidth and conserve power, silicon area, and transistors."
George Gilder Telecosm

> Despite the bubble burst – this is still a driver
  - It will just take longer

The "Network" is a Prime Resource for Large- Scale Distributed System



Integrated SW System Provide the "Glue"

Dynamic optical network as a fundamental Grid service in data-intensive Grid application, to be scheduled, to be managed and coordinated to support collaborative operations

From Super-computer to Super-network

- In the past, computer processors were the fastest part
  - peripheral bottlenecks
- In the future optical networks will be the fastest part
  - Computer, processor, storage, visualization, and instrumentation - slower "peripherals"
- eScience Cyber-infrastructure focuses on computation, storage, data, analysis, Work Flow.
  - The network is vital for better eScience

Cyber-Infrastructure for e-Science: Vast amounts of Data– Changing the Rules of the Game

- **PetaByte storage – Only $1M**
- **CERN - HEP – LHC:**
  - Analog: aggregated Terabits/second
  - Capture: PetaBytes Annually, 100PB by 2008
  - ExaByte 2012
  - The biggest research effort on Earth
- **SLAC BaBar:** PetaBytes
- **Astrophysics:** Virtual Observatories - 0.5PB
- **Environment Science**: Eros Data Center (EDC) – 1.5PB, NASA 15PB
- **Life Science:**
  - Bioinformatics - PetaFlops/s
  - One gene sequencing - 800 PC for a year

Crossing the Peta (10^{15}) Line

- **Storage size, comm bandwidth, and computation rate**
  - Several National Labs have built Petabyte storage systems
  - Scientific databases have exceeded 1 PetaByte
  - High-end super-computer centers - 0.1 Petaflops
    - will cross the Petaflop line in five years
  - Early optical lab transmission experiments - 0.01 Petabits/s
    - When will cross the Petabits/s line?

e-Science example

Application Scenario	Current	Network Issues
Pt – Pt Data Transfer of Multi-TB Data Sets	! Copy from remote DB: Takes ~10 days (unpredictable) ! Store then copy/analyze	! Want << 1 day << 1 hour, ! innovation for new bio-science ! Architecture forced to optimize BW utilization at cost of storage
Access multiple remote DB	! N* Previous Scenario	! Simultaneous connectivity to multiple sites ! Multi-domain ! Dynamic connectivity hard to manage ! Don't know next connection needs
Remote instrument access (Radio-telescope)	! Cant be done from home research institute	! Need fat unidirectional pipes ! Tight QoS requirements (jitter, delay, data loss)

Other Observations:

- **Not Feasible To Port Computation to Data**
- **Delays Preclude Interactive Research: Copy, Then Analyze**
- **Uncertain Transport Times Force A Sequential Process – Schedule Processing After Data Has Arrived**
- **No cooperation/interaction among Storage, Computation & Network Middlewares**

•Dynamic network allocation as part of Grid Workf bw, allows for new scientif ċ experiments that are not possible with today's static allocation

Grid Network Limitations in L3

- **Radical mismatch between the optical transmission world and the electrical forwarding/routing world**
- **Transmit 1.5TB over 1.5KB packet size**
  - ✂→**1 Billion identical lookups**
- **Mismatch between L3 core capabilities and disk cost**
  - **With $2M disks (6PB) can fill the entire core internet for a year**
- L3 networks can't handle these amounts effectively, predictably, in a short time window
  - L3 network provides full connectivity -- major bottleneck
  - **Apps optimized to conserve bandwidth and waste storage**
  - Network does not fit the "e-Science Workflow" architecture

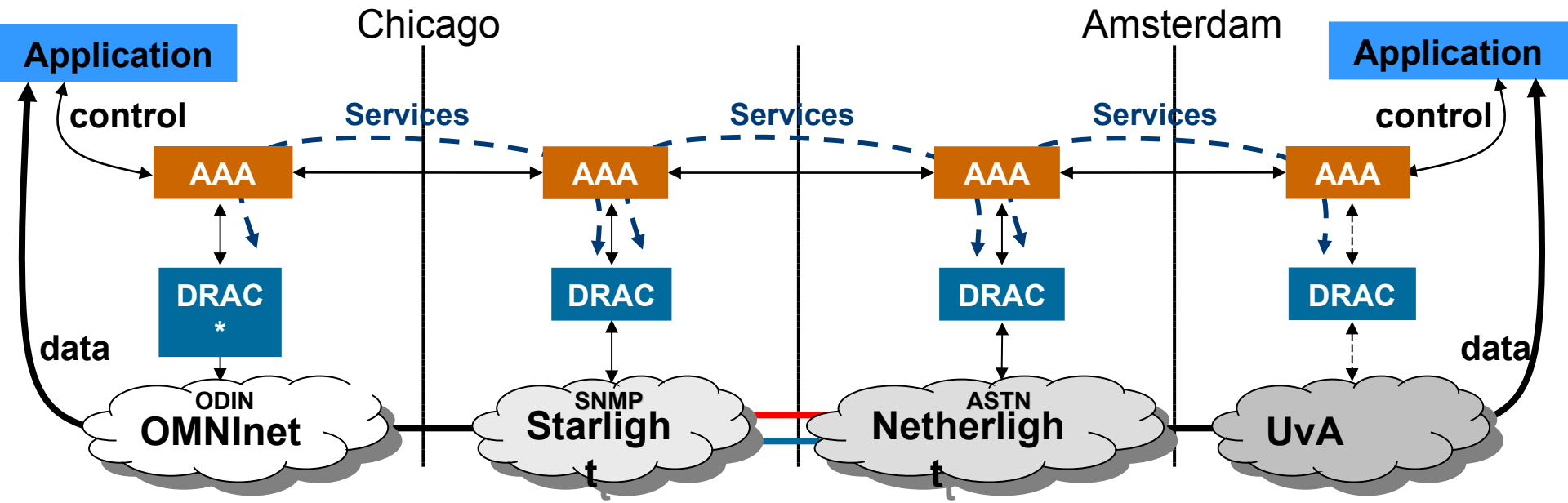
Prevents true Grid Virtual Organization (VO) research collaborations

