Lambda Data Grid

A Grid Computing Platform where Communication Function is in Balance with Computation and Storage

Tal Lavian

Outline of the presentation

- Introduction to the problems
- Aim and scope
- Main contributions
  - Lambda Grid architecture
  - Network resource encapsulation
  - Network schedule service
  - Data-intensive applications
- Testbed, implementation, performance evaluation
- Issue for further research
- Conclusion

Introduction

- Growth of large, geographically dispersed research
  - Use of simulations and computational science
  - Vast increases in data generation by e-Science
- Challenge: Scalability - "Peta" network capacity
- Building a new grid-computing paradigm, which fully harnesses communication
  - Like computation and storage
- Knowledge plane: True viability of global VO

Lambda Data Grid Service

Lambda Data Grid Service architecture interacts with Cyber-infrastructure, and overcomes data limitations efficiently & effectively by:

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

Motivation

- New e-Science and its distributed architecture limitations
- The Peta Line – PetaByte, PetaFlop, PetaBits/s
- Growth of optical capacity
- Transmission mismatch
- Limitations of L3 and public networks for data-intensive e-Science

Three Fundamental Challenges

- Challenge #1: Packet Switching – an inefficient solution for data-intensive applications
  - Elephants and Mice
  - Lightpath cut-through
  - Statistical multiplexing
  - Why not lightpath (circuit) switching?
- Challenge #2: Grid Computing Managed Network Resources
  - Abstract and encapsulate
  - Grid networking
  - Grid middleware for Dynamic Optical Provisioning
  - Virtual Organization (VO) as reality
- Challenge #3: Manage BIG Data Transfer for e-Science
  - Visualization example

Aim and Scope

- Build an architecture that can orchestrate network resources in conjunction with computation, data, storage, visualization, and unique sensors
  - The creation of an effective network orchestration for e-Science applications, with vastly more capability than the public Internet
  - Fundamental problems faced by e-Science research today requires a solution
- Scope
  - Concerns mainly with middleware and application interface
  - Concerns with Grid Services
  - Assumes an agile underlying Optical Network
  - Pays little attention to packet switched networks

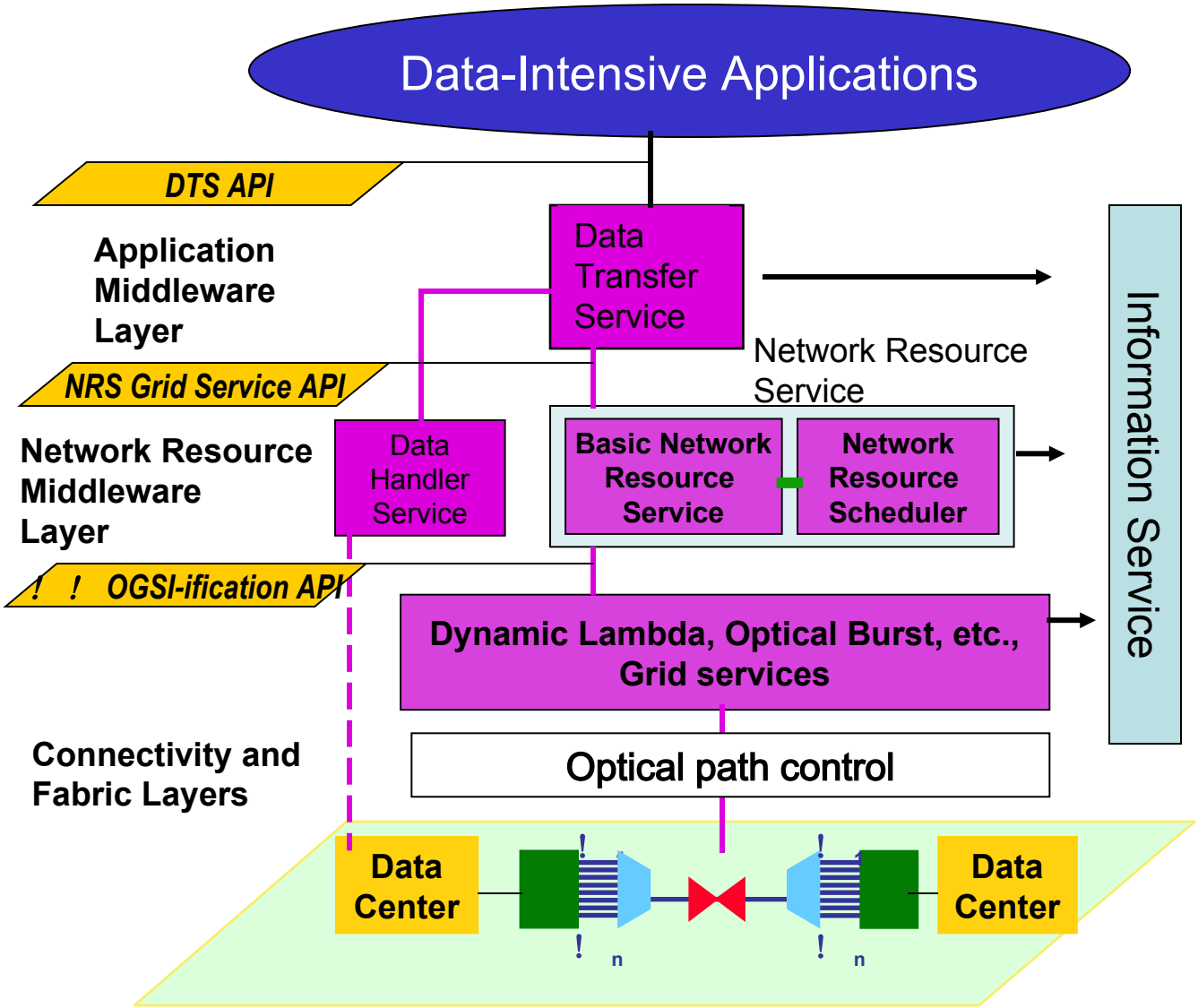
Major Contributions

- Promote the network to a "First Class" resource citizen
- Abstract and encapsulate the network resources into a set of Grid Services
- Orchestrate end-to-end resources
- Schedule network resources
- Design and implement an Optical Grid prototype

Architecture for Grid Network services

- This new architecture is necessary for
  - Deploying Lambda switching in the underlying networks
  - Encapsulating network resources into a set of Grid Network services
  - Supporting data-intensive applications
- Features of the architecture
  - App layer for isolating network service users from complexity of the underlying network
  - Middleware network resource layer for network service encapsulation
  - Connectivity layer for communications

