Intelligent Network Services

Chin Guok

Network Engineering Group

ATLAS Distributed Computing Tier-1/Tier-2/Tier-3 Jamboree

CERN

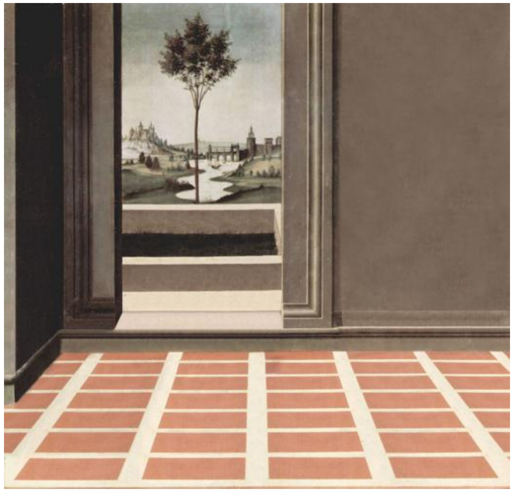
Dec 11, 2012

Networks by themselves are not interesting

abandoned version of fra angelico‘s ‘the annunciation’

abandoned version of leonardo da vinci's 'the last supper', 1495-1498

abandoned version of jacques-louis david‘s ‘oath of the horatii’



abandoned version of sandro botticelli‘s ‘annunciation’

abandoned version of andrea mantegna‘s ‘oculus in the camera degli sposi’

