A Platform for Large-Scale Grid Data Service on Dynamic High-Performance Networks

T. Lavian, D. B. Hoang, J. Mambretti, S. Figueira, S. Naiksatam, N. Kaushik,

I. Monga, R. Durairaj, D. Cutrell, S. Merrill, H. Cohen, P. Daspit, F. Travostino

Presented by Tal Lavian

Topics

- Limitations of Current IP Networks
- Why Dynamic High-Performance Networks and DWDM-RAM?
- DWDM-RAM Architecture
- An Application Scenario
- Testbed and DWDM-RAM Implementation
- Experimental Results
- Simulation Results
- Conclusion

Limitations of Current Network Infrastructures

Packet-Switched Limitation

- ***Packet switching is NOT appropriate for data intensive applications => substantial overhead, delays, CapEx, OpEx***
- ***Limited control and isolation of Network Bandwidth***

Grid Infrastructure Limitation

- ***Difficulty in encapsulating network resources***
- ***Notion of Network resources as scheduled Grid services.***

Why Dynamic High-Performance Networks?

- Support data-intensive Grid applications
- Gives adequate and uncontested bandwidth to an application's burst
- Employs circuit-switching of large flows of data to avoid overheads in breaking flows into small packets and delays routing
- Is capable of automatic end-to-end path provisioning
- Is capable of automatic wavelength switching
- Provides a set of protocols for managing dynamically provisioned wavelengths

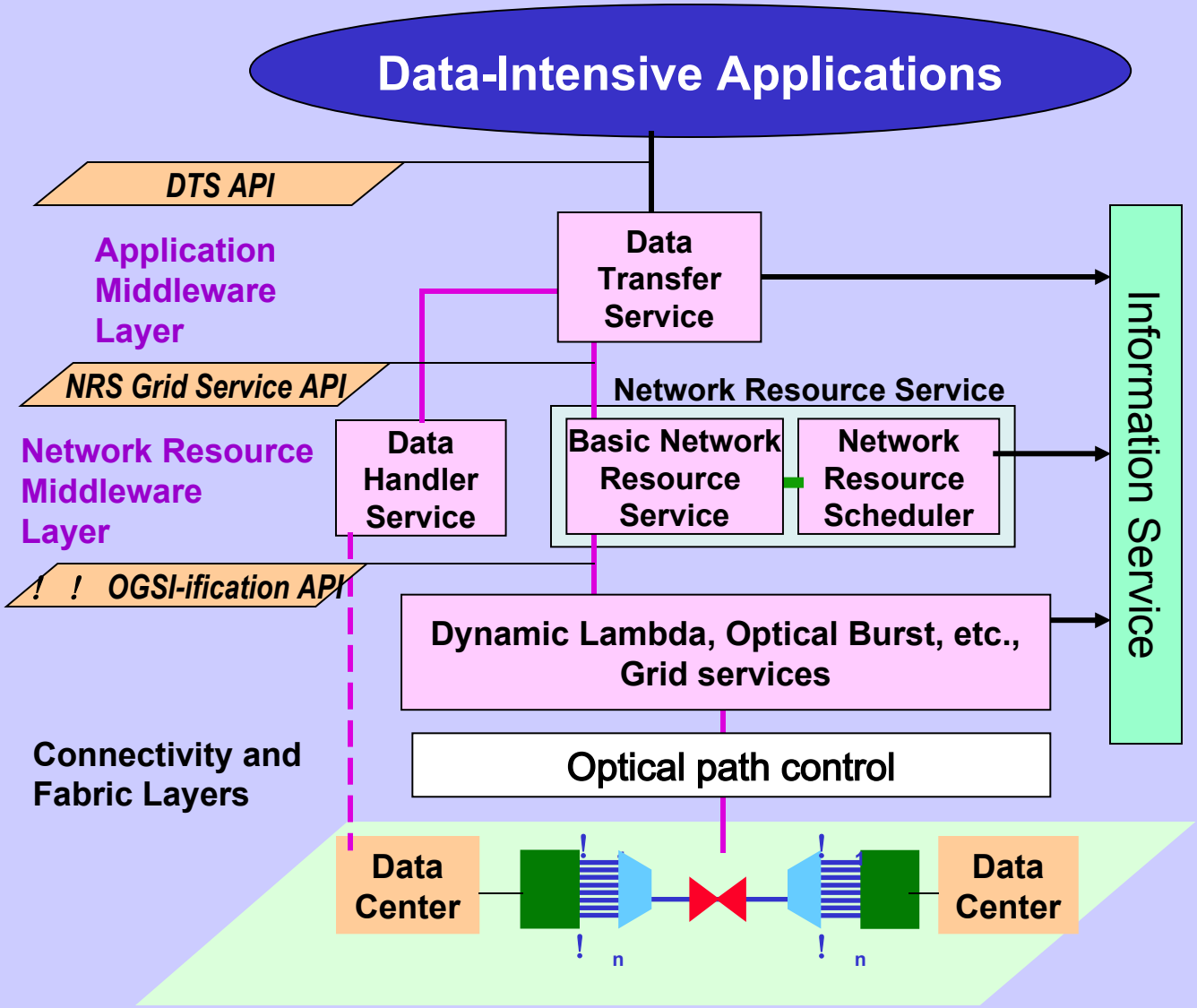
Why DWDM-RAM ?

