Lambda Data Grid

An Agile Optical Platform for Grid Computing and Data-intensive Applications

Focus on BIRN Mouse application

Tal Lavian

Long - ☺ slides = reference only, will not be presented

Feedback & Response

- Issues:
  - Interface between applications and NRS
  - Information that would cross this interface
- Response:
  - eScience: Mouse (Integrating brain data across scales and disciplines. Part of BIRN at SDSC & OptIPuter)
  - Architecture that supports applications with network resource service and application middleware service
  - Detailed interface specification

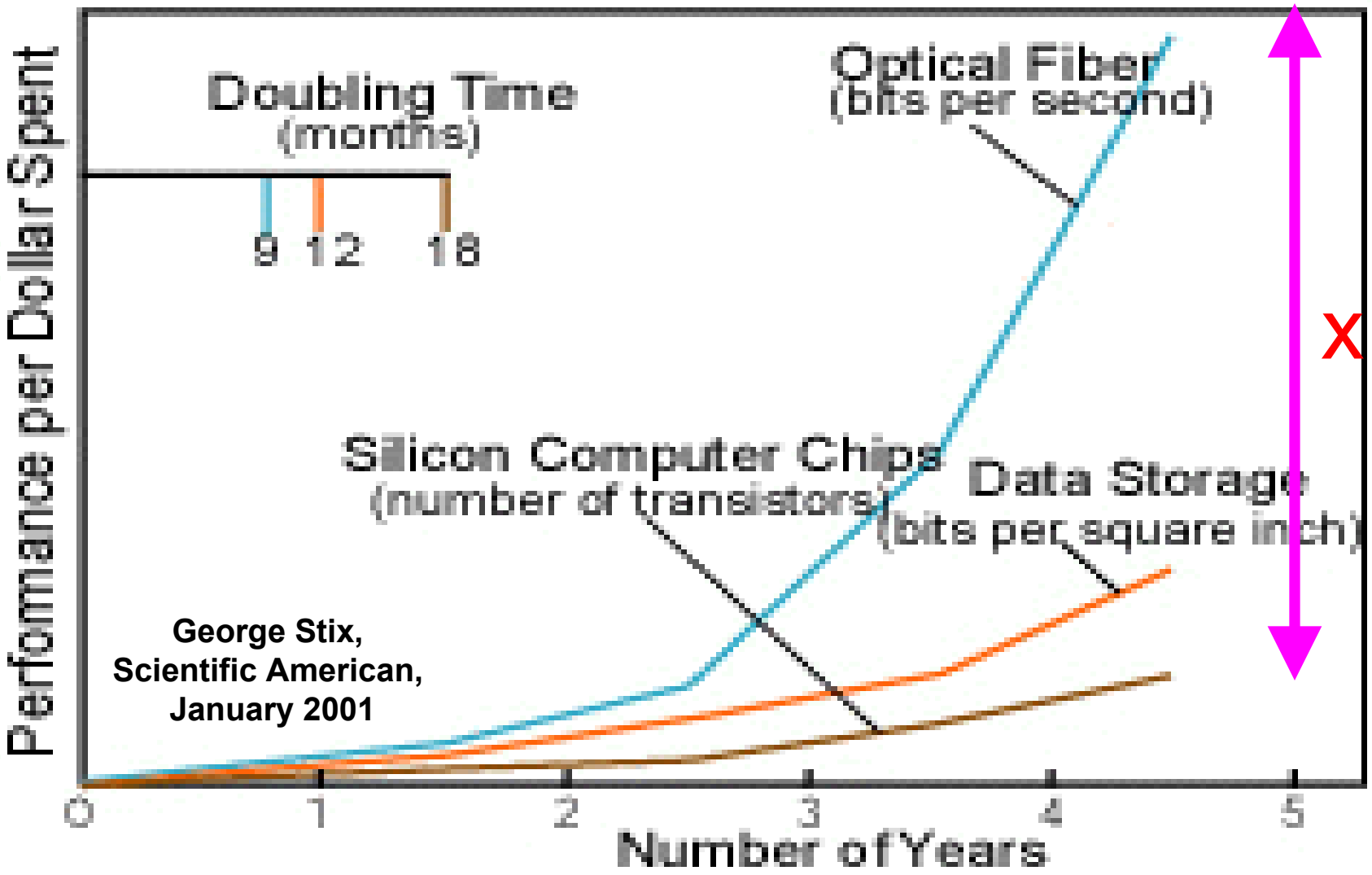
Updates

- Feedback and apps analysis prompted new conceptualization
- Visits: OptIPuter, SLAC, SSRL, SDSC, UCSD BIRN
- Looked at many eSciense compute-data-intensive applications
- Selected BIRN Mouse as a representative example
- Demos – GGF , Super Computing , Globus World
- Concept validation with the research community that:
  "an OGSI-based, Grid Service capable of dynamically controlling end-to-end lightpaths over a real wavelength-switched network"
  - Productive feedback: Ian Foster (used my slide in his Keynote), Carl Kesselman (OptIPuter question), Larry Smarr (proposed BIRN), Bill St. Arnaud, Francine Berman, Tom DeFanti, Cees de Latt

Outline

- Introduction
- The BIRN Mouse Application
- Research Concepts
- Network – Application Interface
- LambdaGrid Features
- Architecture
- Scope & Deliverables

Optical Networks Change the Current Pyramid ☺



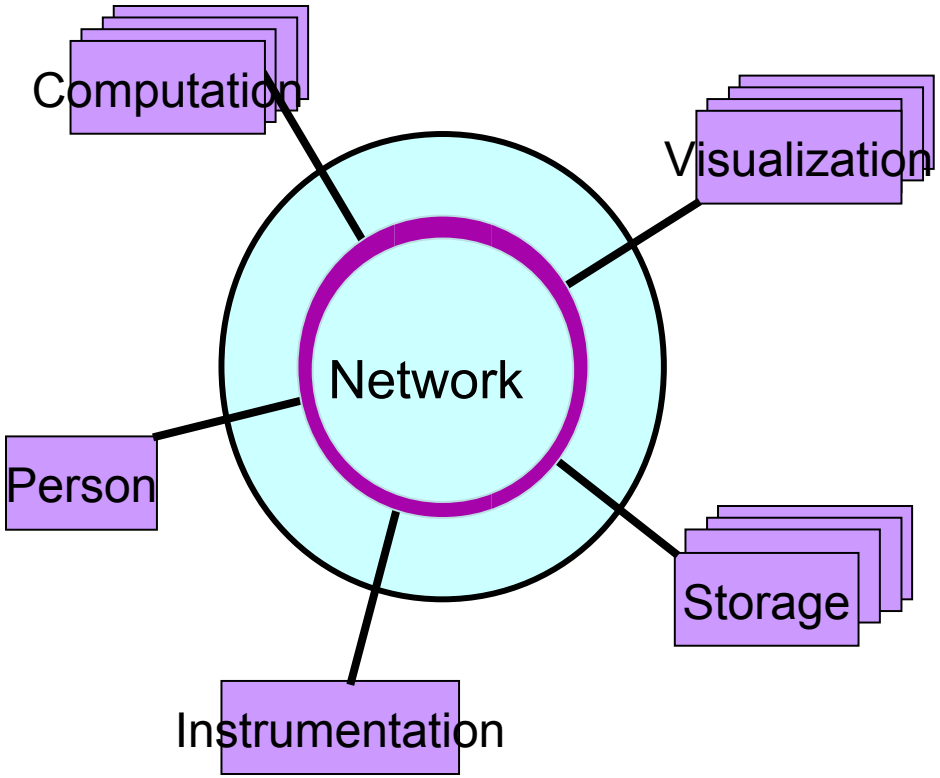
DWDM- fundamental miss-balance between computation and communication