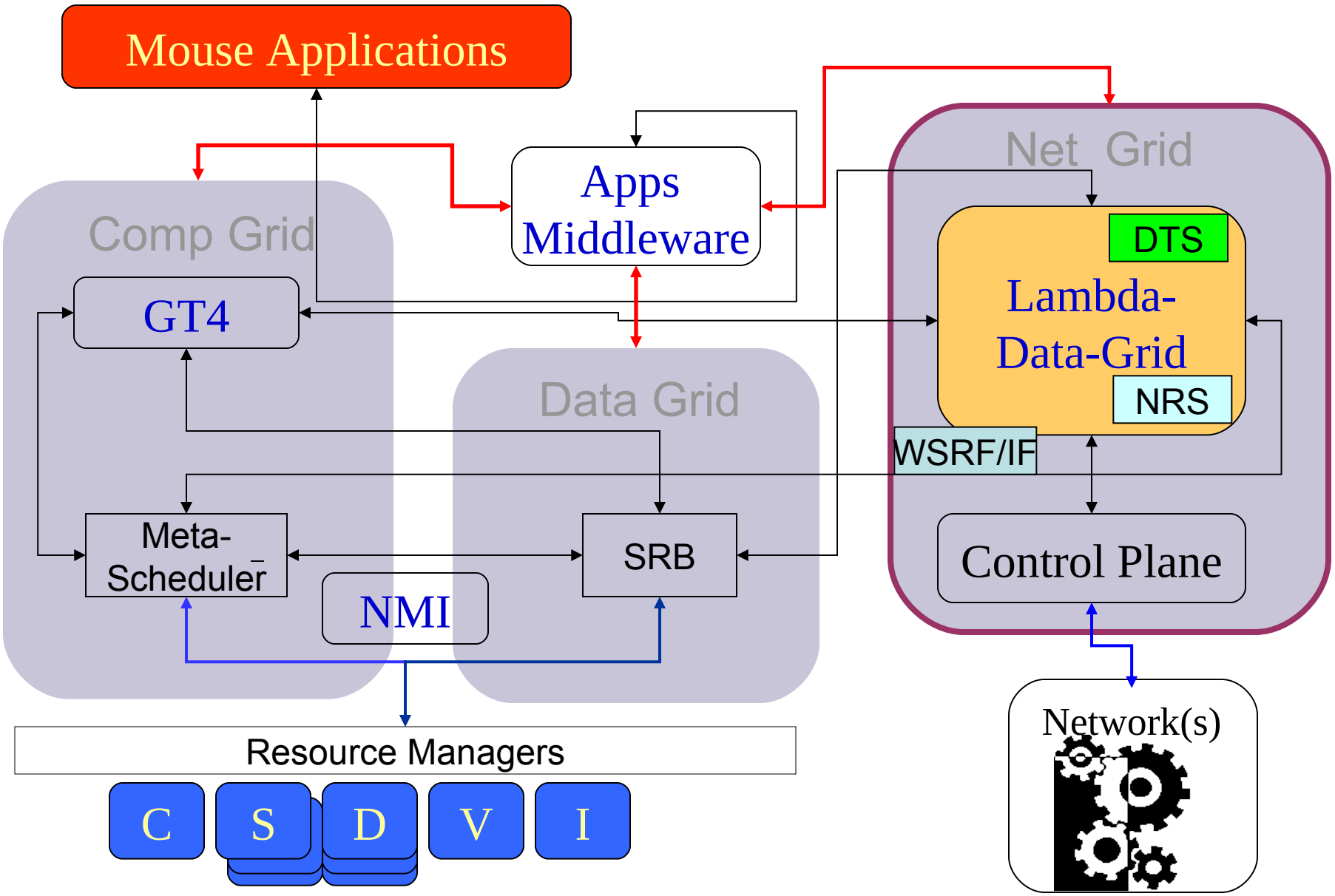
Architecture



Lambda Data Grid Architecture

- Optical networks as a "first class" resource, similar to computation and storage resources
- Orchestrate resources for data-intensive services, through dynamic optical networking
- Date Transfer Service (DTS)
  - presents an interface between the system and an application
  - Client requests – balance resources - scheduling constrains
- Network Resource Service (NRS)
  - Resource management service
- Grid Layered Architecture