- New platform for data intensive (Grid) applications
  - Encapsulates "optical network resources" into a service framework to support dynamically provisioned and advanced data-intensive transport services
  - Offers network resources as Grid services for Grid computing
  - Allows cooperation of distributed resources
  - Provides a generalized framework for high performance applications over next generation networks, not necessary optical end-to-end
  - Yields good overall utilization of network resources

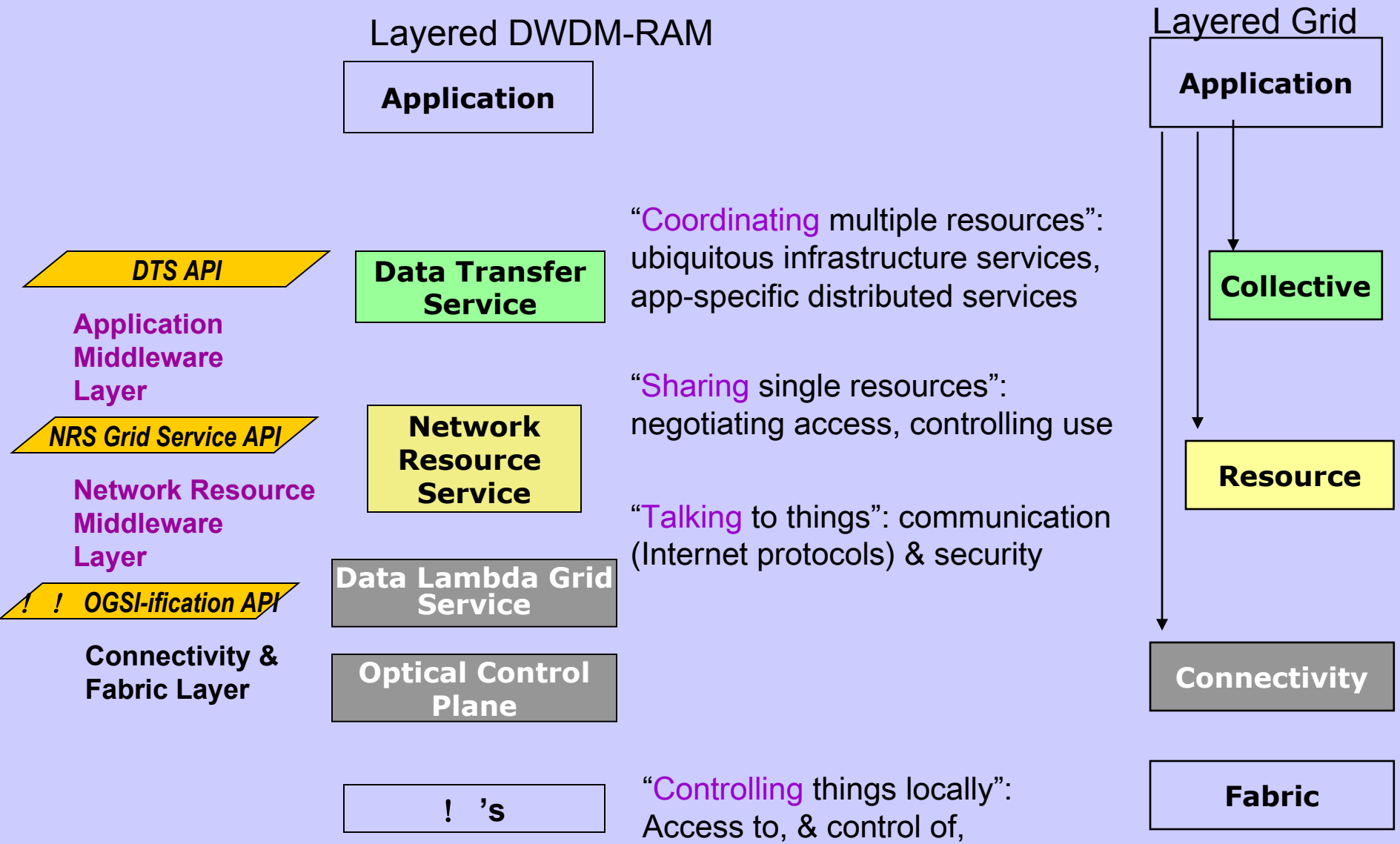
DWDM-RAM

- The generic middleware architecture consists of two planes over an underlying dynamic optical network
  - Data Grid Plane
  - Network Grid Plane
- The middleware architecture modularizes components into services with well-defined interfaces
- DWDM-RAM separates services into 2 principal service layers
  - Application Middleware Layer: Data Transfer Service, Workflow Service, etc.
  - Network Resource Middleware Layer: Network Resource Service, Data Handler Service, etc.
- And a Dynamic Lambda Grid Service over a Dynamic Optical Network

DWDM-RAM Architecture

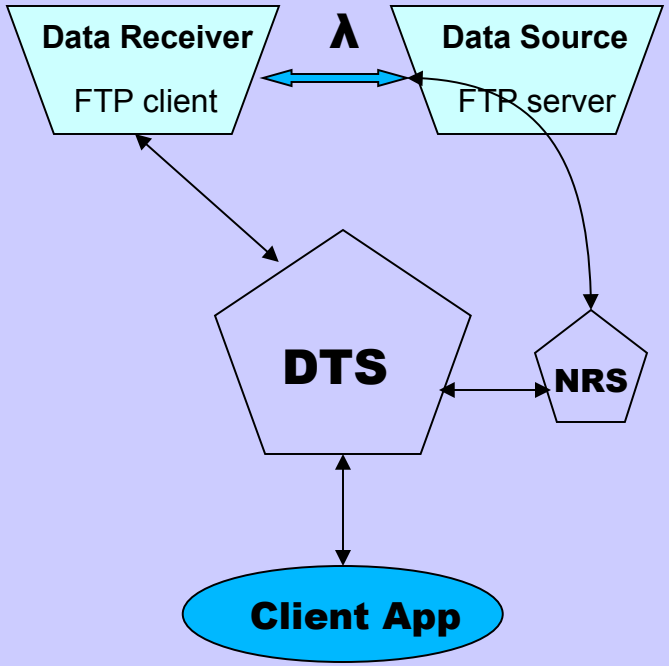


DWDM-RAM vs. Layered Grid Architecture



Data Transfer Service Layer

- Presents an OGSI interface between an application and a system – receives high-level requests, policy-and-access filtered, to transfer named blocks of data
- Reserves and coordinates necessary resources: network, processing, and storage
- Provides Data Transfer Scheduler Service (DTS)
- Uses OGSI calls to request network resources