New Networking Paradigm

"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 (2000)

- Great vision –
  - LambdaGrid is one step towards this concepts
- LambdaGrid –
  - A novel service architecture
  - Lambda as a Scheduled Service
  - Lambda as a prime resource - like storage and computation
  - Change our current systems assumptions
  - Potentially opens new horizon

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
- How can we improve the way of doing eScience?

Outline

- Introduction
- The BIRN Mouse Application
- Research Concepts
- Network – Application Interface
- LambdaGrid Features
- Architecture
- Scope & Deliverables

BIRN Mouse: Example Application

- The Mouse research application at Biomedical Informatics Research Network (BIRN)
  - Studying animal models of disease across dimensional scales to test hypothesis with human neurological disorders
  - Brain disorders studies: Schizophrenia, Dyslexia, Multiple Sclerosis, Alzheimer and Parkinson
  - Brain Morphometry testbed
  - Interdisciplinary, multi-dimensional scale morphological analysis (from genomics to full organs)
- Why BIRN Mouse?
  - illustrate the type of research questions
  - illustrate the type of e-Science collaborative applications for LambdaGrid
  - require analysis of massive amount of data,
  - LambdaGrid can enhance the way of doing the science
  - They are already recognized the current and future limitations,
    - trying to solve it from the storage, computation, visualization and collaborative tools. Working with Grid, NMI, OptIPuter, effective integration

BIRN Mouse Network Types

- Conceptualizing the networking types of eScience research methods of Mouse
  - Data Schlepping Scenario (1-1)
  - Multiple DB (1-N, N-M)
  - Remote Operation Scenario
  - Remote Visualization Scenario

- Details next 4 slides

Data Schlepping Scenario

Mouse Operation

- The "BIRN Workflow" requires moving massive amounts of data:
  - The simplest service, just copy from remote DB to local storage in mega-compute site
  - Copy multi-Terabytes (10-100TB) data
  - Store first, compute later, not real time, batch model

Mouse network limitations:

- Needs to copy ahead of time
- 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 "BIRN Workflow" architecture

Multiple DB, 1-N, N-M

Mouse Operation

- Geographically dispersed massive amount of data need to have connection to mega-computation site
- Multiple (~300) DBs
- Copy data from multiple locations to local storage
  - Or static connection to remote DB (several DBs)
- Middleware needs to know DB ahead of time

Mouse network limitations:

- Similar to data schlepping but in larger scale
- Multi-domain DB (organization, bio-scale)
- Hard to navigate (dynamic network connectivity)
- don't know the next connection needs (location, size, time)

Remote Operation Scenario

Mouse Operation

- One-of-a-kind instrumentation
  - Electron microscope, Photon microscope
    - Real-time feedback (to users), Real-time control,
    - Scheduled operations
    - Data sharing and integration (storage, computation)
  - Collaborative methods, Interdisciplinary operation
- Networks needs
  - Data- Fat pipes in one direction
  - Control - Minimal {jitter, delay, drop} on the other direction
  - Control 1-1, data 1-N
    - Dedicate the links
    - (one drive many view)

Mouse network limitations:

- Need to work at the facility, not from the home research institute
- Store the data locally, then copy, analyze later
- Hard to do real-time feedback from domain experts

Remote Visualization Scenario

Mouse Operation

- Visualization is an essential tool in Mouse analysis
  - OptIPuter uses a cluster (8-32 computers) to drive visualization, remote visualization require 15-60Gbs
  - Allows scientist to visualize and navigate the data
  - Simplifies in-depth information
  - “A picture is worth a thousand words”
  - Collaborative work across disciplines
  - Information sharing

Mouse network limitations:

- Hard to do remote visualization, need copy and analysis
  - L3 network can’t support this type of remote visualization
- Need to break the geographic constraints
- Get the data to the experts
  - Instead of experts to the data

Limitations of Solutions with Current Network Technology

- The BIRN networking is unpredictable, a major bottleneck, specifically over WAN, limit the type, way, data sizes of the biomedical research, prevents true Grid Virtual Organization (VO) research collaborations
- The network model doesn't fit the "BIRN Workflow" model, it is not an integral resource of the BIRN Cyber-Infrastructure

Outline

- Introduction
- The BIRN Mouse Application
- Research Concepts
- Network – Application Interface
- LambdaGrid Features
- Architecture
- Scope & Deliverables

Problem Statement

- Problems
  - Existing packet-switching communications model has not been sufficiently adaptable to meet the challenge of large scale data flows, especially those with variable attributes

  Q?

  Do we need an alternative switching technique?

Problem Statement

- Problems
  - BIRN Mouse often:
    - requires interaction and cooperation of resources that are distributed over many heterogeneous systems at many locations;
    - requires analyses of large amount of data (order of Terabytes-PetaBytes?);
    - requires the transport of large scale data;
    - requires sharing of data;
    - requires to support workflow cooperation model

Q?

Do we need a new network abstraction?

Problem Statement

- Problems
  - BIRN research moves ~10TB from remote DB to local mega-computation in ~10 days (unpredictable). Research would be enhanced with predictable, scheduled data movement, guaranteed in 10 hours (or perhaps 1 hour)
  - Many emerging eScience applications, especially within Grid environments require similar characteristics

Q?

Do we need a network service architecture?

