Traffic Engineering For Dynamically Provisioned Federated Networks

Evolving from Network Services to the "Network as a Resource"

Advanced Scientific Computing Advisory Committee (ASCAC) Meeting
August 14-15, 2012

Tom Lehman
University Southern California
Information Sciences Institute (USC/ISI)

Chin Guok
Energy Sciences Network (ESnet)

U.S. DEPARTMENT OF ENERGY
Office of Science

Presentation Outline

- **Traffic Engineering For Dynamically Provisioned Federated Networks Today**
  - How we got here as a result of past ASCR Research projects
- **Traffic Engineering For Dynamically Provisioned Federated Networks Tomorrow**
  - Building on past projects
  - Evolving from Network Services to Network as a Resource
- **Beyond Dynamic Network Provisioning**
  - Intelligent Networking
  - Software Defined Networking (OpenFlow)

Past Research Projects Impact on Today's Production Networking

- **The Internet as designed is a best-effort infrastructure but High-end science applications require**
  - Predictable and guaranteed performance
  - 100x end-to-end performance
  - Multiple-domain coordination
- **Some of the past research…**
  - 2003: ASCR funds Ultra-Science Network to prototype dynamic provisioning of circuits
  - 2003: NSF funds DRAGON to research multi-domain dynamic provisioning of circuits
  - ***2004: ASCR funds ESnet to develop on-demand dedicated bandwidth circuit reservation system (OSCARS)***
  - 2006: ASCR funds Hybrid MLN project to enhance OSCARS with multi-domain capabilities

Past Research Projects Impact on Today's Production Networking

The impact…

- **OSCARS**
  - Deployed as a production service in ESnet since mid 2007
  - About 50% of ESnet's total traffic is now carried via OSCARS circuits
  - Adopted by SciNet since SC09 (1999) to manage network bandwidth resources for demos and bandwidth challenges
  - Integral in ESnet winning the Excellence.gov "Excellence in Leveraging Technology" award in 2009
  - Received the Internet2 IDEA award in 2011
  - Adopted by LHC to support Tier 0 – Tier 1 and Tier 1 – Tier 2 transfers
  - Currently deployed in over 20 networks world wide including wide-area backbones, regional networks, exchange points, local-area networks, and testbeds
  - Adopted by NSF DYNES project which will result in over 40 more OSCARS deployments

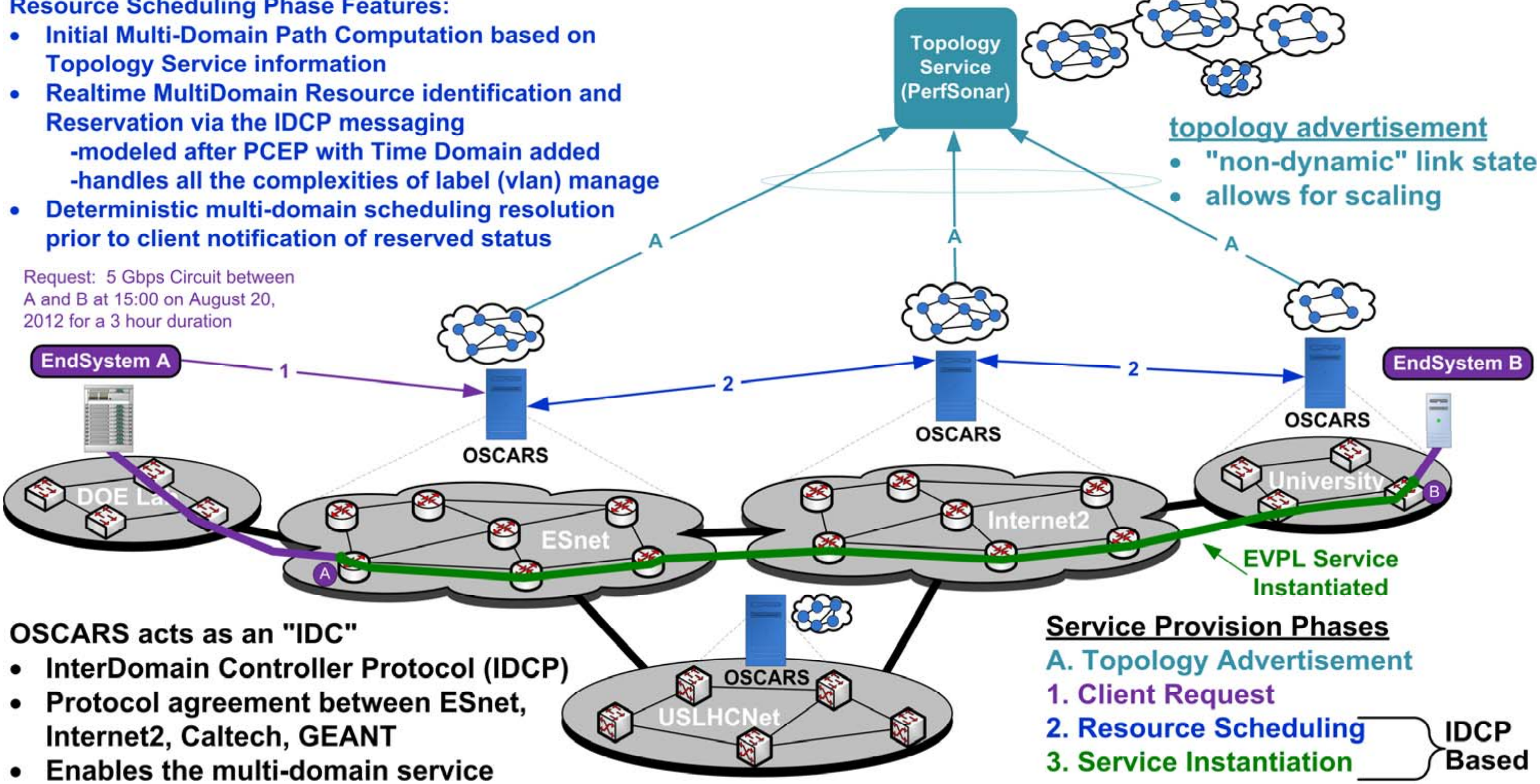
OSCARS today and its Impact on ESnet and R&E Networks

The "service" is Ethernet Virtual Private Line (EVPL) with dedicated bandwidth

Different networks use different technologies to instantiate (MPLS, SONET, Native Ethernet, WDM)

Resource Scheduling Phase Features:

- Initial Multi-Domain Path Computation based on Topology Service information
- Realtime MultiDomain Resource identification and Reservation via the IDCP messaging
  - -modeled after PCEP with Time Domain added
  - -handles all the complexities of label (vlan) manage
- Deterministic multi-domain scheduling resolution prior to client notification of reserved status



topology advertisement

- "non-dynamic" link state
- allows for scaling

Request: 5 Gbps Circuit between A and B at 15:00 on August 20, 2012 for a 3 hour duration

OSCARS acts as an "IDC"

- InterDomain Controller Protocol (IDCP)
- Protocol agreement between ESnet, Internet2, Caltech, GEANT
- Enables the multi-domain service

Service Provision Phases

A. Topology Advertisement
1. Client Request
2. Resource Scheduling
3. Service Instantiation

IDCP Based

Generalizing OSCARS for Heterogeneous and Federated Networking

Project Title

Advanced Resource Computation for Hybrid Service and TOpology NEtworks (ARCHSTONE)

Objectives

- Multi-Layer Network Control
- Multi-Domains control and provisioning
- Intelligent Network Services for Science Applications

PIs

- Tom Lehman - USC/ISI (lead)
- Chin Guok – LBL/ESnet
- Nasir Ghani - UNM

ARCHSTONE
Vision Statement and Motivations

- **Multi-Layer Networking**
  - Networks are really Multi-Layer. Today from a dynamic control and service provision perspective the layers are treated independent and separately

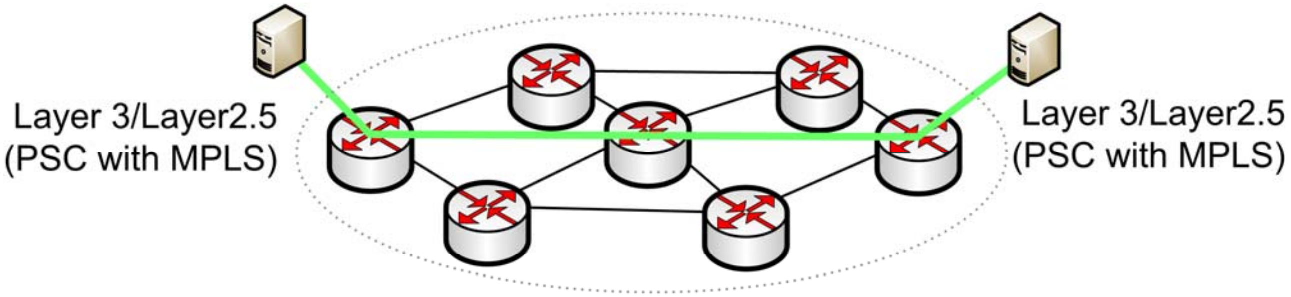
For service provision we should treat all the network layers in a holistic and integrated fashion