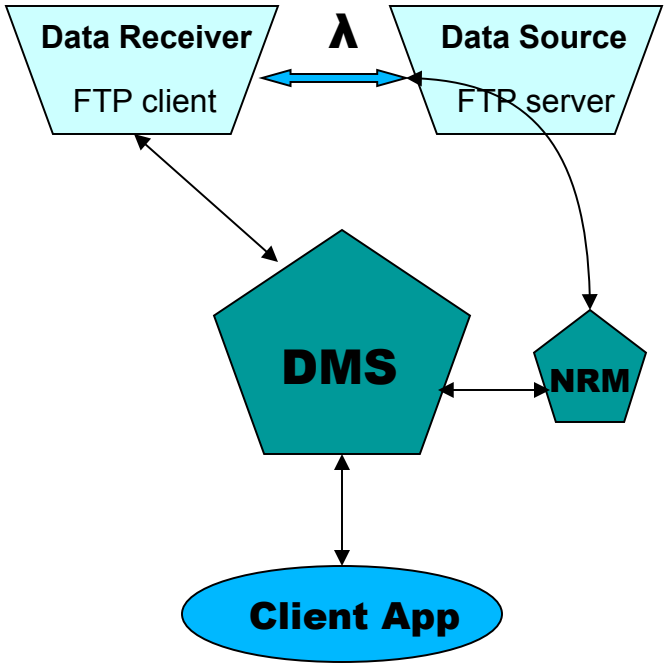
BIRN Mouse Example



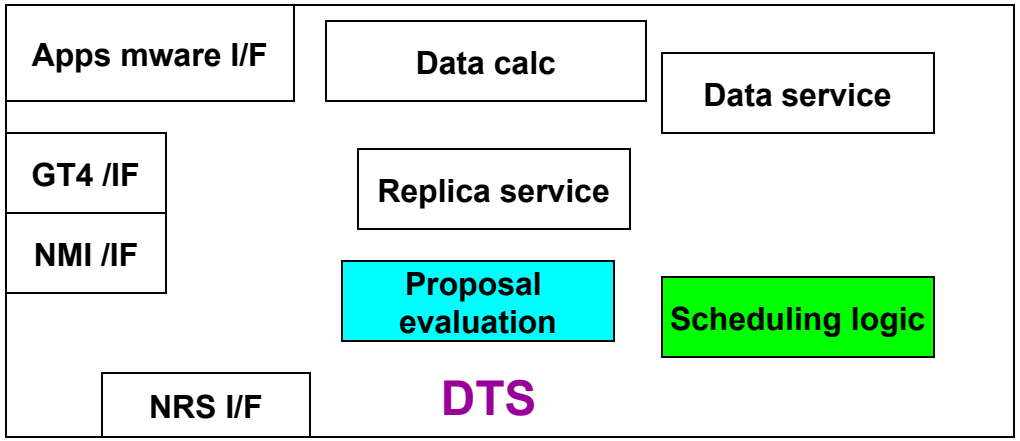
Network Resource Encapsulation

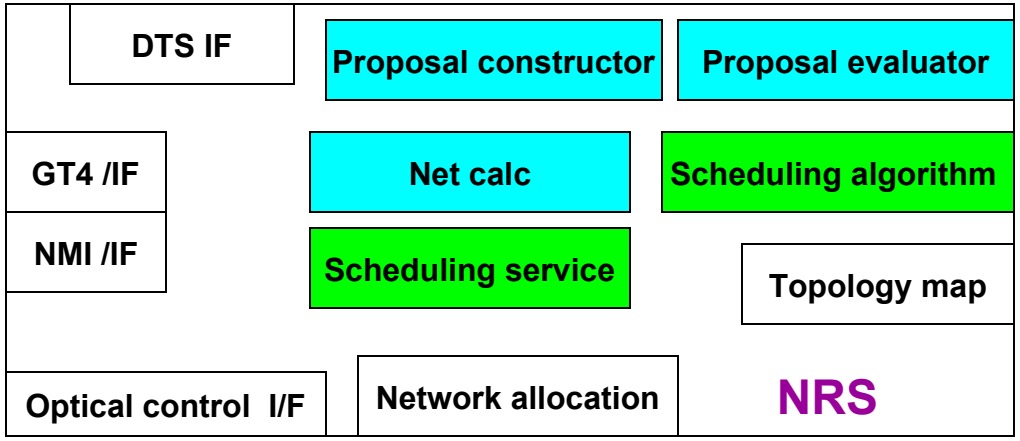
- To make network resource a "first class resource" like CPU and storage resources that can be scheduled
- Encapsulation is done by modularizing network functionality and providing proper interfaces

Data Management Service



DTS - NRS


Apps mware I/F
Data calc
Data service
GT4 /IF
Replica service
NMI /IF
Proposal evaluation
Scheduling logic
NRS I/F
DTS


DTS IF
Proposal constructor
Proposal evaluator
GT4 /IF
Net calc
Scheduling algorithm
NMI /IF
Scheduling service
Topology map
Optical control I/F
Network allocation
NRS

NRS Interface and Functionality

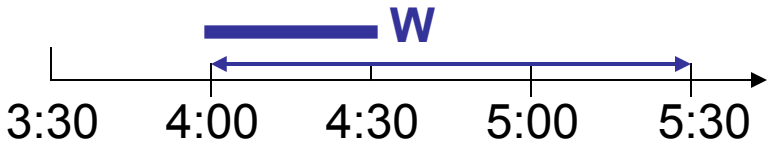
```
// Bind to an NRS service:
NRS = lookupNRS(address);
//Request cost function evaluation
request = {pathEndpointOneAddress,
           pathEndpointTwoAddress,
           duration,
           startAfterDate,
           endBeforeDate};
ticket = NRS.requestReservation(request);
// Inspect the ticket 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();
```

Network schedule service – an example of use

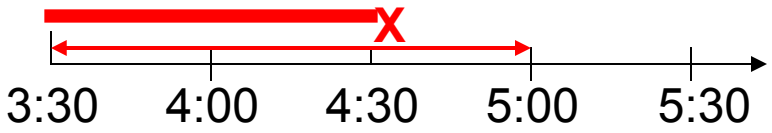
- Encapsulate it as another service at a level above the basic NRS

Example: Lightpath Scheduling

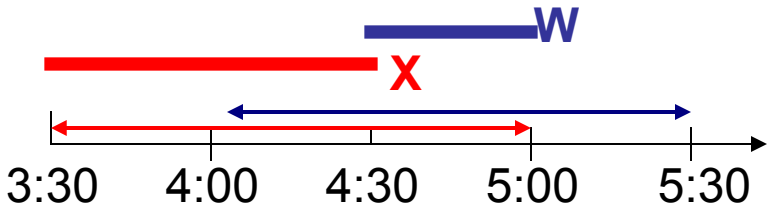
- Request for 1/2 hour between 4:00 and 5:30 on Segment D granted to User W at 4:00



- New request from User X for same segment for 1 hour between 3:30 and 5:00



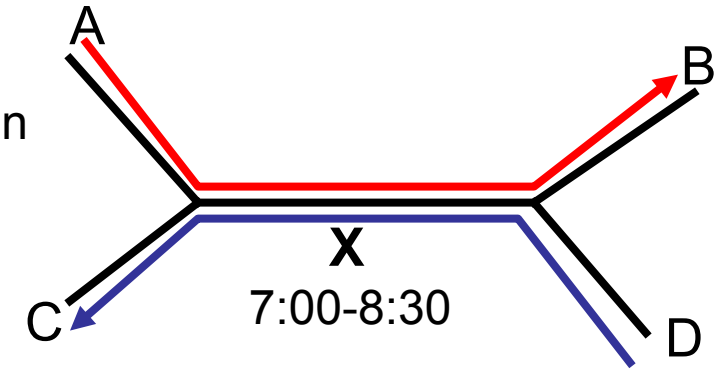
- Reschedule user W to 4:30; user X to 3:30. Everyone is happy.



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

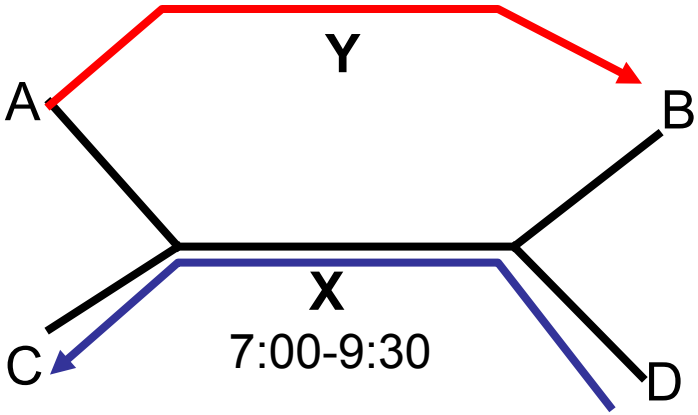
Scheduling Example - Reroute

- 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



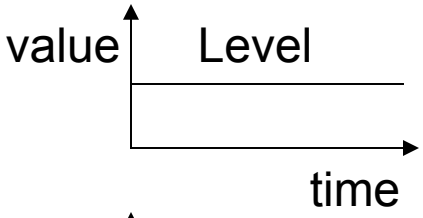
- New request for 2 hours between nodes C and D between 7:00 and 9:30 This route needs to use Segment X to be satisfied

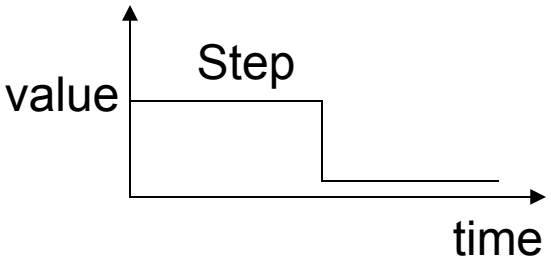
- Reroute the first request to take another path through the topology to free up Segment X for the 2nd request. Everyone is happy

