2014 ALL-HANDS MEETING

EXECUTIVE DIRECTOR'S WELCOME AND OPENING STATEMENTS
TRIPTI SINHA

SEPTEMBER 25, 2014

MID-ATLANTIC CROSSROADS

Welcome to our participants, partners and guests!

- Higher Education
- Federal Labs and Agencies
- Non-Profit Agencies
- Corporations
- Partners
- Guests

Introductions

MID-ATLANTIC CROSSROADS

All Hands Meeting
September 25, 2014
College Park Marriott Hotel & Conference Center
3501 University Boulevard, East
Hyattsville, Maryland 20783
2nd Floor – Room 2110

9:15am	**Breakfast**	
10:00am	**Executive Director's Welcome and Opening Statements**	Tripti Sinha *Executive Director, MAX*
10:10am	**Keynote Address**	Eric Denna *CIO & VP of Information Technology* *University of Maryland*
10:30am	**Executive Director's Address**	Tripti Sinha
11:00am	**Refreshment Break**	
11:15am	**Innovation and Advanced Services**	Jarda Flidr *Director of Services, MAX*
11:45am	**MAX – BYTEGRID Partnership**	Tripti Sinha Don Goodwin *Executive Vice President, BYTEGRID*
12:15pm	**Lunch**	
1:15pm	**Sponsored Research Projects**	Tom Lehman *Director of Research, MAX*
2:00pm	**MAX Innovation Sandbox: Student Spotlight**	Christian Johnson
2:15pm	**Refreshment Break**	
2:30pm	**Participants Forum**	Tripti Sinha to facilitate
3:30pm	**Closing Remarks**	Tripti Sinha

Eric Denna

CIO and Vice President of Information Technology

University of Maryland

2014 ALL-HANDS MEETING

EXECUTIVE DIRECTOR'S ADDRESS

TRIPTI SINHA

SEPTEMBER 25, 2014

Since we last met - MAX Focused on Thematic Activities

Network Refresh	• Upgrading the MAX 100G footprint
New service pricing model	• Implementing MAX's new pricing model on July 1, 2014
Architecting a Cyberplatform	• Solving complex problems with the integration of storage, compute and networking
SDN Strategy	• Deeper focus on SDN and creating MAX's SDN roadmap
Strategic Partnerships	• Establishing strategic and synergistic partnerships

MOTIVATORS

Yesterday

Today

Tomorrow?

- Regional Cooperation
- Bandwidth
- R&E Networking

- Enable domain sciences
- Innovate
- Integrate innovations

Operative Words

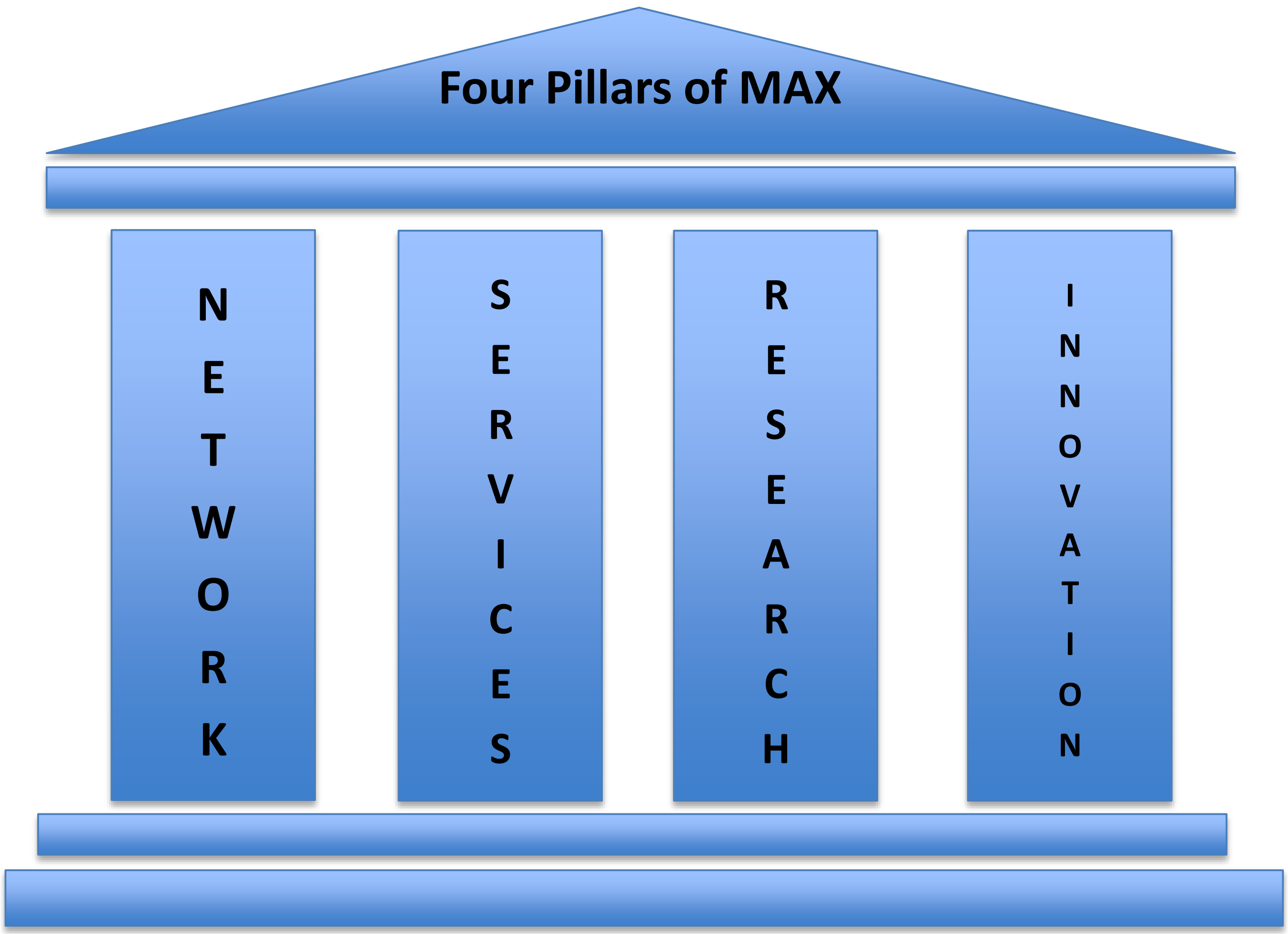
Advanced Regional Internetworking for Higher Education and Research

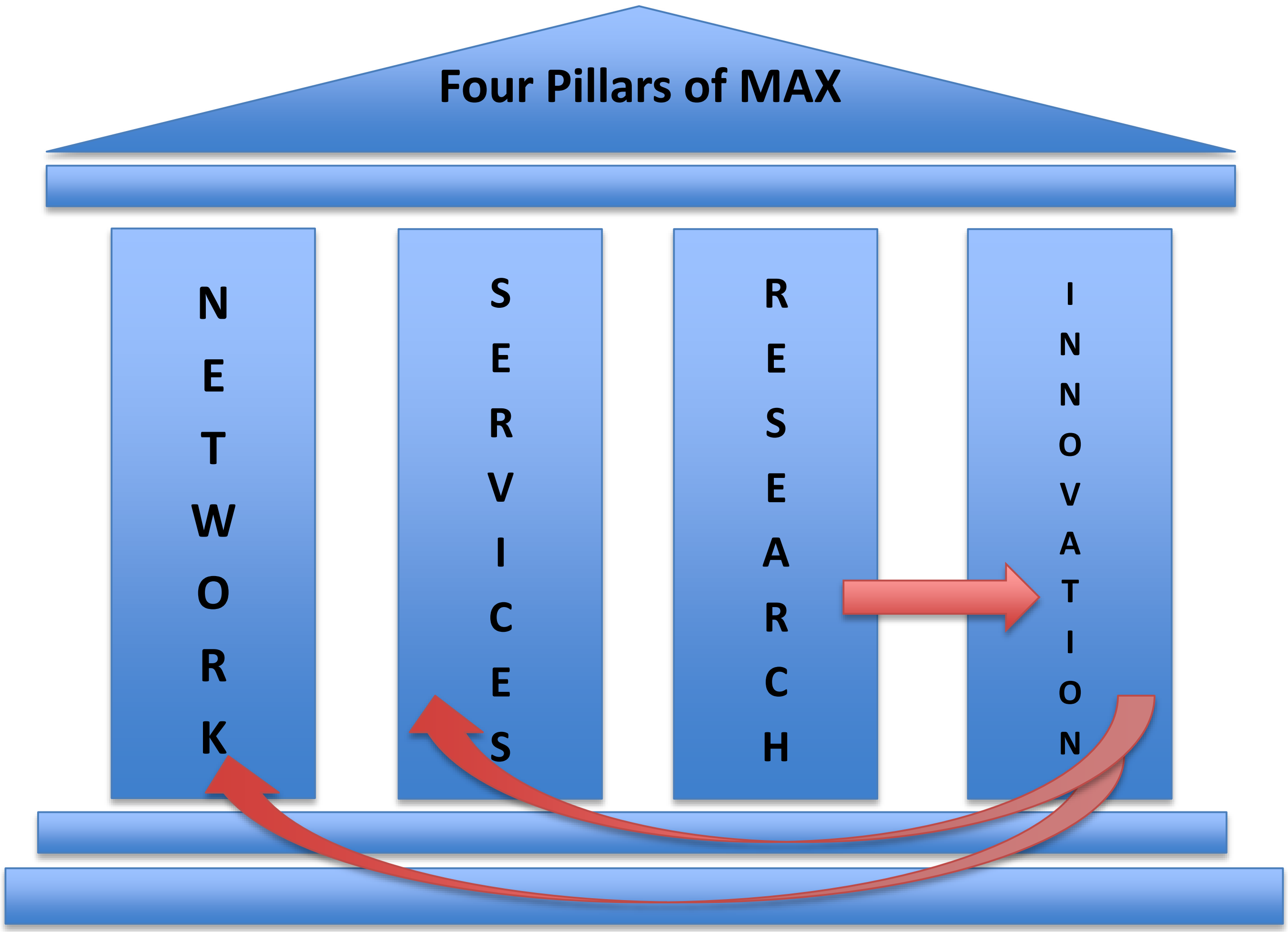
Applied Cyber Innovation for Higher Education and Research

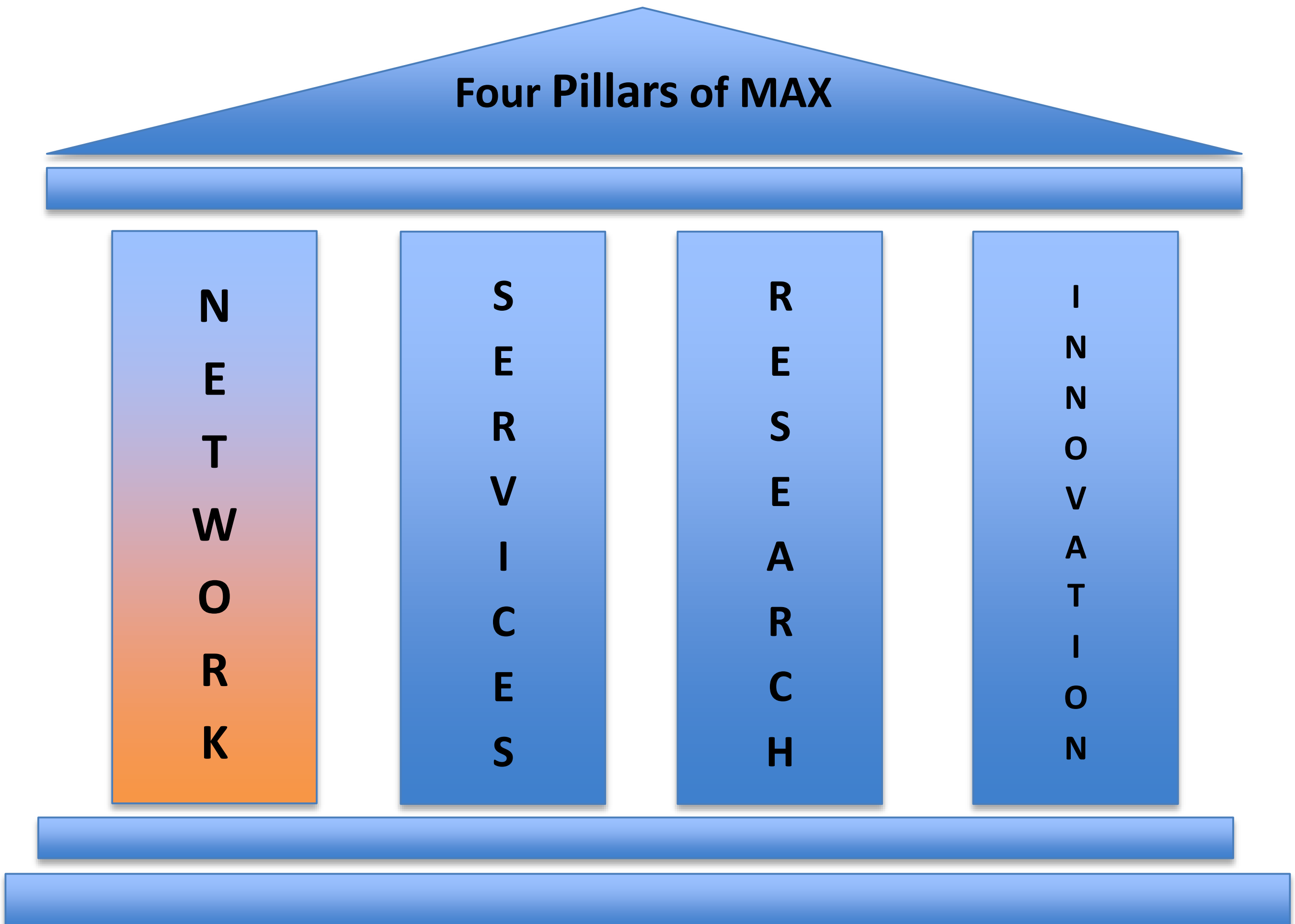
Today's world is complex!

