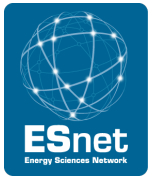
Advanced Resource Computation for Hybrid Service and TOpology NEtworks (ARCHSTONE)

DOE NGNS PI Meeting, Mar 18-20, 2013
Emeryville, CA

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

Chin Guok
Energy Sciences Network (ESnet)

Nasir Ghani
University of New Mexico



U.S. DEPARTMENT OF ENERGY
Office of Science

Personnel

- **USC/ISI**
  - Tom Lehman
  - Xi Yang
- **ESnet**
  - Chin Guok
  - Eric Pouyoul
  - Inder Monga
  - Vangelis Chaniotakis
  - Bharath Ramaprasad (UMass)
- **UNM**
  - Nasir Ghani
  - Feng Gu
  - Kaile Liang

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

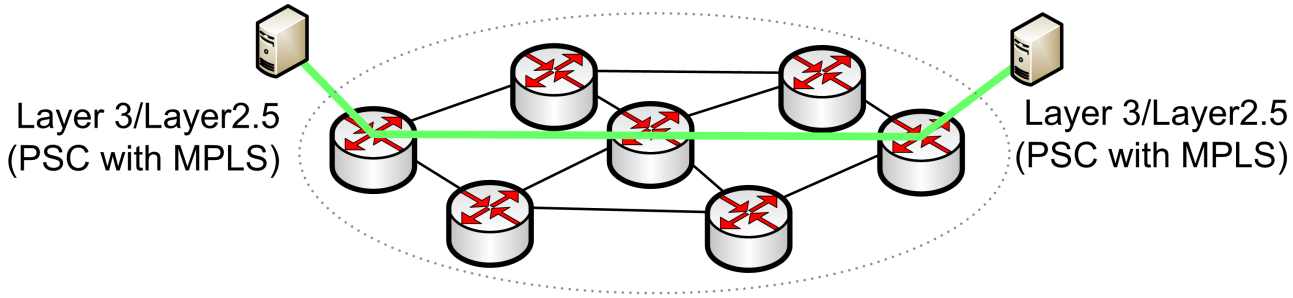
For service provision we should treat all the network layers in a holistic and integrated fashion

- **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)

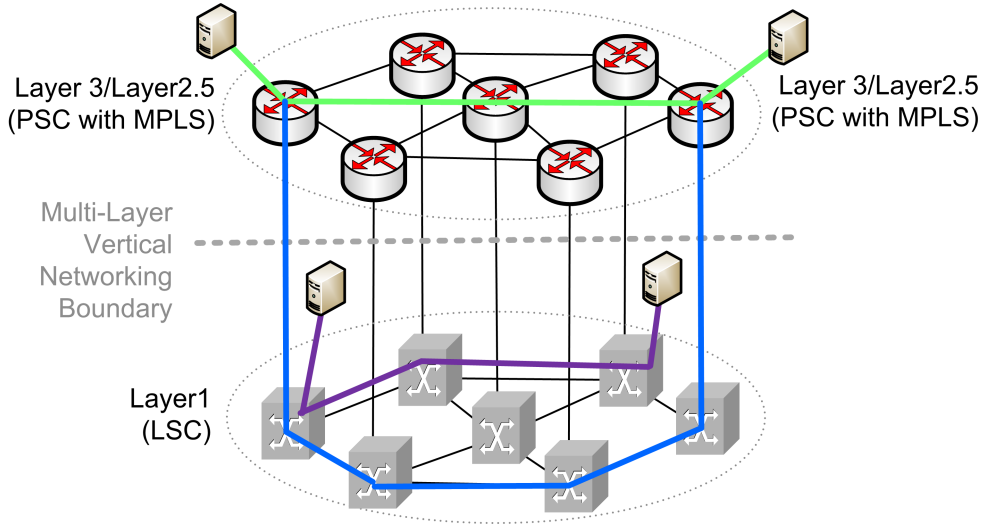
The Network needs to be available to application workflows as a first class resource in this ecosystem

Multi-Layer Networks

- Today our dynamic provisioning systems only see single layer



- But the networks are really multi-layer

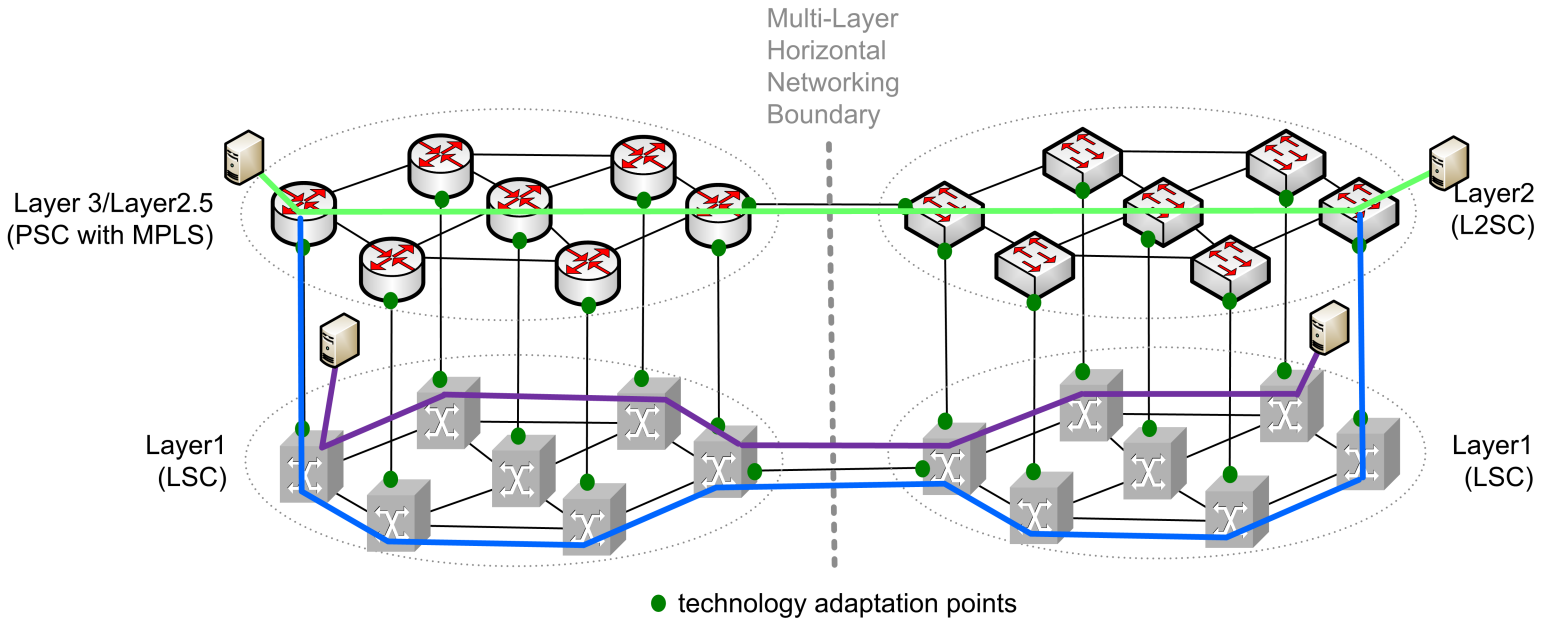


would like to:

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

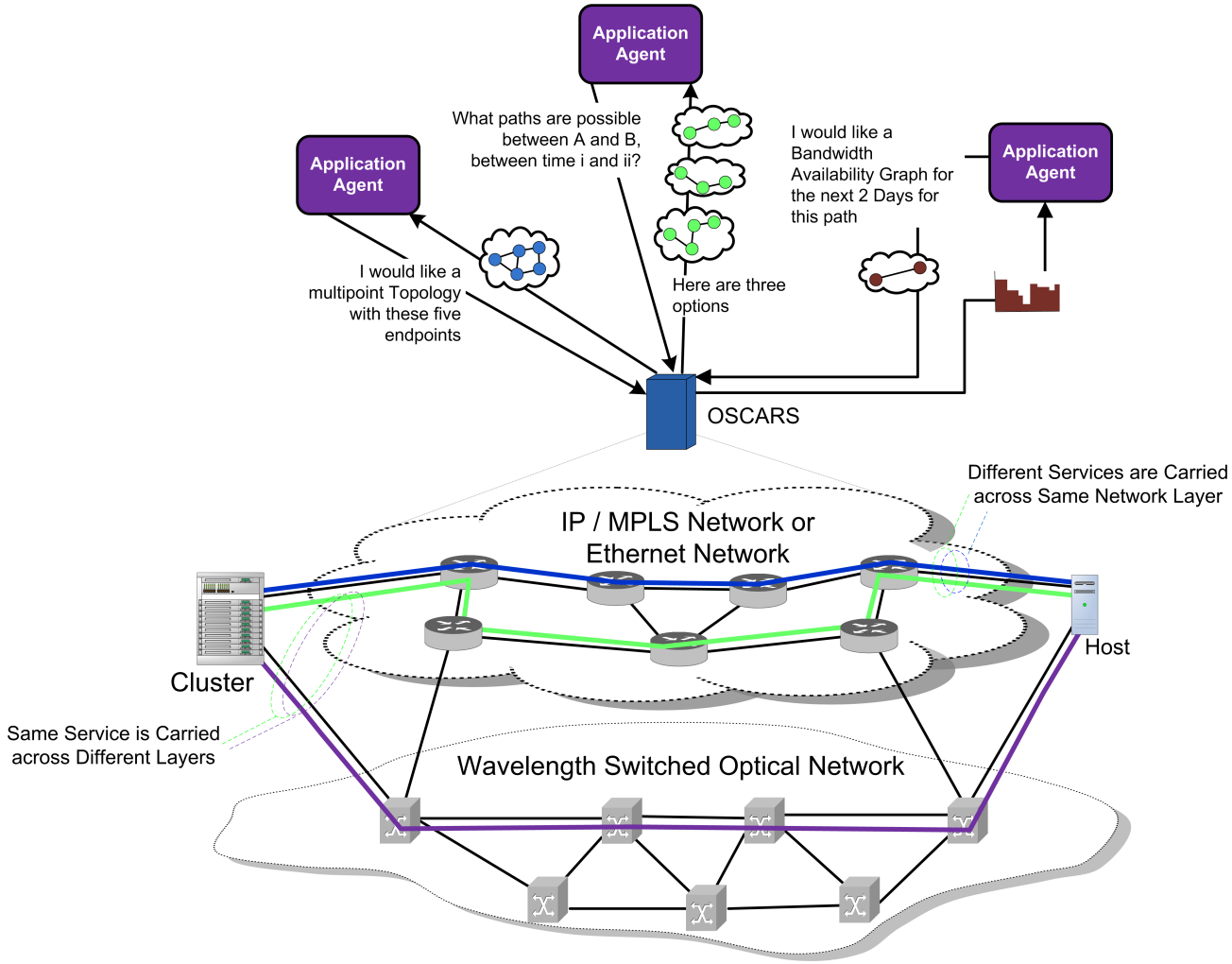
Multi-Layer Networks

- 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"**

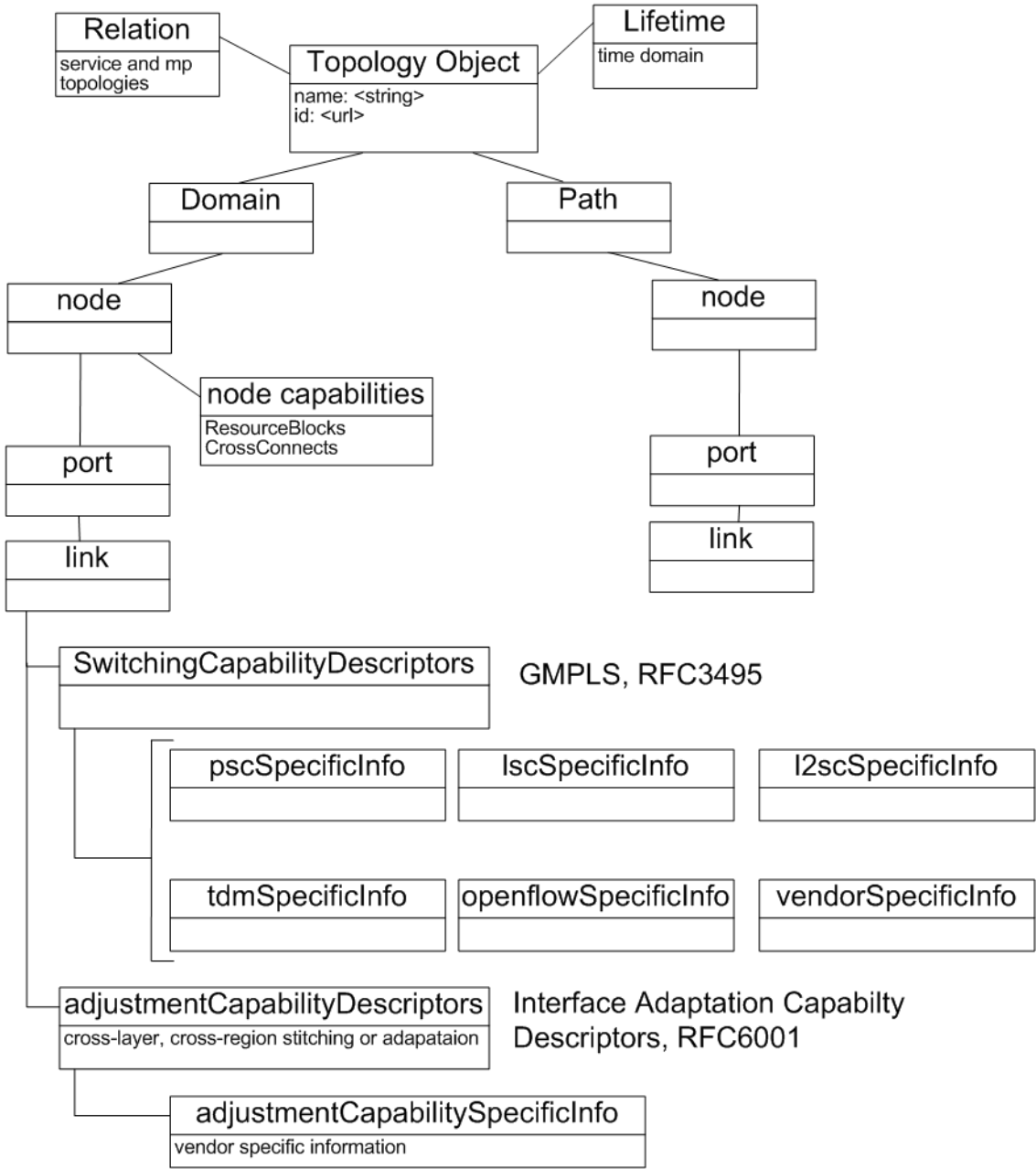


What are the Main Challenges?