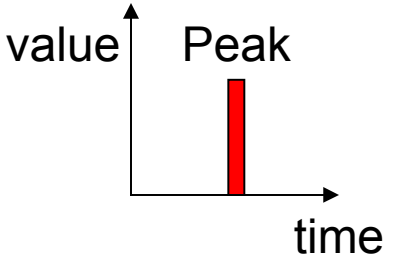


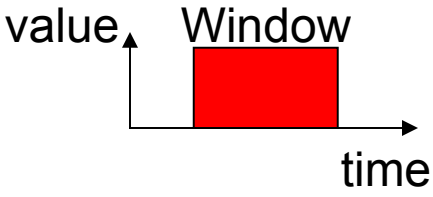
Route allocated; new request comes in for a segment in use; 1st route can be altered to use different path to allow 2nd to also be serviced in its time window

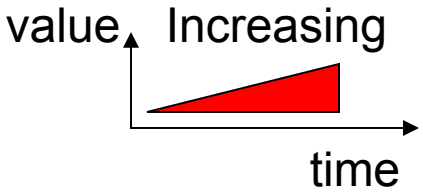
Scheduling - Time Value

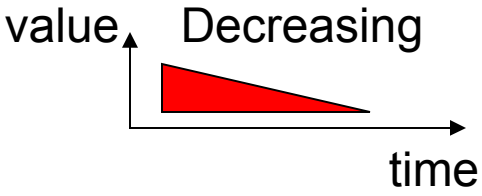


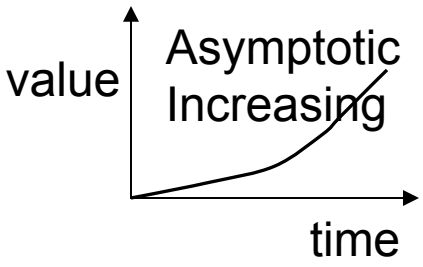


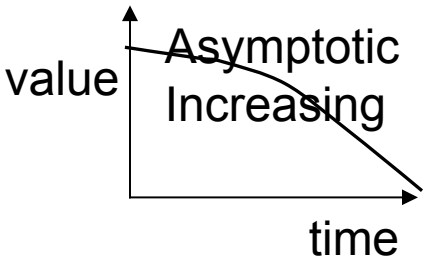




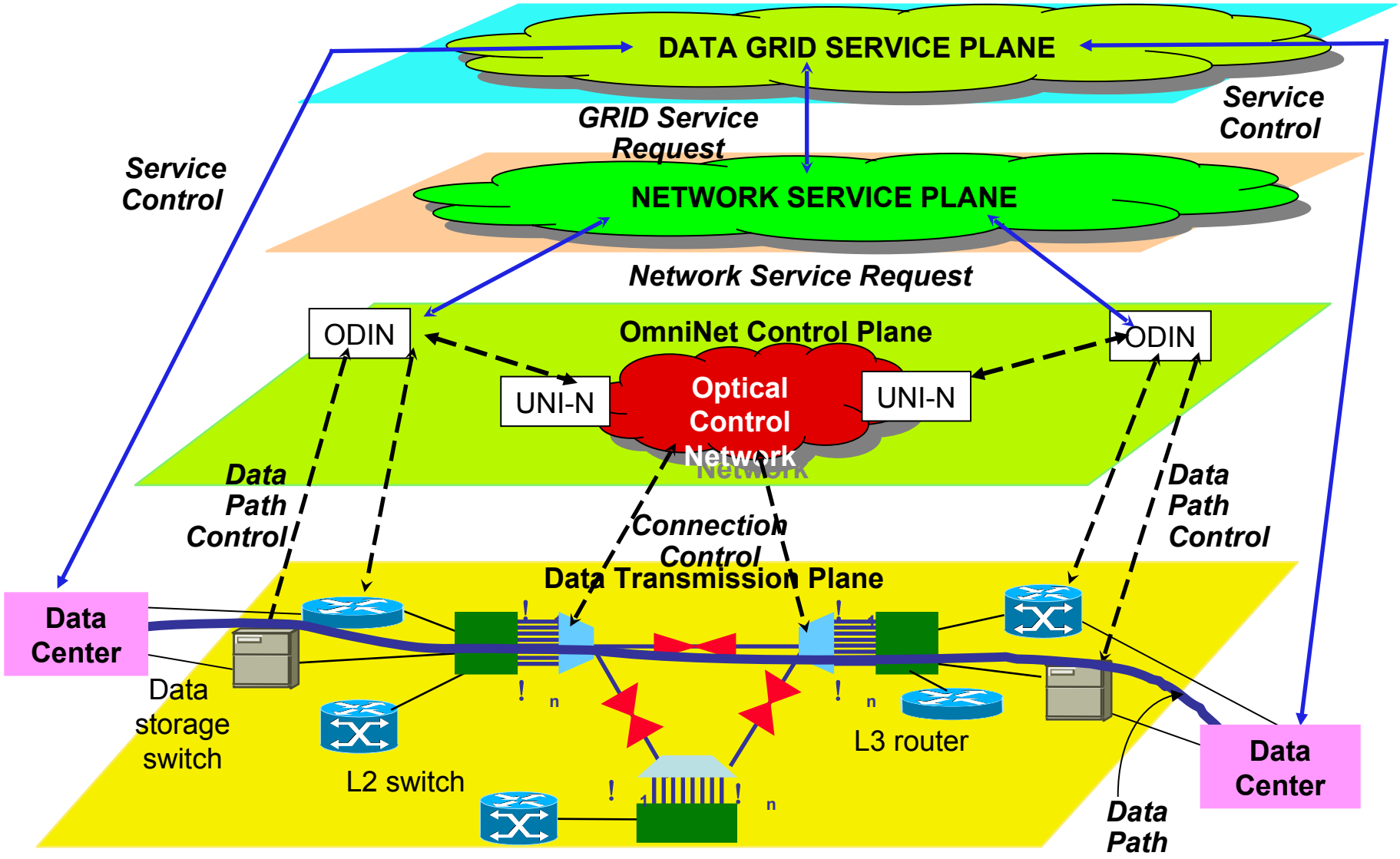




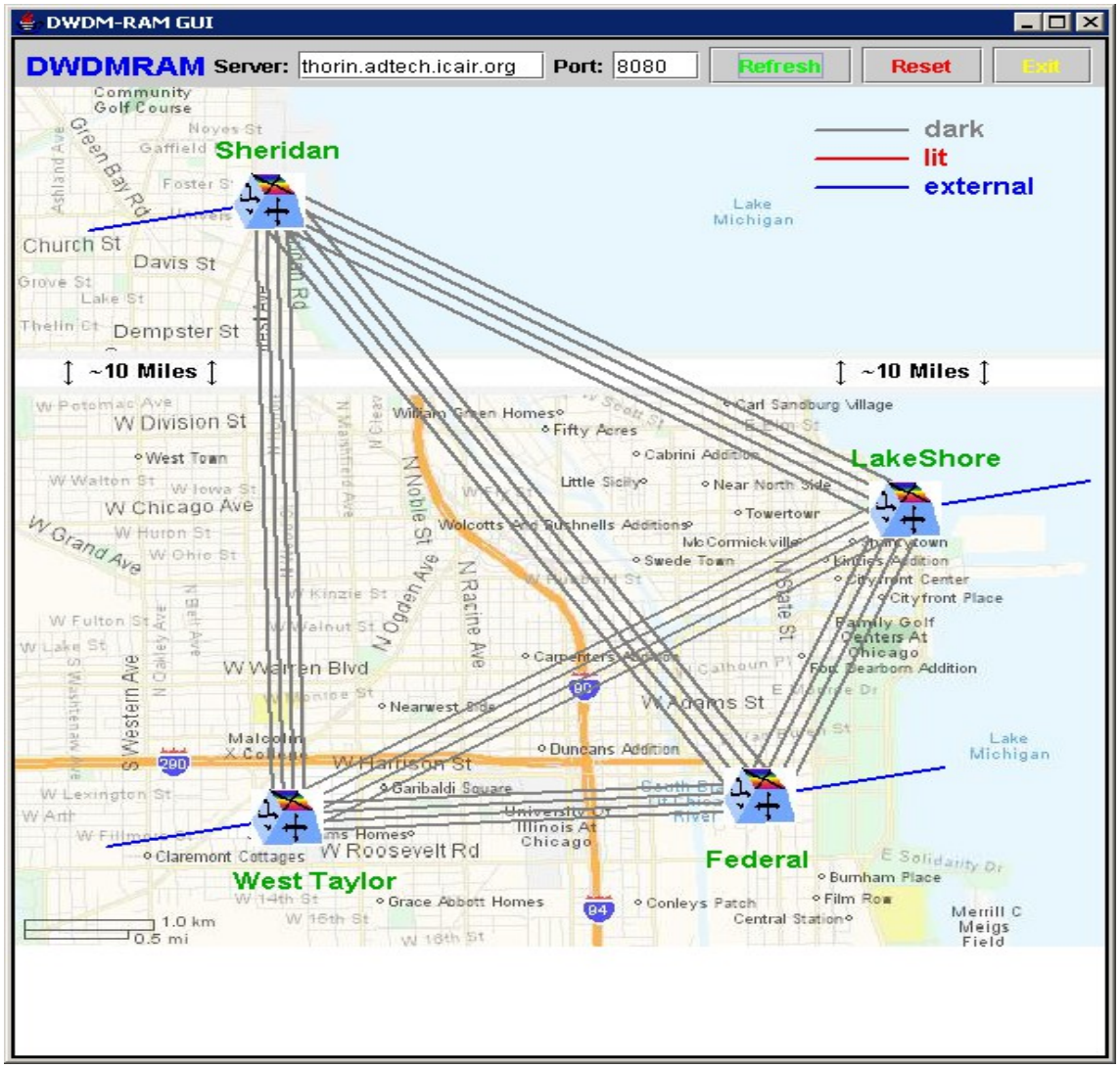




Service Control Architecture



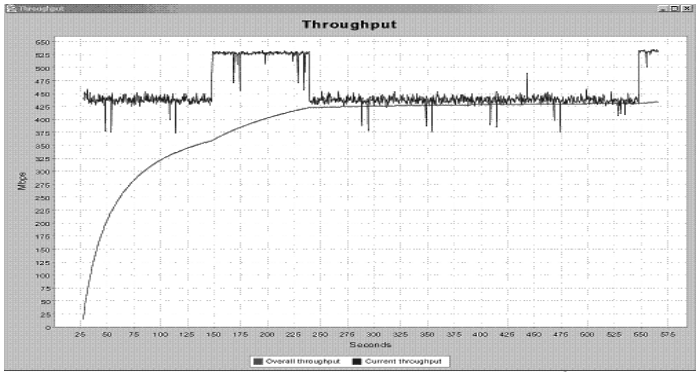
OMNI-View Lightpath Map



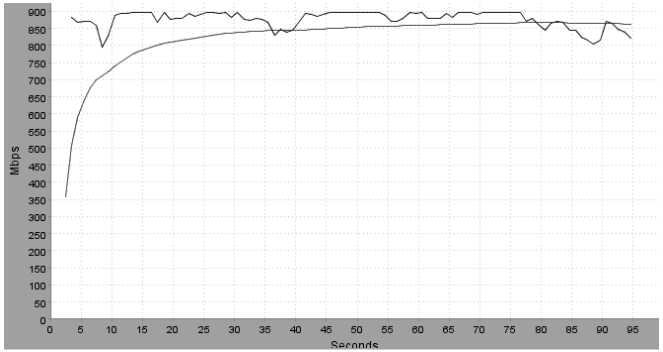