BIRN Network Limitations

- **Optimized to conserve bandwidth and waste storage**
  - Geographically dispersed data
  - Data can scale up 10-100 times easily
- L3 networks can't handle multi-terabytes efficiently and cost effectively
- Network does not fit the "BIRN Workflow" architecture
  - Collaboration and information sharing is hard
- Mega-computation, not possible to move the computation to the data (instead data to the computation site)
- Not interactive research, must first copy then analyze
  - Analysis locally, but with strong limitations geographically
  - Don't know a head of time where the data is
    - Can't navigate the data interactively or in real time
    - **Can't "Webify"** the information of large volumes
- No cooperation/interaction between the storage and network middleware(s)

Proposed Solution

- Switching technology: Lambda switching for data-intensive transfer
- New abstraction: Network Resource encapsulated as a Grid service
- New middleware service architecture: LambdaGrid service architecture

Proposed Solution

- LambdaGrid Service architecture interacts with BIRN Cyber-infrastructure, and overcome BIRN 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

Goals of Investigation

- Explore a new type of infrastructure which manages codependent storage and network resources
- Explore dynamic wavelength switching, based on new optical technologies
- Explore protocols for managing dynamically provisioned wavelengths
- Encapsulate “optical network resources” into the Grid services framework to support dynamically provisioned, data-intensive transport services
- Explicit representation of future scheduling in the data and network resource management model
- Support a new set of application services that can intelligently schedule/re-schedule/co-schedule resources
- Provide for large scale data transfer among multiple geographically distributed data locations, interconnected by paths with different attributes.
- To provide inexpensive access to advanced computation capabilities and extremely large data sets

New Concepts

- The “Network” is NO longer a Network
  - but a **large scale Distributed System**
- Many-to-Many vs. Few-to-Few
- Apps optimized to waste bandwidth
- Network as a Grid service
- Network as a scheduled service
- Cloud bypass
- New transport concept
- New cooperative control plane

Research Questions (Dist Comp)

- Statistically multiplexing works for small streams, but not for mega flows. New concept for **multiplexing by scheduling**
  - What will we gain from the new scheduling model?
  - evaluate design tradeoffs
- Current apps designed and optimized to conserve bandwidth
  - If we provide the LambdaGrid service architecture, how will this change the design and the architecture of eScience data-intensive workflow?
  - What are the tradeoffs for applications to waste bandwidth?
- The batch & queue model does not fit networks. manual scheduling (email)–not efficient
  - What is the right model for adding the network as a Grid service?
  - network as integral Grid resource
  - Overcome Grid VO networking limitations

Research Questions (Dist Comp)

Assuming Mouse VO with 500TB over 4 remote DB, connected to one mega-compute site, where apps don't know ahead of time the data subset that is required.

- Network interaction with other services
  - The "right" way to hide network complexity from the application, and (contradict) interact with the network internals? The model for black vs. white box (or hybrid)
  - flexibility by negotiation schedule/reschedule in an optimal way
    - opens new set of interesting questions
    - New set of network optimization

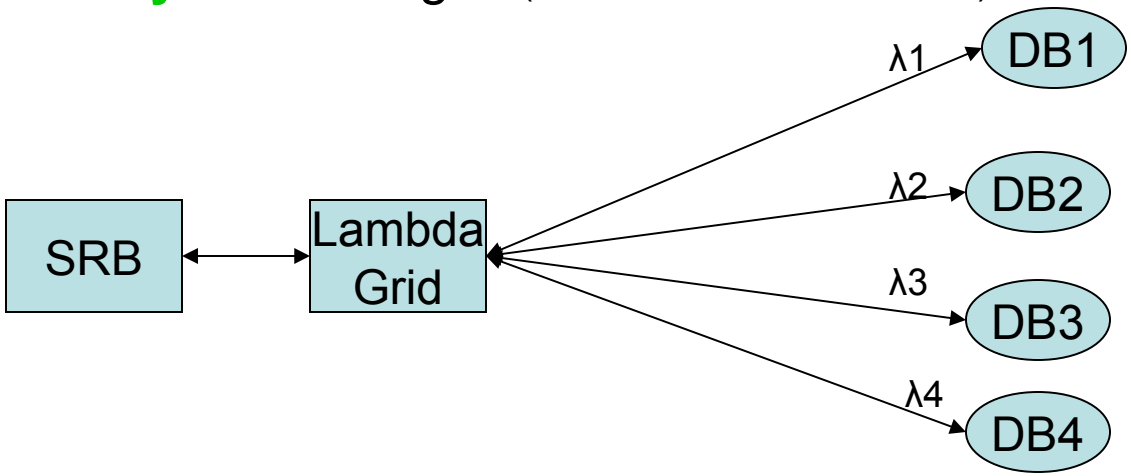
- No control plane to the current Internet
  - centralized control plane for subset network
  - Cooperative control planes (eSciece & optical networking)
  - Evaluation of the gain (measurable) from such interaction

Research Questions (Net)

- IP service best for **many-to-many** (~100M) small files (~100k-10MB) . LambdaGrid services best for **few-to-few** (~100) huge storage (~TeraBytes-PetaBytes)
- Cloud bypass
  - 100TB over the Internet, break down the system
  - Offload Mega-flows from the IP cloud to the optical cloud. Alternate rout for Dynamic Optical network
    - What are the design characteristics for cloud bypass?
- TCP does not fit mega-flows (well documented)
  - new proposals, XCP, SABUL, FAST, Tsunami…, fits the current Internet model. However, does not take advantage of dynamic optical network
    - {distinct characteristics - fairness, collision, drop, loss, QoS…}
  - What are the new design principle for cooperative protocols {L1+L4} ?
  - What to expose and in what interface to L3/L4?
- Economics – expensive WAN static optical links, and L3 WAN ports
  - Given dynamic optical links, and the proposed Service Architecture, what are the design tradeoffs resulted in affordable solution?

Research Questions (Storage)

- BIRN's Storage Resource Broker (SRB)
  - Hide the physical location of DB
    - Concepts of remote FS (RFS)
    - Based on the analysis search in other DB
- SRB Lambda
  - Tight interaction between SRB and LambdaGrid
  - Optical control as a File System interface
  - **The "Right Lambda" to the "Right DB" at the "Right time"**
  - **"Webfy"** MRI images (mouse click ~1GB, ~2 sec)



Outline

- Introduction
- The BIRN Mouse Application
- Research Concepts
- Network – Application Interface
- LambdaGrid Features
- Architecture
- Scope & Deliverables

Information Crossing NRS-Apps Interface

The information is handled by the apps middleware as a running environment for the application, and communicating with LambdaGrid as a whole

- SA, DA, Protocol, SP, DP
- Scheduling windows (time constrains)
  - {Start-after, end-before, duration}
- Connectivity: {1-1, 1-N, N-1, N-N, N-M}
- Data Size (calc bandwidth & duration)
- Cost (or cost function)
- Type of service
  - {Data Schlepping, Remote Operation, Remote Visualization }
    - Middleware calc {bandwidth, delay, Jitter, QoS… }
    - Translate into Lambda service
- WS-Agreement
  - ID {user, application, data}, {credential , value, priorities, flexibility, availability}
  - Handle for returning output (scheduled plane, etc) renegotiation
  - Handle back for feedback, notification

Interface Details

SA, DA, Protocol, SP, DP

Stream binding identifier

{Src addr, Dest addr, Protocol, Src port, Dest Port)

Scheduling Window:

Window of time that the applications need the network.

{start time, end time, and duration}. Start and end can be " * " (don't care)

The network will try to schedule the service under these constrains.

The network middleware will reply with the allocated time and this can be shifted and renegotiated. The multiplexing will be done on data scheduling.