Data
Deluge!

Big Science
Domains

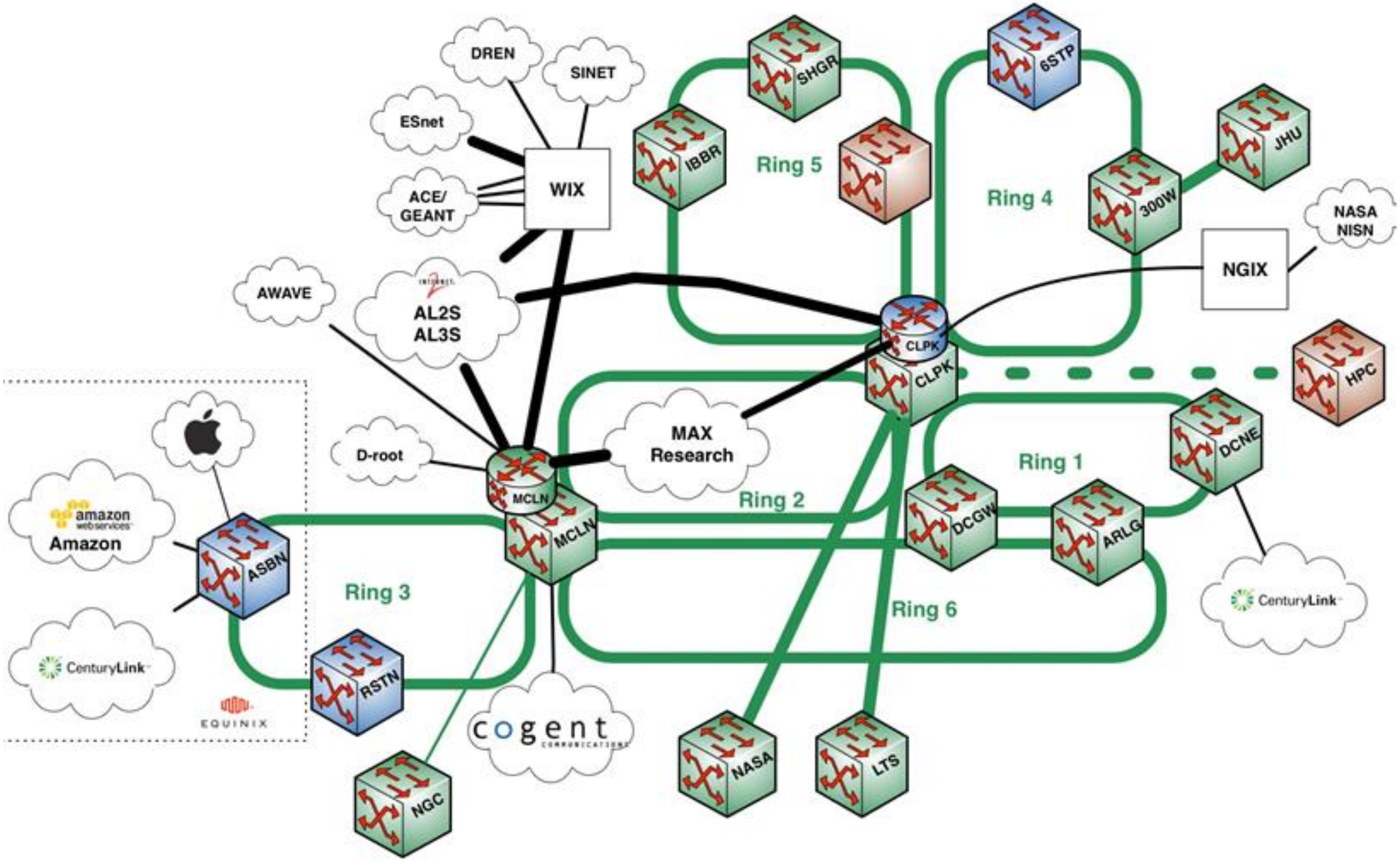

Four Pillars of MAX
NETWORK
SERVICES
RESEARCH
INNOVATION


Four Pillars of MAX
NETWORK
SERVICES
RESEARCH
INNOVATION

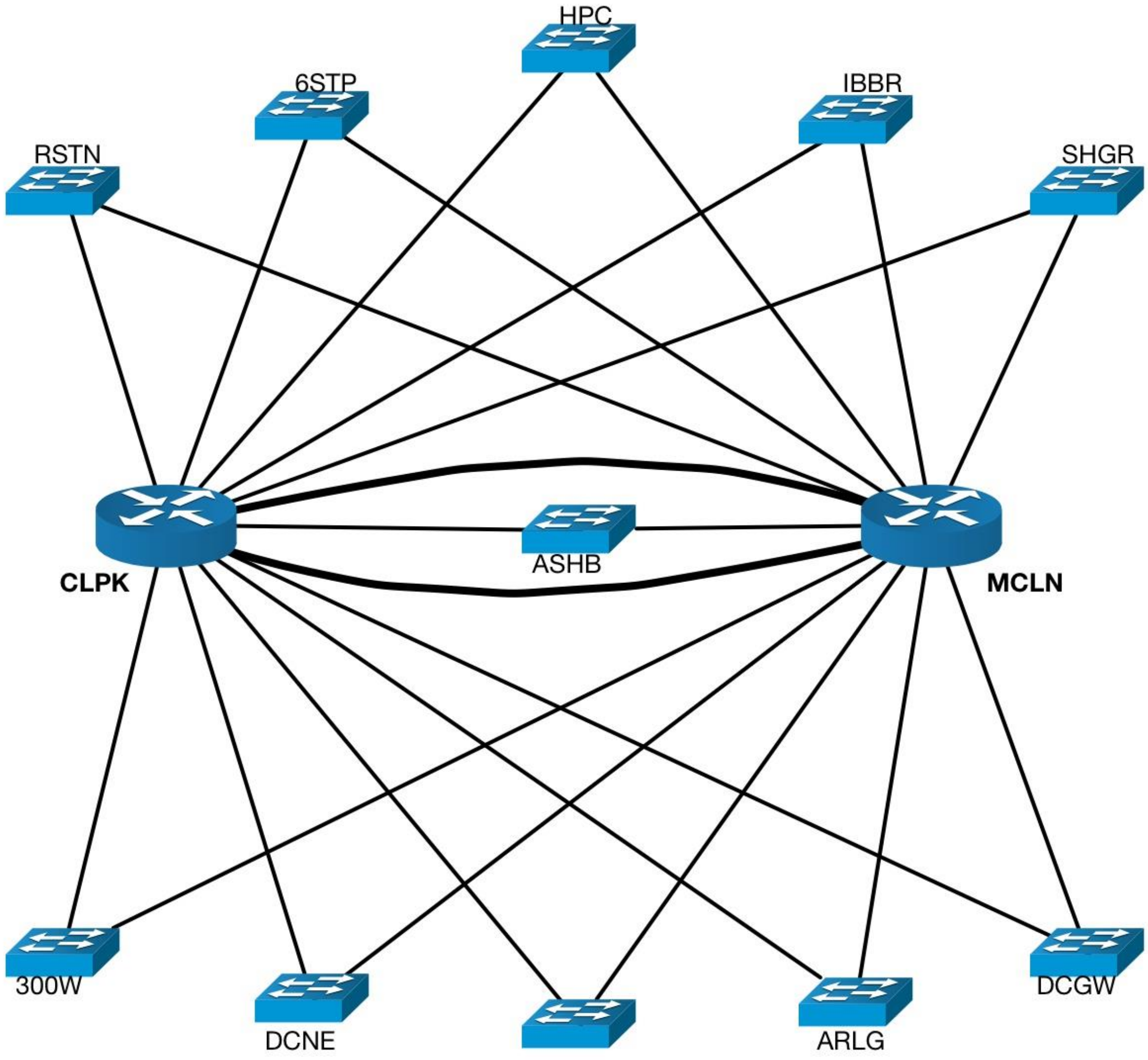

Four Pillars of MAX
NETWORK
SERVICES
RESEARCH
INNOVATION

Network Points of Presence

MAX Network Refresh



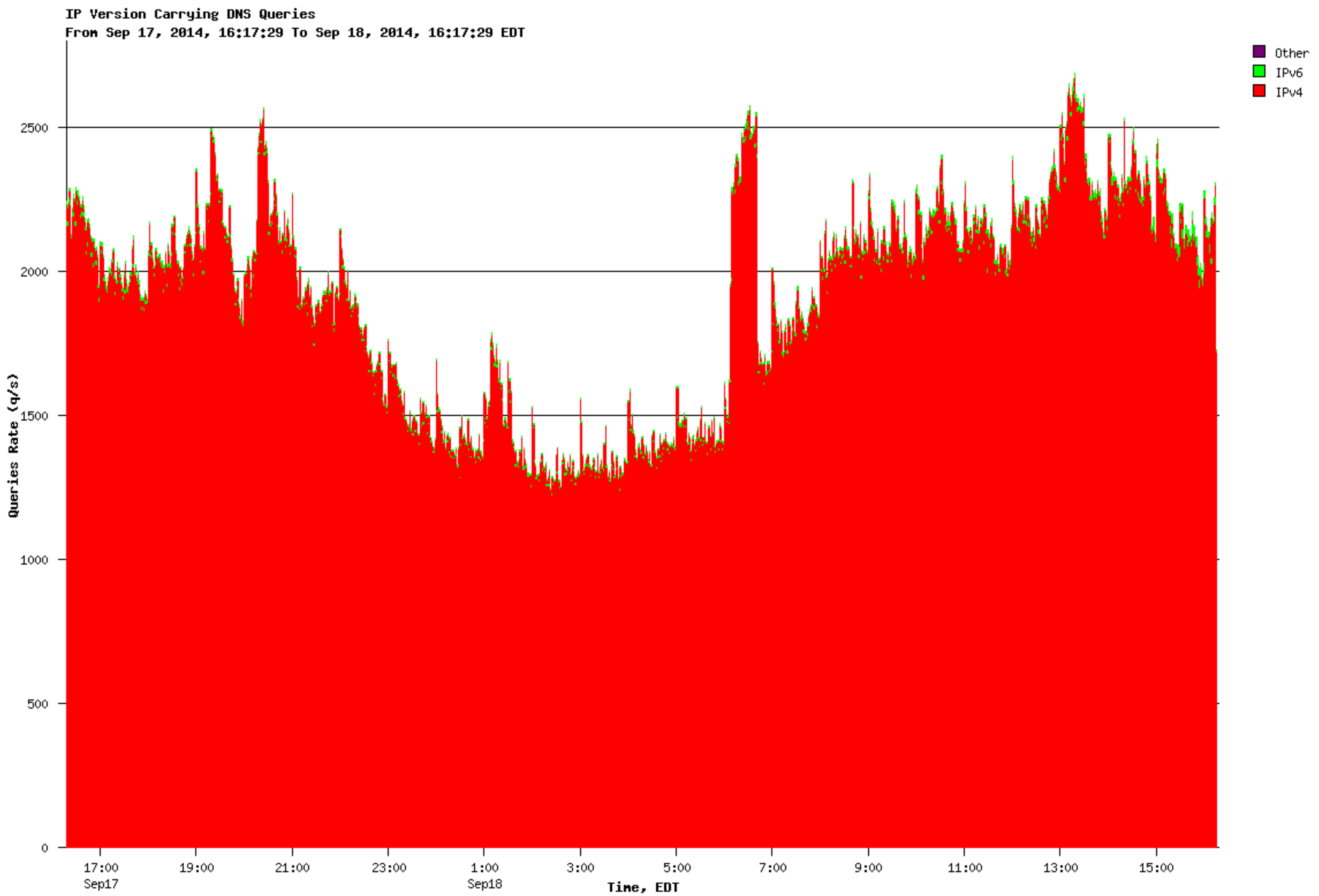
MAX Network – layer 2 view



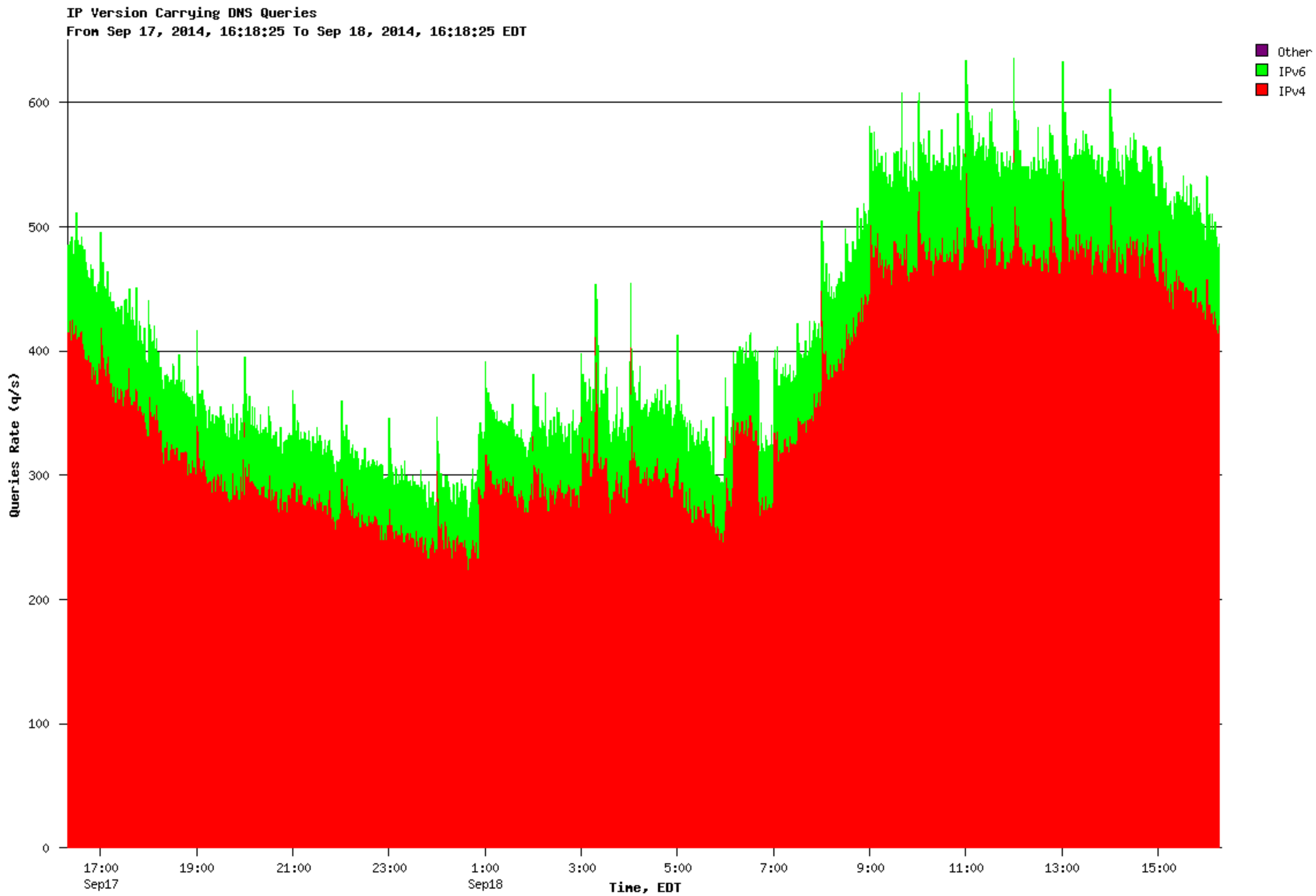
d.root.servers.net

- In 1988, the University of Maryland was selected to serve the root of the Domain Name System by operating D-root.
- Root servers are the foundation of global DNS services.
- DNS is a hierarchical lookup system.
- 12 distinct operators operate 13 root services (A thru M).
- Root servers are anycasted.
- D-root currently has over 58 sites, 94 instances in 34 countries
- One instance of d-root lives in the heart of the MAX network.

Queries by Node (DNSMON-MCVA-Ext-IPtype)

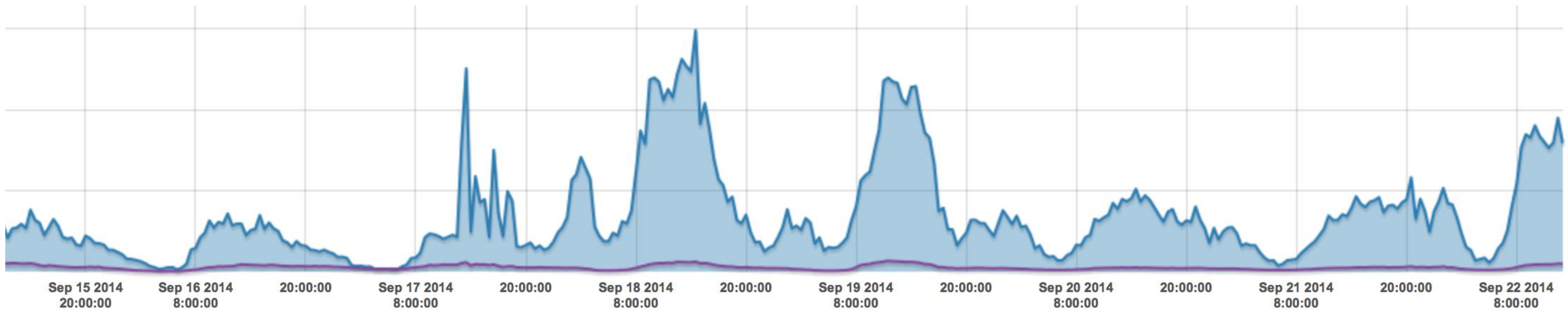


Queries by Node (DNSMON-MCVA-MAX-IPtype)