Experiments

1. Proof of concept between four nodes, two separate racks, about 10 meters
2. Grid Services - dynamically allocated 10Gbs Lambdas over four sites in the Chicago metro area, about 10km
3. Grid middleware - allocation and recovery of Lambdas between Amsterdam and Chicago, via NY and Canada, about 10,000km

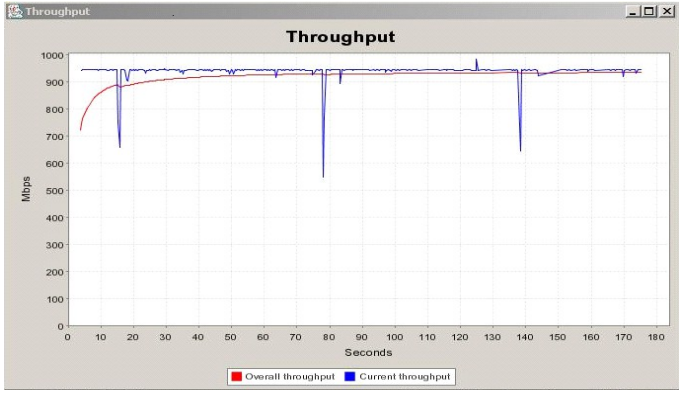
Results and Performance Evaluation



30 GB – Over OMNInet mem-to-mem



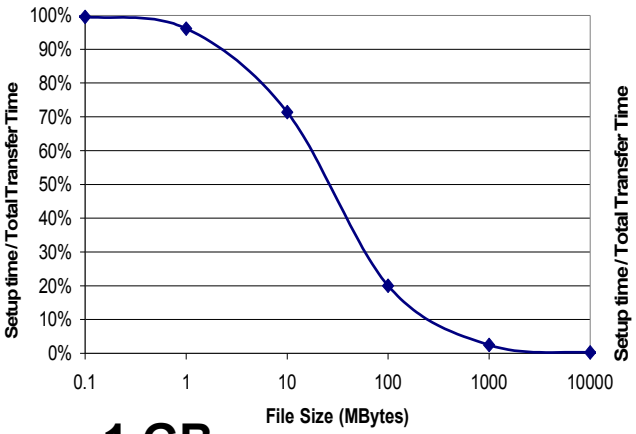
10 GB – Mem-to-mem –one rack



20 GB - Effective 920 Mbps

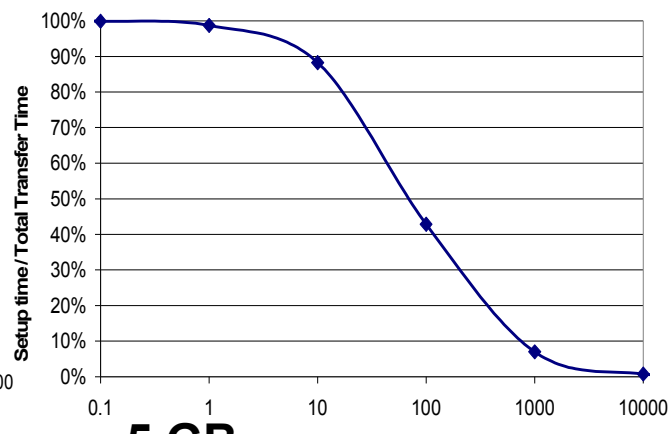
Results and Performance Evaluation
Overhead is Insignificant