- **Network Services vs the "Network as a Resource”**
  - The Next Generation of Advanced Networked Applications will require more "flexible control", "scheduling", and "deterministic performance" across all the resources in their ecosystem
  - This will require integration and co-scheduling across Network, Middleware, and Application level resources (compute, storage, domain specific instruments)

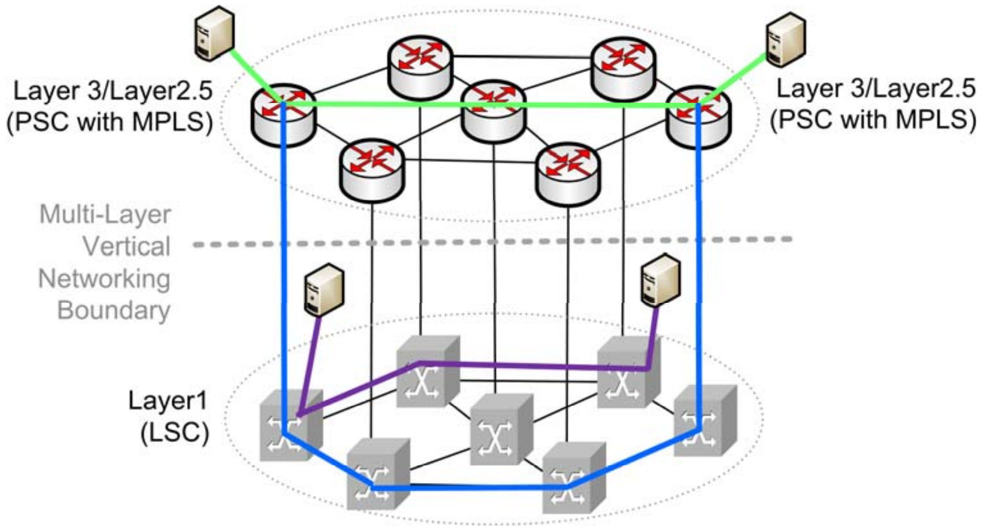
The Network needs to be available to application workflows as a first class resource in this ecosystem

Multi-Layer Networks

- Today our dynamic provisioning systems only see single layer



- But the networks are really multi-layer

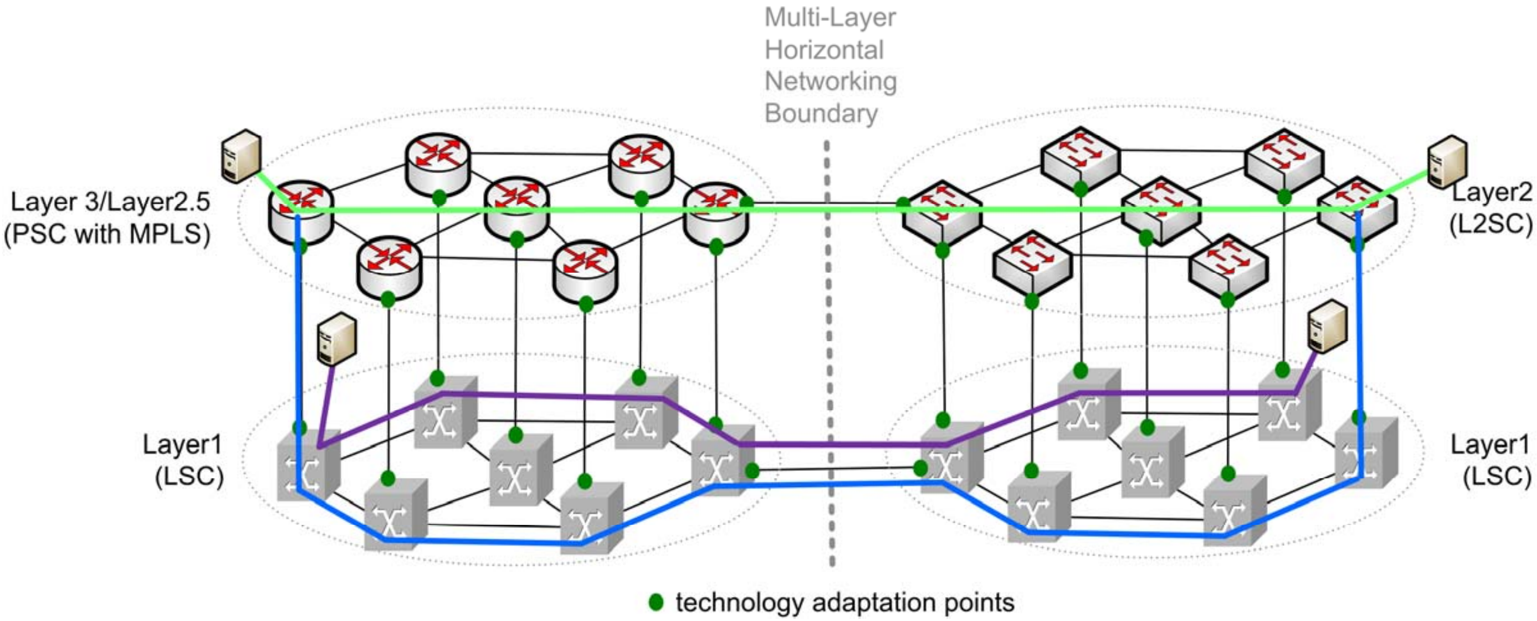


would like to:

- Provision services at lower layer to create a topology element at the higher layer (link between routers)
- Offer services directly at the lower layer

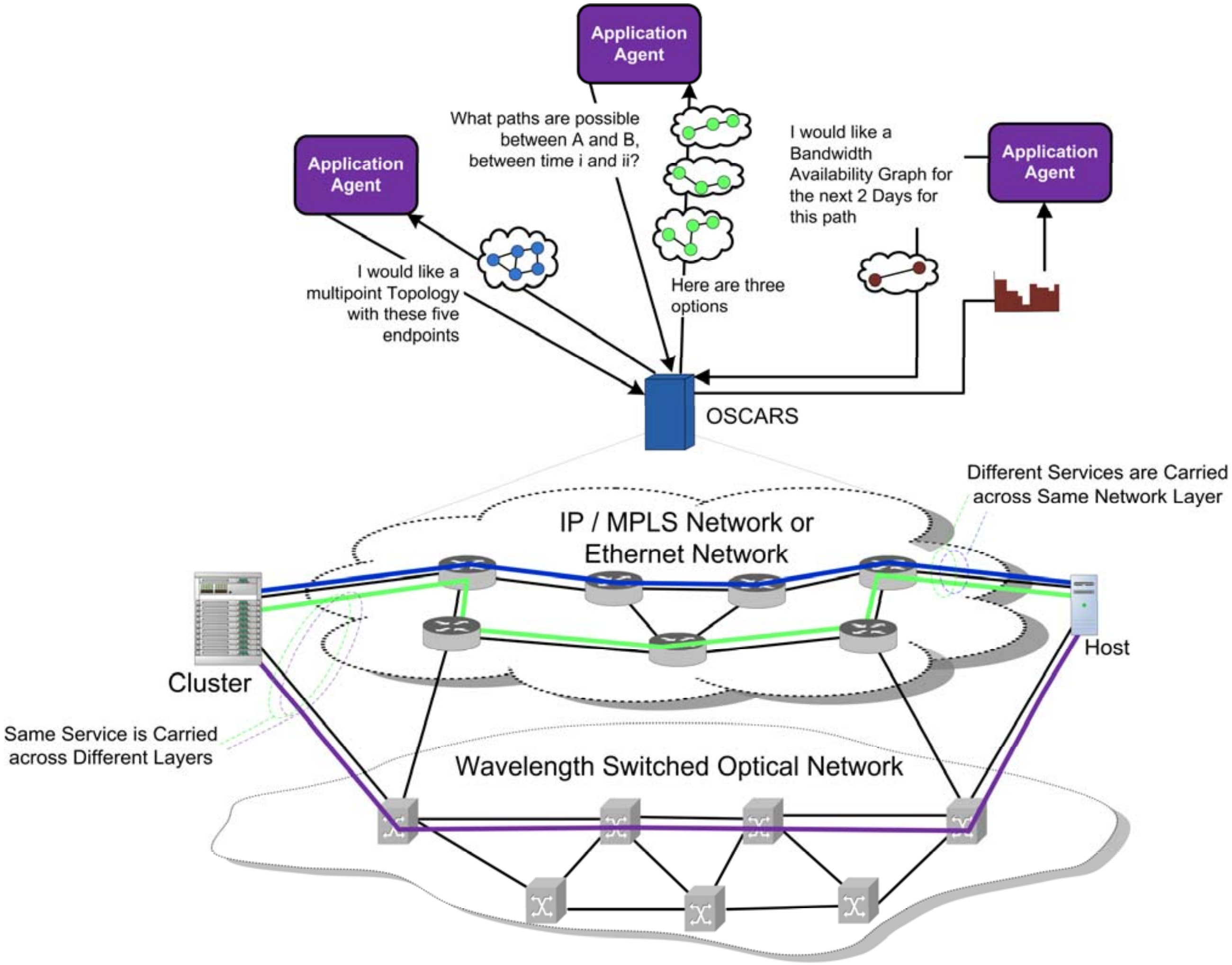
Multi-Layer Networks

- **would also like to do this on a multi-domain basis**



The Network as a Resource for Application Workflows

- The network needs to be able to respond to "What is Possible?" and "What do you recommend" questions
  - today the application must say "provision this specific path at this specific time"
- These are referred to as "Intelligent Network Services"