The middleware will resolve conflicts and allocate the available time slot

Example: allocate 2 hours, start after 5pm, end before 11pm

Bandwidth:

What is the bandwidth requirements

Example: allocate 10Gbs half duplex, 100Mbs the opposite direction

Interface details (cont)

Data Size:

Data service require copy the data, regardless the bandwidth or the duration. The data needs to be copied, the system will take care on how. When the data is copied, send a notification. The system can adapt the network dynamically based on new constrains, future expected availability and optimizations

Example: copy 10TB from source to destination

Allocation start with 1Gbs, changed to 10Gbs, and finished with 4Gbs

Connectivity:

Multiple connection request. What is the type of connectivity {1-1, 1-N, N-1, N-N, N-M},

Examples: Connect {1-1} to a remote DB,

or Connect to all of these DBs: 1-N

Interface details (cont)

Type of Service:

What is the type of operation?

- {Data Schlepping, Remote Operation, Remote Visualization }
  - Middleware calc {bandwidth, delay, Jitter, QoS… }
  - Translate into Lambda service

Example: {Remote operation}

dedicated 10Gbs Lambda, one direction

control back – small bandwidth, min delay, min jitter, top QoS

- High cost for the requested time window

Example: Visualization

5Gbs, min jitter, ok high delay, ok high loss

Interface details (cont)

Cost:

What is the cost or the cost function of this operation?

Example: premium 30% between 8:00-5:00

50% premium for less than 2 hours advanced reservation

Min cost 24 hours in advance,

WS-Agreement:

The agreement characteristics between the request and the data service

Example:

need Ψ SLA, critical apps, important data, tight coupled rescheduling to the Φ computing, ω storage and availability of φ data, hard until midnight, flexible afterwards, ready for negotiation, willing to pay $f(\zeta)$,

Outline

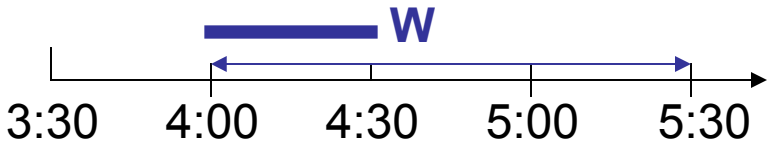
- Introduction
- The BIRN Mouse Application
- Research Concepts
- Network – Application Interface
- LambdaGrid Features
- Architecture
- Scope & Deliverables

Features

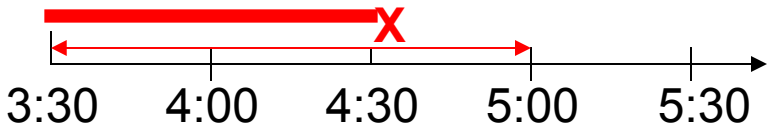
- Dynamic Optical Underlying network
  - Dynamic lambda switching
  - All optical switches
- Endpoint to Endpoint Lambda paths
  - Segments
  - Data paths
- Network Resource Grid Service
  - Encapsulation of network resources into a Grid service
- Scheduled network service
  - Schedule lambda path resources to satisfy multiple complex requests
- Application Middleware Abstraction
  - Hiding networks details from application

Example: Lightpath Scheduling

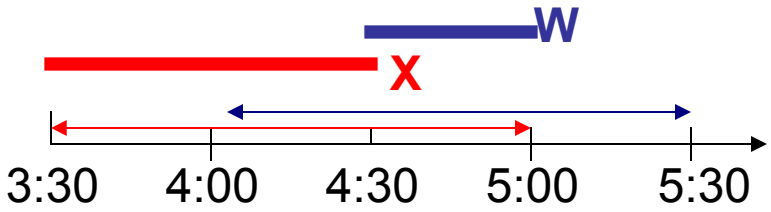
- Request for 1/2 hour between 4:00 and 5:30 on Segment D granted to User W at 4:00



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



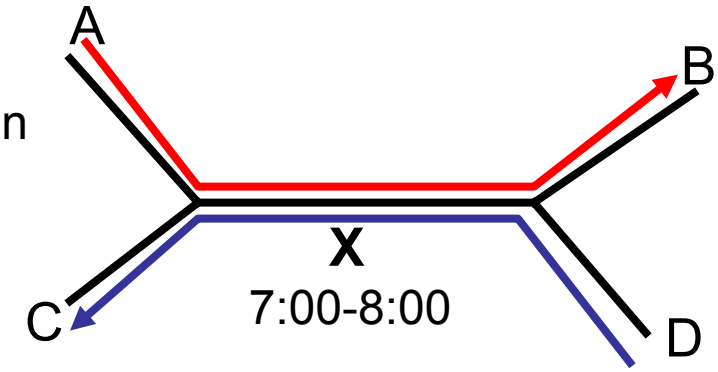
- Reschedule user W to 4:30; user X to 3:30. Everyone is happy.



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

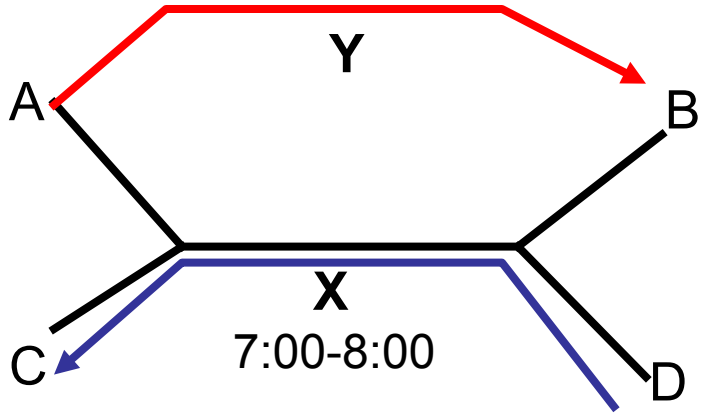
Scheduling Example - Reroute

- 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 thru 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