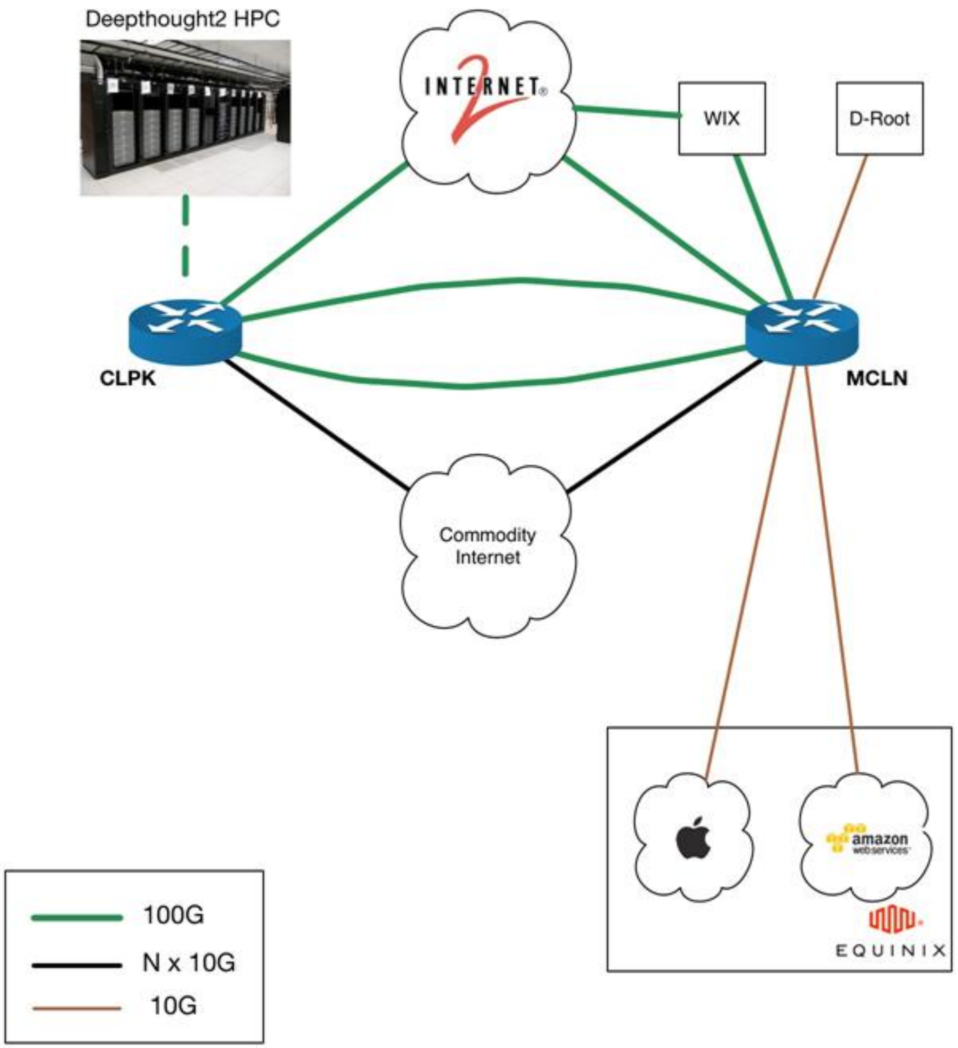


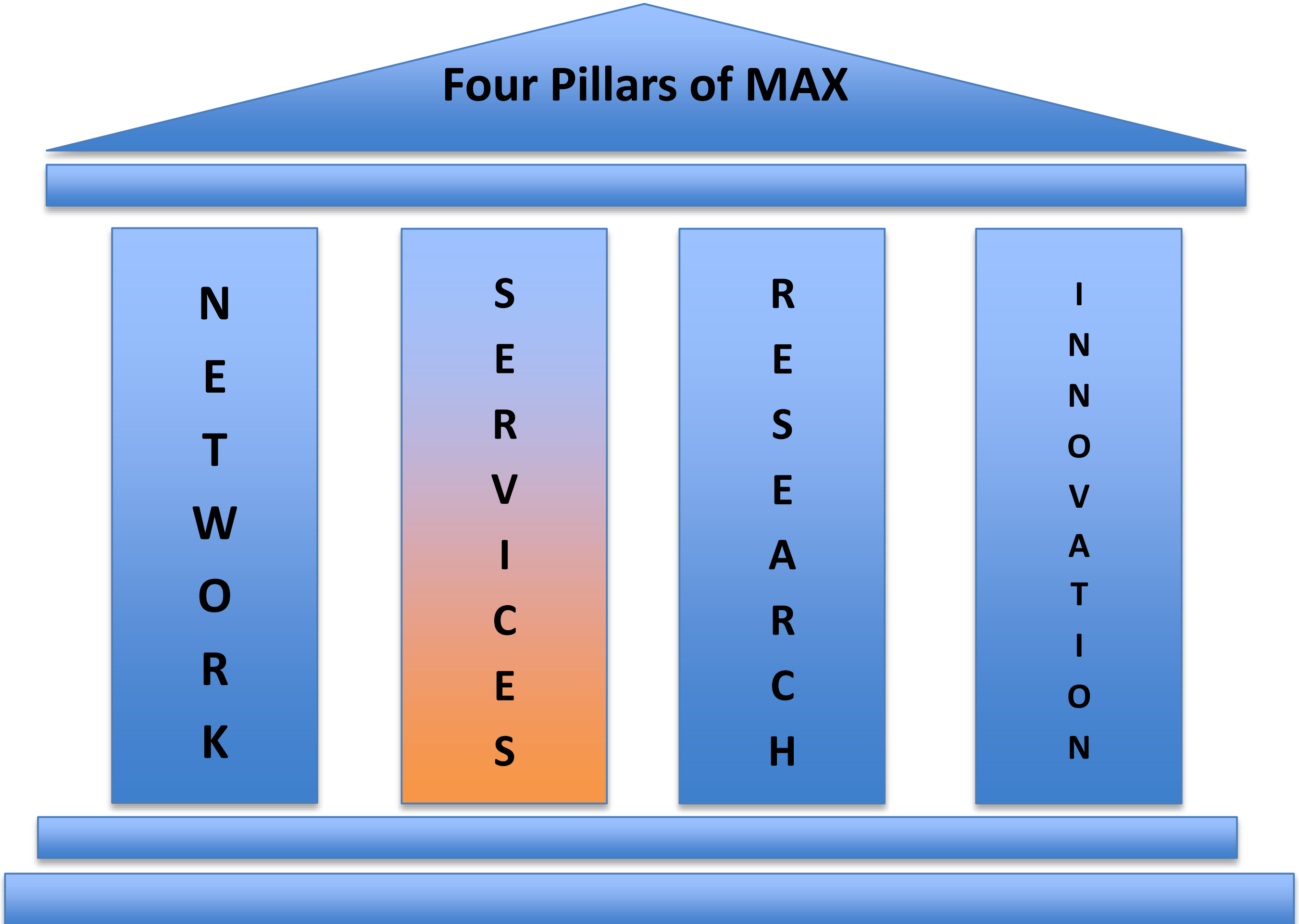
MAX peers with Apple CDN at Equinix

- Peering is on a 10G port
- MAX network capable of handling high demand situations (like IOS 8 rollout)



MAX Network Resources




Four Pillars of MAX
NETWORK
SERVICES
RESEARCH
INNOVATION

Services

Principles

- Meet the needs of MAX's participant community
- Innovation and advanced services for Research and Education
- Cost recovery and sustainability based business model

Service Suite

- WDM and Ethernet Transport
- IP Routed Services
  - R&E
  - Commodity
- Advanced Services
  - AWS
  - Research Network
- Co-location Services
- Exchange Points (WIX, NGIX)

MID-ATLANTIC CROSSROADS

MAX Services & Fee Structure Implemented July 1, 2014

Mid-Atlantic Crossroads (MAX) Services
Participation Fee
MAX Participation Fee
Layer 3 – IP Routed (R&E) Service
1G
10G
100G
Layer 2 – Ethernet Transport Service
1G
10G
Layer 1 – DWDM Transport Service
10G
100G

Mid-Atlantic Crossroads (MAX) Services
IP Commodity Routes
Commercial Providers
TR-CPS
Advanced Services
MAX AWS Direct Connect
Research Network Connection
MAX Platinum Service
Access to multiple services
Other Services
Rack Colocation Space
Machine/Virtual Machine Hosting
Remote Hands

MAX Services & Fee Structure Implemented July 1, 2014

Mid-Atlantic Crossroads (MAX) Services
Washington International Exchange (WIX)
10G
100G
Next Generation Internet Exchange (NGIX)
1G-10G

HPC – Deepthought2

The University of Maryland's Cyberinfrastructure Center, a new facility to enhance the university's advanced research capabilities, opened in January 2014. The Cyberinfrastructure Center is home to Deepthought2, a new high-performance computing cluster launched in May 2014. It also offers space for colocation of departmental research computing equipment. www.it.umd.edu/CC

Cyberinfrastructure Center

Offers Research Computing Resources

UMD's new Cyberinfrastructure Center is located in approximately **9,000** square feet of leased space in the Rivertech Building at 5700 Rivertech Court in the university's M Square research park.

1,800 square feet of floor space in the data facility is dedicated to colocation of college and department research computing assets.

The colocation facility was developed with the needs of campus researchers in mind and provides environmental and physical security controls.

The new center's electrical supply offers both **UPS** and **generator** service for reliability.

A staging area is available for preparing equipment to be installed in the colocation area.