Bence Hajdu's Abandoned Paintings

However when integrated with scientific workflows…

original of fra angelico's 'the annunciation', 1450

original of leonardo da vinci's 'the last supper', 1495-1498

original of jacques-louis david's 'oath of the horatii', 1784

original of sandro botticelli's 'annunciation', 1489-1490

original of andrea mantegna's 'oculus in the camera degli sposi', 1473

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)

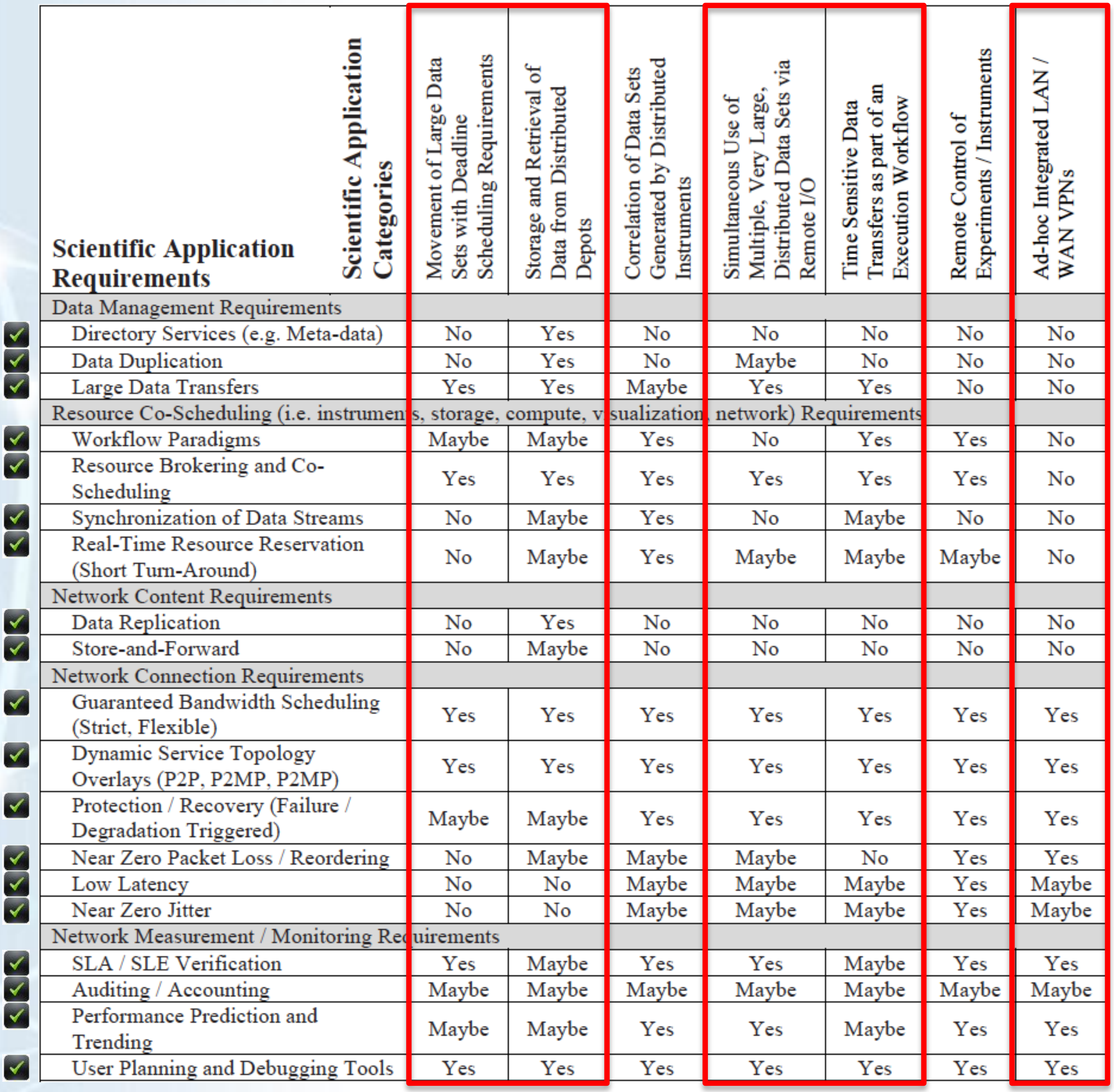
What Are Some Workflow Drivers (3/4)