Outline

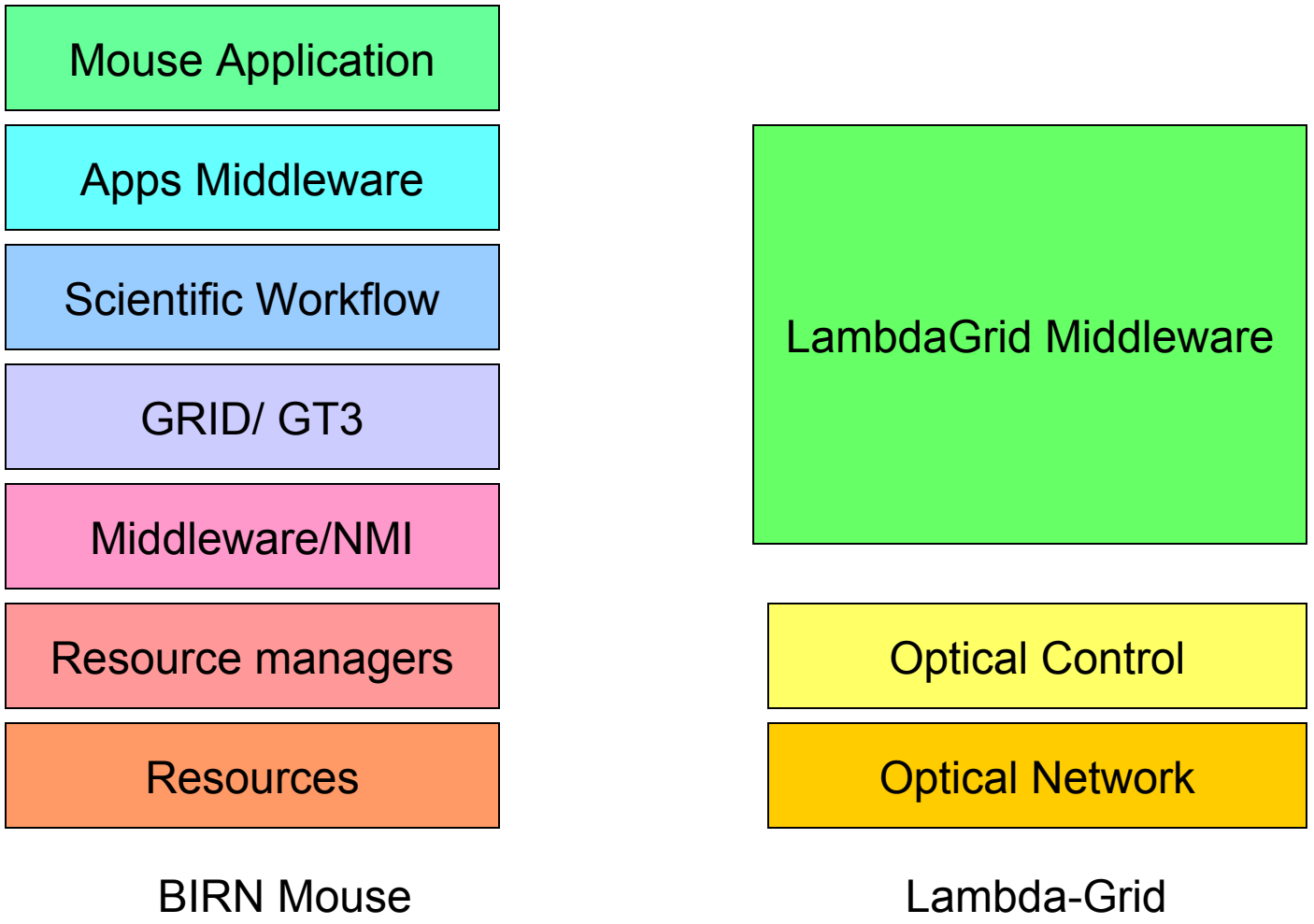
- Introduction
- The BIRN Mouse Application
- Research Concepts
- Network – Application Interface
- LambdaGrid Features
- Architecture
- Scope & Deliverables

Architecture

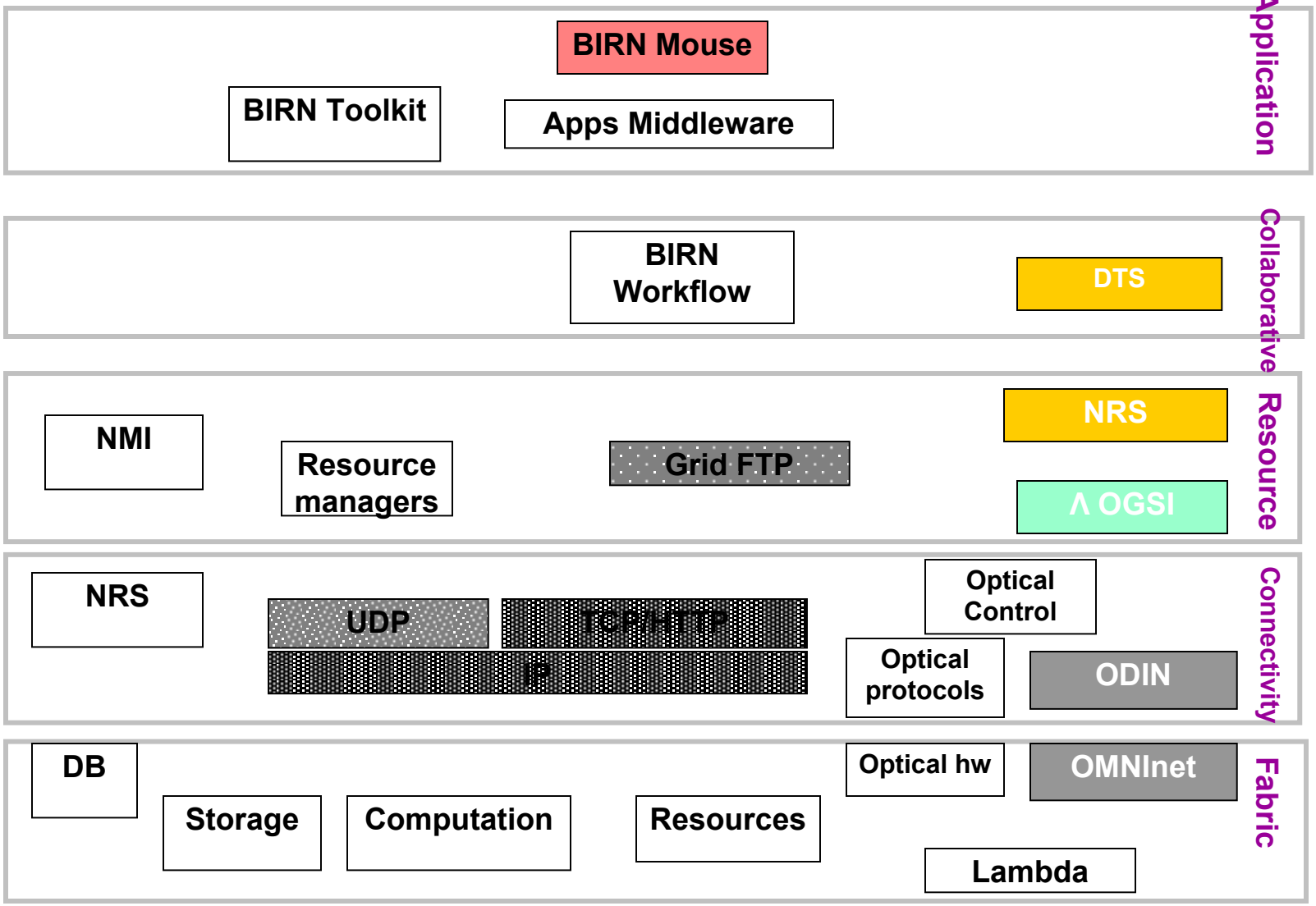
LambdaGrid Service architecture that interacts with BIRN Cyber-infrastructure, NMI, and GT3 and include:

- Encapsulation of “optical network resources” into the Grid services framework part of OGSA and with an OGSI, to support dynamically provisioned data-intensive transport services
- Data Transfer Service (DTS) and Network Resource Service (NRS) that interacts with other middleware and optical control plane
- Cut-through mechanism on edge device that allows mega flows to bypass the L3 cloud into Dynamic Optical Network