- **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 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

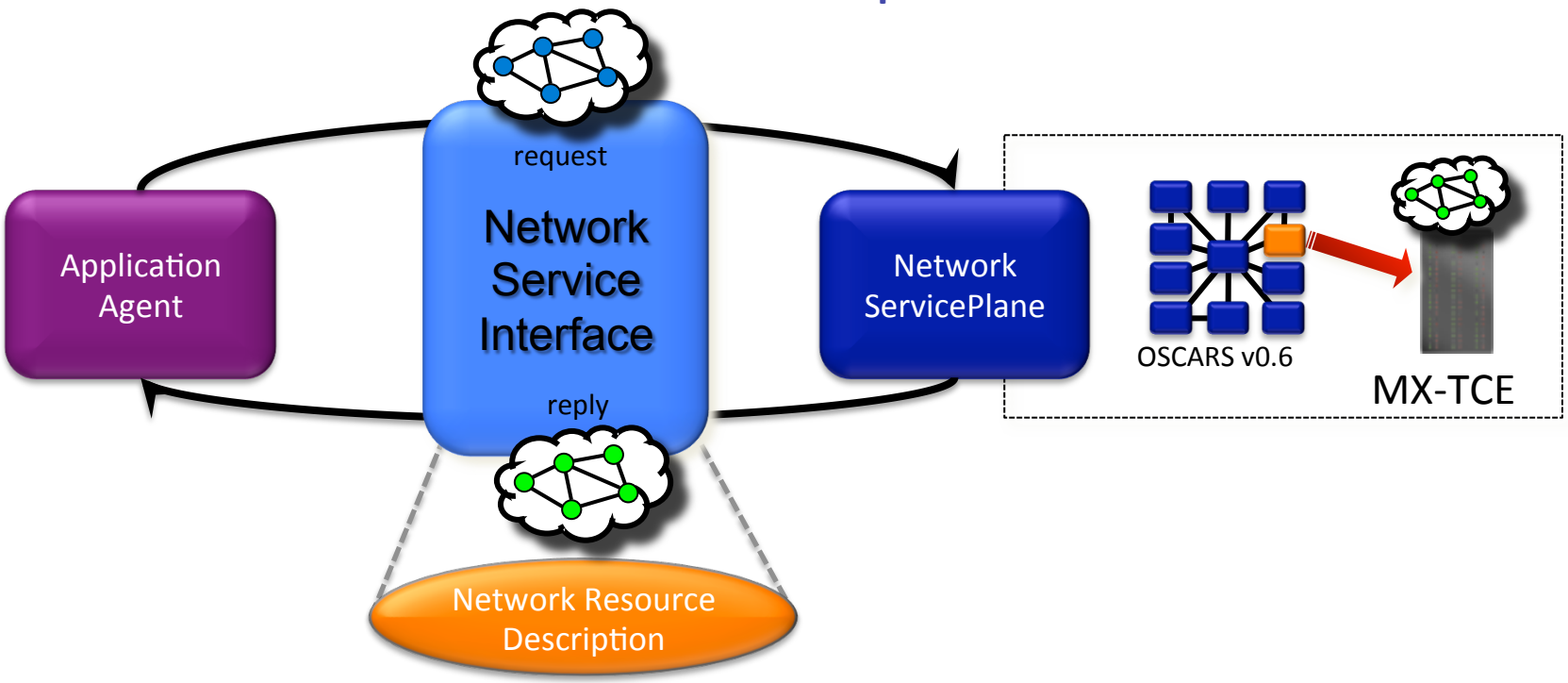


ARCHSTONE Network Schema Extensions

- **Detailed ARCHSTONE Schema Extensions available here:**
  - http://archstone.east.isi.edu/twiki/bin/view/ARCHSTONE/Software
- **Topology Schema**
- **Example Network Advertisement and Path Description Schema**
- **Example Service Topologies (Request and Reply)**
  - Point-2-Point Service Topology
  - Simple-MultiPoint Service Topology
  - Bridged-MultiPoint Service Topology
  - Meshed-MultiPoint Service Topology