Setup time = 2 sec, Bandwidth=100 Mbps



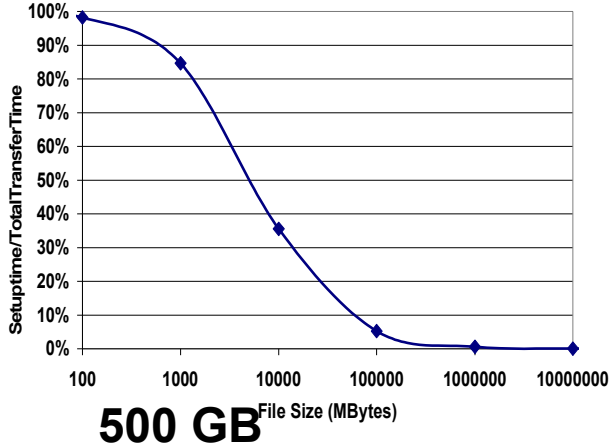
1 GB

Setup time = 2 sec, Bandwidth=300 Mbps



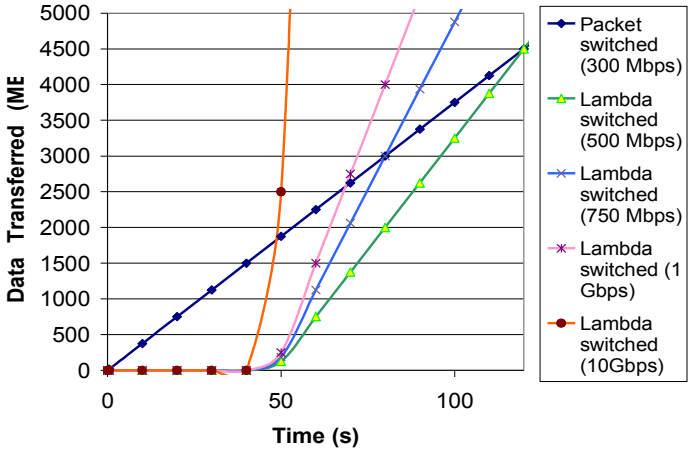
5 GB

Setup time = 48 sec, Bandwidth=920 Mbps

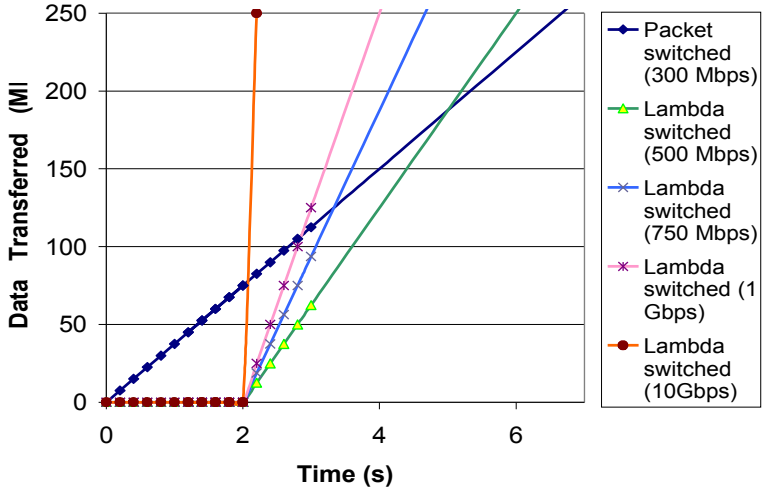


500 GB

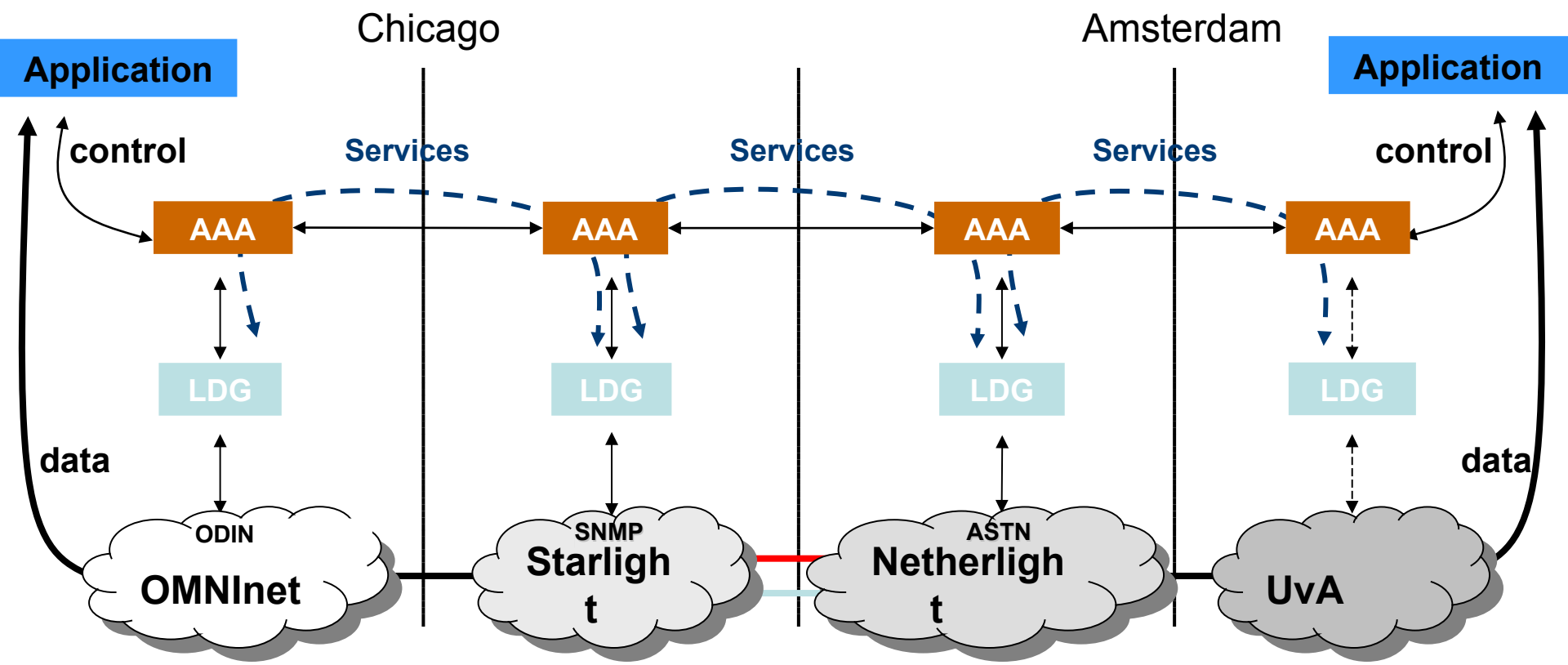
Optical path setup time = 48 sec



Optical path setup time = 2 sec

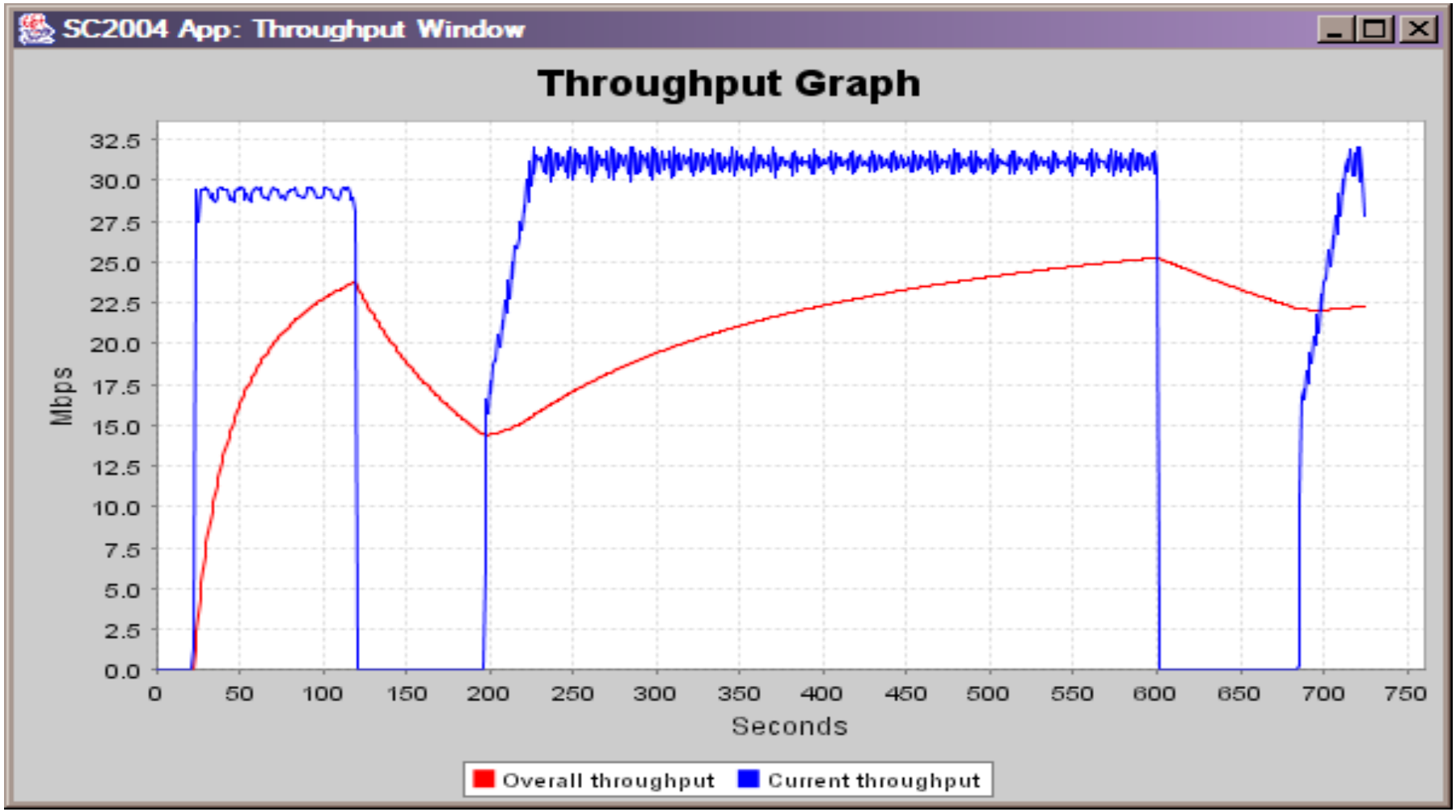


Super Computing CONTROL CHALLENGE



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

From 100 Days to 100 Seconds



Discussion: What I Have Done

- Deploying optical infrastructure for each scientific institute or large experiment would be cost prohibitive, depleting any research budget
- Unlike the Internet topology of “many-to-many”
  - “few-to-few” architecture
- LDG acquires knowledge of the communication requirements from applications, and builds the underlying cut-through connections to the right sites of an e-Science experiment
- New optimization to waste bandwidth
  - Last 30 years – bandwidth conservation
  - Conserve bandwidth – waste computation (silicon)
  - New idea – waste bandwidth

Discussion

- Lambda Data Grid architecture yields data-intensive services that best exploits 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

Conclusion - Promote the network to a first class resource citizen

- The network is no longer a pipe; it is a part of the Grid computing instrumentation
- it is not only an essential component of the Grid computing infrastructure but also an integral part of Grid applications
- Design of VO in a Grid computing environment is accomplished and lightpath is the vehicle
  - allowing dynamic lightpath connectivity while matching multiple and potentially conflicting application requirements, and addressing diverse distributed resources within a dynamic environment

Conclusion - Abstract and encapsulate the network resources into a set of Grid services

- Encapsulation of lightpath and connection-oriented, end-to-end network resources into a stateful Grid service, while enabling on-demand, advanced reservation, and scheduled network services
- Schema where abstractions are progressively and rigorously redefined at each layer
  - avoids propagation of non-portable implementation-specific details between layers
  - resulting schema of abstractions has general applicability

Conclusion- Orchestrate end-to-end resource

- A key innovation is the ability to orchestrate heterogeneous communications resources among applications, computation, and storage
  - across network technologies and administration domains

Conclusion- Schedule network resources

- (wrong) Assumption that the network is available at all times, to any destination
  - no longer accurate when dealing with big pipes
- Statistical multiplexing will not work in cases of few-to-few immense data transfers
- Built and demonstrated a system that allocates the network resources based on availability and **scheduling of full pipes**

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 ways 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 Large Scale Distributed Computing
  - Resource co/allocation and optimization with storage and computation
  - Grid system architecture
  - Enables new horizons for network optimization and Lambda scheduling
  - The network a white box – optimal scheduling and algorithms