After initial set-up fees, there are **no recurring charges** associated with colocation space and power in the Cyberinfrastructure Center.

Deepthought2

Better Supports UMD Researchers

UMD paid about **$4.2 million** for **Deepthought2**, which has a processing speed of about **300 teraflops.**

The new supercomputer can complete between **250 trillion** and **300 trillion** operations per second.

Deepthought2 has a **petabyte** (1 million gigabytes) of storage as well as a very high-speed internal network.

Deepthought2 is the equivalent of **10,000** laptops working together. It has **2,000** times the storage of an average laptop and an internal network that is **50** times faster than broadband. This is the type of compute power needed to solve urgent scientific and societal problems.

Based on current rankings, Deepthought2 is expected to rank as one of the top high-performance computing clusters among U.S. universities and as one of the top 500 clusters in the world. **TOP 500** in the World

Deepthought2 Facilitates Mass Processing and Big Data Analysis:

- Studying the formation of the first galaxies
- Simulating fire and combustion for fire protection advancements
- Probing the causes of multi-drug resistance in bacteria to help develop better antibiotics
- Understanding how the universe evolved

UMD Partners

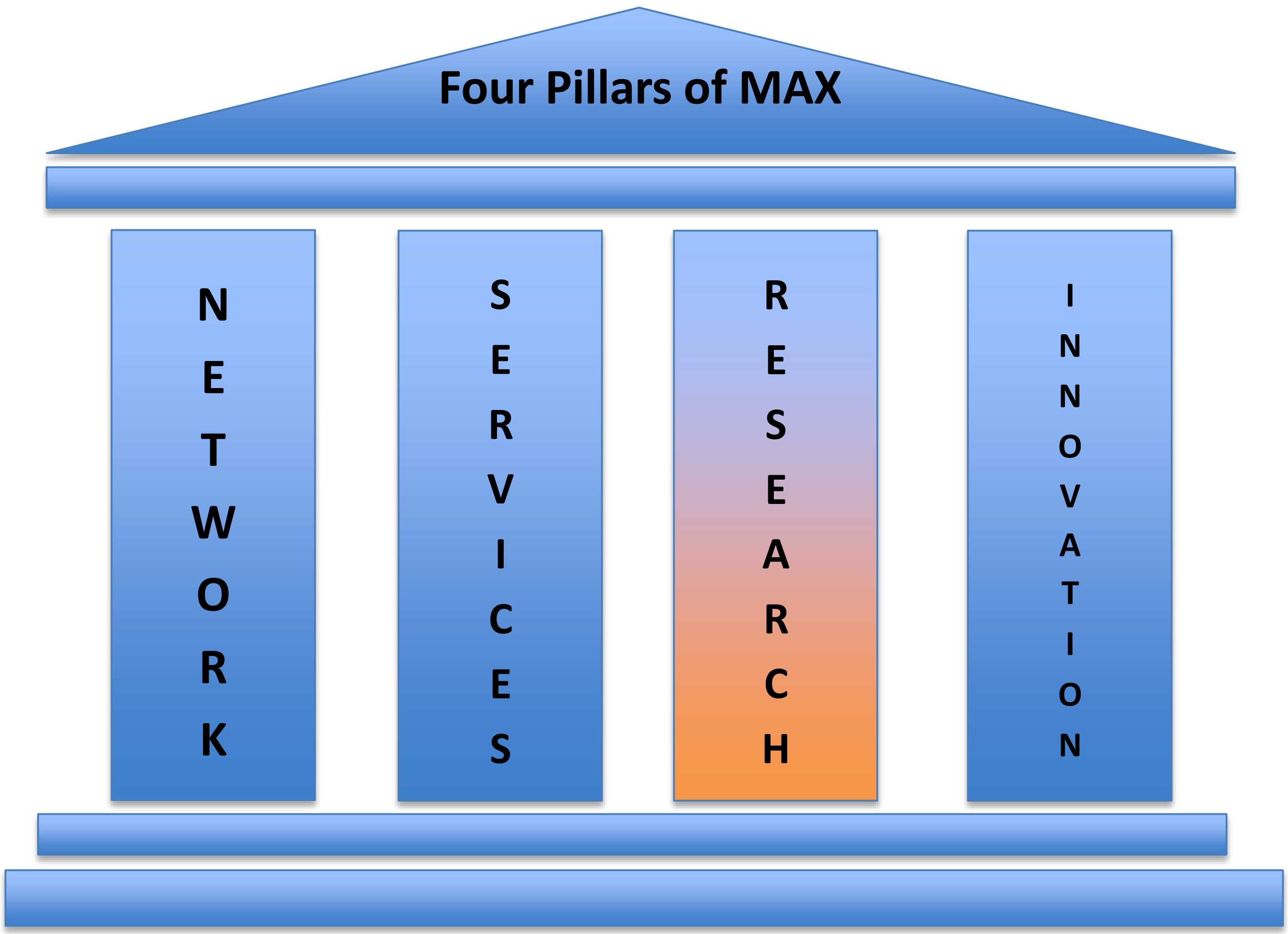
- A. James Clark School of Engineering
- College of Behavioral and Social Sciences
- College of Computer, Mathematical, and Natural Sciences
- Division of Administration and Finance
- Division of Information Technology

Expanding our research computing assets helps UMD researchers further contribute to solving major societal challenges, answering complex scientific questions, and advancing human welfare.

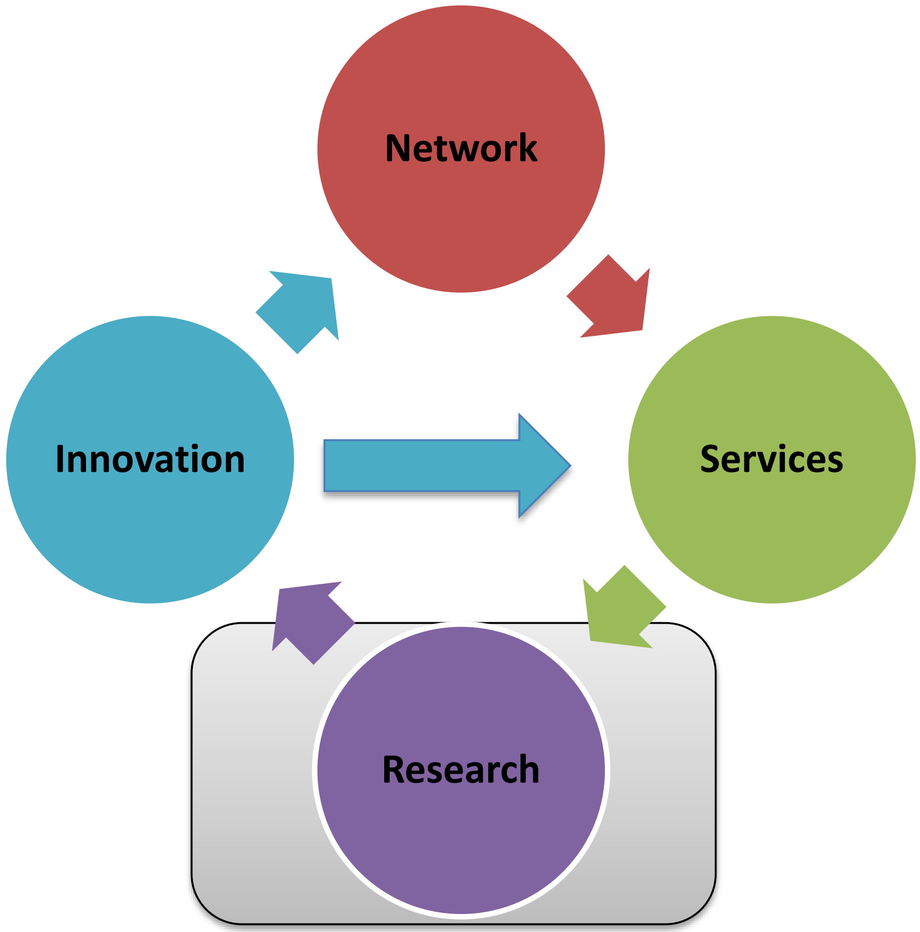
UNIVERSITY OF MARYLAND

New Services in the next year

Mid-Atlantic Crossroads (MAX) Services
HPC Offering
Data Center
Security


Four Pillars of MAX
NETWORK
SERVICES
RESEARCH
INNOVATION

The cycle of innovation and advanced services



MAX Focus on Thematic Activities

Network Refresh	• Upgrading the MAX 100G footprint
New service pricing model	• Implementing MAX's new pricing model on July 1, 2014
Architecting a Cyberplatform	• **Solving complex problems with the integration of storage, compute and networking**
SDN Strategy	• **Deeper focus on SDN and creating MAX's SDN roadmap**
Strategic Partnerships	• Establishing strategic and synergistic partnerships