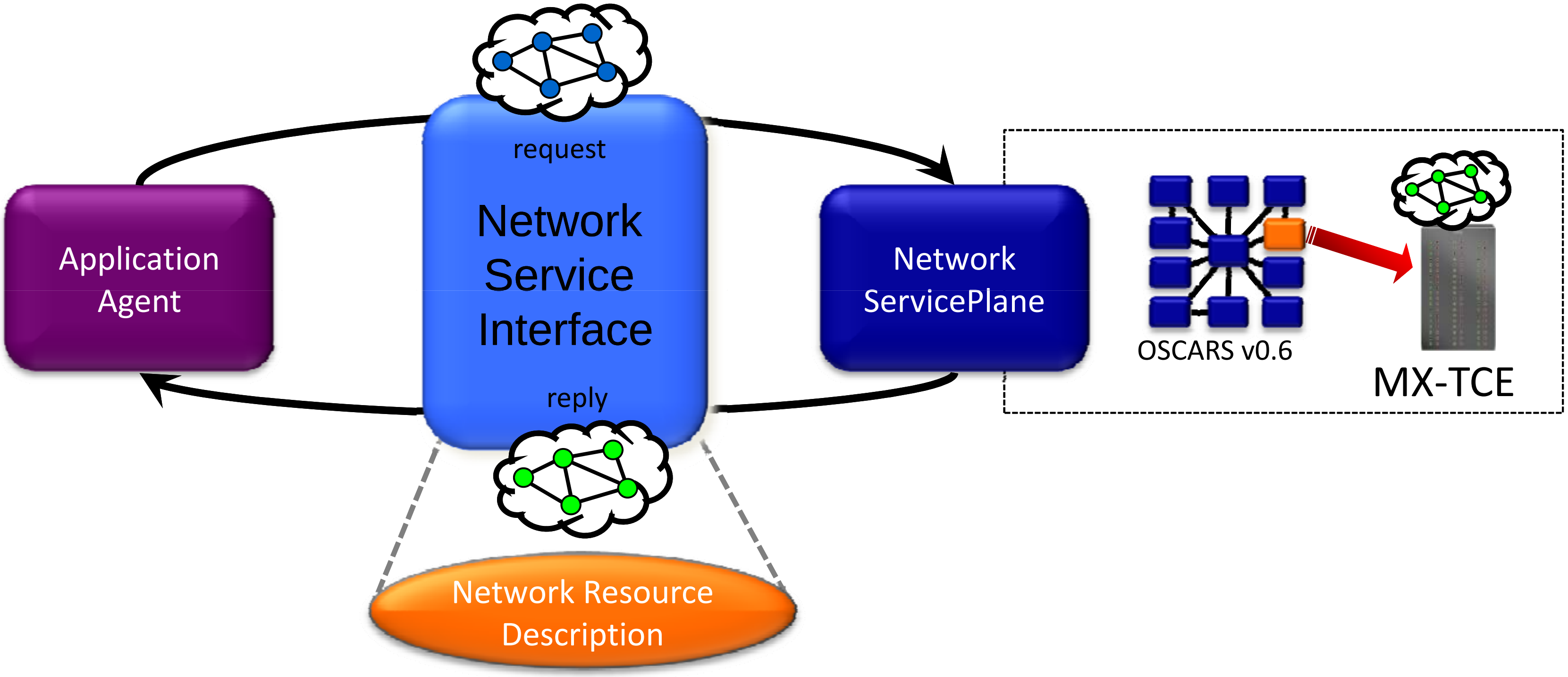

Application Agent
Application Agent
Application Agent
What paths are possible between A and B, between time i and ii?
I would like a Bandwidth Availability Graph for the next 2 Days for this path
I would like a multipoint Topology with these five endpoints
Here are three options
OSCARS
Different Services are Carried across Same Network Layer
IP / MPLS Network or Ethernet Network
Cluster
Host
Same Service is Carried across Different Layers
Wavelength Switched Optical Network

What are the Main Challenges?

- **Multi-Layer Network Control**
  - Routing domains are different between the layers, i.e., topology and state information is not shared across layer boundaries
  - Vendor unique functions and capabilities must be understood
  - The result of multi-layer control is we have Dynamic Topologies instead of Dynamic Services. This can create instability in the network if not managed properly.
- **Intelligent Network Services**
  - Resource computation in response to open-ended questions can be complex and processing intensive
  - Since we are limiting ourselves to "scheduled" services, this will help
  - For single domain, we can have a single state aware entity. But for multi-domain we will likely need a two-phase commit type of process.
- **A common capability in the form of Multi-Constraint Resource Computation is needed to enable both of these capabilities**
- **Multi-domain topology sharing and multi-domain messaging also presents challenges, but not to the degree of computation**

ARCHSTONE Architecture Components

- **Advanced Network Service Plane and Network Service Interface**
  - "Request Topology" and "Service Topology" concepts
  - Common Network Resource Description schema
  - Formalization of the Application to Network interactions
- **Multi-Dimensional Topology Computation Element (MX-TCE)**
  - High Performance computation with flexible application of constraints
  - Multi-Constraint Topology Computation is the main challenge to enable OSCARS to become Multi-Layer Network Aware and to provide Intelligent Network Services
- **Use OSCARS v0.6 as base infrastructure and development environment**



Multi-Dimensional Topology Computation

- **Topology computation is an advanced path computation process which is an order of magnitude more complex in the constraint and network graph dimensions**
- **Traffic Engineering Constraints are categorized for subsequent treatment in the multi-stage computation process:**
  - Prunable constraints: including bandwidth, switching type, encoding type, service times and policy-induced exclusion etc.
  - Additive constraints: including path length, latency and linear optical impairments (e.g. dispersion) etc.
  - Non-additive constraints: including optical wavelength continuity, Ethernet VLAN continuity and non-linear optical impairments (e.g. cross-talk) etc.
  - Adaptation constraints: conditions for traffic to traverse across layers (i.e. cross-layer adaptation), or to modify some of the above constraints into relaxed or more stringent forms (e.g. wavelength or VLAN conversion).

Multi-Dimensional Topology Computation

- **The following computation techniques were evaluated:**
  - Constrained Shortest Path First (SPF)
  - Constrained Breadth First Search (BSF)
  - Graph Transformation
    - Label-Layer Graph Transformation Technique
    - Channel Graph Transformation Technique
  - Heuristic Search Solution
- **Evaluated multiple combinations of these approaches**
  - C-BSF constrained BSF search solution
  - K-Shortest Path (KSP) heuristic search solution
  - Graph transformation based KSP heuristic search solution
- **Initial Conclusion: We settled on an multi-stage KSP (heuristic) with ordering criteria for initial implementation**
- **Future services may require other techniques**

Cross-Layer Constrained Search Solution

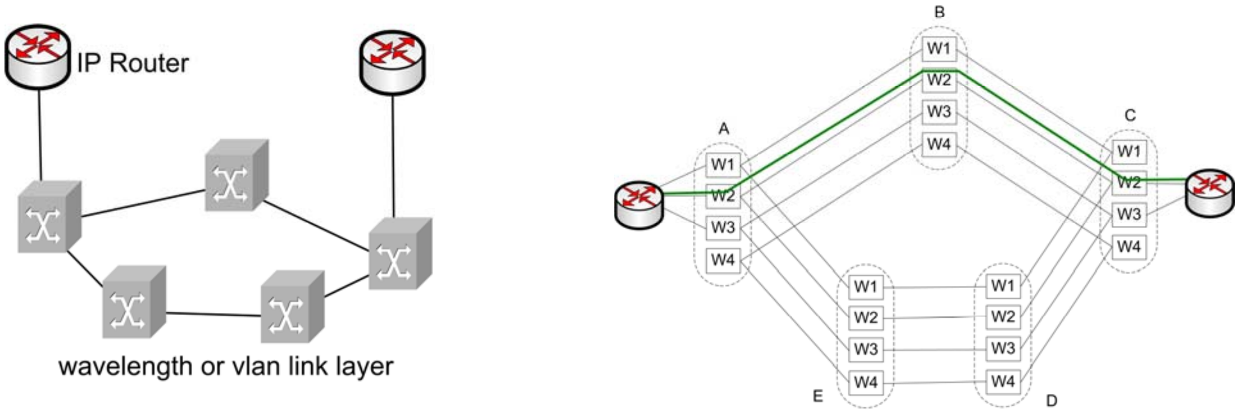
- **Applying full TE constraints when search procedure proceeds**
  - Search procedure can be based any modified SPF
  - Largely expanded search space compared to simple SPF
  - May or may not be exhaustive as some search branches can be trimmed
- **A Constrained Breadth First Search (C-BSF) implementation**
  - Handling TE constraints
    - Prunable constraints and additive constraints such as bandwidth and path length.
    - Cross-layer adaptation constraints:
    - Wavelength continuity constraints:
  - Extra logic
    - Loop avoidance logic
    - Parallel link handling logic:
  - Additions to complexity
    - Unlike a basic BFS that only visit each node and link once, C-BFS has to reenter some nodes and links multiple times.
    - Each search hop needs a constant number of stack operations for restoring and preserving the search scene at the head node.