Network Resource Service Layer

- Provides an OGSI-based interface to network resources
- Provides an abstraction of "communication channels" as a network service
- Provides an explicit representation of network resources scheduling model
- Enables capabilities for dynamic on-demand provisioning and advance scheduling
- Maintains schedules and provisions resources in accordance with the schedule

The Network Resource Service

- On Demand
  - Constrained window
  - Under-constrained window
- Advance Reservation
  - Constrained window
    - Tight window, fits the transference time closely
  - Under-constrained window
    - Large window, fits the transference time loosely
    - Allows flexibility in the scheduling

Dynamic Lambda Grid Service

- Presents an OGSI interface between the network resource service and the network resources of the underlying network
- Establishes, controls, and deallocates complete paths across both optical and electronic domains
- Operates over a dynamic optical network

An Application Scenario

A High Energy Physics group may wish to move 100 Terabytes data block from a particular run or set of events at an accelerator facility to its local or remote computational machine farm for extensive analysis

- Client requests: "Copy data X to the local store on machine Y after 1:00 and before 3:00."

- Client receives a "ticket" which describes the resultant scheduling and provides a method for modifying and monitoring the scheduled job

An Application Scenario (cont'd)

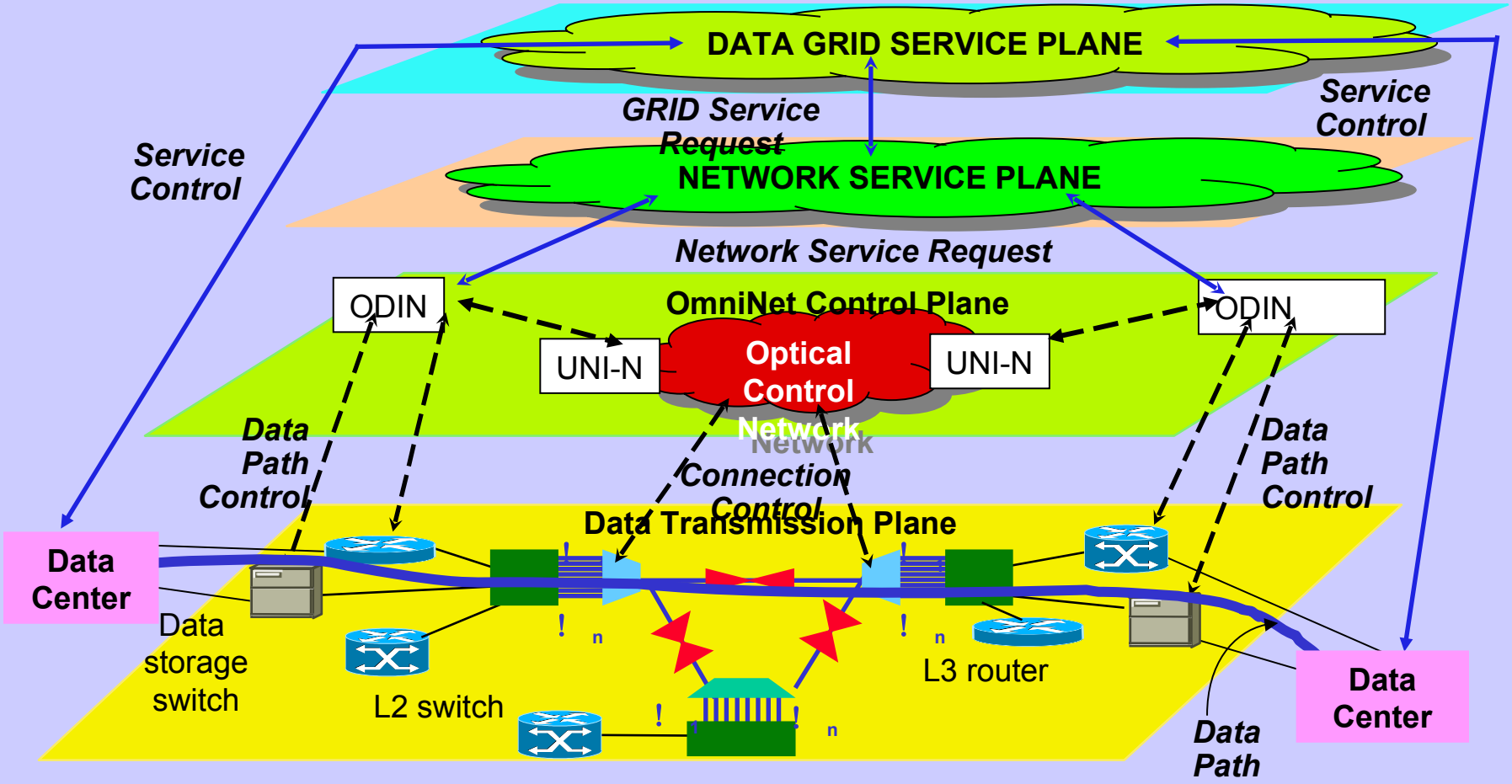
- At application level: Data Transfer Scheduler Service creates a tentative plan for data transfers that satisfies multiple requests over multiple network resources distributed at various sites
- At middleware level: A network resource schedule is formed based on the understanding of the dynamical lightpath provisioning capability of the underlying network and its topology and connectivity
- At resource provisioning level: Actual physical optical network resources are provisioned and allocated at the appropriate time for a transfer operation
- Data Handler Service on the receiving node is contacted to initiate the transfer
- At the end of the data transfer process, the network resources are de-allocated and returned to the pool