Lambda Grid Service

Need for Lambda Grid Service architecture that interacts with Cyber-infrastructure, and overcome data limitations efficiently & effectively by:

- treating the "network" as a primary resource just like "storage" and "computation"
- treat the "network" as a "scheduled resource"
- rely upon a massive, dynamic transport infrastructure: Dynamic Optical Network

Super Computing CONTROL CHALLENGE



* Dynamic Resource Allocation Controller

- finesse the control of bandwidth across multiple domains
- while exploiting scalability and intra- , inter-domain fault recovery
- thru layering of a novel SOA upon legacy control planes and NEs

CANARIE NETWORKS > COLLABORATION > RESULTS RÉSEAUX > COLLABORATION > RÉSULTATS

GigaPort

glif

iCAIR

INTERNET2 www.internet2.edu

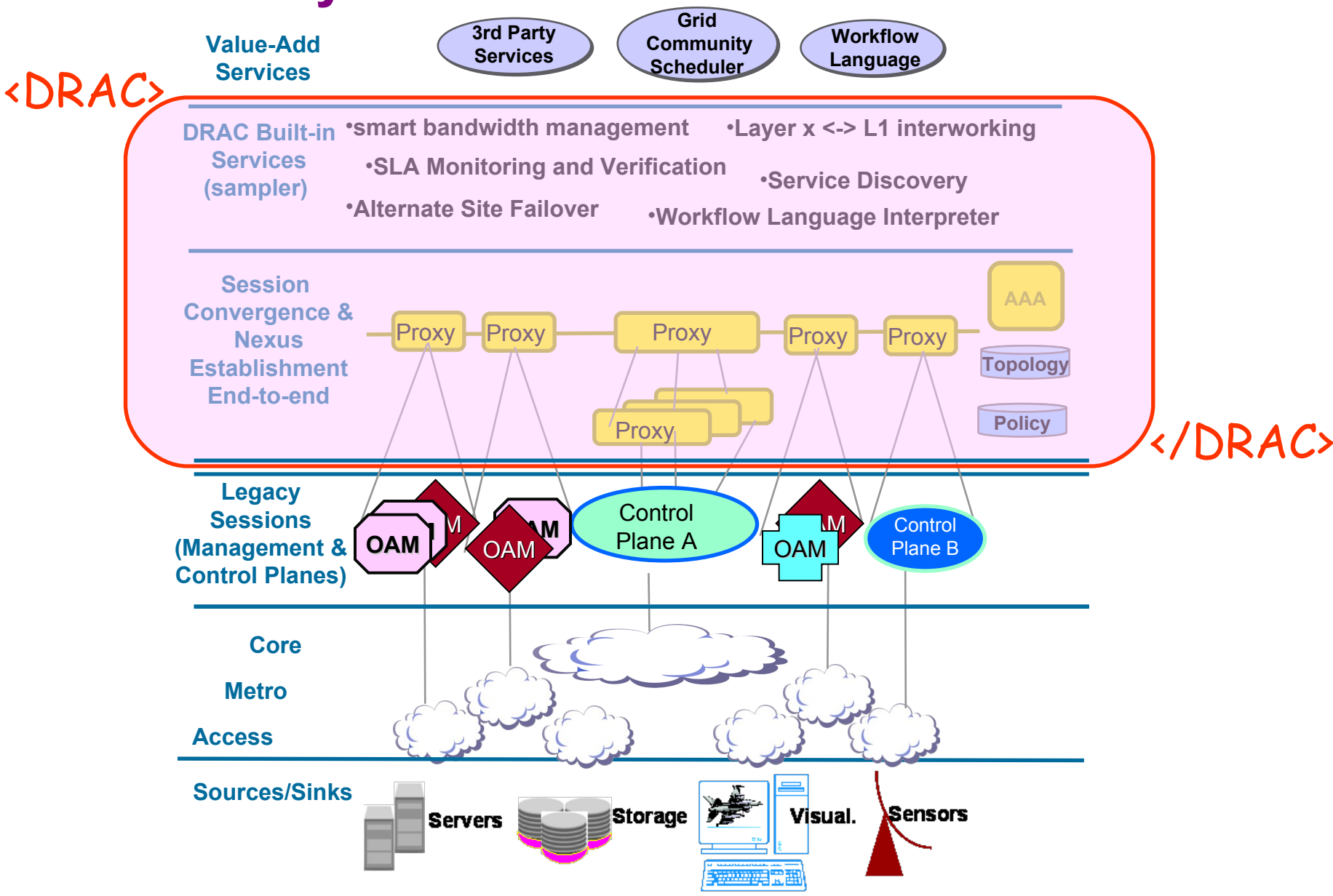
NL Light

STARLIGHT SM

SURFnet

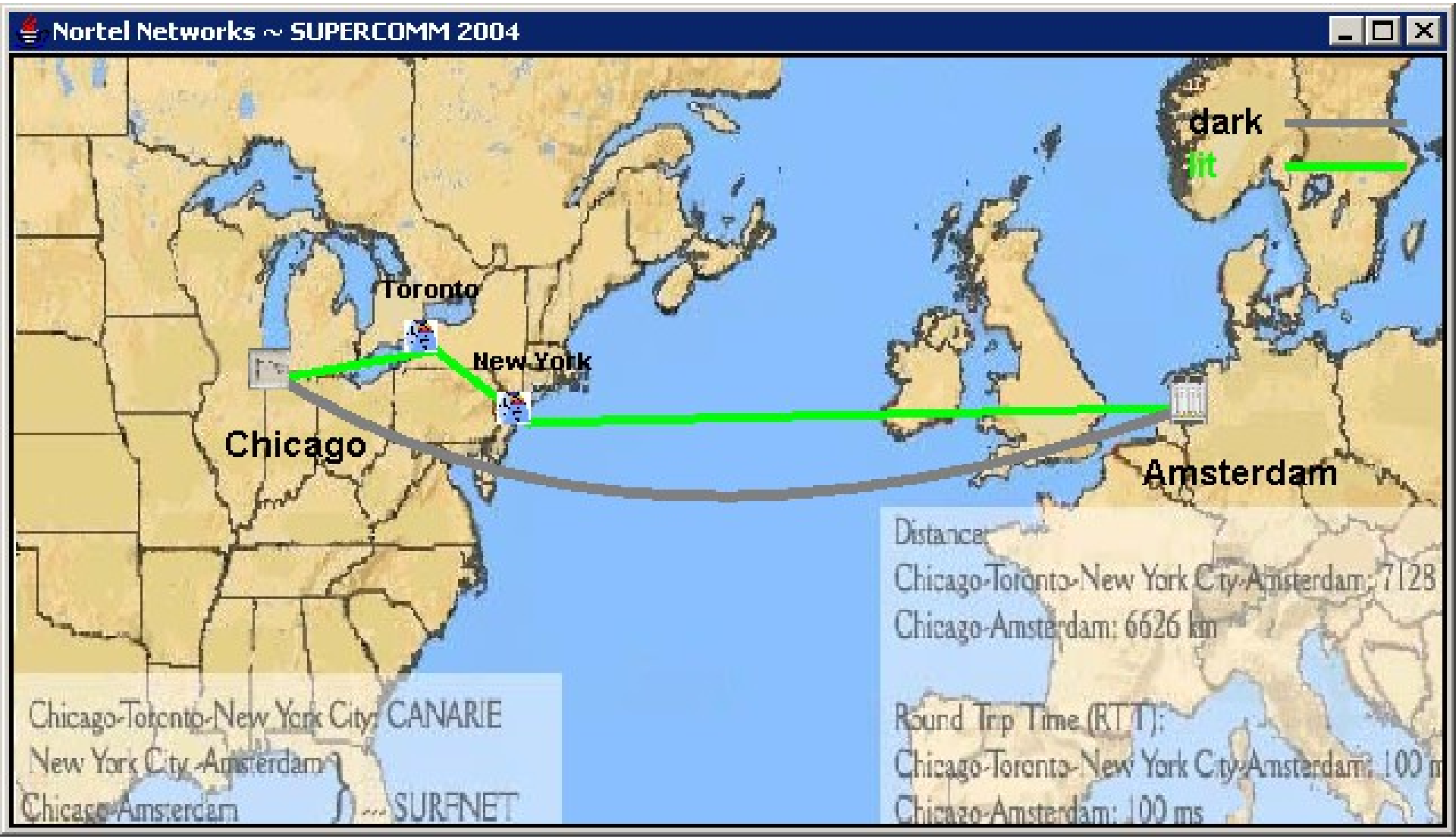
UvA Universiteit van Amsterdam

Bird's eye View of the Service Stack

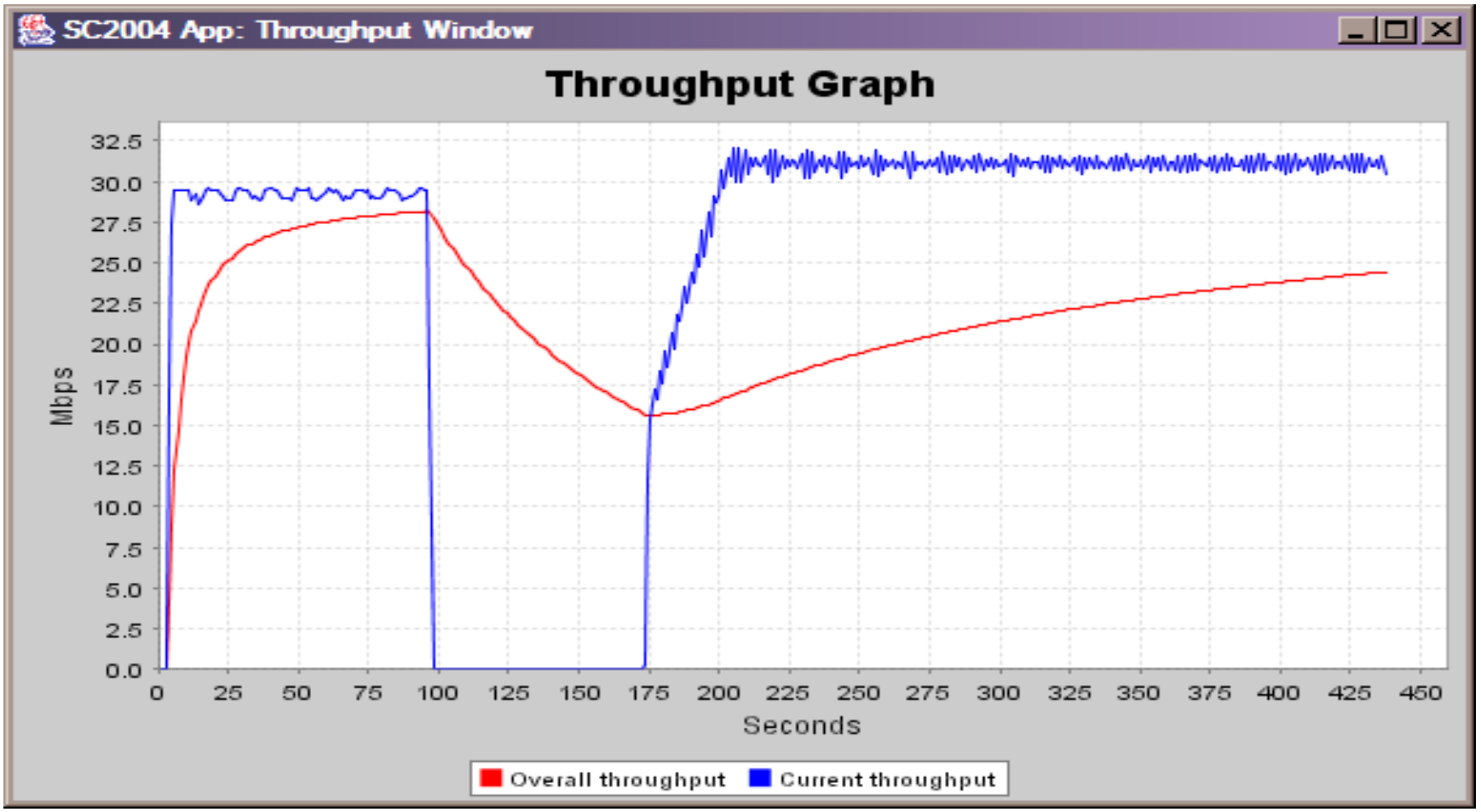


Fail over From Rout-D to Rout-A

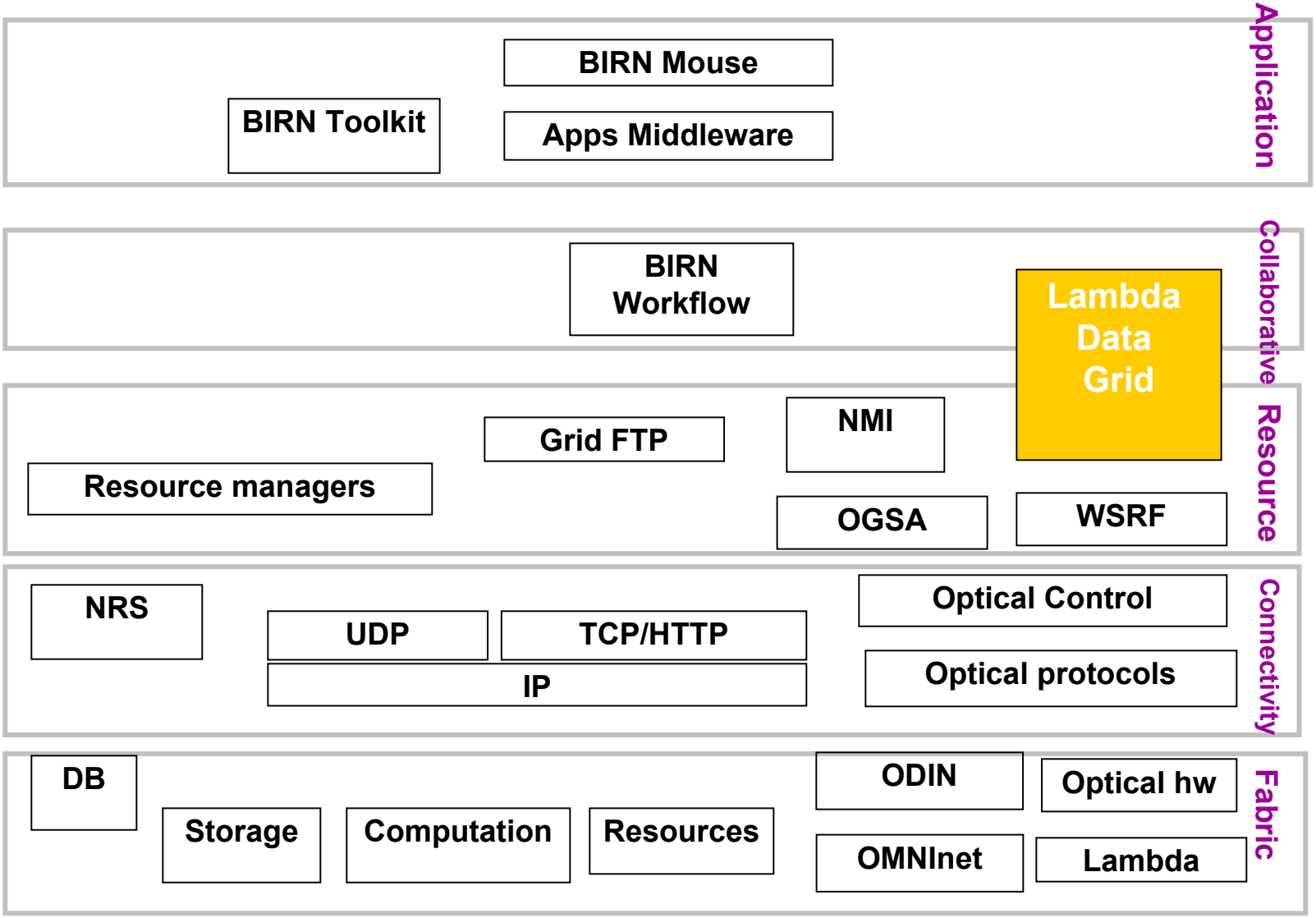
(SURFnet Amsterdam, Internet-2 NY, CANARIE Toronto, Starlight Chicago)