Graph Transformation Solution

- **Unlike Constrained Search, this solution does not conduct path computation on the network graph of the original topology.**
- **Instead, it first transforms the network graph into a new form that can take some constraints into the graph construction.**
  - Part of TE constraints are embedded in graph
  - Search procedure only applies the remaining TE constraints
  - When a path is found with any simplified search procedure, the graph-transformed constraints have already been included in the resulting path.
  - While some constraints are removed from the search procedure, graph transformation/construction introduces other computation needs.
  - Well constructed graph can reduce overall complexity.

Graph Transformation Solution – Label-Layer Graph Technique

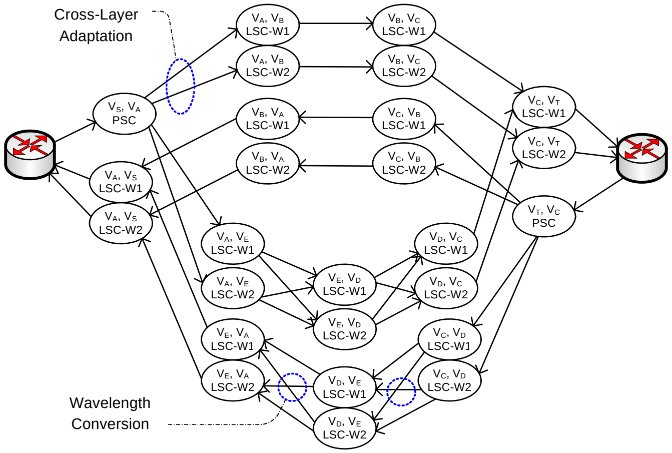
- Handling general data channel continuity constraints.
- A data channel could be an Ethernet VLAN, TDM timeslot or wavelength in the data plane.
- Each data channel is noted by a label and the network topology is split into a number of label layers.
- Data channels of the same label are grouped into a graph layer.



Example: label-layer graph transformation for a 7-node, 4-wavelength IP-over-WDM network.

Graph Transformation Solution – Channel Graph Technique

- **Handling general adaptation constraints**
  - A channel graph is the **dual** of the network graph.
  - It translates each link triplet <*head, tail, switching_capability*> into a node and add an edge between two constructed nodes <*v1, v2, swcap1*> and <*v2, v3, swcap2*> if the switching capability swcap1 on link <*v1, v2*> can be adapted to switching capability swcap2 on link <*v2, v3*>.
  - For **cross-layer adaptation**, *switching type* and *encoding type* are included in the *switching_capability* parameter vector.
  - For **wavelength conversion**, wavelength ID is included in the *switching_capability* parameter vector.

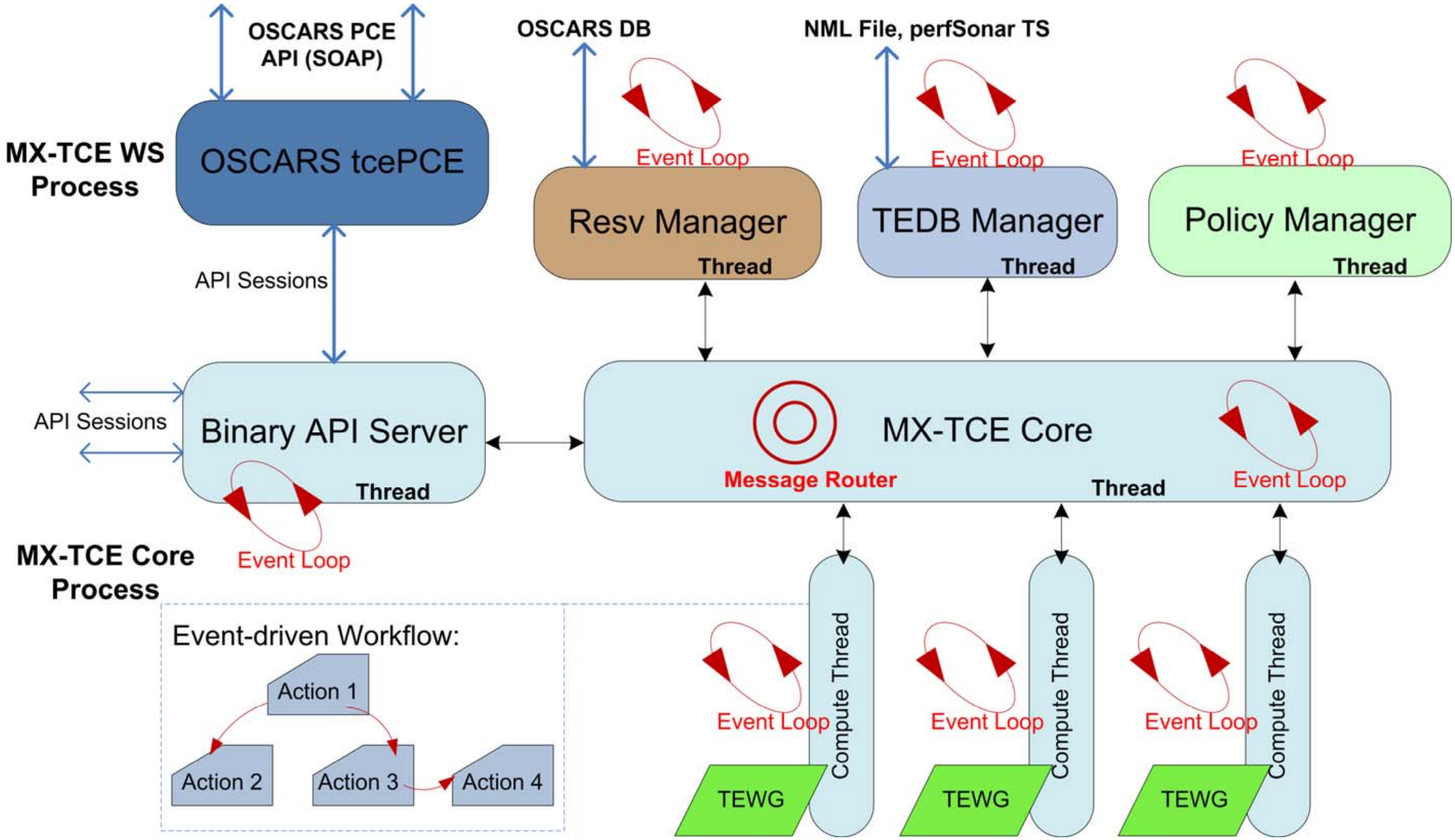


- Example: Original link (S→A) is transformed into channel graph node [S, A, <PSC, Packet>].
- Original link (A→B) into channel graph node [A, B, <LSC, Packet, w_1+w_2>].
- Channel graph link ([S, A, <PSC, Packet>] → [A, B, <LSC, Packet, w1+w2>]) is created for adaptation between IP and WDM layers at node A.

Heuristic Search Solution

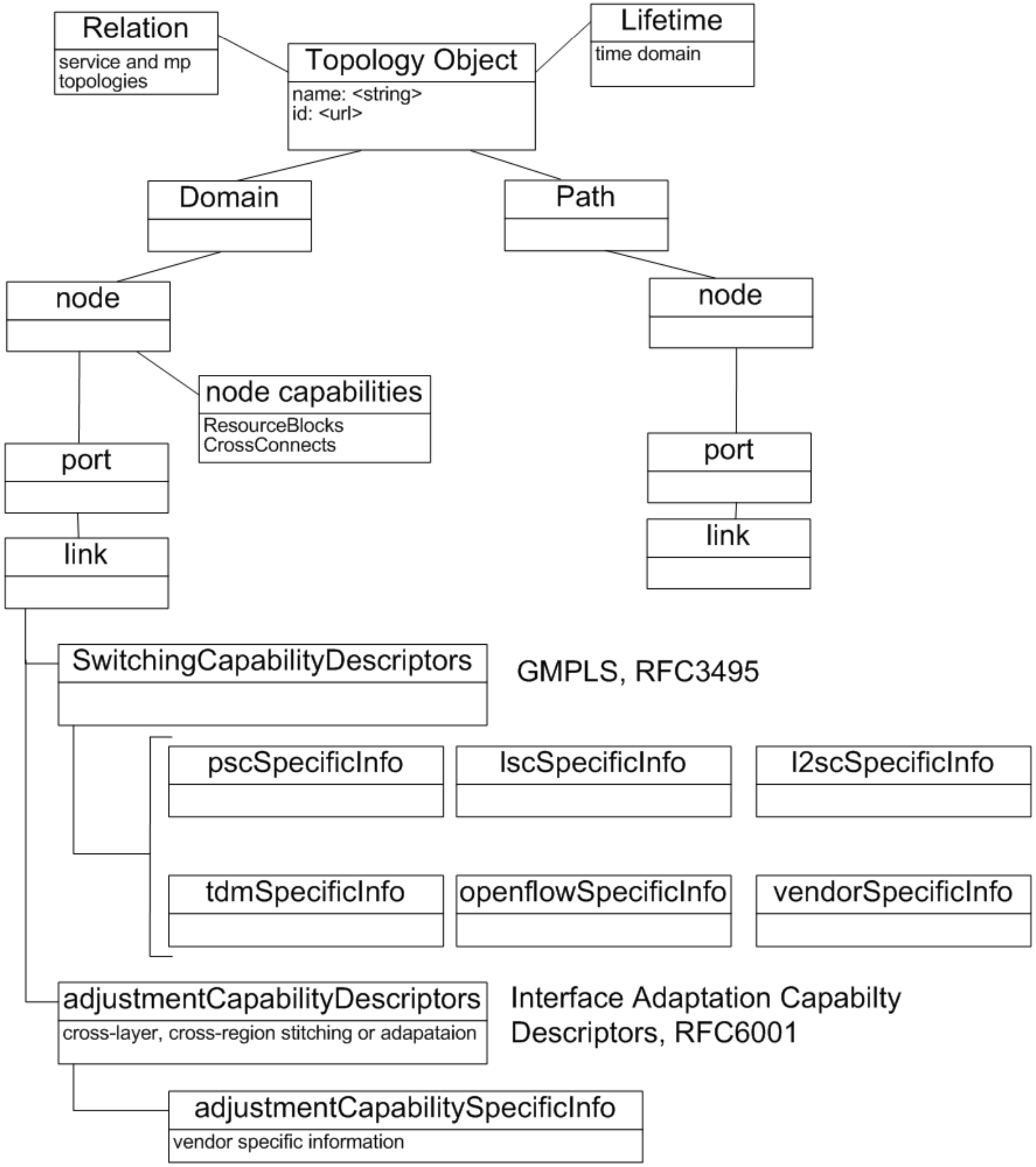
- **Constrained Search and Graph Transformation may not be sufficient to fully address the high complexity.**
  - The search space have not been reduced to a degree that scalability is no longer an issue for even very large networks.
- **Heuristic search solution may be necessary when network scales to very large size.**
  - The basic idea is to trade off reduced search space for sub-optimal paths.
  - Heuristic search can be combined with Constrained Search and applied on the original network topology. Or it can be combined with graph transformation techniques and applied on a transformed network graph.
- **Techniques such as K-Shortest Path (KSP) search have been studied and found effective.**