MAX Sponsored Research

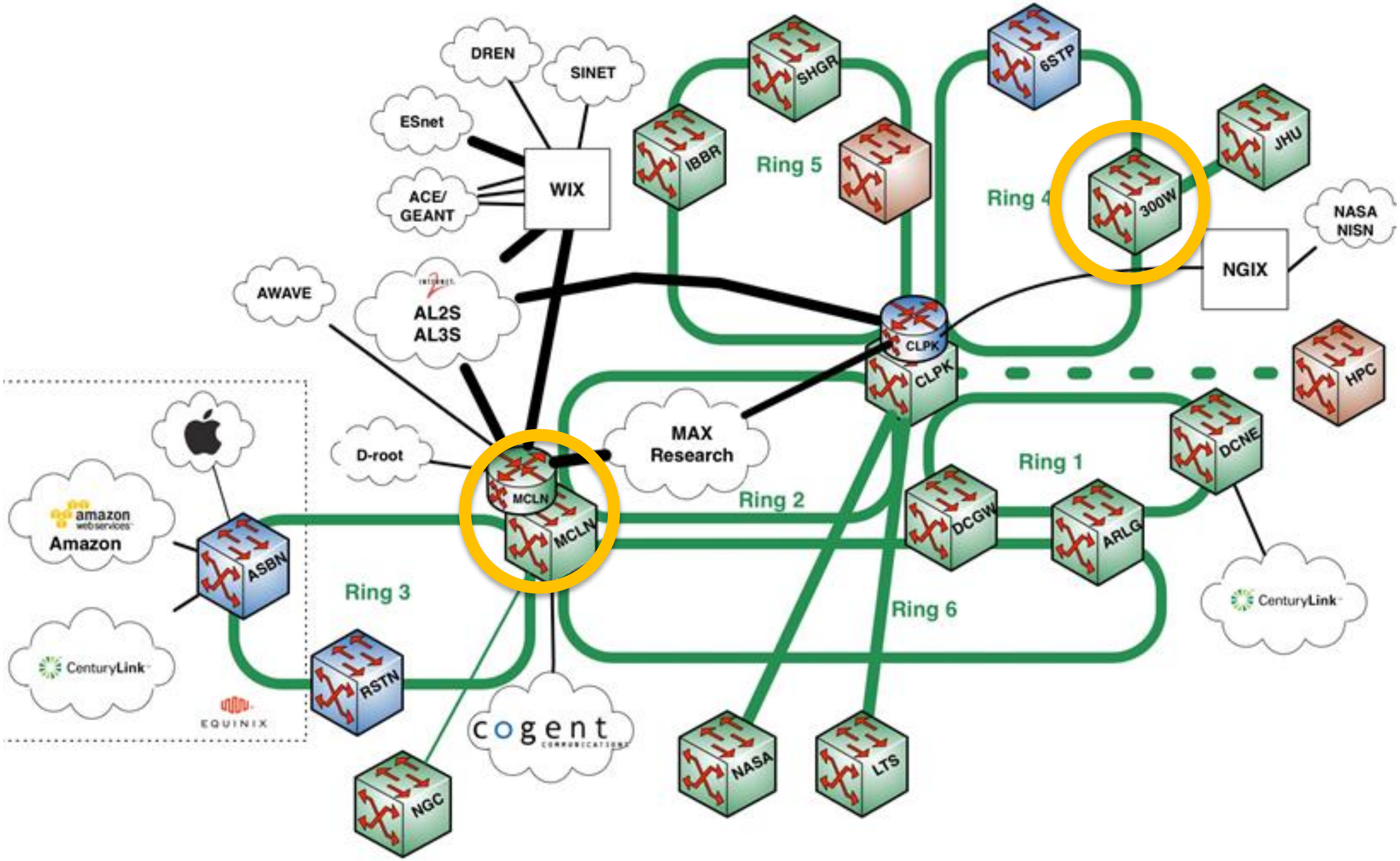
SDNX, HPCDNA,GENI,JHU100G

DoD

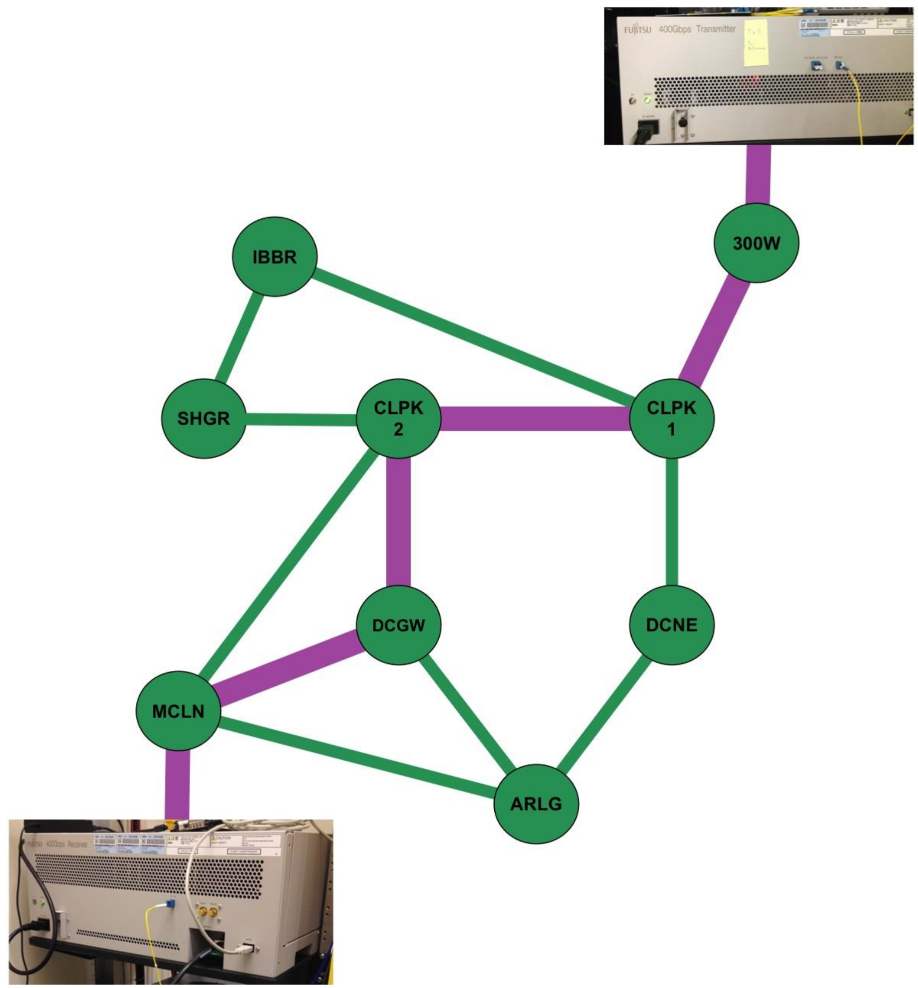
NetSurvive

RAINS

MAX-Fujitsu 400 Gbps and 800 Gbps Field Trial



MAX-Fujitsu 400 Gbps and 800 Gbps Field Trial

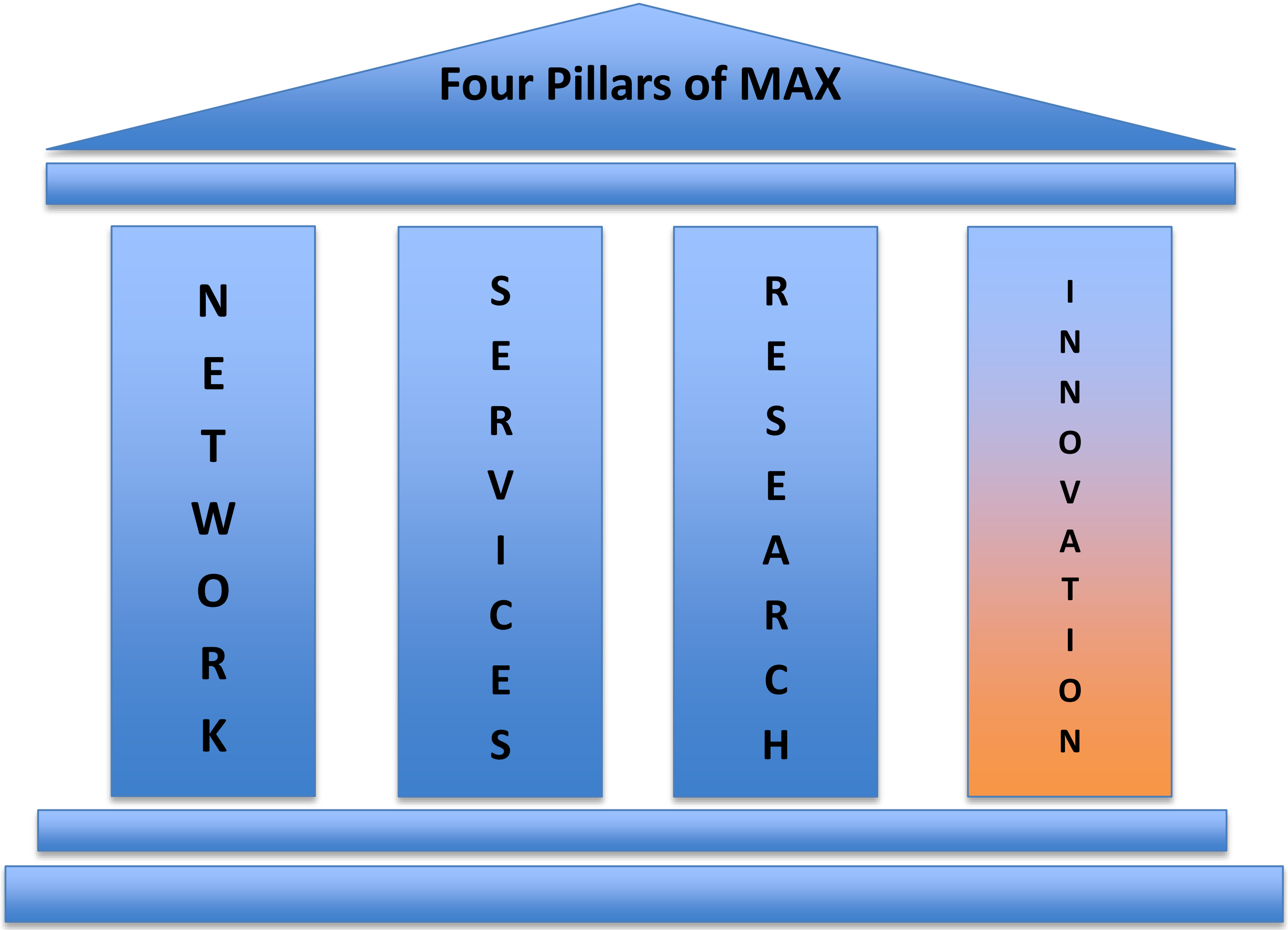


MAX-Fujitsu 400 Gbps and 800 Gbps Field Trial

Successful transmission of data at rates of 400 Gbps and 800 Gbps → **Reveals future of terabit networking capabilities**

- **Data transferred over MAX's optical network from Baltimore, MD, to McLean, VA.**
- First-ever trial demonstrating Fujitsu's super-channel capabilities on a deployed network, which allow higher speeds on the existing installed base of equipment.
- Fujitsu FLASHWAVE® 9500 Packet Optical Networking Platform (Packet ONP) transmitted data with a 25% improvement in channel spacing over conventional dense wavelength division multiplexing (DWDM) – greatly increases network utilization without requiring any physical adjustments to the MAX network infrastructure.
- The field trial demonstrated several key technical advancements which could lead to the next generation of optical transmission.
- This dramatic increase in network speed will help scientists across the mid-Atlantic minimize the limitations of geographic distance and maximize the demands of science applications in order to expedite the transmission of data.