30k feet view: BIRN Cyber-infrastructure



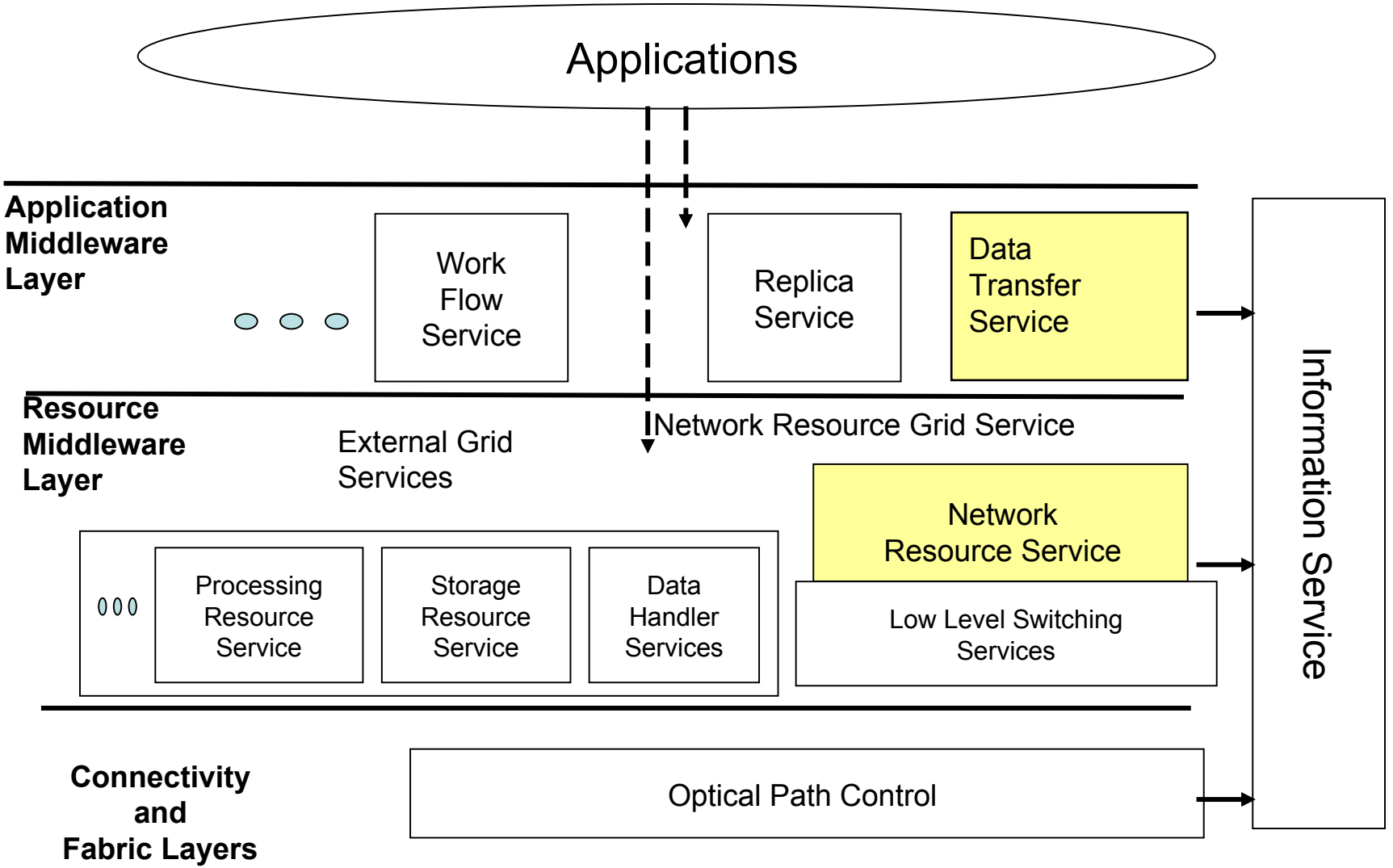
Layered Architecture



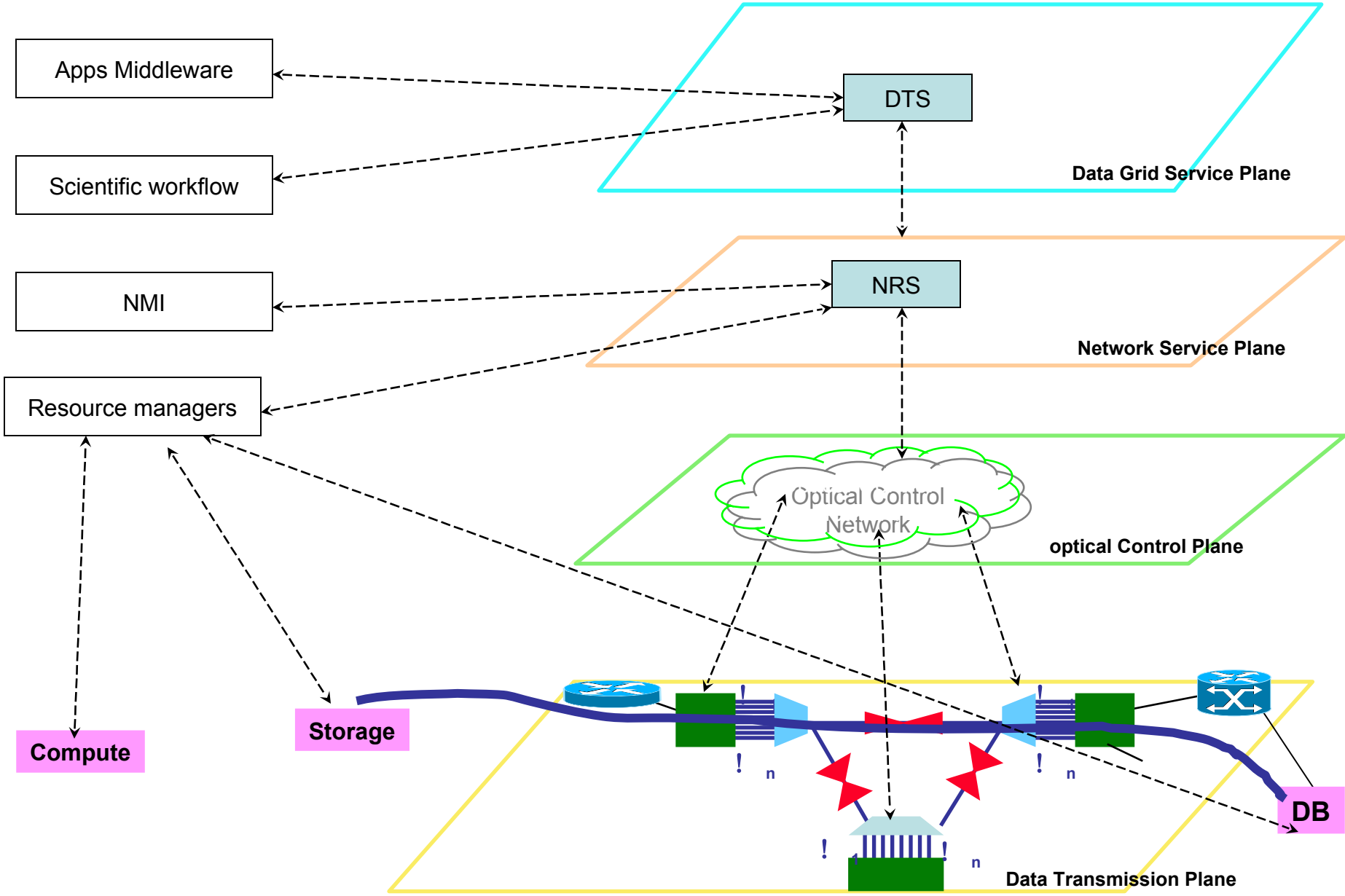
Grid Layered Architecture

My view of the layers

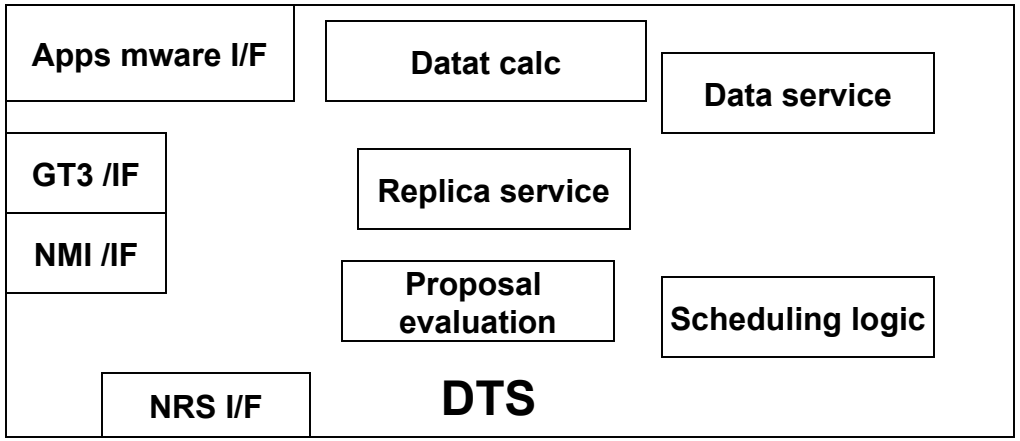
Generalized Architecture

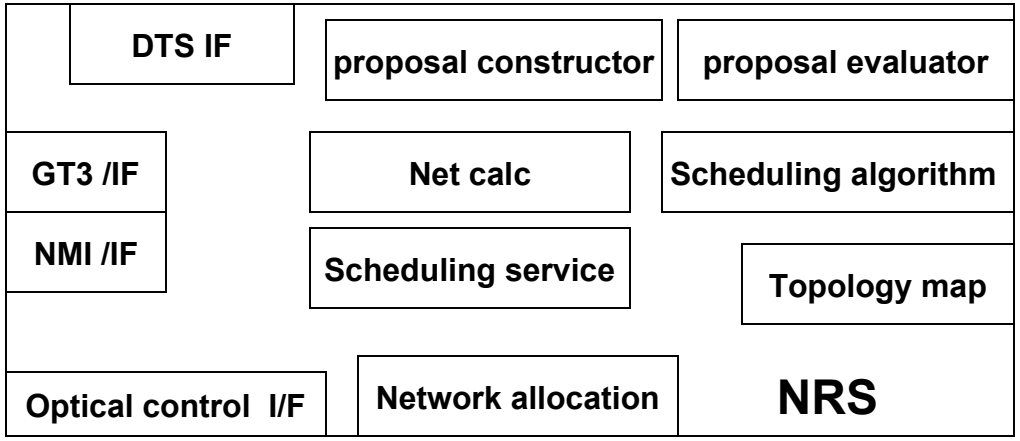


Control Interactions

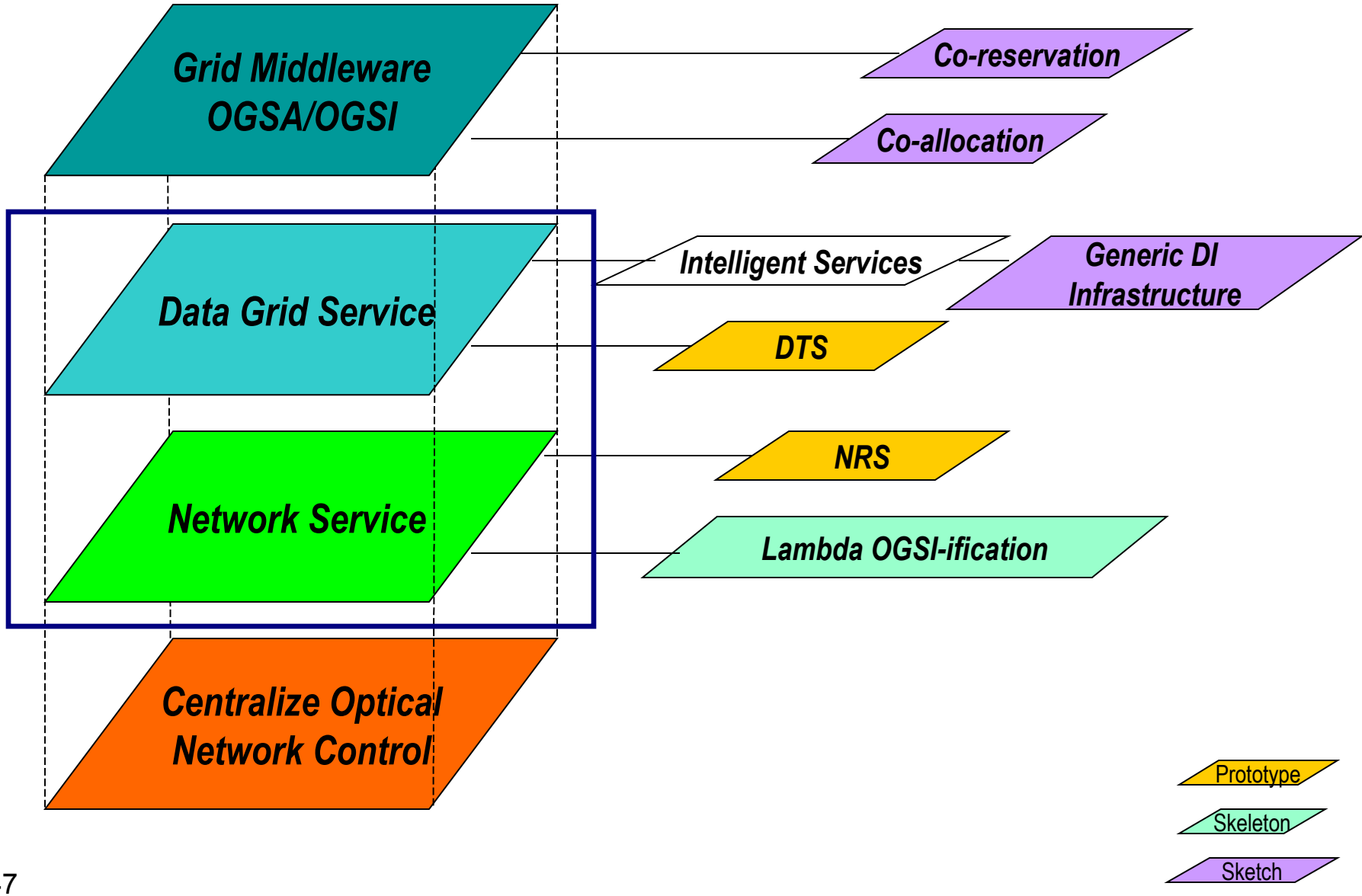


DTS - NRS

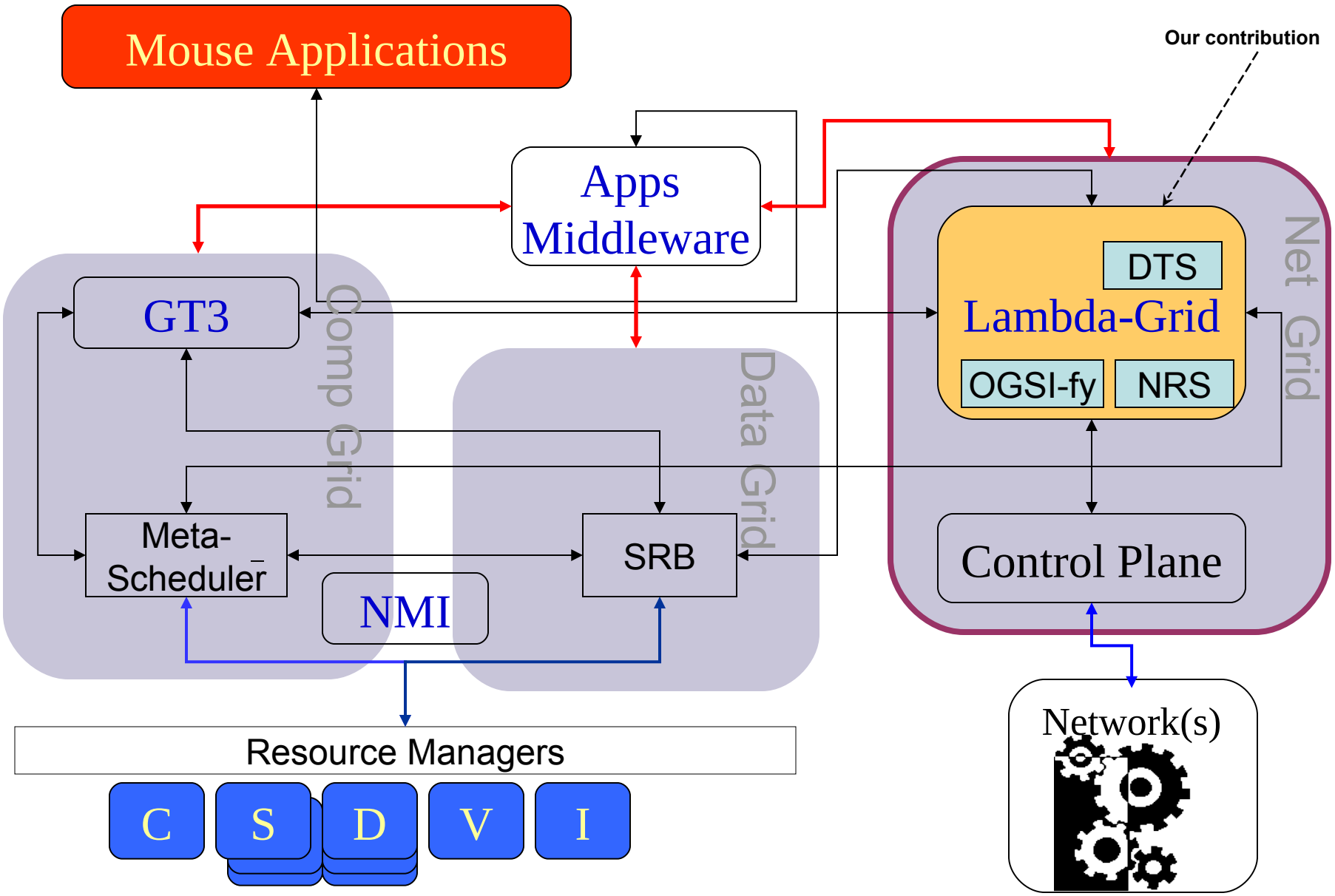

Apps mware I/F
Datat calc
Data service
GT3 /IF
Replica service
NMI /IF
Proposal evaluation
Scheduling logic
NRS I/F
DTS


DTS IF
proposal constructor
proposal evaluator
GT3 /IF
Net calc
Scheduling algorithm
NMI /IF
Scheduling service
Topology map
Optical control I/F
Network allocation
NRS

Layered Architecture



Overall System



Outline

- Introduction
- The BIRN Mouse Application
- Research Concepts
- Network – Application Interface
- LambdaGrid Features
- Architecture
- Scope & Deliverables

Scope of the Research