MX-TCE Architecture and Implementation



ARCHSTONE Network Schema Extensions

- **Extensions to OSCARS v0.6**
- **Added features for:**
  - multi-layer topologies
  - multi-point topologies
  - requests in the form of a "service-topology"
  - vendor specific features
  - technology specific features
  - node level constraints
- **Result is a schema "Superset" to what OSCARS v0.6 has now uses**
  - schema with ARCHSTONE extensions will be backward compatible with current OSCARS operations



ARCHSTONE Summary

- **Network "Service Plane" formalization**
  - Composable Network Service architecture
  - ARCHSTONE Network Service Interface as client entry point
- **Extensions to OSCARS Topology and Provisioning Schemas to enable:**
  - multi-layer topologies
  - multi-point topologies
  - requests in the form of a "service-topology"
  - vendor specific features
  - technology specific features
  - node level constraints
- **MX-TCE (Multi-Dimensional Topology Computation Engine)**
  - Computation Process and Algorithms
- **Enable a New class of Network Services referred to as "Intelligent Network Services"**
  - clients can ask the network "what is possible?" questions
  - can ask for "topologies" instead of just point-to-point circuits

Relationship of our Research to other Internet Development Activities

- **There are other advanced network research activities underway; software defined networking, OpenFlow, clouds, network as a service. Our view of the relationship between our work and these is:**
  - These are tools or mechanisms that will provide more options and features with respect to making things happen in the network
  - This will facilitate our creation of a Network ServicePlane with Intelligent Network Services
  - We are focused on developing the intelligence to use these tools, not the tools themselves
  - Our objective is to utilize every vendor and open source feature we can find, and concentrate on value-added features and intelligence
  - We believe we must develop some complexity to make things simple
- **The core and difficult issues for the ServicePlane will remain even after new tools are developed;**
  - heterogeneous technologies and control planes
  - multiple control and policy domains
  - multi-constraint resource computations
  - need for flexible interaction with application workflows
  - maintenance of service states

Related Activities Funded by ASCR

ASCR Next Generation Networks for Science Research Projects

COMMON: Coordinated Multi-Layer Multi-Domain Optical Network – (09/2010-08/2013)

Vinod Vokkarane – University of Massachusetts Dartmouth

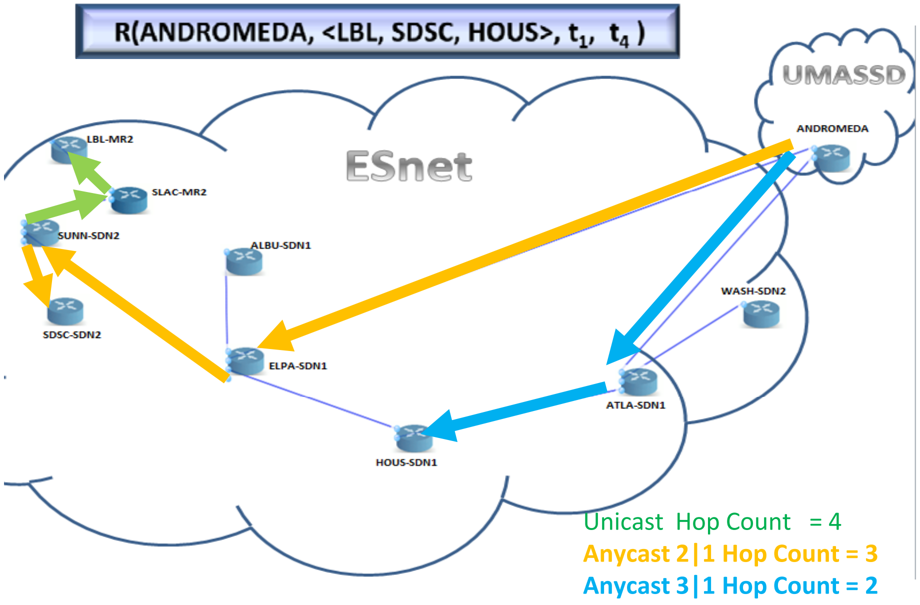
Project Goals

1. To design and implement a production-ready anycast and multicast services on the existing OSCARS framework.
2. To design and implement survivability techniques on OSCARS.
3. To provide for user-profile based access to network resources, while provisioning connection requests.

Current Accomplishments

1. Deployment ready anycast service on OSCARS v0.6 available.
2. Designed and Implemented Multicast overlay service and dedicated path protection service across a single-domain network on OSCARS v0.6.
3. Developed a What-If OSCARS tool for providing user-profile based services.

Multi-Domain Anycast



Impacts on DOE's Mission

Provide DOE scientific community with ability to:

(a) Allow for destination-agnostic service hosting on large-scale networks.
(b) Use a multicast service and increase the service acceptance.
(c) Users are protected from link failures.
(d) Providing user-profile based services.

Hybrid Network Traffic Engineering System

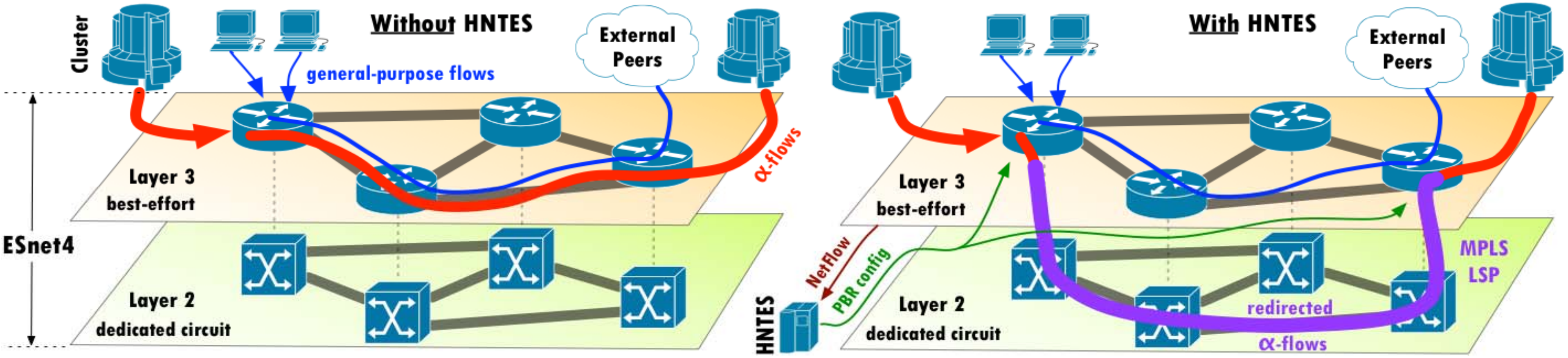
Lead PI: Malathi Veeraraghavan, Univ. of Virginia, Co- PI: Chris Tracy, LBNL/ESnet

Goals:

- *Research and development of a hybrid network traffic engineering system that combines best –effort IP traffic with dynamically provisioned circuit traffic into integrated traffic carried over a common network infrastructure.*
- *Prototype and demonstrate the resulting hybrid network traffic engineering capability on a testbed for possible adoption by ESnet*

- *Designed and implemented an automatic offline alpha flow identification algorithm (HNTES v2.0)*
- *Tested on ESnet, and used to collect NetFlow data from four routers for May-Nov. 2011*
- *Analyzed NetFlow data and showed that this offline mechanism is highly effective (91% of bytes generated by alpha flows in bursts would have been redirected) for BNL PE router*
- *Preliminary experiments for flow redirection completed on ANI testbed*

This project, if completed successfully, has potential to be adopted by ESnet to combine DOE's Science Data Network (SDN) and ESnet that are currently managed an operated as two separate infrastructures. This will significantly reduce the operational cost and will provide guaranteed end-to-end differentiated services to high-end science applications.