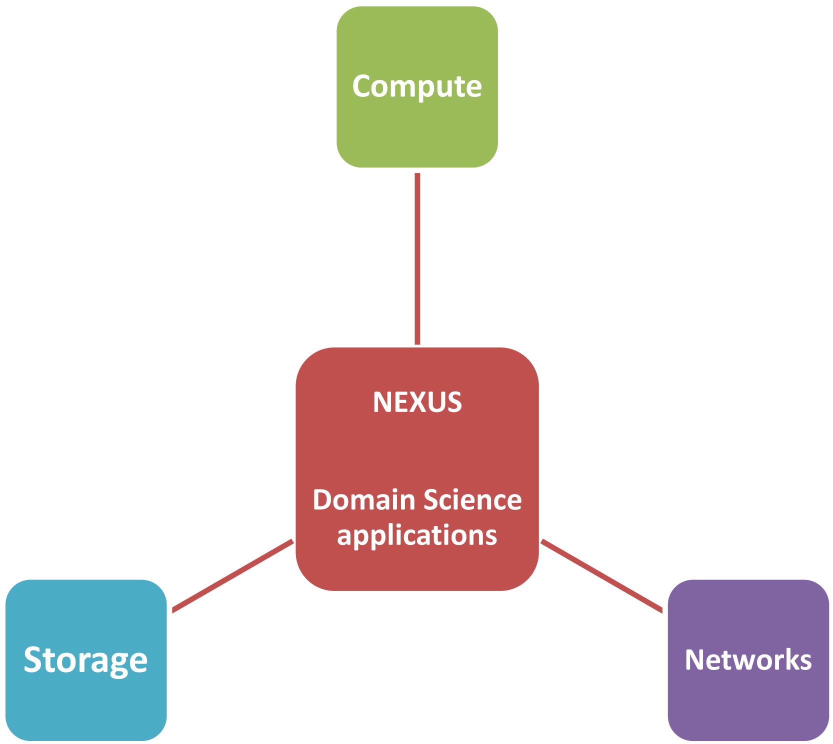
**All of these advancements enable a much higher utilization of costly fiber infrastructure and maximize the bandwidth available for demanding R&E applications!*


Four Pillars of MAX
NETWORK
SERVICES
RESEARCH
INNOVATION

Innovation

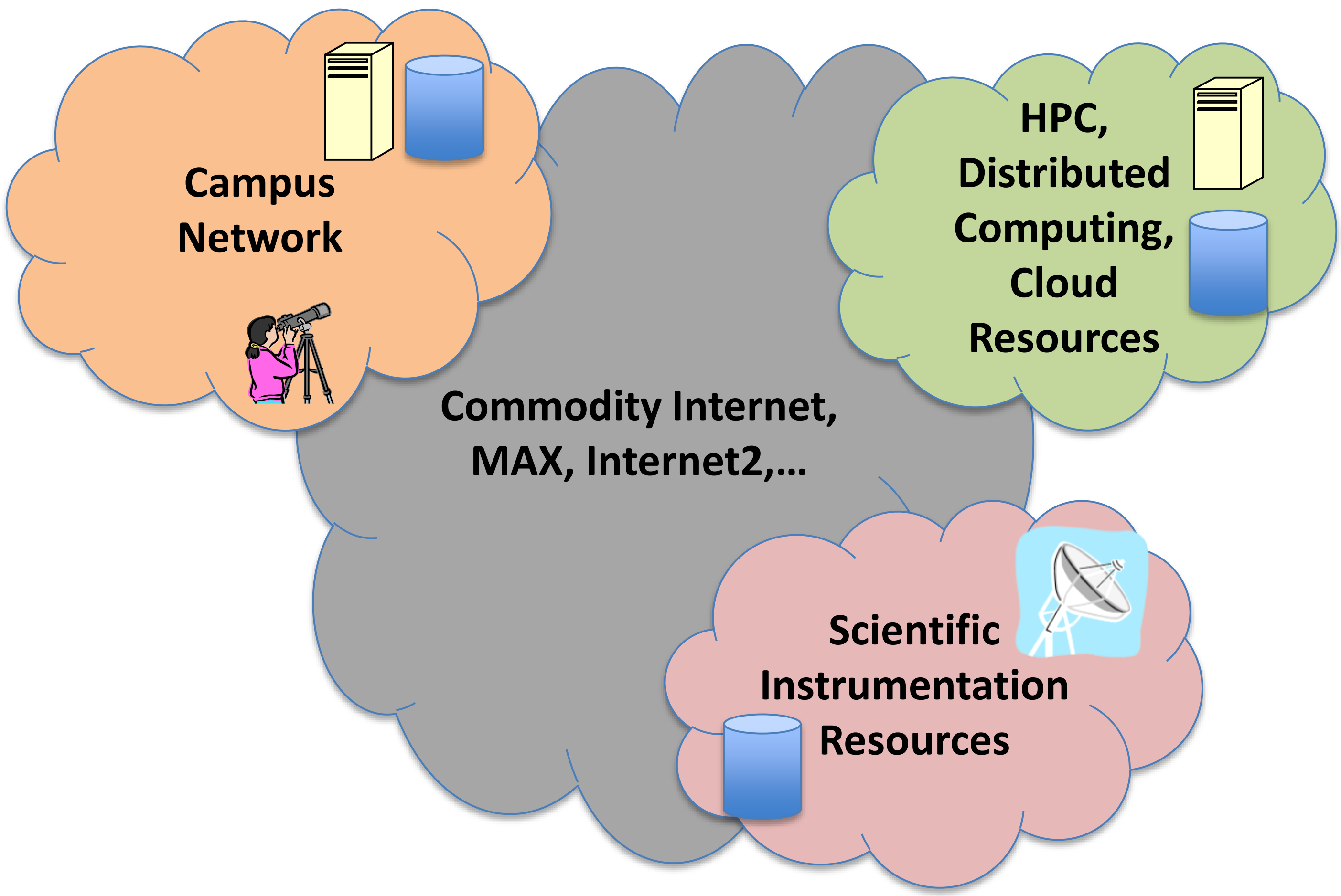
The Holy Triad

The Domain Science says, "Don't just connect me but compute me, store me and transport me."

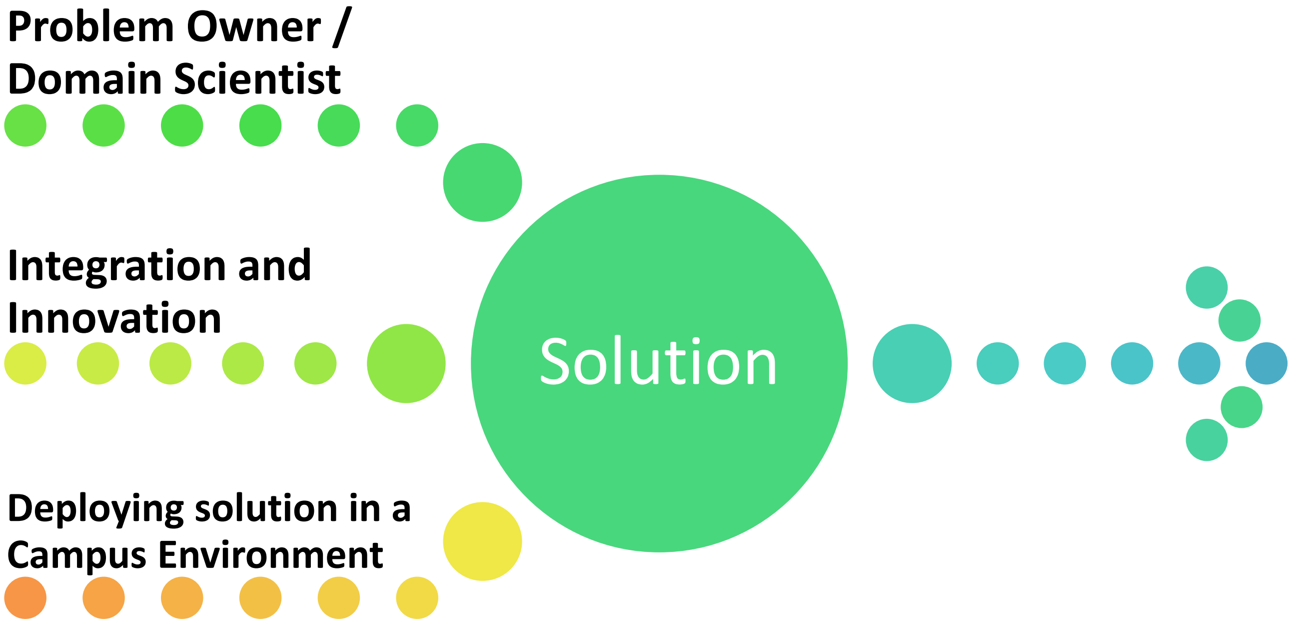


- SDNX
  - Application and SDN integration technology
  - Well engineered and optimally positioned network related service exchange point
- HPCDNA – flexible coupling of application specific data sets with high performance compute and networking

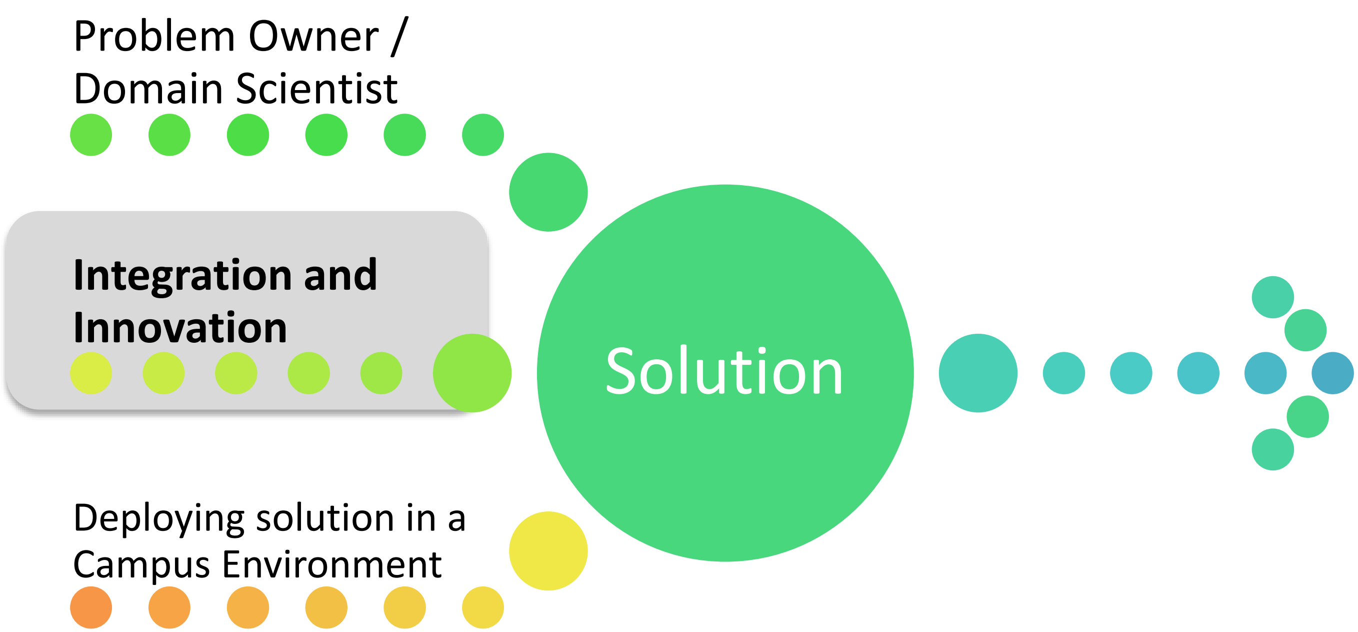
The Problem Space (of the domain scientist)



Three-Pronged Solution for the Problem Space



Integration and Innovation Prong



Close Examination of Four Use cases at Maryland

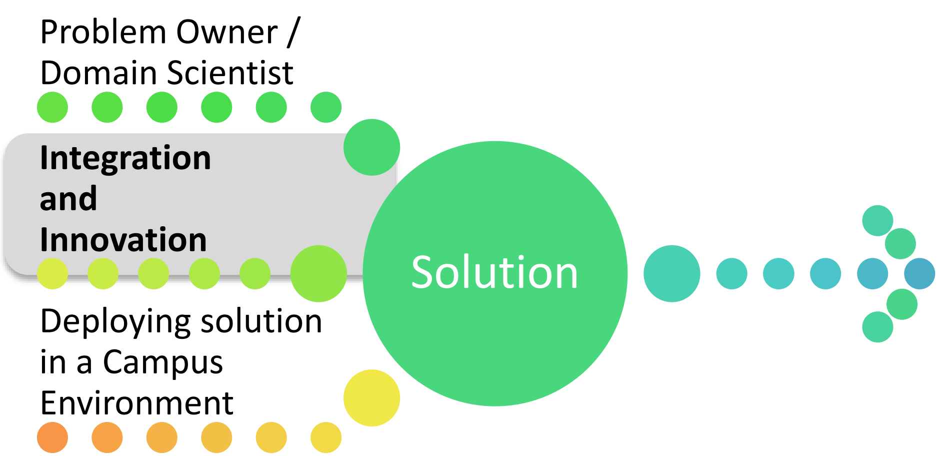
Data-Intensive Research Use Cases

- Imagery from NASA satellites
- Data from telescopes
- Scientific instruments
- Massive scale simulations

Before-and-After MSX

BEFORE	AFTER
• The data were transferred physically on 2-TB hard drives from Maui to Maryland. • By the end of the survey in March 2014, the team expected to collect hundreds of terabytes of data, which would take approximately **9 months** to download.	• The full download time is reduced to only **1.5 days**. • The team is able to **in-line process the sets in near-real time as the data flows in**, rather than downloading, storing, and processing (which alleviates the need for local infrastructure upgrades).