- **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

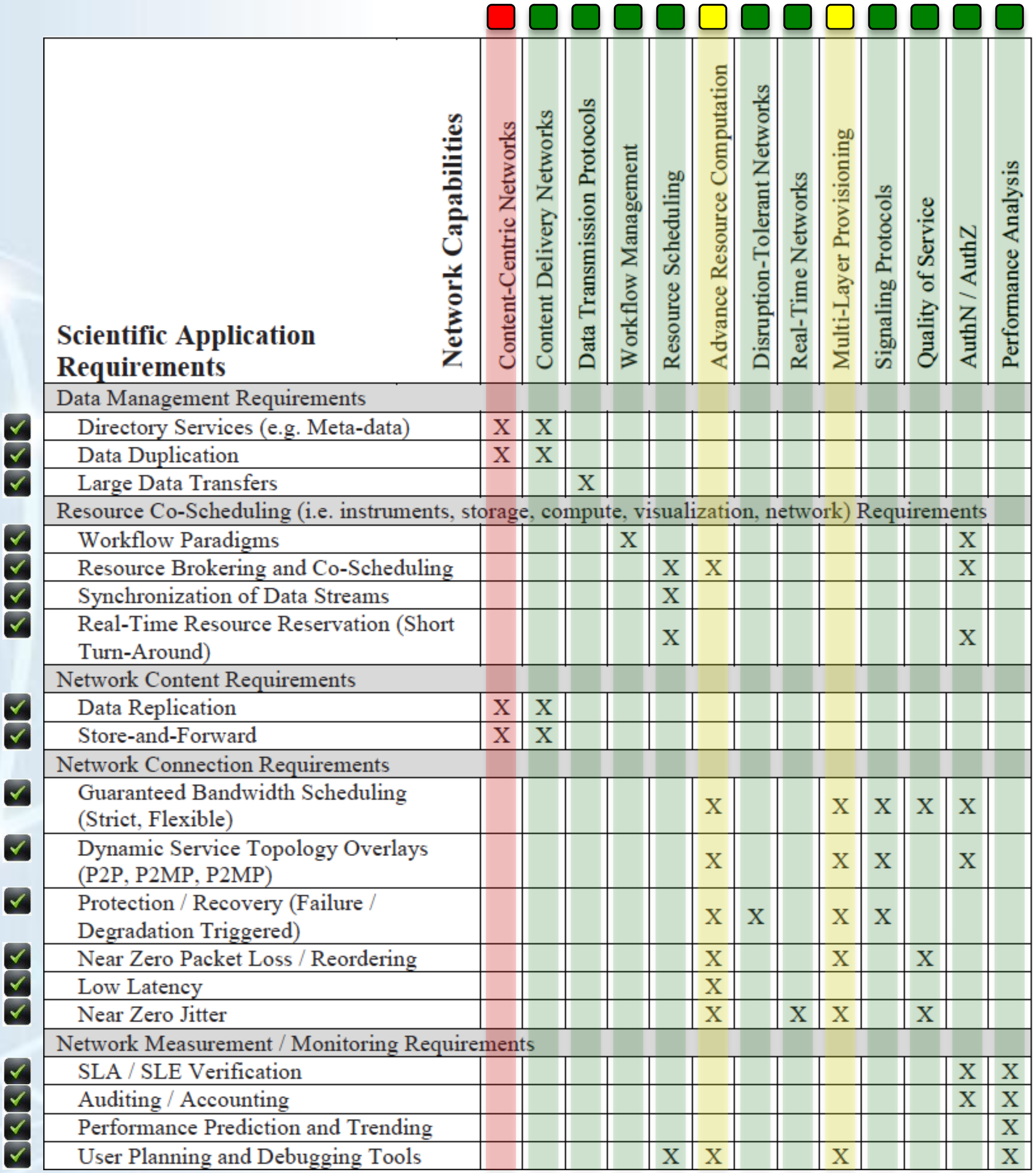


	Scientific Application Requirements Scientific Application Categories	Movement of Large Data Sets with Deadline Scheduling Requirements	Storage and Retrieval of Data from Distributed Depots	Correlation of Data Sets Generated by Distributed Instruments	Simultaneous Use of Multiple, Very Large, Distributed 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