Transatlantic Lambda reservation



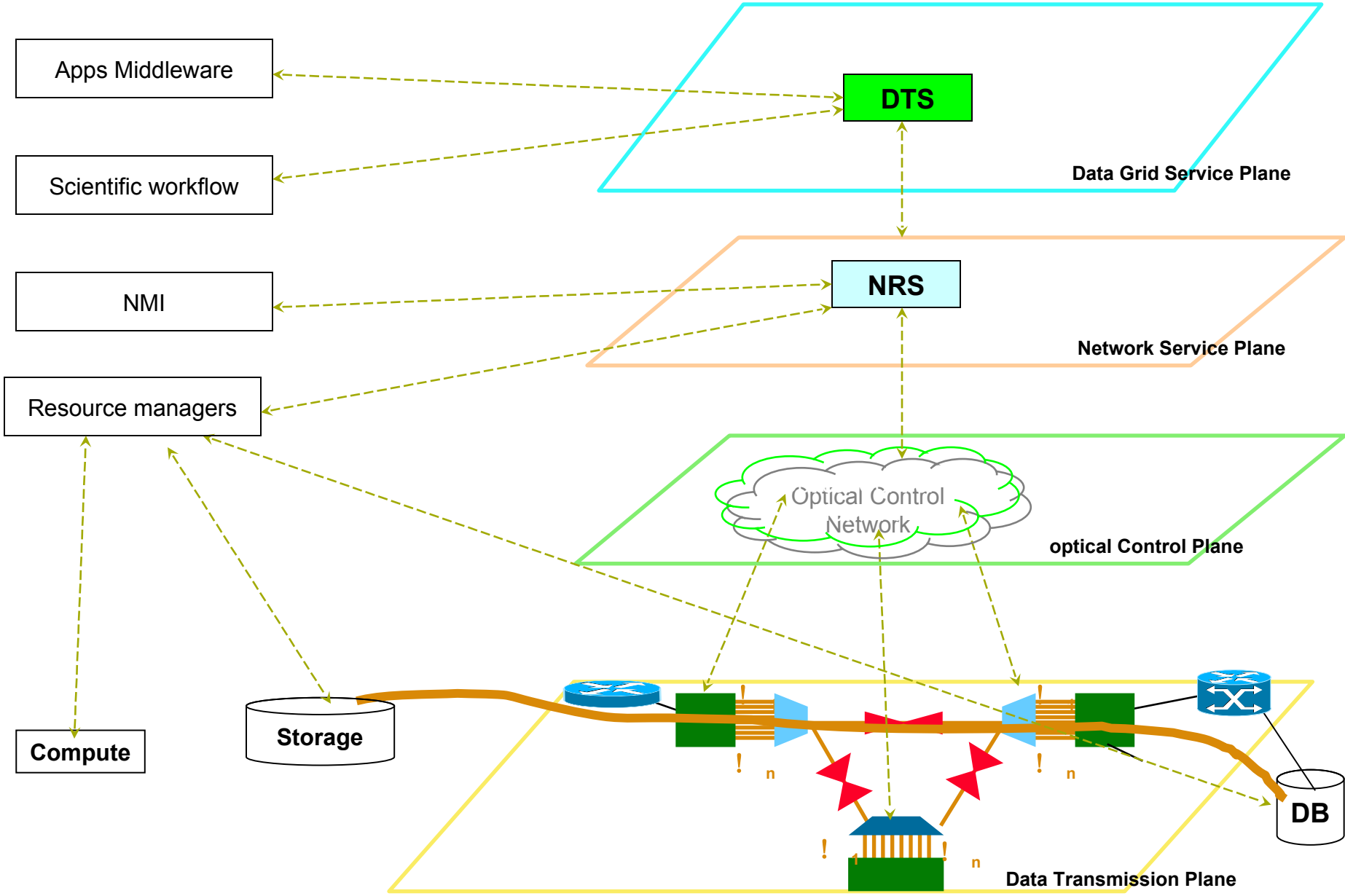
Layered Architecture



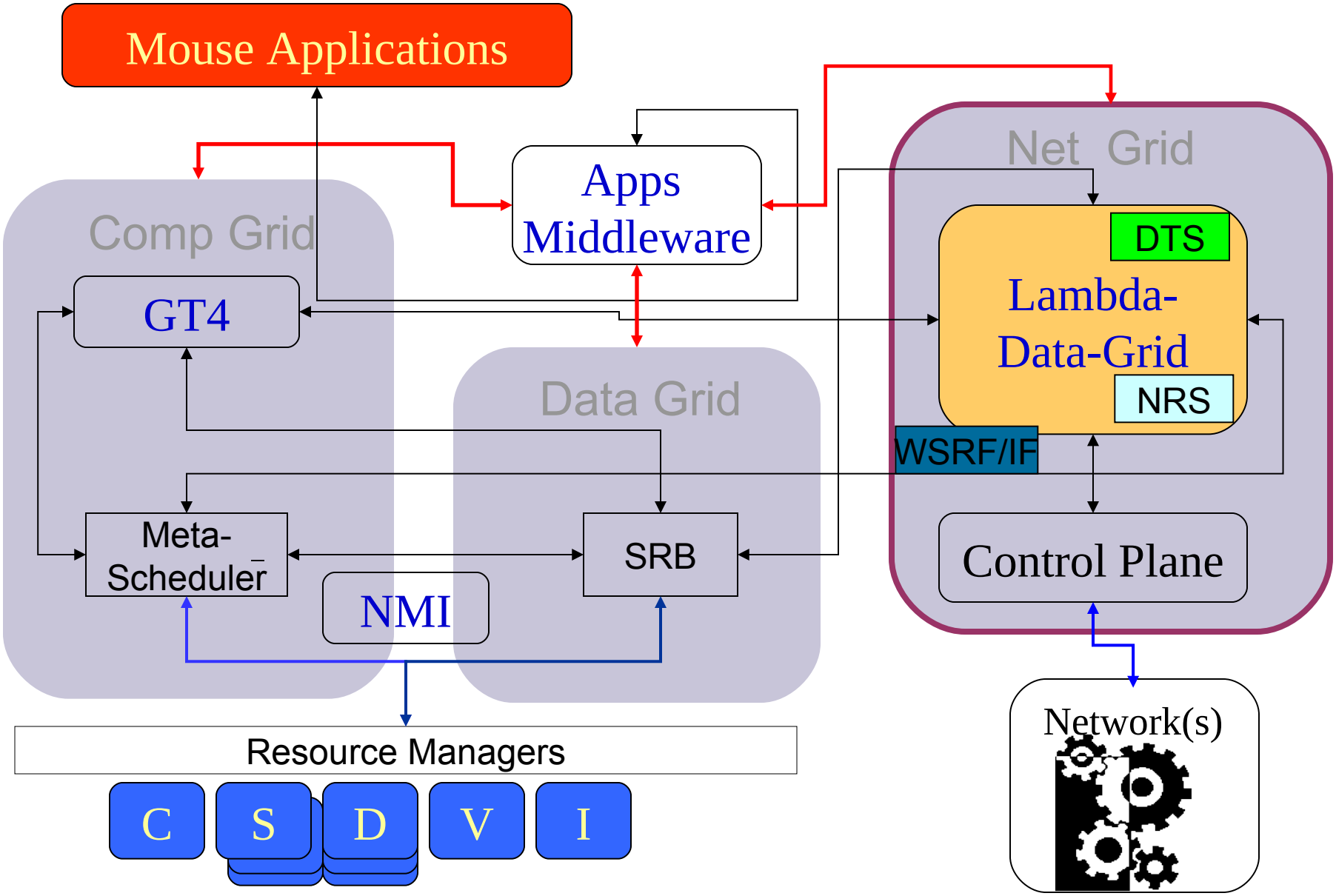
Grid Layered Architecture

CONNECTION

Control Interactions



BIRN Mouse Example



Summary

- Cyber-infrastructure – for emerging e-Science
- Realizing Grid Virtual Organizations (VO)
- Lambda Data Grid
- Communications Architecture in Support of Grid Computing
- Middleware for automated network orchestration of resources and services
- Scheduling and co-scheduling or network resources

Back-up

Generalization and Future Direction for Research

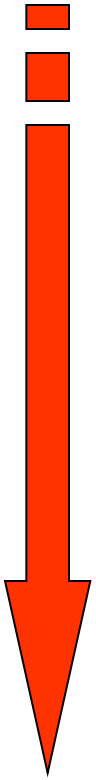
> Need to develop and build services on top of the base encapsulation