ASCR Next Generation Networks for Science Research Projects

End Site Control Plane System – (10/1/2009-9/30/2012)

Phil DeMar (FNAL), Dantong Yu (BNL), Martin Swany (Univ of Delaware)

Project Goals:

- *Develop network service to facilitate site use of circuit services*
  - *Accept and process user/app requests for circuit services*
  - *Initiate reservation, setup, & teardown of WAN circuit services (ie., OSCARs)*
  - *Configure local network infrastructure for use of circuits*
- *Monitor local segments of end-to-end path*

Current Accomplishments:

- *End-to-end path model and corresponding information schema completed*
- *Network model for generic configuration of site-specific local infrastructure completed*
- *Local infrastructure configuration module (LDC) in prototype*
  - *Evaluating potential OpenFlow interaction*
- *Local path segment monitoring capability developed (ESCPScope)*
- *Prototype system developed and functionality demonstrated*

Impacts on DOE's Mission:

- *Supports DOE strategic networking direction toward deployment and use of data circuits for high-impact, large-scale science data movement*
- *Provides critical component that ties together end-sites and WAN to achieve end-to-end QoS guarantees for high-impact data flows*
- *Complements DOE R&D efforts in wide-area network support services (i.e. OSCARS)*

ASCR Next Generation Networks for Science Research Projects

VNOD: Virtual Network On Demand – (10/1/2010-9/30/2013)

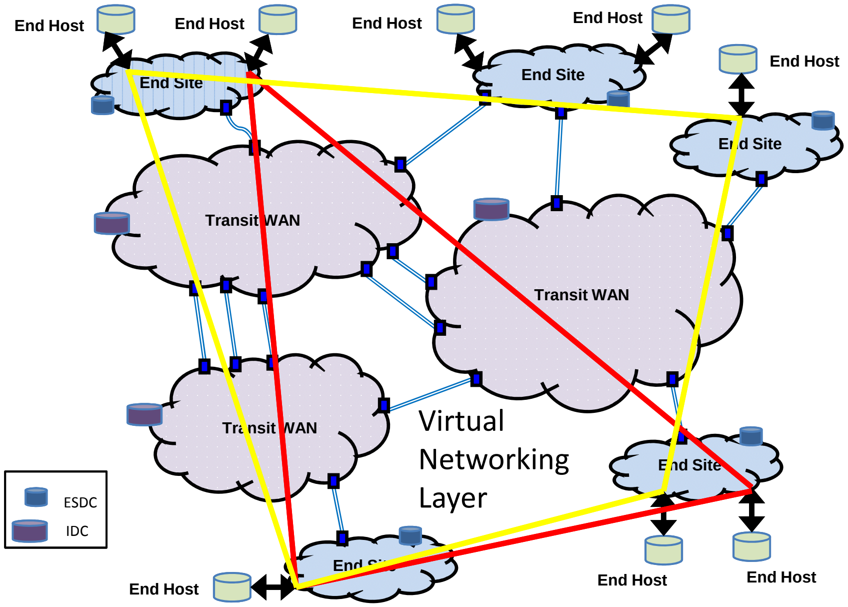
Dimitrios Katramatos and Dantong Yu, BNL

Goals

- Build an on-demand network virtualization infrastructure for data-intensive scientific applications/workflows spanning multiple end-sites
- Intelligently form virtual network domains (ViNets) encompassing multiple end-sites by leveraging end-to-end virtual path technology
- Enable scientific teams to use high-speed connections effectively and efficiently

Accomplishments

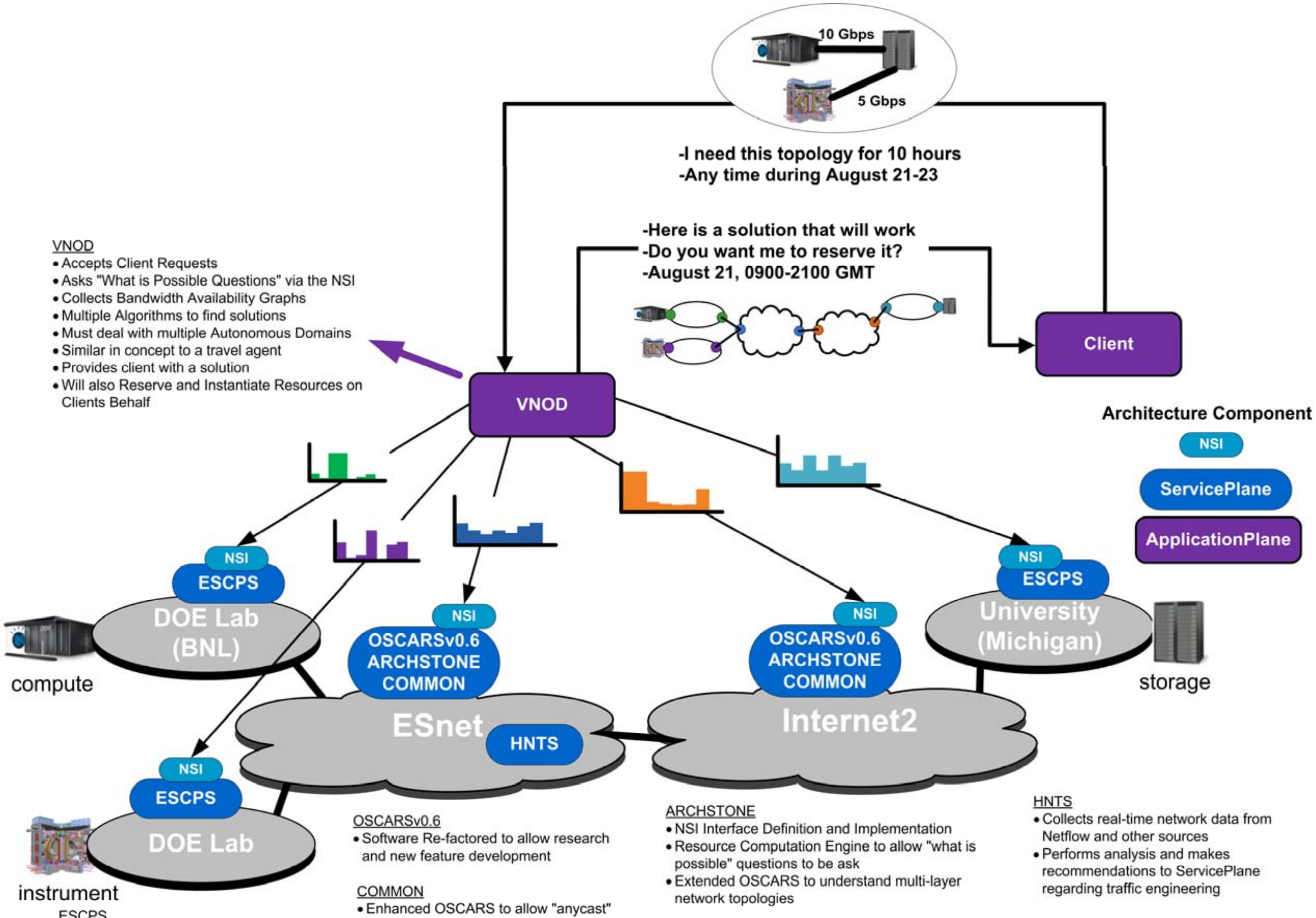
- Development of scheduling framework for accommodating multiple requests in a multiple end-site / multiple network domain environment
- Development of prototype system
- Publications:
  - *Design and Implementation of an Intelligent End-to-End Network QoS System*, to appear ACM/IEEE Supercomputing 2012
  - *Virtual Network On Demand: Dedicating Network Resources to Distributed Scientific Workflows*, in Proceedings of DIDC Workshop, ACM HPDC 2012
  - *End-to-End Network QoS via Scheduling of Flexible Resource Reservation Requests*, in Proceedings of ACM/IEEE Supercomputing 2011



Impact

- Provides an end-to-end network virtualization layer which can overlay multiple virtual networks, tailor-made to the needs of users/application communities, over the physical network infrastructure
- Offers scientists an easy way to utilize cutting-edge virtualization technology by providing a center for defining, establishing, and managing virtual networks
- Improves efficiency of applications by providing true end-to-end QoS between end hosts
- Provides the network scheduling component of a wider scope resource co-scheduler

Building a ProtoType Network ServicePlane with Intelligent Network Services



VNOD
- Accepts Client Requests
- Asks "What is Possible Questions" via the NSI
- Collects Bandwidth Availability Graphs
- Multiple Algorithms to find solutions
- Must deal with multiple Autonomous Domains
- Similar in concept to a travel agent
- Provides client with a solution
- Will also Reserve and Instantiate Resources on Clients Behalf