ESnet

🟥 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***

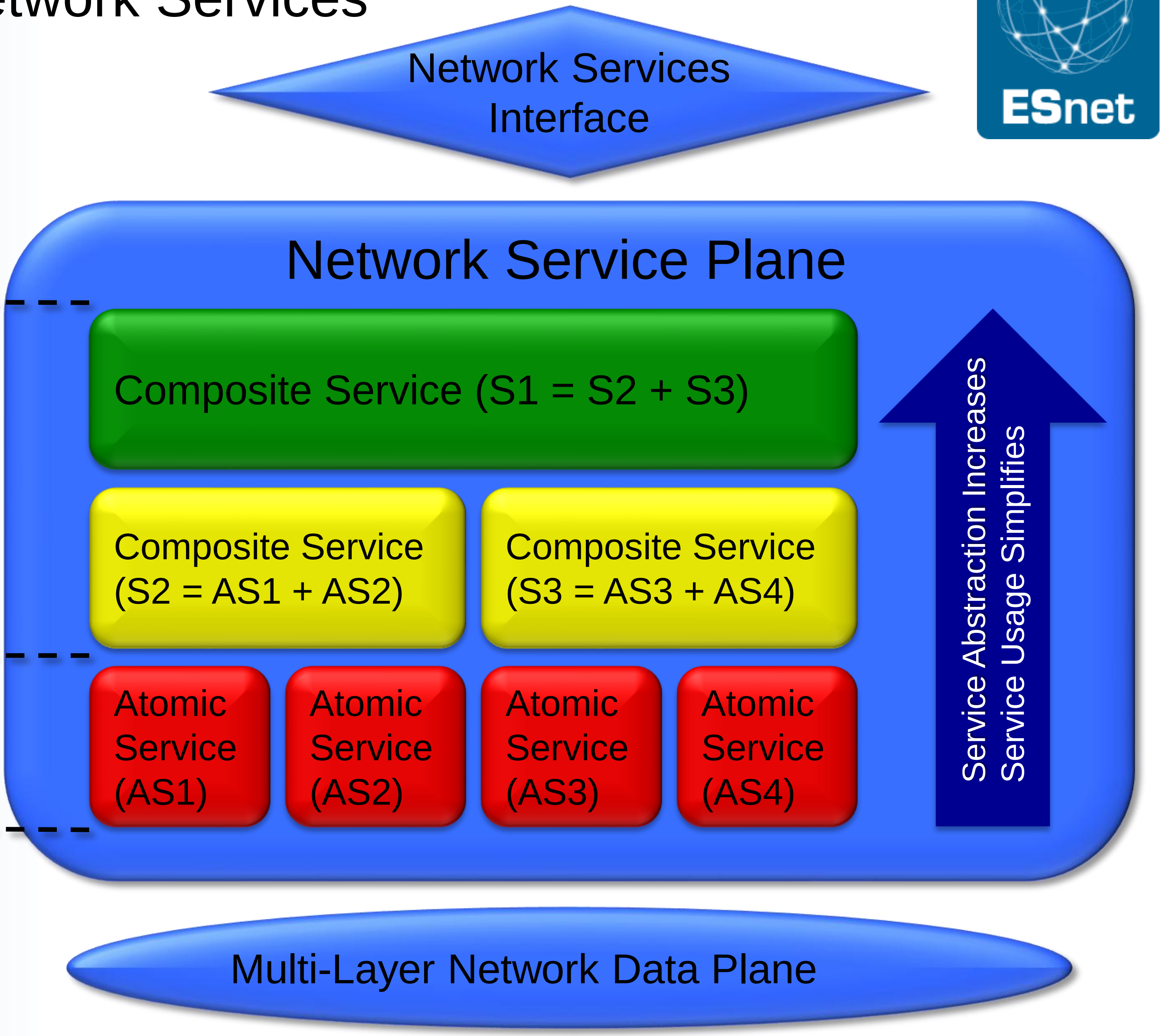
Network Capability Highlights (2/2)

- **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***

Building Network Capabilities using Atomic and Composite Network Services

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

Atomic services used as building blocks for composite services



Network Services Interface

Network Service Plane

Composite Service (S1 = S2 + S3)

Composite Service (S2 = AS1 + AS2)

Composite Service (S3 = AS3 + AS4)

Atomic Service (AS1)

Atomic Service (AS2)

Atomic Service (AS3)

Atomic Service (AS4)

Service Abstraction Increases

Service Usage Simplifies

Multi-Layer Network Data Plane

Examples of Atomic Services

Resource Discovery Service to determine resources and orientation

Resource Computation Service to determine possible resources based on multi-dimensional constraints