NRS Interface and Functionality

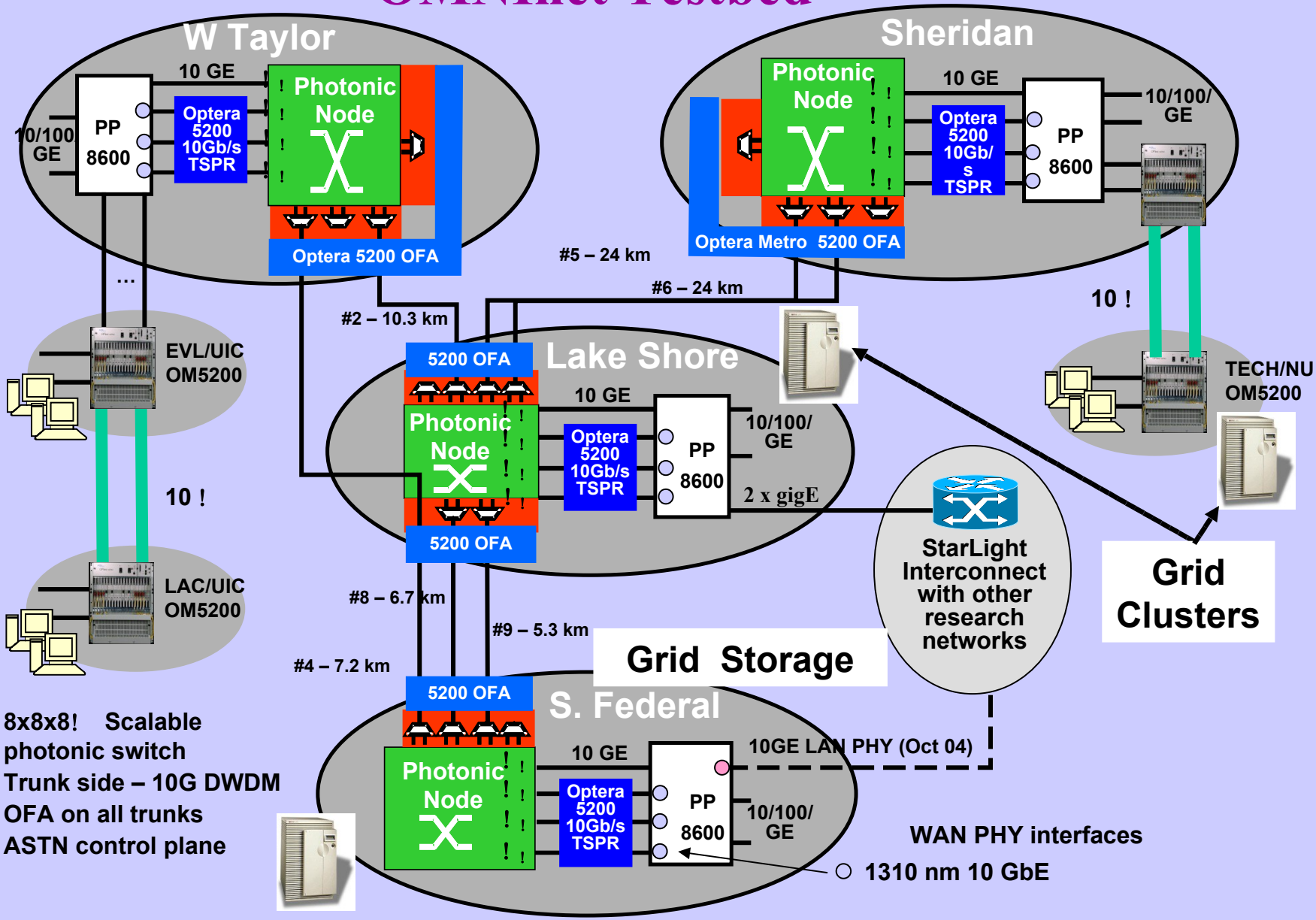
```
// Bind to an NRS service:
NRS = lookupNRS(address);
//Request cost function evaluation
request = {pathEndpointOneAddress,
           pathEndpointTwoAddress,
           duration,
           startAfterDate,
           endBeforeDate};
ticket = NRS.requestReservation(request);
// Inspect the ticket to determine success, and to find
the currently scheduled time:
ticket.display();
// The ticket may now be persisted and used
from another location
NRS.updateTicket(ticket);
// Inspect the ticket to see if the reservation's scheduled time has changed, or
verify that the job completed, with any relevant status information:
ticket.display();
```

Testbed and Experiments

- Experiments have been performed on the OMNInet
  - End-to-end FTP transfer over a 1Gbps link



OMNInet Testbed

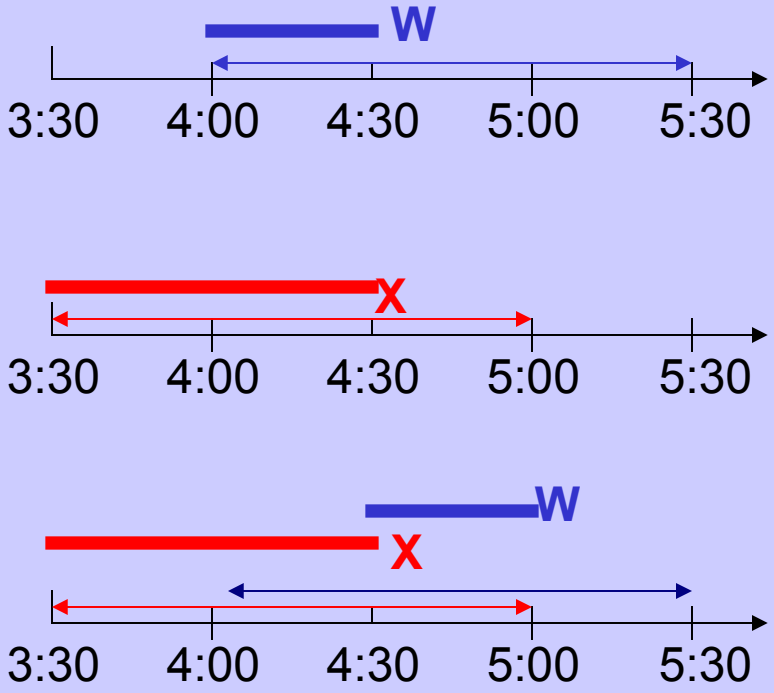


- 8x8x8! Scalable photonic switch
- Trunk side – 10G DWDM
- OFA on all trunks
- ASTN control plane