ESCPS
- Responsible for DOE Lab Sites Network Resources
- Should responds to VNOD request for "what is possible" queries

OSCARSv0.6
- Software Re-factored to allow research and new feature development

COMMON
- Enhanced OSCARS to allow "anycast" and "manycast" provisioning actions
- Path computation to allow "what is possible" questions to be ask

ARCHSTONE
- NSI Interface Definition and Implementation
- Resource Computation Engine to allow "what is possible" questions to be ask
- Extended OSCARS to understand multi-layer network topologies

HNTS
- Collects real-time network data from Netflow and other sources
- Performs analysis and makes recommendations to ServicePlane regarding traffic engineering

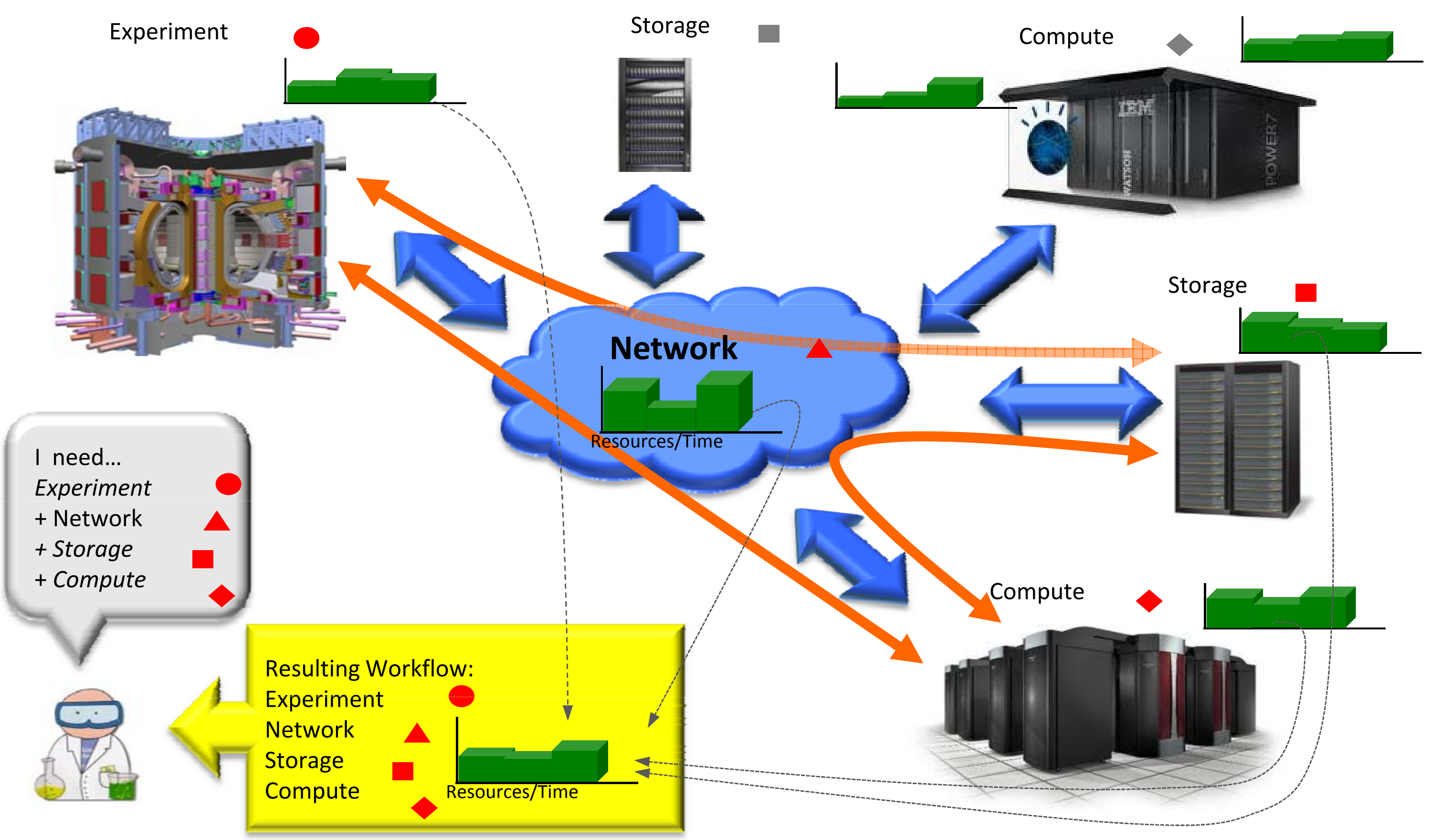
Thank-you

Extras

Application Workflow Integration

A key focus is on technology development which allow networks to participate in application workflows

The Network needs to be available to application workflows as a first class resource in this ecosystem



Experiment ●

Storage ■

Compute ◆

Network ▲

Resources/Time

Storage ■

Compute ◆

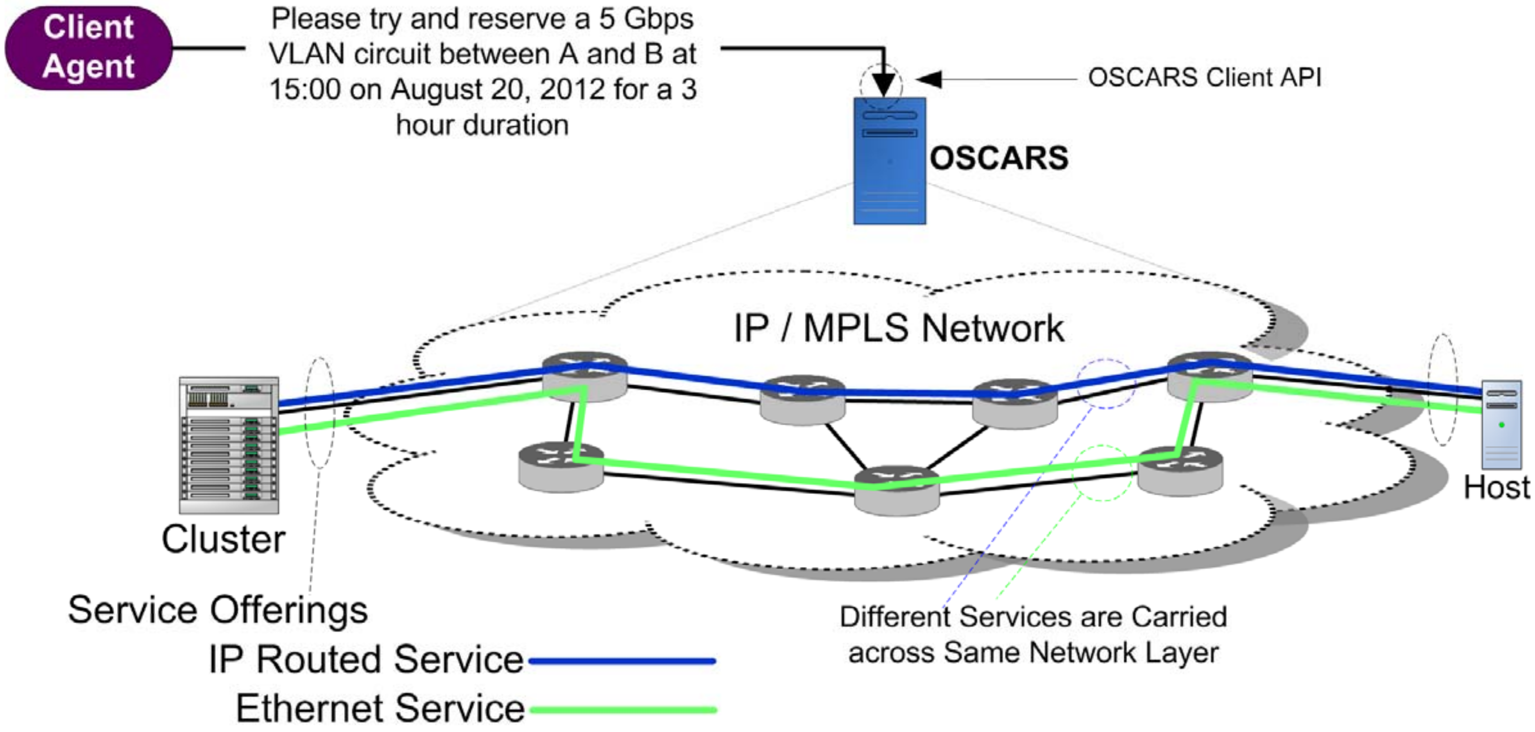
I need…
Experiment ●
+ Network ▲
+ *Storage* ■
+ *Compute* ◆

Resulting Workflow:
Experiment ●
Network ▲
Storage ■
Compute ◆

Resources/Time

Looking again at Network Service Today

- **Advanced Guaranteed Bandwidth Dynamic Network Services are available today on ESnet via OSCARS**
  - OSCARS API is simple to use with a basic service offering



Please try and reserve a 5 Gbps VLAN circuit between A and B at 15:00 on August 20, 2012 for a 3 hour duration

Different Services are Carried across Same Network Layer

- **Also Inter-domain and multi-technology capable**
- **This is a great "Service", but does not bring the network to the level of a "Resource"**