> Lambda Grid concept can be generalized to other eScience apps which will enable new way of doing scientific research where bandwidth is “infinite”

> The new concept of network as a scheduled grid service presents new and exciting problems for investigation:

- New software systems that is optimized to waste bandwidth
  - Network, protocols, algorithms, software, architectures, systems
- Lambda Distributed File System
- The network as a Large Scale Distributed Computing
- Resource co/allocation and optimization with storage and computation
- Grid system architecture
- enables new horizon for network optimization and lambda scheduling
- The network as a white box, Optimal scheduling and algorithms

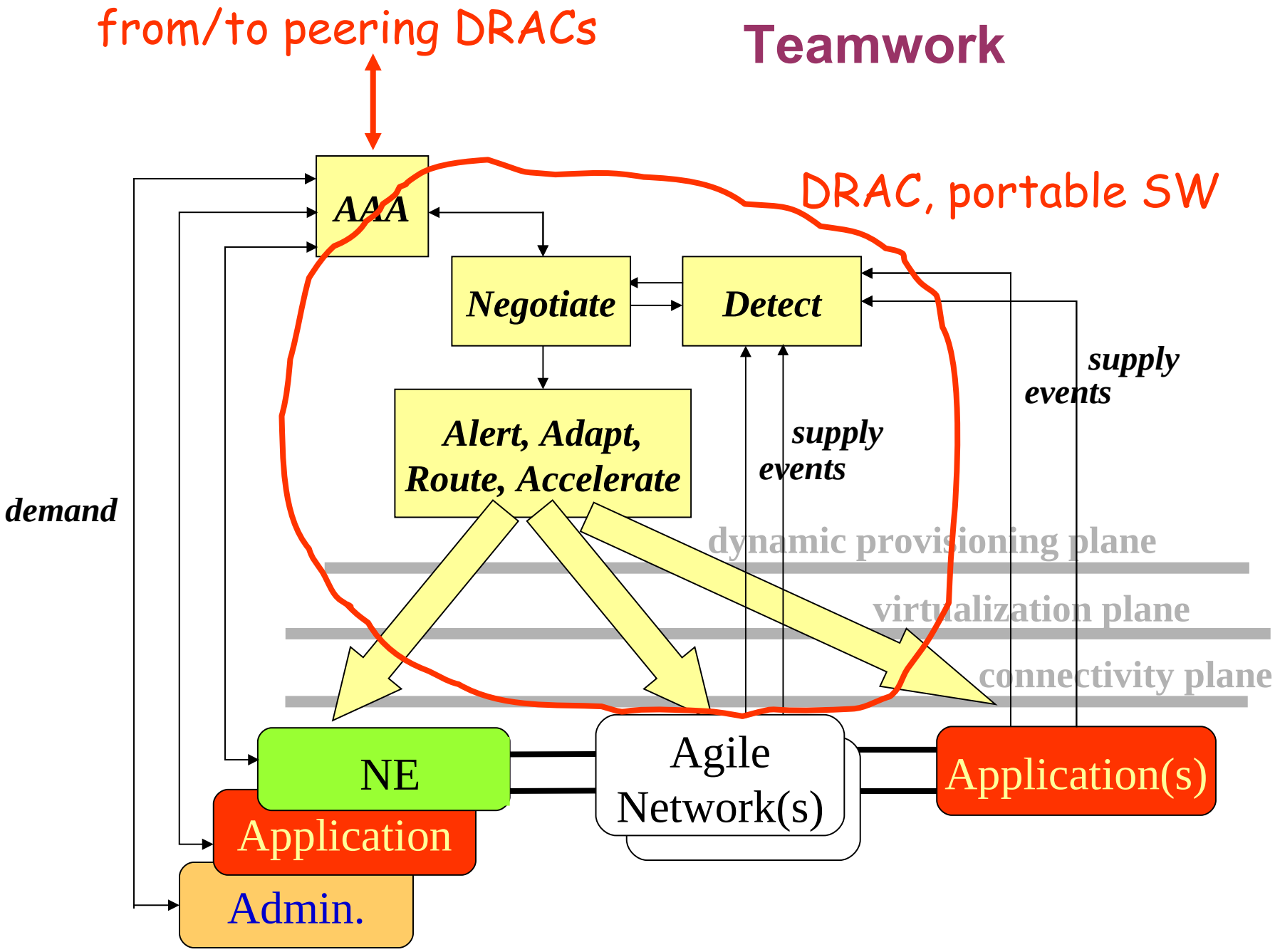
Enabling new degrees of App/Net coupling