The Network Resource Scheduler Service

Under-constrained window

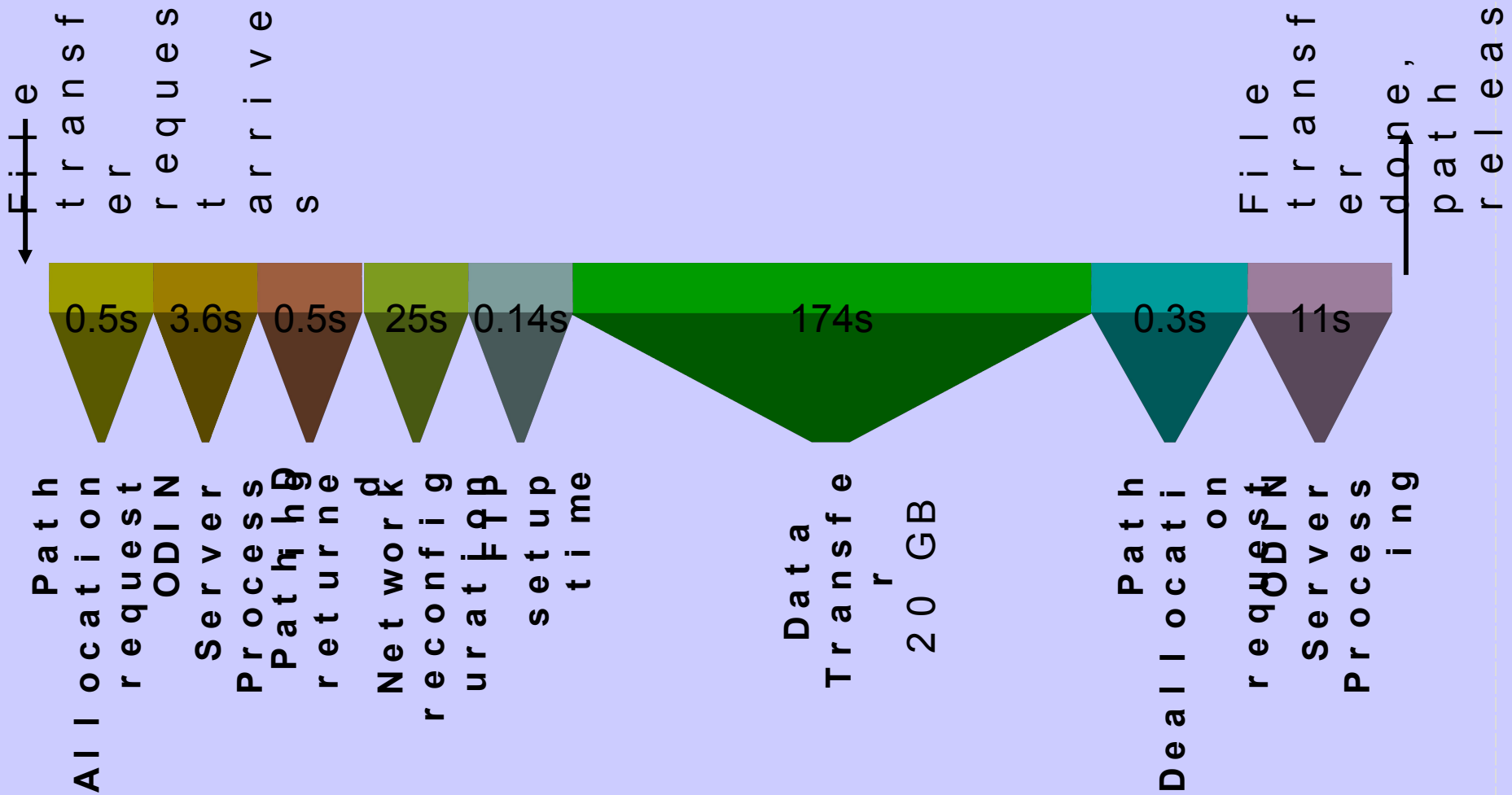
- Request for 1/2 hour between 4:00 and 5:30 on Segment D granted to User W at 4:00
- New request from User X for same segment for 1 hour between 3:30 and 5:00
- Reschedule user W to 4:30; user X to 3:30. Everyone is happy.



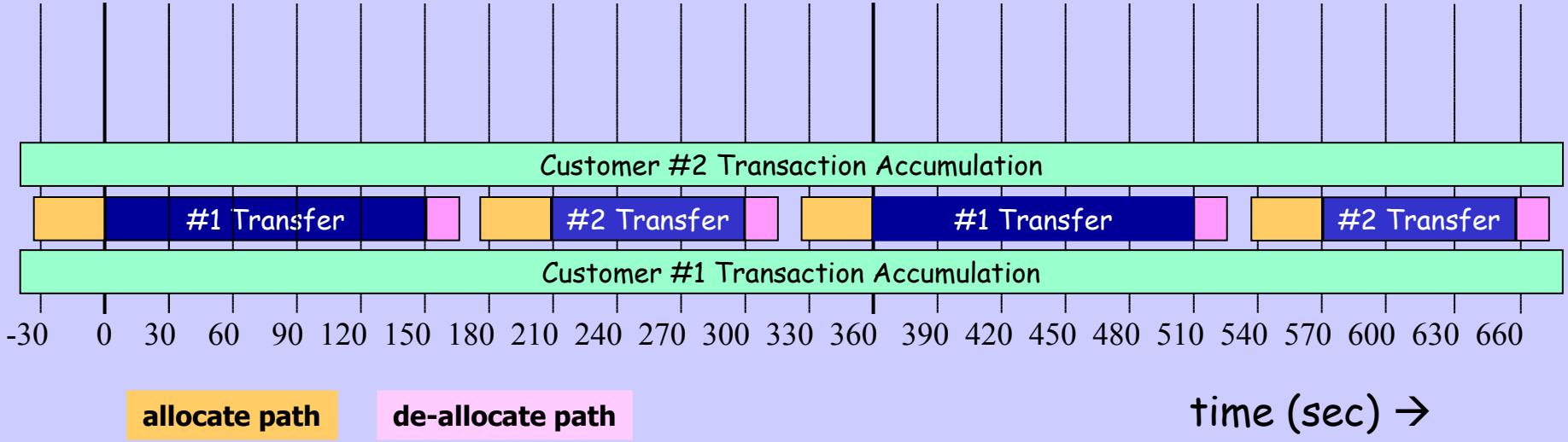
Route allocated for a time slot; new request comes in; 1st route can be rescheduled for a later slot within window to accommodate new request