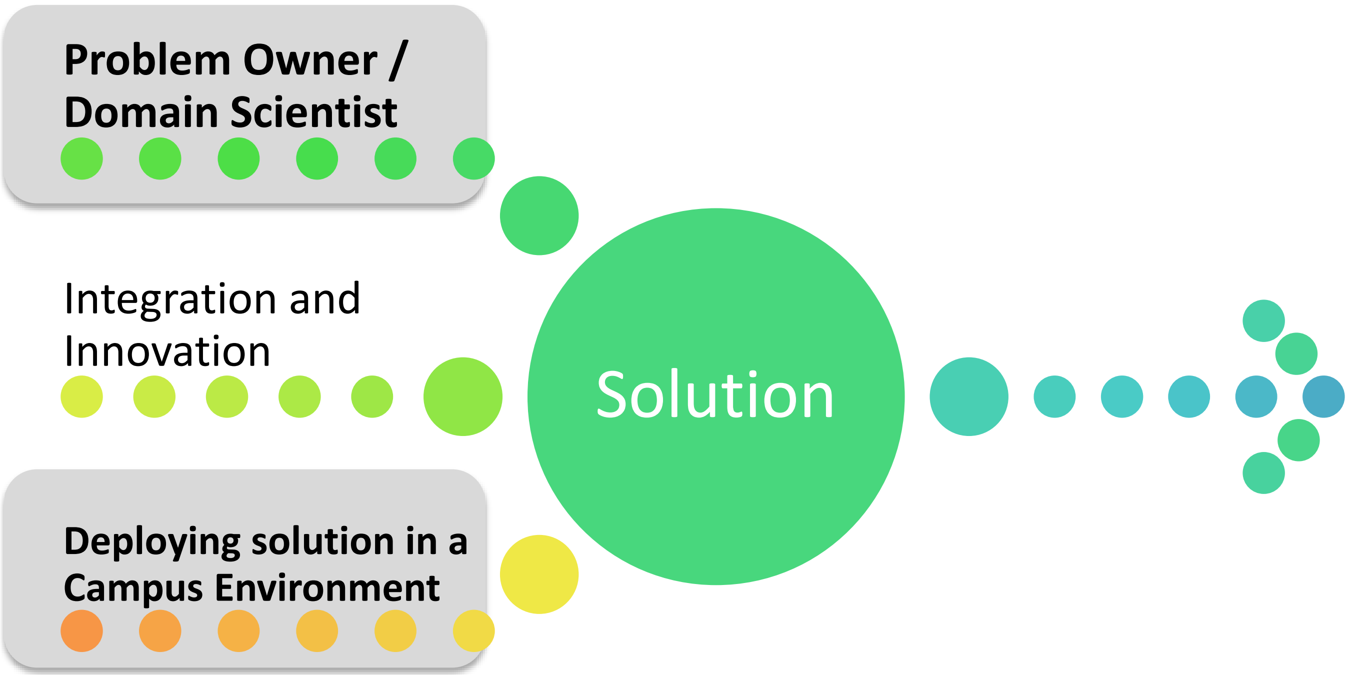
Three-Pronged Solution – Integration and Innovation is complex!



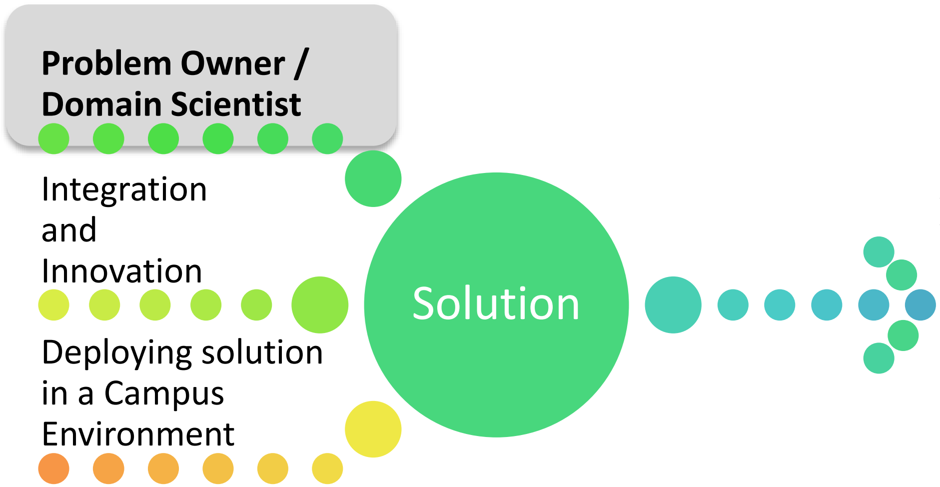
The integration and innovation effort culminates in high returns with non-trivial challenges:

- One size does not fit all! Every distinct problem requires a custom solution.
- Very, very labor intensive.
- No control of environment beyond one's span of control.
- End-to-end solution is only as good as the end that you do NOT control.
- Well resourced connectivity can have zero impact at solving a problem when an ill-defined (i.e. not well known) end-point is a critical element in the solution.

Three-Pronged Solution has Two Major Challenges!



Three-Pronged Solution has Two Major Challenges!



Human factor elements cannot be underestimated. We encountered three types of problem owners:

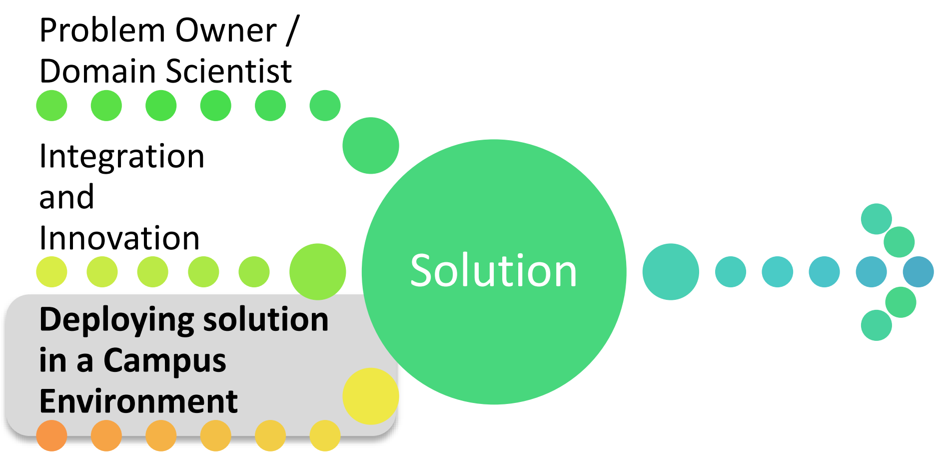
- Know what they want but **do not** know how to get there (i.e how to use cyber resources effectively). External dependency (remote instrument). Very driven.

- Know what they want and **do know** how to get there . Well-defined workflow. Insist on complete control of IP. Self-reliant and therefore limited by their own resources. Very driven.

- Know what they want. **Do not** want to change status quo. Rate of scientific discovery satisfactory.

Note: All three types are tightly coupled with cyber engineers – very, very labor intensive!

Three-Pronged Solution has Two Major Challenges!

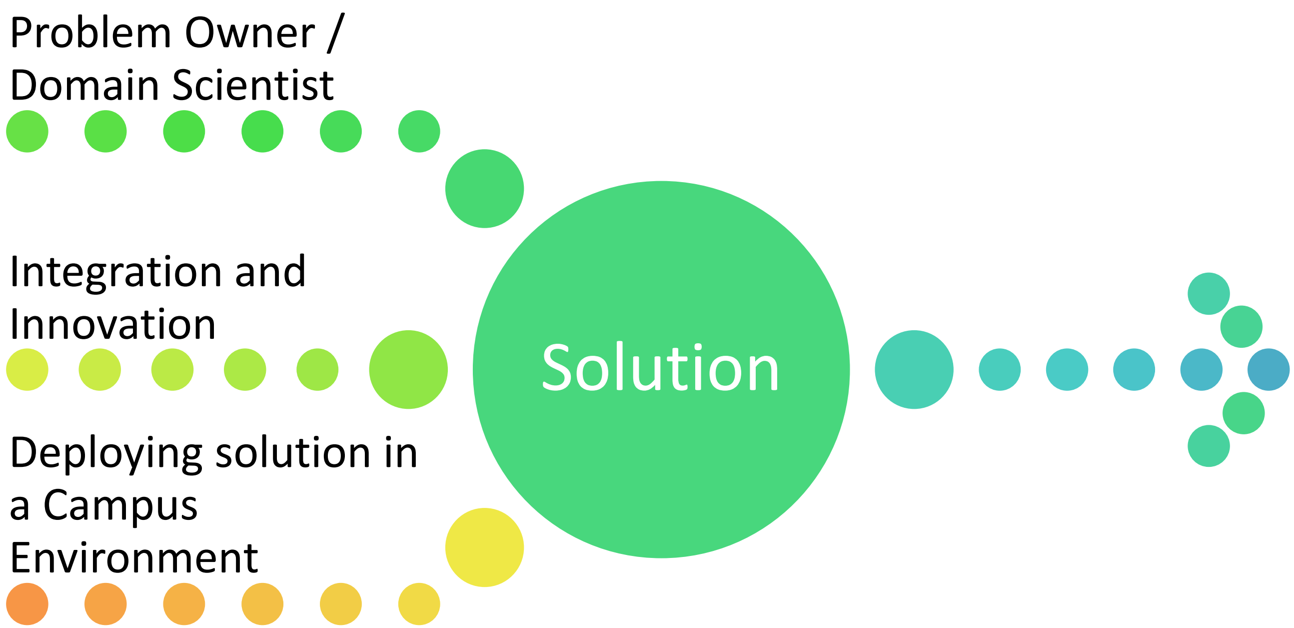


Campus Deployment Challenge:

- The network is the platform for many services for multiple constituents.
- Risk must be minimized. Attack vectors mitigated. **Cyber security is a very high priority** for the campus.
- Many network services - like intrusion prevention and firewall services – introduce network performance degradation which is not optimal for scientific research.
- Well-engineered **network edge points** that bypass security measures and traffic policies in an isolated topology in the campus network are desirable.