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**



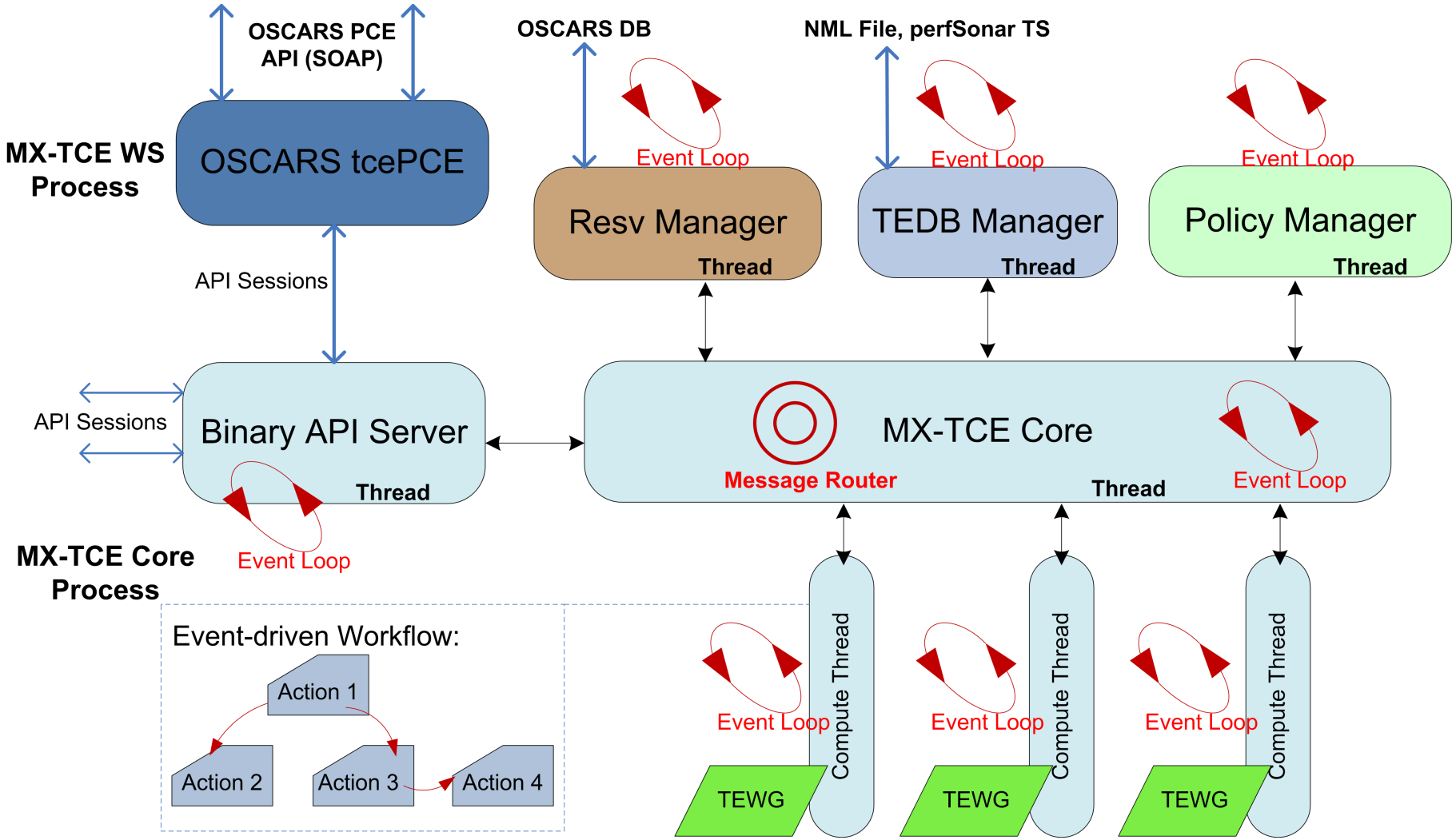
Multi-Dimensional Topology Computation

- **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).

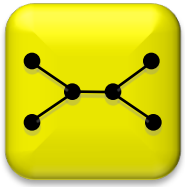
Multi-Dimensional Topology Computation

- **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**

MX-TCE Architecture and Implementation

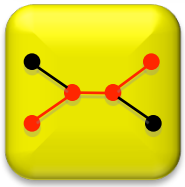


Atomic Services Examples



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

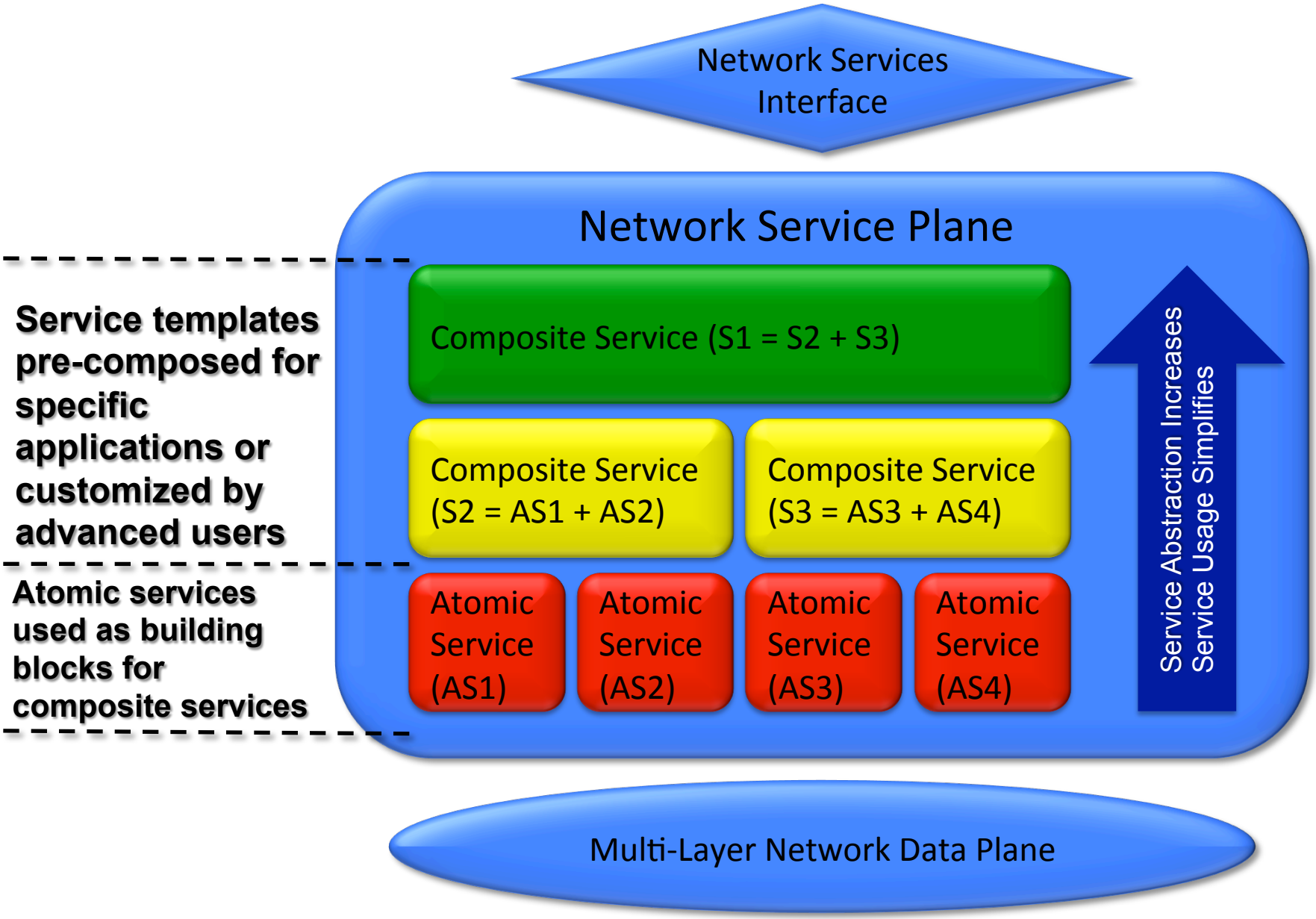
Security Service (e.g. encryption) to ensure data integrity