20GB File Transfer

Initial Performance measure: End-to-End Transfer Time



Transaction Demonstration Time Line
6 minute cycle time

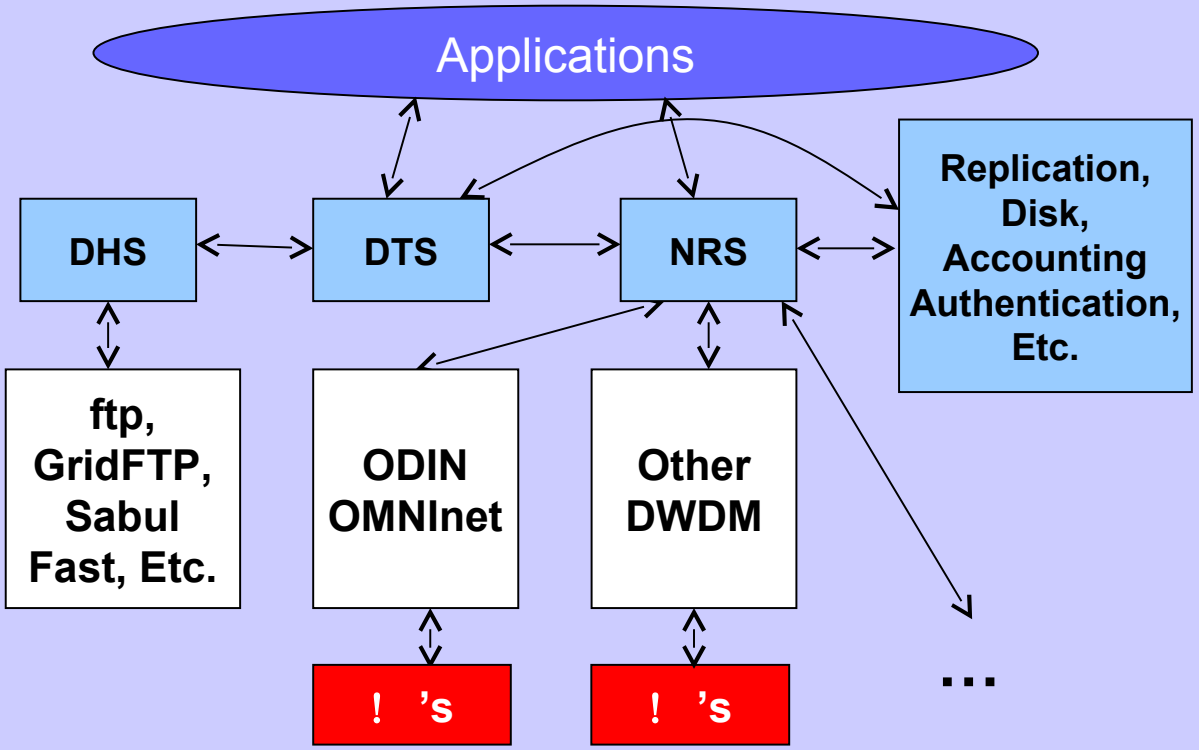


Conclusion

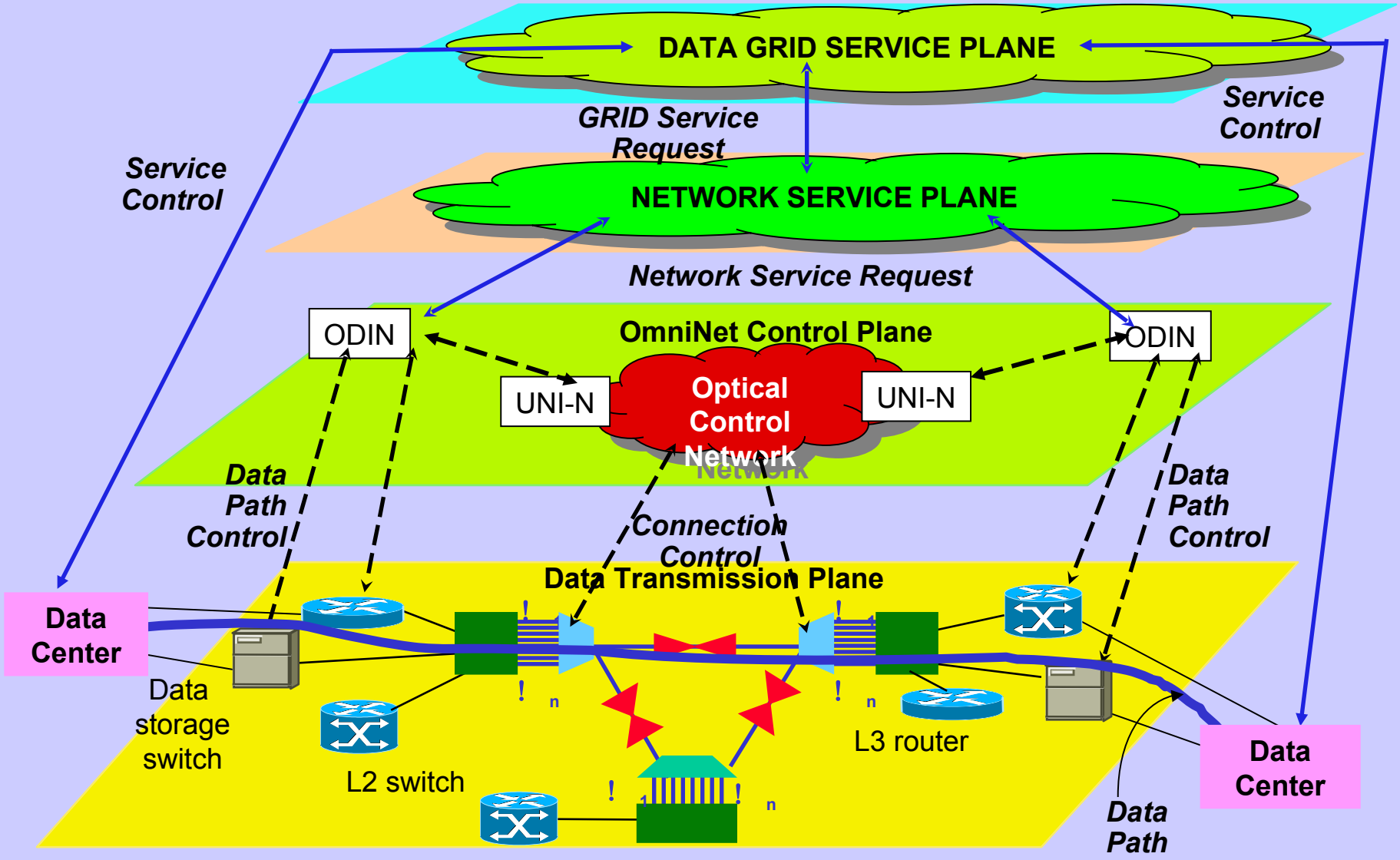
- The DWDM platform forges close cooperation between data intensive Grid applications and network resources
- The DWDM-RAM architecture yields Data Intensive Services that best exploit Dynamic Optical Networks
- Network resources become actively managed, scheduled services
- This approach maximizes the satisfaction of high-capacity users while yielding good overall utilization of resources
- The service-centric approach is a **foundation for new types of services**

Back up slides

DWDM-RAM Prototype Implementation



DWDM-RAM Service Control Architecture



Application Level Measurements

File size:	20 GB
Path allocation:	29.7 secs
Data transfer setup time:	0.141 secs
FTP transfer time:	174 secs
Maximum transfer rate:	935 Mbits/sec
Path tear down time:	11.3 secs
Effective transfer rate:	762 Mbits/sec

The Network Resource Service (NRS)

- Provides an OGSI-based interface to network resources