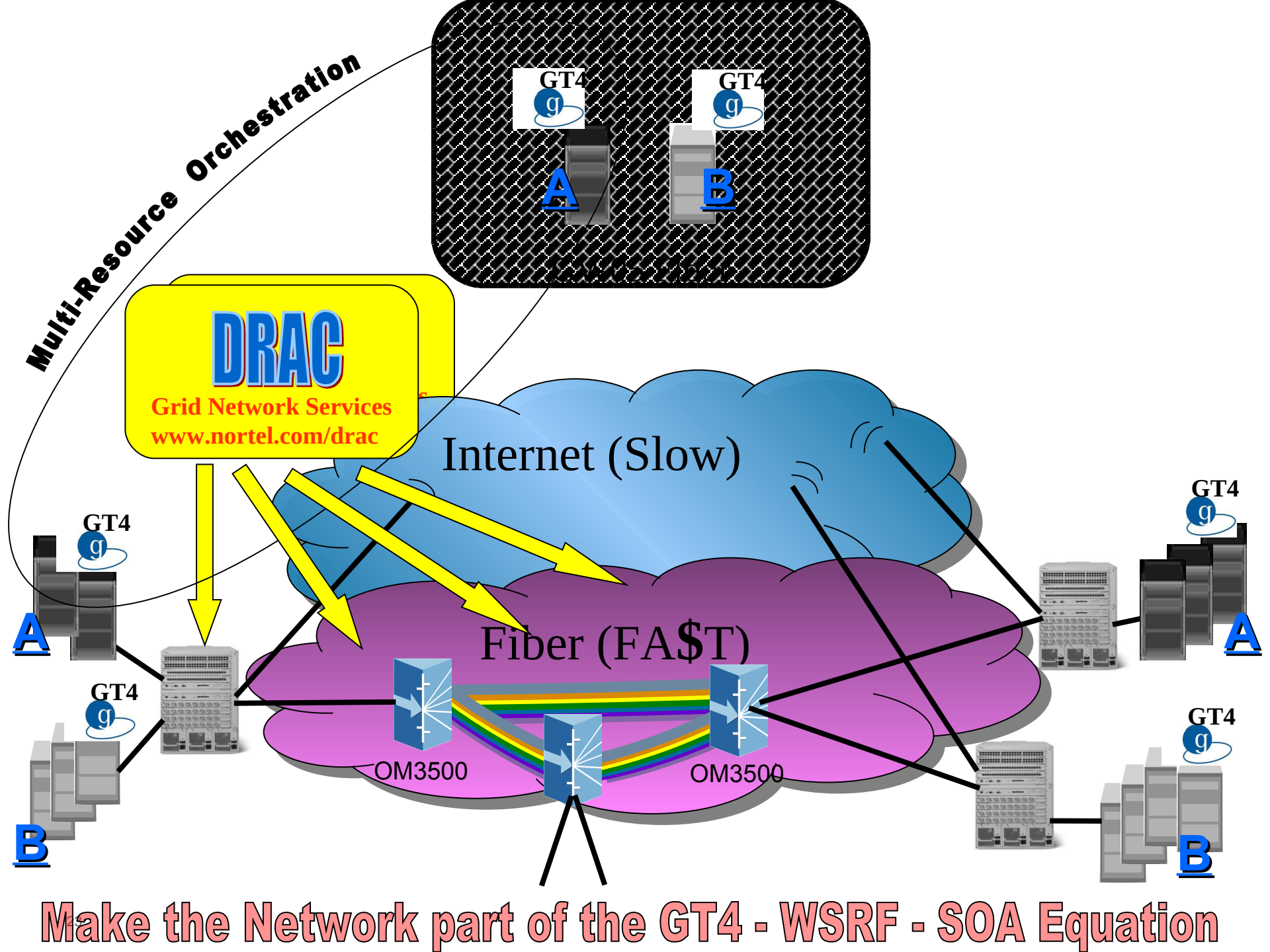


- Optical Packet Hybrid
  - Steer the herd of elephants to ephemeral optical circuits (few to few)
  - Mice or individual elephants go through packet technologies (many to many)
  - Either application-driven or network-sensed; hands-free in either case
  - Other impedance mismatches being explored (e.g., wireless)
- Application-engaged networks
  - The application makes itself known to the network
  - The network recognizes its footprints (via tokens, deep packet inspection)
  - E.g., storage management applications
- Workflow-engaged networks
  - Through workflow languages, the network is privy to the overall “flight-plan”
  - Failure-handling is cognizant of the same
  - Network services can anticipate the next step, or what-if’s
  - E.g., healthcare workflows over a distributed hospital enterprise