Thank You_

The Future is Bright

_

- Imagine the next 10 years
- There are more questions than answers

Vision

- Lambda Data Grid provides the knowledge plane that allows e-Science applications to orchestrate enormous amounts of data over a dedicated Lightpath
  - Resulting in the true viability of global VO
- This enhances science research by allowing large distributed teams to work efficiently, utilizing simulations and computational science as a third branch of research
  - Understanding of the genome, DNA, proteins, and enzymes is prerequisite to modifying their properties and the advancement of synthetic biology

BIRN 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)	! Can't be done from home research institute	! Need fat unidirectional pipes ! Tight QoS requirements (jitter, delay, data loss)

Other Observations:

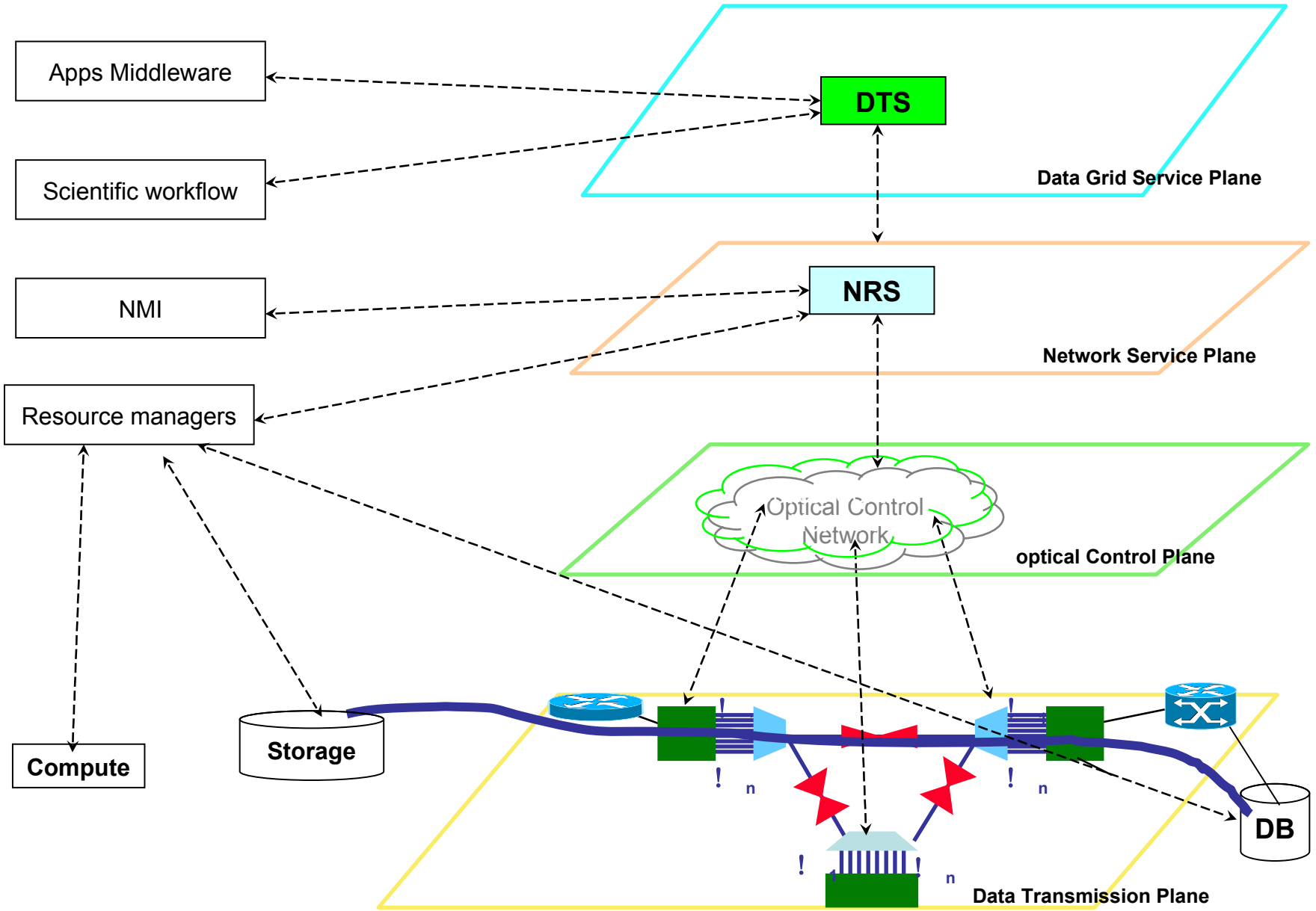
- **Not Feasible To Port Computation to Data**
- **Delays Preclude Interactive Research: Copy, Then Analyze**
- **Uncertain Transport Times Force A Sequential Process – Schedule Processing After Data Has Arrived**
- **No cooperation/interaction among Storage, Computation & Network Middlewares**

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

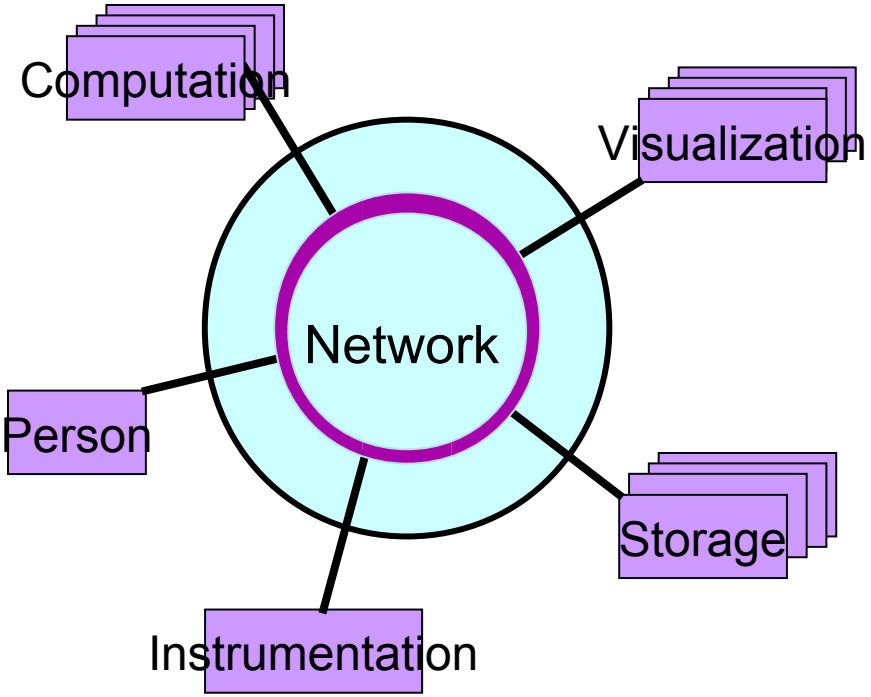
–

Backup Slides

Control Interactions



New Idea - 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

New Idea-
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

Conclusion

- New middleware to manage dedicated optical network
  - Integral to Grid middleware
- Orchestration of dedicated networks for e-Science use only
- Pioneer efforts in encapsulating the network resources into a Grid service
  - accessible and schedulable through the enabling architecture
  - opens up several exciting areas of research