- Request parameters
  - Network addresses of the hosts to be connected
  - Window of time for the allocation
  - Duration of the allocation
  - Minimum and maximum acceptable bandwidth (future)

The Network Resource Service

- Provides the network resource
  - On demand
  - By advance reservation
- Network is requested within a window
  - Constrained
  - Under-constrained

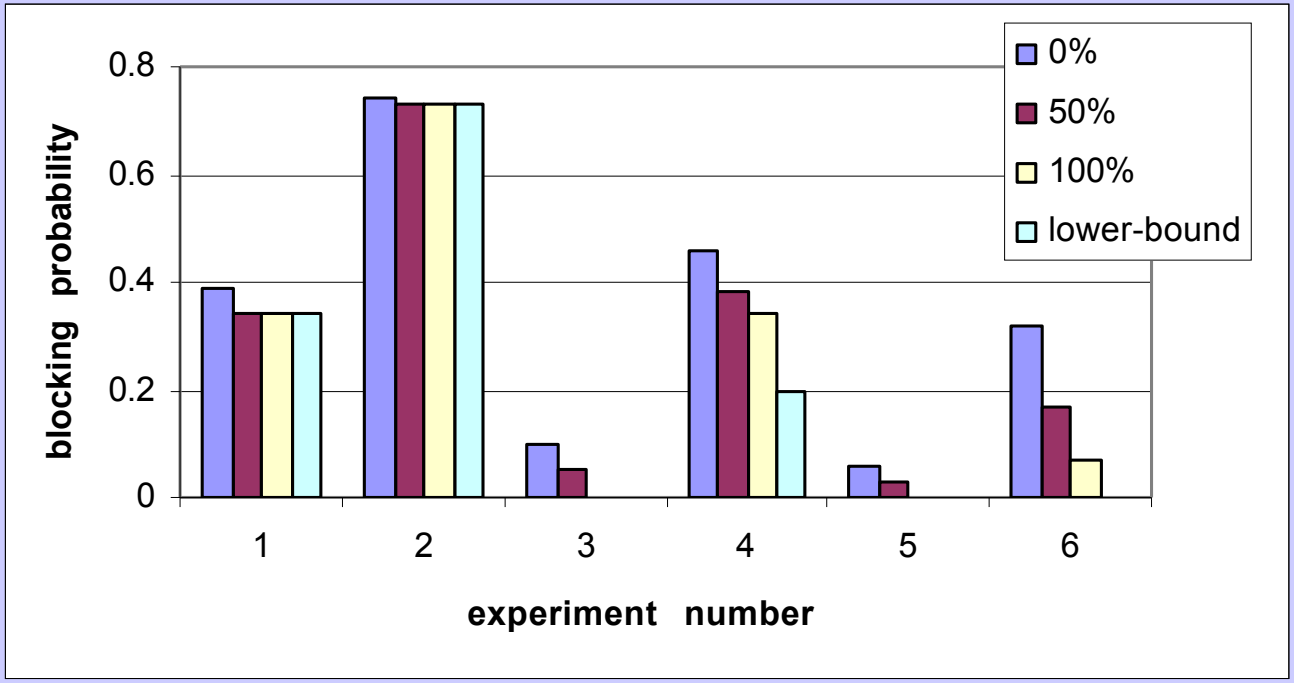
OMNInet Testbed

- Four-node multi-site optical metro testbed network in Chicago -- the first 10GigE service trial when installed in 2001
- Nodes are interconnected as a partial mesh with lightpaths provisioned with DWDM on dedicated fiber.
- Each node includes a MEMs-based WDM photonic switch, Optical Fiber Amplifier (OFA), optical transponders, and high-performance Ethernet switch.
- The switches are configured with four ports capable of supporting 10GigE.
- Application cluster and compute node access is provided by Passport 8600 L2/L3 switches, which are provisioned with 10/100/1000 Ethernet user ports, and a 10GigE LAN port.
- Partners: SBC, Nortel Networks, iCAIR/Northwestern University