DRAC - Dynamic Resource Allocation Controller

Teamwork

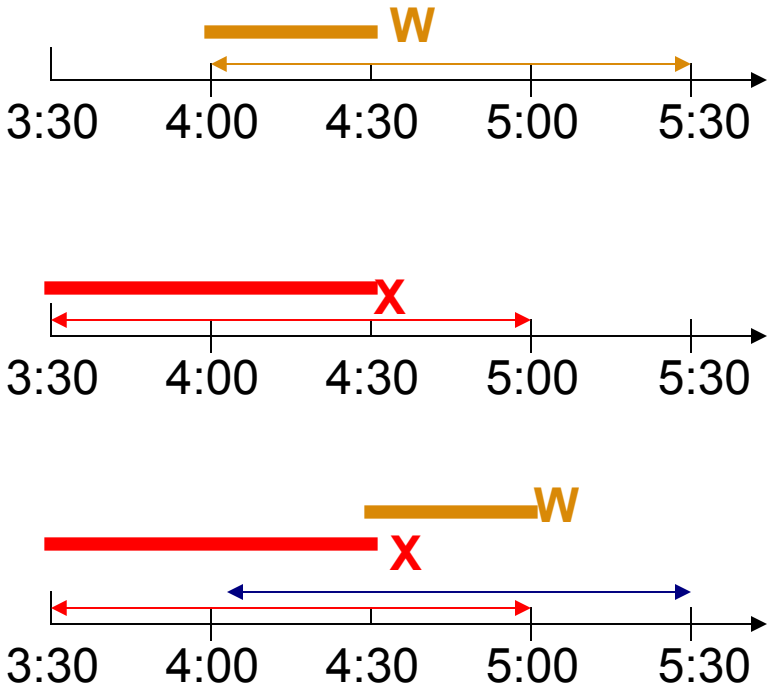

from/to peering DRACs
DRAC, portable SW
AAA
Negotiate
Detect
Alert, Adapt,
Route, Accelerate
supply
events
supply
events
demand
dynamic provisioning plane
virtualization plane
connectivity plane
NE
Application
Admin.
Agile
Network(s)
Application(s)



Make the Network part of the GT4 - WSRF - SOA Equation

Example: Lightpath Scheduling

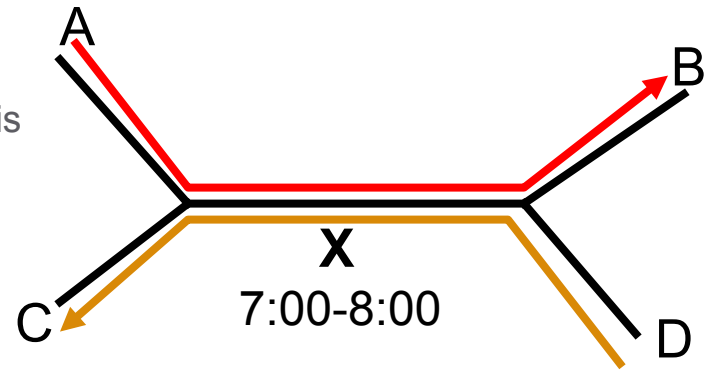
- Request for 1/2 hour between 4:00 and 5:30 on Segment D granted to User W at 4:00
- New request from User X for same segment for 1 hour between 3:30 and 5:00
- Reschedule user W to 4:30; user X to 3:30. Everyone is happy.



Route allocated for a time slot; new request comes in; 1st route can be rescheduled for a later slot within window to accommodate new request

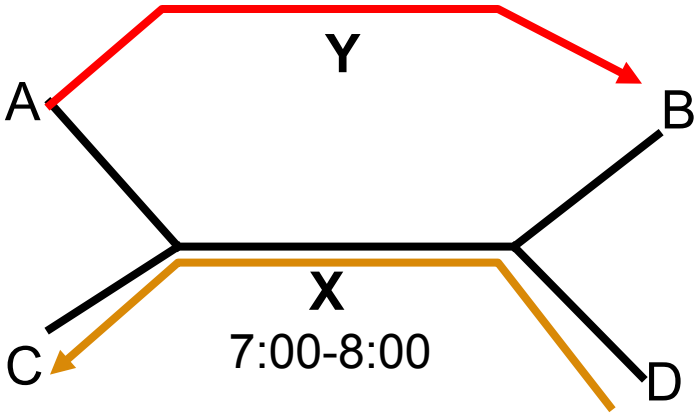
Scheduling Example - Reroute

- Request for 1 hour between nodes A and B between 7:00 and 8:30 is granted using Segment X (and other segments) is granted for 7:00



- New request for 2 hours between nodes C and D between 7:00 and 9:30 This route needs to use Segment E to be satisfied

- Reroute the first request to take another path thru the topology to free up Segment E for the 2nd request. Everyone is happy



Route allocated; new request comes in for a segment in use; 1st route can be altered to use different path to allow 2nd to also be serviced in its time window

Some key folks checking us out at our booth, GlobusWORLD ‘04, Jan ‘04

Ian Foster, Carl Kesselman, Larry Smarr