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

Examples of Composite Network Services

LHC: Resilient High Bandwidth Guaranteed Connection

- Connection
- Protection
- Measurement
- Monitoring

Reduced RTT Transfers: Store and Forward Connection

- Connection
- Store and Forward
- Restoration
- Monitoring

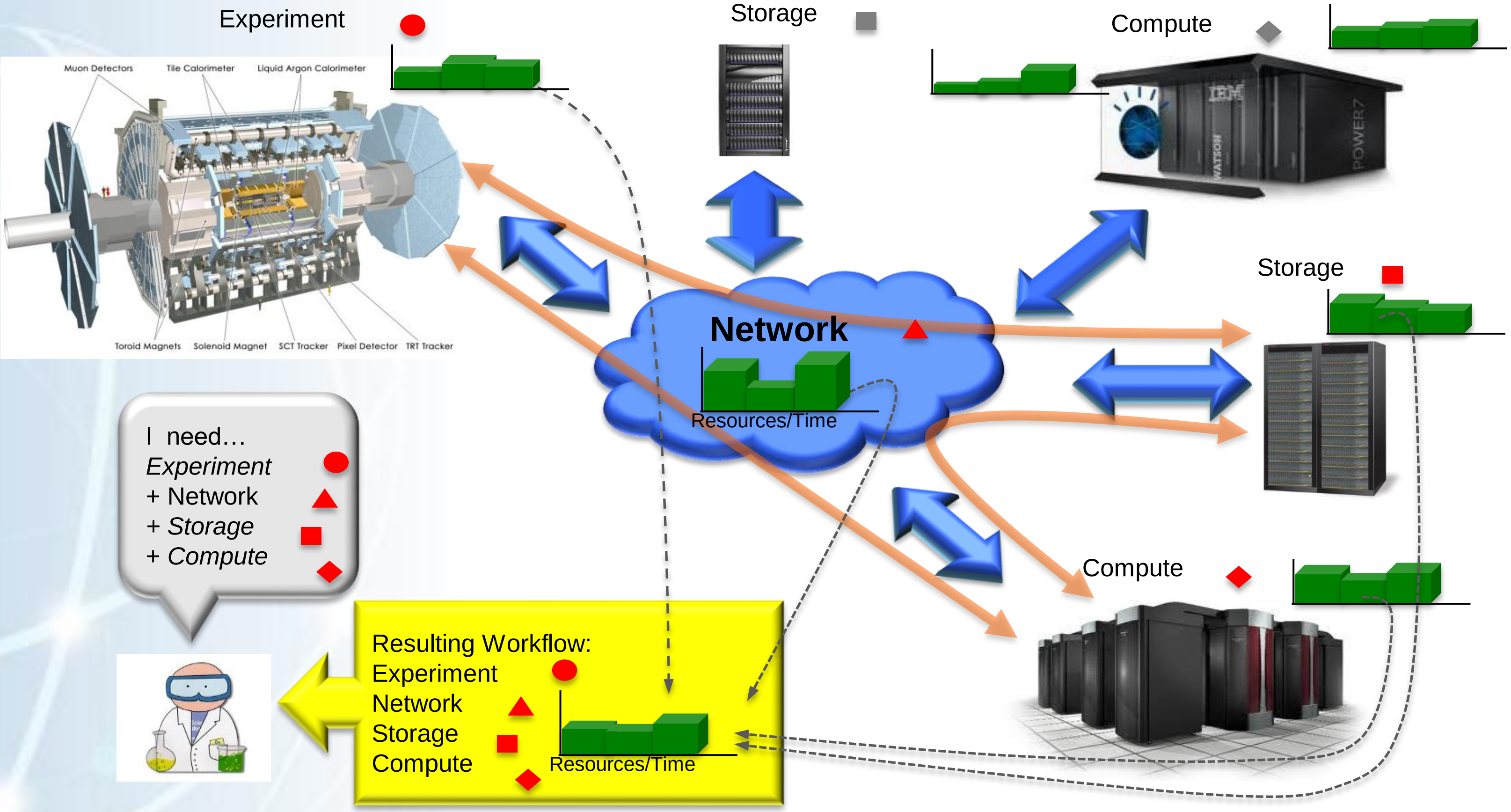
Protocol Testing: Constrained Path Connection

- Resource Discovery
- Resource Computation
- Connection
- Measurement

Compossible Network Services are nice, but what's missing?

Conclusion: Application Workflow Integration is Critical!

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



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

Resulting Workflow:
Experiment ●
Network ▲
Storage ■
Compute ◆

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

Thoughts / Questions?

"The Stone Age did not end because humans ran out of stones. It ended because it was time for a re-think about how we live"

- William McDonough, architect