Optical Dynamic Intelligent Network Services (ODIN)

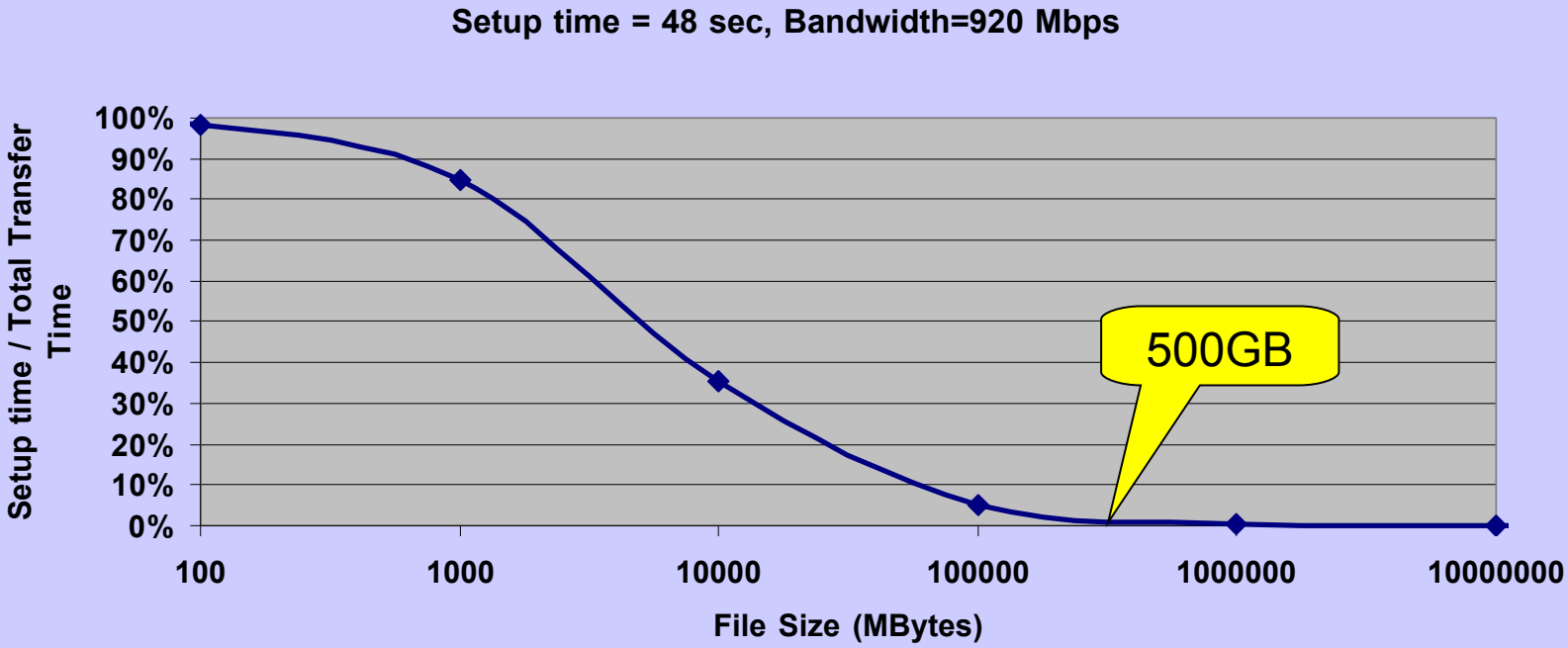
- Software suite that controls the OMNInet through lower-level API calls
- Designed for high-performance, long-term flow with flexible and fine grained control
- Stateless server, which includes an API to provide path provisioning and monitoring to the higher layers

Blocking probability
Under-constrained requests



Overheads - Amortization

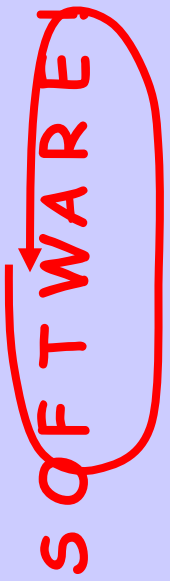
When dealing with data-intensive applications, overhead is insignificant!



Grids urged us to think End-to-End Solutions

Look past boxes, feeds, and speeds

Apps such as Grids call for a complex mix of:



Bit-blasting

Finesse (*granularity of control)*

\+ Virtualization (*access to diverse knobs*)

\+ Resource bundling (*network AND …*)

\+ Multi-Domain Security (*AAA to start*)

\+ Freedom from GUIs, human intervention

Our recipe is a software-rich symbiosis of Packet and Optical products

Optical Abundant Bandwidth Meets Grid

The Data Intensive App Challenge:

Emerging data intensive applications in the field of HEP, astro-physics, astronomy, bioinformatics, computational chemistry, etc., require extremely high performance and long term data flows, scalability for huge data volume, global reach, adjustability to unpredictable traffic behavior, and integration with multiple Grid resources.

Response: DWDM-RAM

An architecture for data intensive Grids enabled by next generation dynamic optical networks, incorporating new methods for lightpath provisioning. **DWDM-RAM** is designed to meet the networking challenges of extremely large scale Grid applications. Traditional network infrastructure cannot meet these demands, especially, requirements for intensive data flows