Store and Forward Service to enable caching capability in the network

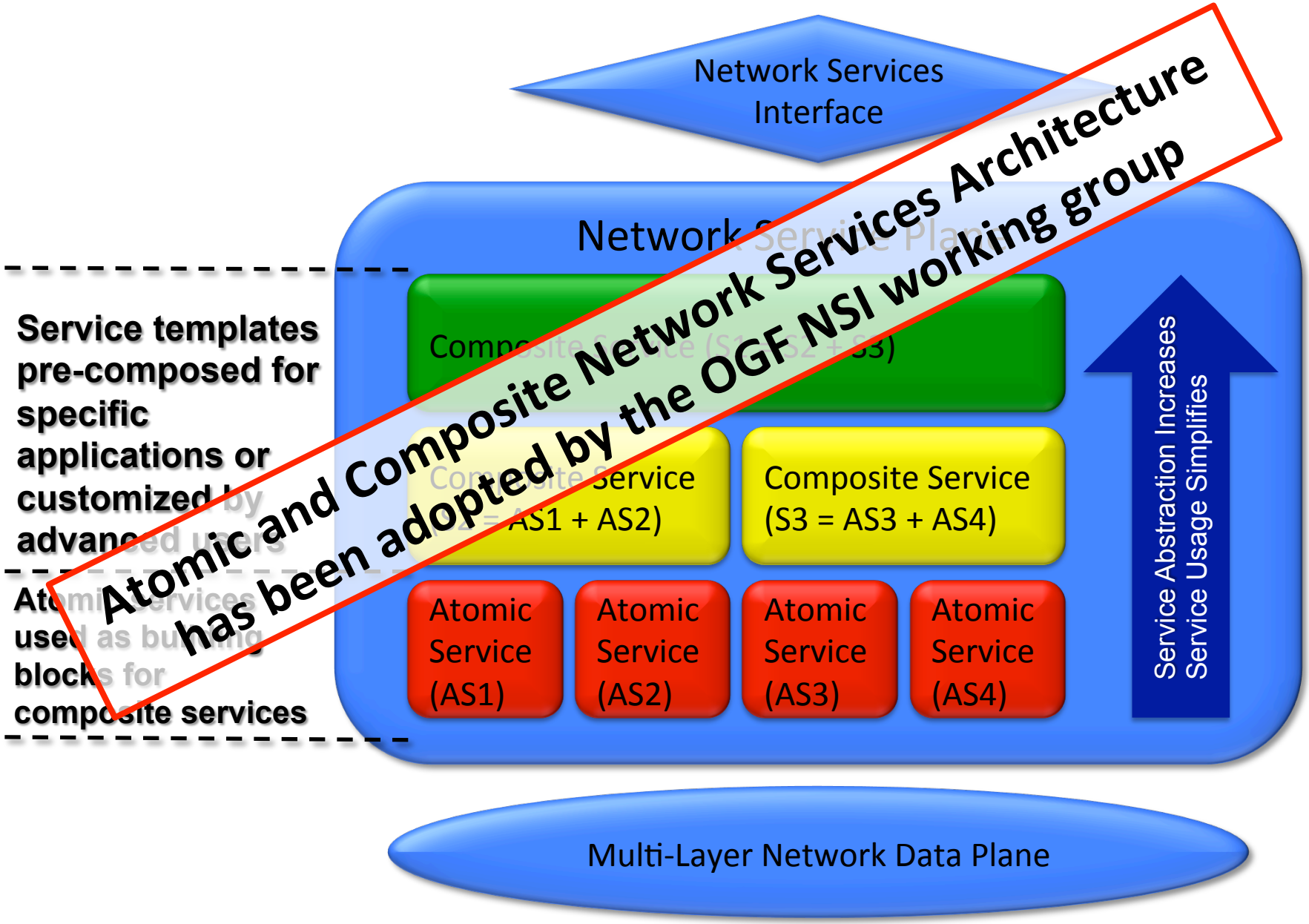
Measurement Service to enable collection of usage data and performance stats

Monitoring Service to ensure proper support using SOPs for production service

Atomic and Composite Network Services Architecture



Atomic and Composite Network Services Architecture



ARCHSTONE Accomplishment Summary

- **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
- **Network "Service Plane" formalization**
  - Composable Network Service architecture
  - ARCHSTONE Network Service Interface as client entry point
- **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

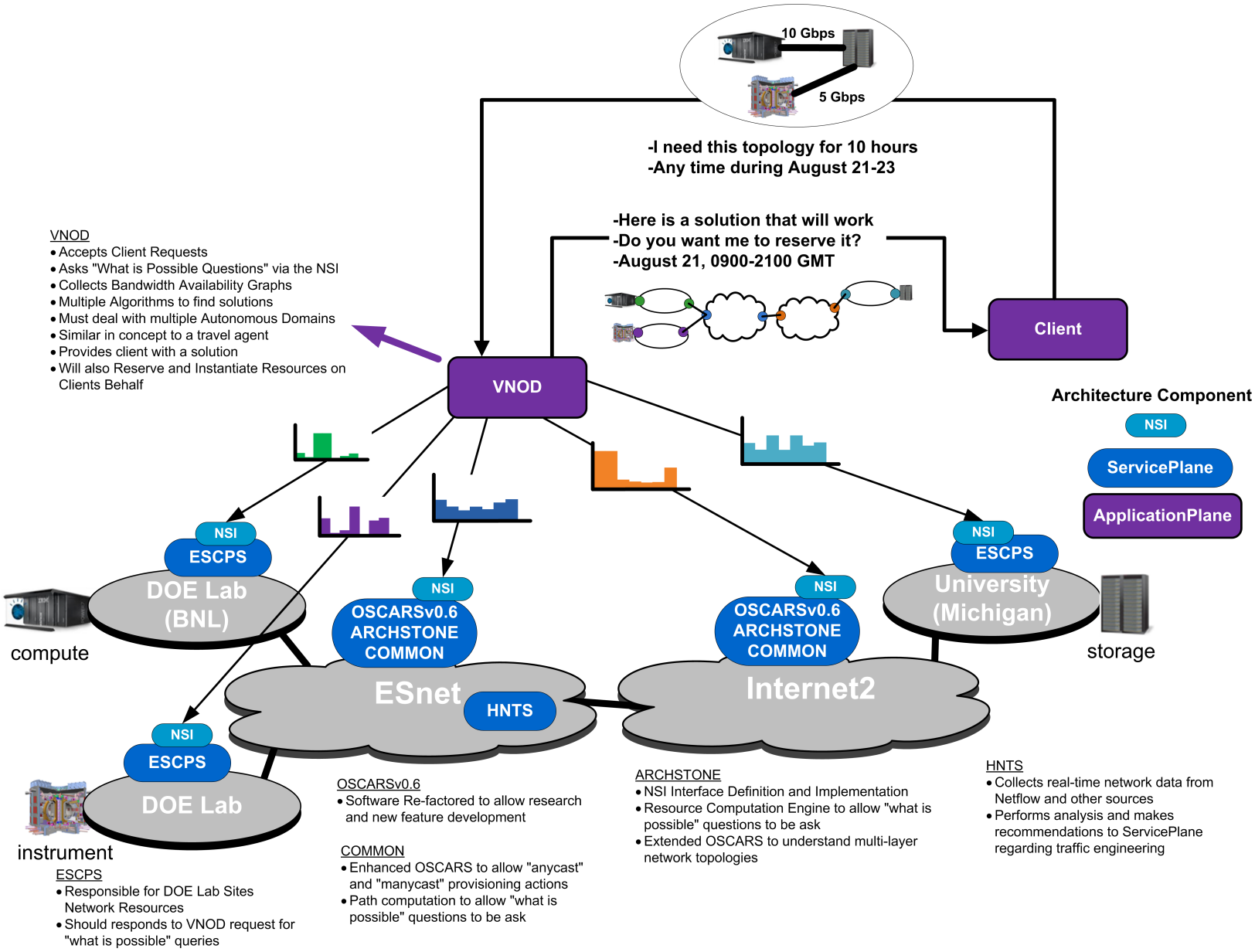
ARCHSTONE Document and Software Links

- **ARCHSTONE Multi-X Topology Computation Element (MX-TCE)**
  - http://archstone.east.isi.edu ===> Software
- **ARCHSTONE Final Report and other Documents**
  - http://archstone.east.isi.edu ===> Documentation
- **OSCARS v0.6 ARCHSTONE Branch**
  - https://oscars.es.net/repos/oscars/branches/archstone/
- **OSCARS v0.6 Modular PCE Implementation (part of the Production Trunk)**
  - https://oscars.es.net/repos/oscars/trunk/

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

Building a ProtoType Network ServicePlane with Intelligent Network Services



Intelligent Network Services Document

Available at: http://archstone.east.isi.edu/twiki/pub/ARCHSTONE/Documents/intelligent-network-services-v1.2.pdf