The Network as a Resource

- **Toward these goals we have developed an architecture to realize the Network as a Resource**
- **There are three key architectural components**
  - **Network Service Interface (NSI)**: a well defined interface that applications can use to plan, schedule, and provision
  - **Network ServicePlane**: a set of systems and processes that are responsible for providing services to users and maintaining state on those services
  - **Intelligent Network Services**: a set of ServicePlane capabilities that allow other processes to interact with the network in a workflow context

Atomic Services Examples

Topology Service to determine resources and orientation

Resource Computation Service* to determine possible resources based on multi-dimensional constraints (*MX-TCE)

Connection Service to specify data plane connectivity

Protection Service to enable resiliency through redundancy

Restoration Service to facilitate recovery

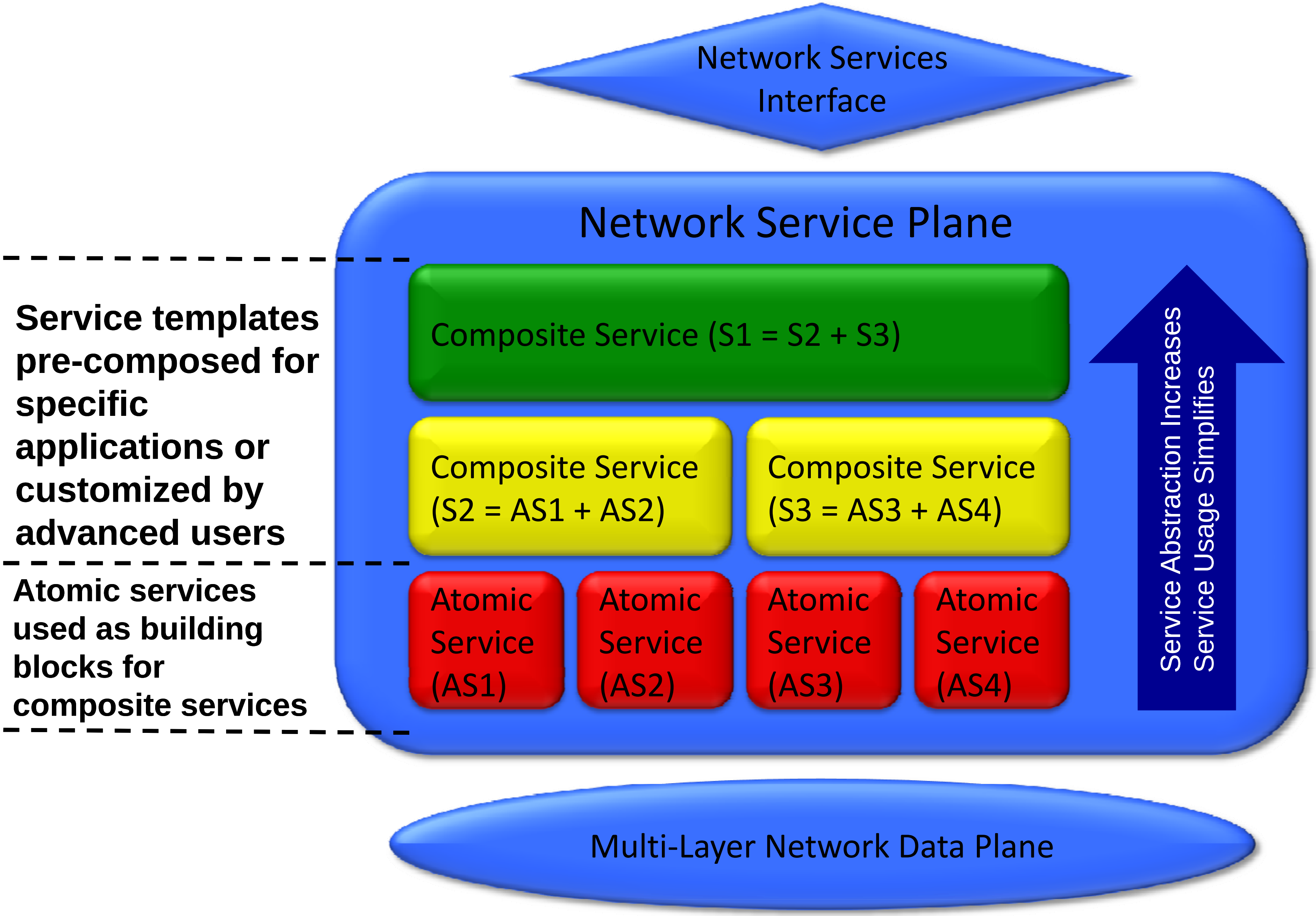
Security Service (e.g. encryption) to ensure data integrity

Store and Forward Service to enable caching capability in the network

Measurement Service to enable collection of usage data and performance stats

Monitoring Service to ensure proper support using SOPs for production service

Atomic and Composite Network Services Architecture



Network Services Interface

Network Service Plane

Service templates pre-composed for specific applications or customized by advanced users

Composite Service (S1 = S2 + S3)

Composite Service (S2 = AS1 + AS2)

Composite Service (S3 = AS3 + AS4)

Atomic services used as building blocks for composite services

Atomic Service (AS1)

Atomic Service (AS2)

Atomic Service (AS3)

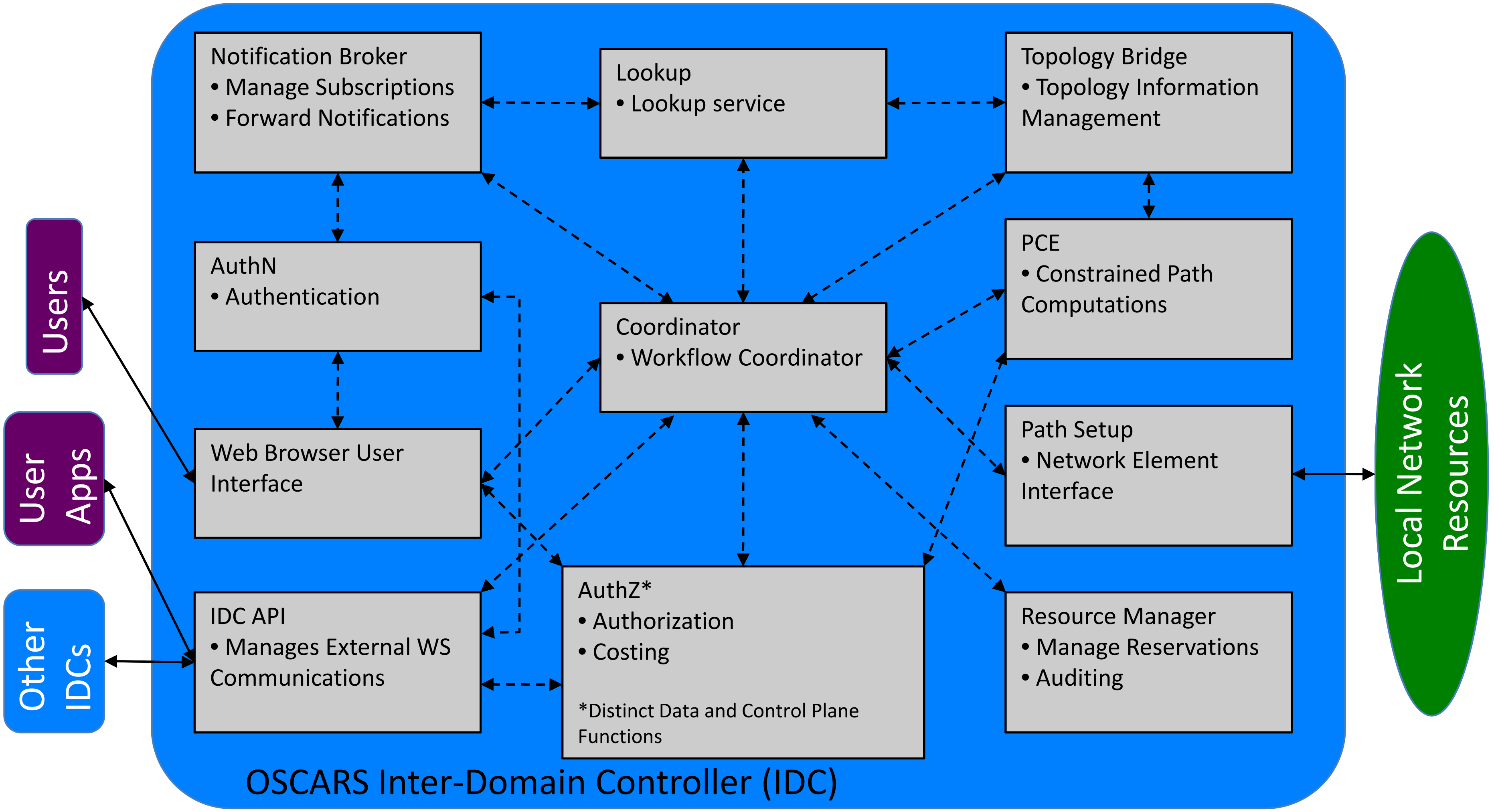
Atomic Service (AS4)

Service Abstraction Increases

Service Usage Simplifies

Multi-Layer Network Data Plane

Modularization of OSCARS



OSCARS (v0.6) was re-factoring in order to provide a platform for research and development into next generation network capabilities