- Investigate the feasibility of encapsulating lightpath as an OGSI Grid Service
  - Identify the characteristics to present the lightpath as a prime resource like computation and storage
  - Define and design the “right” framework & architecture for encapsulating lightpath
  - Demonstrate network Grid service concept
- Identify the interactions between storage, computation and networking middleware
  - Understand the concept in interactions with NMI, GT3, SRB
  - Collaborate with the BIRN Mouse Cyber-infrastructure research

Out of scope:

- Scheduler, scheduling algorithms, network optimization
- Photonics, hardware, optical networking, optical control, protocols, GT3, NMI, SRB, BIRN components
- Develop it for any commercial optical networking or commercial hardware

Deliverables

- Initial investigation
- Build a testbed
- "Proof-of-concept {++}" of LambdaGrid service architecture
- Demonstrate one aspect of BIRN Mouse application with the proposed concepts
- Prototype Lambda as an OGSI Grid Service
- Develop DTS and NRS, a Grid scheduling service
- Propose a service architecture and generalize the concept with other eScience projects

Pro Forma Timeline

- Phase 1
  - Continue to develop the *Lambda Data Grid* prototype
  - Build BIRN basic testbed
    - incorporate Lambda as a service,
    - measure/analyze the performance and scaling behavior
- Phase 2
  - Develop an OGSI wrapper to Lambda service
    - integrate as part of OGSA, interact with OptIPuter
  - Interface to NMI and SRB
  - Analyze the overall performance, incorporate the enhancements
- Phase 3
  - Extract generalized framework for intelligent services and applications
  - Incorporate experience from the Grid research community
  - Measure, optimize, measure, optimize, …

Generalization and Future Direction for Research

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

Thank You_

The Future is Bright

_

- Imagine the next 5 years
- There are more questions than answers