What Are Some Workflow Drivers (1/4)

- **Movement of Large Data Sets with Deadline Scheduling Requirements**
  - Motivation
    - Big science generates big data that need to be moved between the experiment, compute, and storage resources
  - Core Requirements
    - Advance co-scheduling of network and storage resources
    - Protection and/or recovery to prevent data loss and re-transmission delays
    - Interim storage to mitigate temporary service interruptions
  - Example Applications
    - Large Hadron Collider (LHC) (High Energy Physics)
    - Belle II (High Energy Physics)
    - 4th Generation Light Sources (Basic Energy Sciences)
- **Time Sensitive Data Transfers as part of an Execution Workflow**
  - Motivation
    - Deterministic distributed workflow execution
  - Core Requirements
    - Strict co-scheduling to ensure all components of the workflow pipeline is online
    - Fault tolerance, ability to specify alternatives in the event of errors
  - Example Applications
    - Large Hadron Collider (LHC) (High Energy Physics)
    - International Fusion Experimental (ITER) (Fusion Energy)
    - 3rd Generation Light Sources (APS, ALS, LCLS, NSLS, SSRL) (Basic Energy Sciences)

What Are Some Workflow Drivers (2/4)

- **Simultaneous Use of Multiple, Very Large, Distributed Data Sets via Remote I/O**
  - Motivation
    - Real-rime access to large data sets with limited or no local storage
  - Core Requirements
    - Dynamic network service topologies with real-time networking
    - Co-scheduling of resources to access data sets at different locations
    - Close to zero packet loss and reordering to prevent performance collapse
  - Example Applications
    - Large Hadron Collider (LHC) (High Energy Physics)
    - Square Kilometer Array (SKA) (Astrophysics)
    - Systems Biology Applications (Genomics, Metabolomics, Proteomics, etc) (Biological and Environmental Research)
    - Atmospheric Radiation Measurement Program (ARM) (Biological and Environmental Research)
- **Ad-hoc Integrated LAN/WAN VPNs**
  - Motivation
    - Implement complex or unique routing policies on a private (multi-domain) network substrate
  - Core Requirements
    - Dynamic network service topologies (overlays) with predictable characteristics to accommodate end-to-end service level consistency
    - Resiliency to mitigate outages in participating network domains
  - Example Applications
    - Large Hadron Collider (LHC) (High Energy Physics)
    - 4th Generation Light Sources (Basic Energy Sciences)
    - Systems Biology Applications (Genomics, Metabolomics, Proteomics, etc) (Biological and Environmental Research)

What Are Some Workflow Drivers (3/4)

- **Storage and Retrieval of Data from Distributed Depots**
  - Motivation
    - Load balancing of multiple concurrent data transfers, bringing data closer to where is it needed
  - Core Requirements
    - Dynamic network service topologies (overlays) with replication capabilities
    - Resource management and optimization algorithms to determine “best” depot to retrieve data
  - Example Applications
    - Large Hadron Collider (LHC) (High Energy Physics)
    - Earth System Grid Federation (ESGF) (Biological and Environmental Research)
    - Systems Biology Knowledgebase (KBase) (Biological and Environmental Research)
- **Remote Control of Experiments/Instruments**
  - Motivation
    - Support real-time requirements of distributed collaborations
  - Core Requirements
    - Real-time networking for predictable network behavior
    - Low/zero jitter and low latency
  - Example Applications
    - 3rd Generation Light Sources (APS, ALS, LCLS, NSLS, SSRL) (Basic Energy Sciences)
    - International Fusion Experimental (ITER) (Fusion Energy)

What Are Some Workflow Drivers (4/4)

- **Correlation of Data Sets Generated by Distributed Instruments**
  - Motivation
    - Real-time coordination of data streams from distributed instruments
  - Core Requirements
    - Dynamic network service topologies with real-time networking for predictable network behavior
    - Strict scheduling of network resources to facilitate data movement when observation is in progress
    - Close to real-time resource reservations (short turn-around) if observations are transient
    - Protection and/or recovery to prevent loss of observation data
  - Example Applications
    - Very Long Baseline Interferometry (VLBI) (Astrophysics)
    - Square Kilometer Array (SKA) (Astrophysics)
    - Multi-Modal Experimental Analysis (Basic Energy Sciences)

Summary of Science Applications and Requirements

Scientific Application Requirements / Scientific Application Categories	Movement of Large Data Sets with Deadline Scheduling Requirements	Storage and Retrieval of Data from Distributed Depots	Correlation of Data Sets Generated by Distributed Instruments	Simultaneous Use of Multiple, Very Large, Distributed Data Sets via Remote I/O	Time Sensitive Data Transfers as part of an Execution Workflow	Remote Control of Experiments / Instruments	Ad-hoc Integrated LAN / WAN VPNs
Data Management Requirements							
✓ Directory Services (e.g. Meta-data)	No	Yes	No	No	No	No	No
✓ Data Duplication	No	Yes	No	Maybe	No	No	No
✓ Large Data Transfers	Yes	Yes	Maybe	Yes	Yes	No	No
Resource Co-Scheduling (i.e. instruments, storage, compute, visualization, network) Requirements							
✓ Workflow Paradigms	Maybe	Maybe	Yes	No	Yes	Yes	No
✓ Resource Brokering and Co-Scheduling	Yes	Yes	Yes	Yes	Yes	Yes	No
✓ Synchronization of Data Streams	No	Maybe	Yes	No	Maybe	No	No
✓ Real-Time Resource Reservation (Short Turn-Around)	No	Maybe	Yes	Maybe	Maybe	Maybe	No
Network Content Requirements							
✓ Data Replication	No	Yes	No	No	No	No	No
✓ Store-and-Forward	No	Maybe	No	No	No	No	No
Network Connection Requirements							
✓ Guaranteed Bandwidth Scheduling (Strict, Flexible)	Yes	Yes	Yes	Yes	Yes	Yes	Yes
✓ Dynamic Service Topology Overlays (P2P, P2MP, P2MP)	Yes	Yes	Yes	Yes	Yes	Yes	Yes
✓ Protection / Recovery (Failure / Degradation Triggered)	Maybe	Maybe	Yes	Yes	Yes	Yes	Yes
✓ Near Zero Packet Loss / Reordering	No	Maybe	Maybe	Maybe	No	Yes	Yes
✓ Low Latency	No	No	Maybe	Maybe	Maybe	Yes	Maybe
✓ Near Zero Jitter	No	No	Maybe	Maybe	Maybe	Yes	Maybe
Network Measurement / Monitoring Requirements							
✓ SLA / SLE Verification	Yes	Maybe	Yes	Yes	Maybe	Yes	Yes
✓ Auditing / Accounting	Maybe	Maybe	Maybe	Maybe	Maybe	Maybe	Maybe
✓ Performance Prediction and Trending	Maybe	Maybe	Yes	Yes	Maybe	Yes	Yes
✓ User Planning and Debugging Tools	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Workflows that would be of most interest to HEP

In all cases, network measurement and monitoring were a requirement for services beyond best effort.

Table of Network Capabilities to Support Science Requirements

Scientific Application Requirements / Network Capabilities	Content-Centric Networks	Content Delivery Networks	Data Transmission Protocols	Workflow Management	Resource Scheduling	Advance Resource Computation	Disruption-Tolerant Networks	Real-Time Networks	Multi-Layer Provisioning	Signaling Protocols	Quality of Service	AuthN / AuthZ	Performance Analysis
Data Management Requirements													
✓ Directory Services (e.g. Meta-data)	X	X											
✓ Data Duplication	X	X											
✓ Large Data Transfers			X										
Resource Co-Scheduling (i.e. instruments, storage, compute, visualization, network) Requirements													
✓ Workflow Paradigms				X								X	
✓ Resource Brokering and Co-Scheduling					X	X						X	
✓ Synchronization of Data Streams					X								
✓ Real-Time Resource Reservation (Short Turn-Around)					X							X	
Network Content Requirements													
✓ Data Replication	X	X											
✓ Store-and-Forward	X	X											
Network Connection Requirements													
✓ Guaranteed Bandwidth Scheduling (Strict, Flexible)						X			X	X	X	X	
✓ Dynamic Service Topology Overlays (P2P, P2MP, P2MP)						X			X	X		X	
✓ Protection / Recovery (Failure / Degradation Triggered)						X	X		X	X			
✓ Near Zero Packet Loss / Reordering						X			X		X		
✓ Low Latency						X							
✓ Near Zero Jitter						X		X	X		X		
Network Measurement / Monitoring Requirements													
✓ SLA / SLE Verification												X	X
✓ Auditing / Accounting												X	X
✓ Performance Prediction and Trending													X
✓ User Planning and Debugging Tools					X	X			X				X

- Conceptual or prototype
- Beta or early deployment
- Matured or ubiquitous deployments

"Above the network" services and functions

Functions and services within the network

Network supporting functions or services

Network Capability Highlights (1/2)

- **Content-Centric Networks (CCN)**
  - Users can request data without any knowledge of it's location
  - The name of the content sufficiently describes the information and captures it's ontology, provenance, and locality
  - Requires fundamental changes in today's network infrastructure to support this
  - ***May prove to be a disruptive technology model, but it is at least 5-10 years out***
- **Content Delivery Networks (CDN)**
  - Essentially storage in the network
  - In today's deployments (e.g. Akamai), the model revolves around small data sets that are typically short-lived (e.g. hot list)
  - ***No one has tried CDNs with BIG data (anyone here interested to try?)***
- **Data Transmission Protocols**
  - Congestion control mechanisms of "conventional" TCP stacks cannot keep up with large bandwidth pipes (e.g. 40G, 100G)
  - ***Alternatives, such as InfiniBand and RoCE require bandwidth guarantees to function optimally***

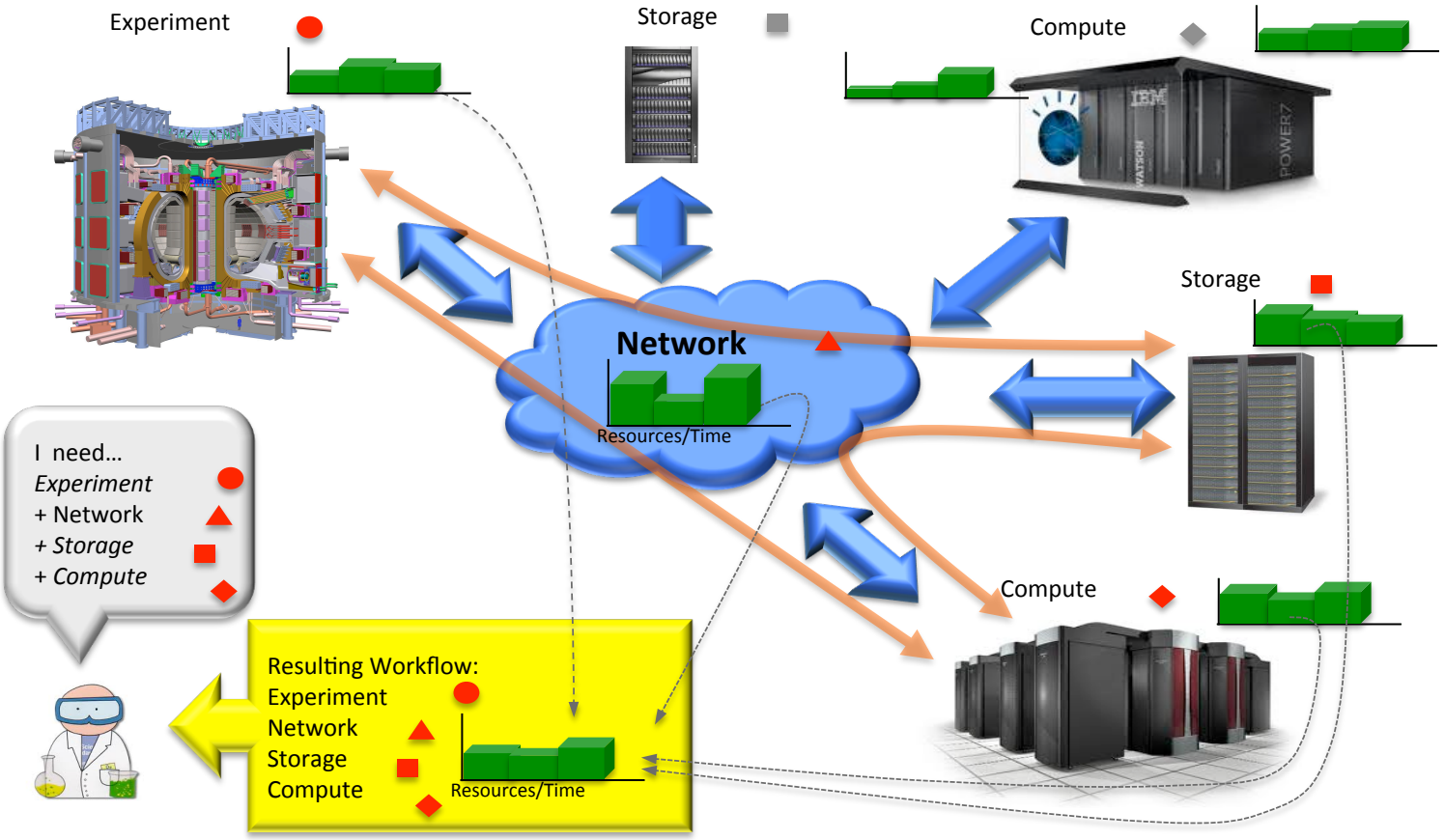
Network Capability Highlights (2/2)

- **Resource Scheduling**
  - Scheduling of experiments, compute, and storage resources is common place
  - Networks services are moving beyond best-effort and are offering scheduling capabilities (e.g. OSCARS)
  - ***Co-scheduling of ALL resources (e.g. experiment, compute, storage, network) is necessary to make the workflow run smoothly!***
- **Advance Resource Computation**
  - This is a non-trivial task, especially for complex workflows
  - Exchange of resource information (e.g. manifest) is necessary to determine co-availability
  - ***Negotiation and/or "What if" functions must be developed to help with planning and reduce rejection rate*** *(e.g. ARCHSTONE Research Project)*
- **Multi-Layer Provisioning**
  - Can provide better network transport determinism by eliminating unnecessary higher layer transport devices from the traffic path
  - ***Detailed information of the network's capability is necessary to determine the appropriate layer and adaptation/de-adaptation points for the traffic path***

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



Future Work

Future Work

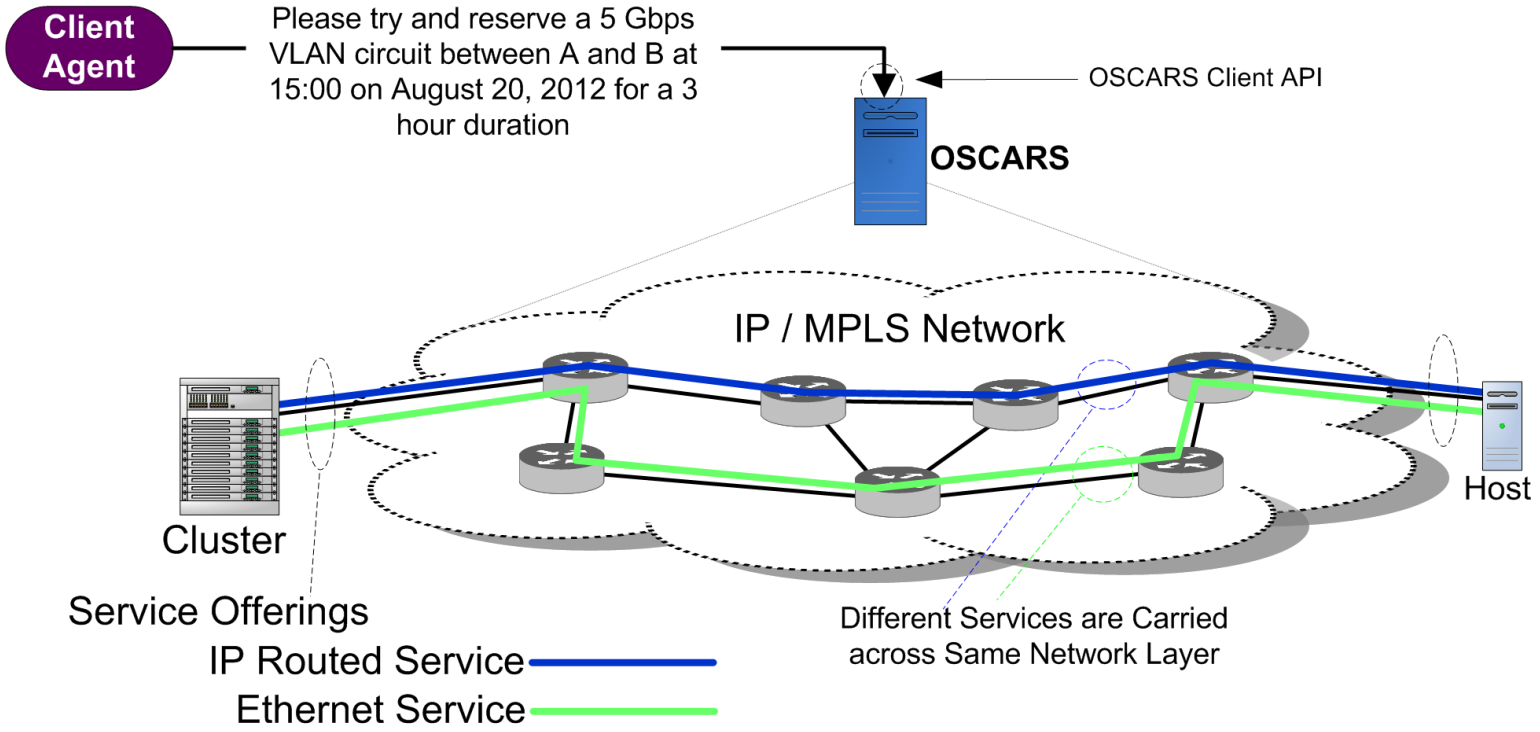
- **Expansion of Services**
  - Develop of new Atomic Services
  - Investigate algorithms for better topology computation (path finding)
  - Leverage SDN trends to enhance Intelligent Network Services
- **Extension of Service Models**
  - Extend service models to include new technologies (e.g. OpenFlow)
  - Push for adoption of protocols into Standards bodies (e.g. OGF)
- **Integration with Applications**
  - Design/build appropriate interfaces and atomic services for integration with application workflows

Thank-you

Extras

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



- **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

Cross-Layer Constrained Search Solution

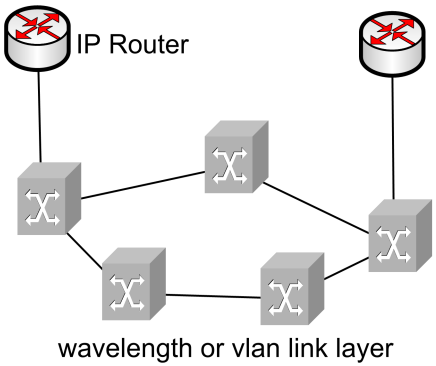
- **Applying full TE constraints when search procedure proceeds**
  - Search procedure can be based any modified SPF
  - Largely expanded search space compared to simple SPF
  - May or may not be exhaustive as some search branches can be trimmed
- **A Constrained Breadth First Search (C-BSF) implementation**
  - 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.

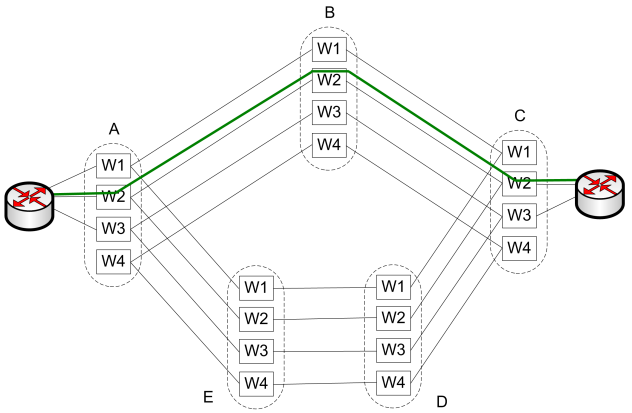
Graph Transformation Solution

- **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.

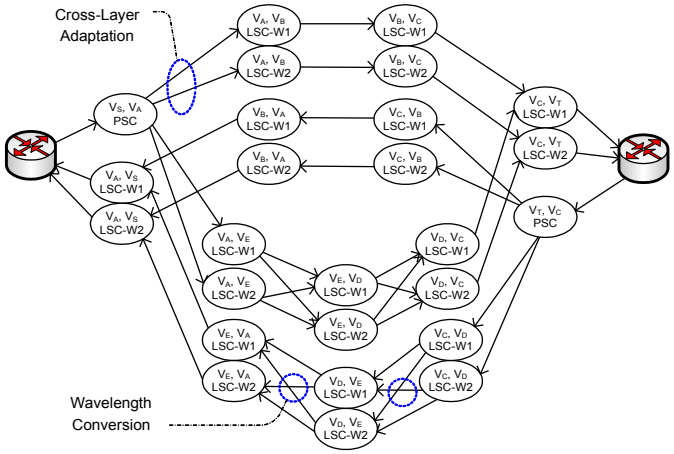




Example: label-layer graph transformation for a 7-node, 4-wavelength IP-over-WDM network.

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.**