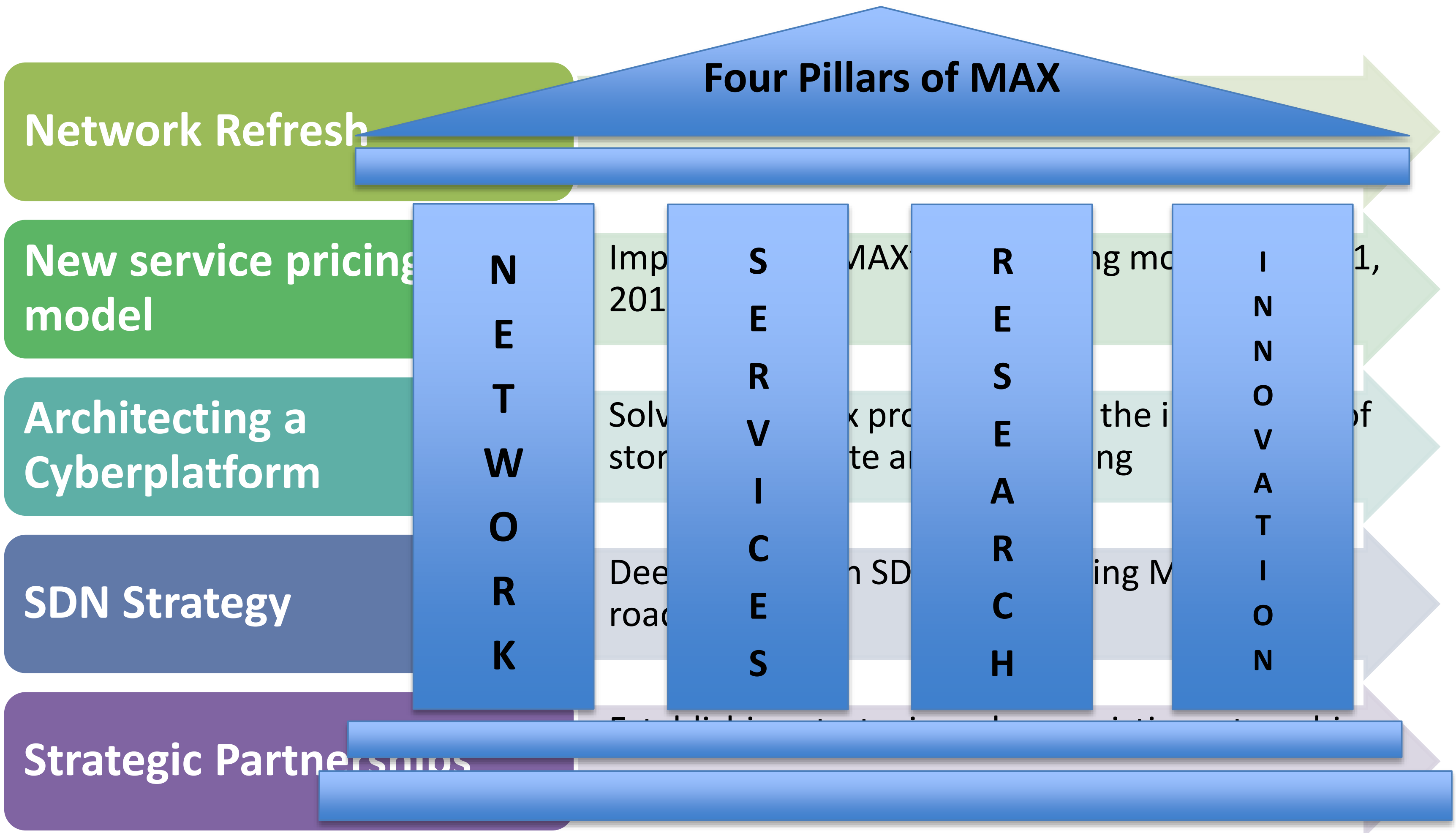
Challenges in Integration and Innovation within the Campus Environment

Conclusion: There are many challenges in finding a solution for the problem space within a campus environment



Extremely complex problem space. You will not see the complexity of the problem until you get into the problem! - Xi Yang, MAX Senior Scientist

MAX Focus on Thematic Activities



Questions?

MID-ATLANTIC CROSSROADS

Innovation and Advanced Services

Jarda Flidr

Director of Services

- AWS
- MultiService eXchange
- HPC – Cloud Integration

Definition

- Current Services (L1, L2, L3 transport)
  - Edge-agnostic
    - data movement from anywhere to anywhere
- <u>Advanced Services</u>
  - Edge-aware
    - Network Services are an integral part of bigger-scope, specific solutions
      - Well-defined destinations
      - Ecosystem of Storage, Compute, and Data sources

What are we trying to do?

- Enabling users and their applications
  - Well-engineered paths to major destinations
    - AWS
    - HPC Clusters
  - Well-engineered network edge colocation
    - High-Performance Virtualization
  - Network, Compute, and Data optimization
    - Domain Science Application integration
    - Science instrument integration

AWS SERVICES

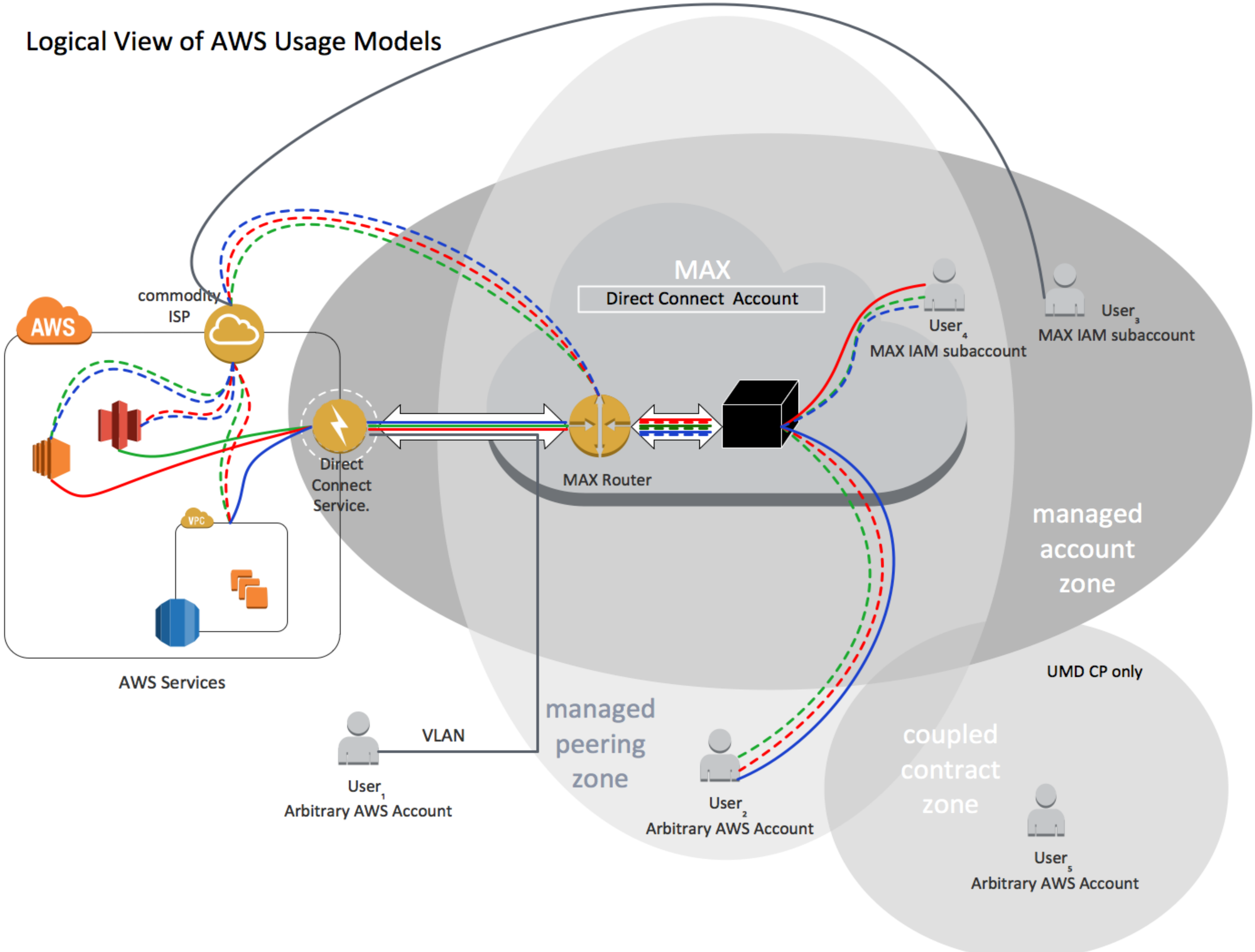
AWS Services – Scope

- Managed Peering
  - Based on AWS service *Direct Connect*
  - More than *Direct Connect*: Layer 2, brokered BGP peering, dynamic provisioning
- Account Management
- Migration Services

Managed Peering Overview

- Physical:
  - Dedicated Network Connection
    - Cross connect at MAX Equinix POP in Ashburn, VA
- L2 configuration
  - Multiple Public or Private Virtual Interfaces (VLANs)
  - Controlled by API
  - One VLAN per AWS account
- L3 configuration
  - BGP Peering
    - All Amazon East Region routes either direct (Layer 2 path through MAX) or brokered (MAX maintains the BGP adjacency on behalf of Customer)
- Benefits
  - Discounted data pricing
  - Dedicated path
  - Private BGP peering for VPC integration

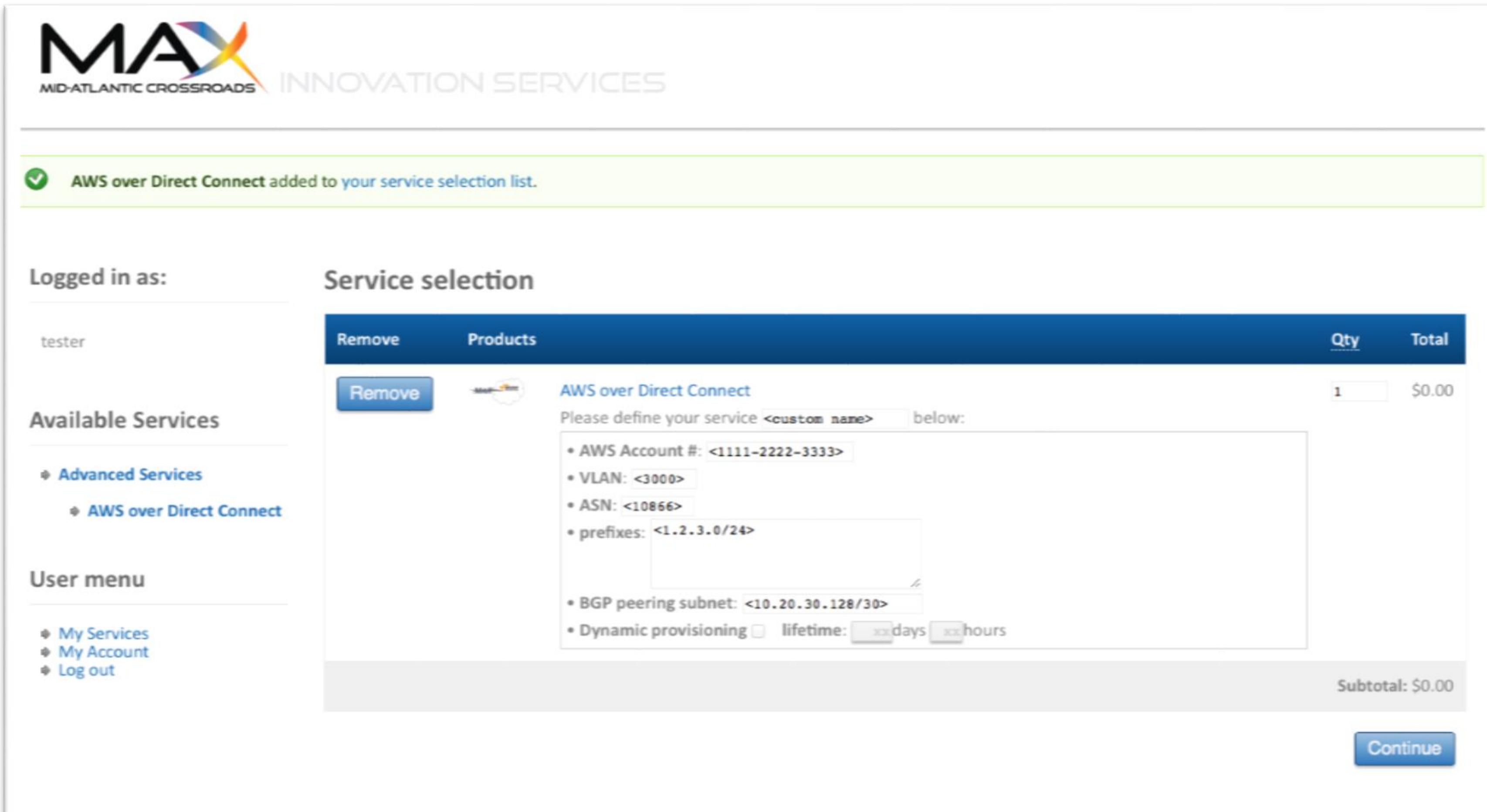
Logical View of AWS Usage Models



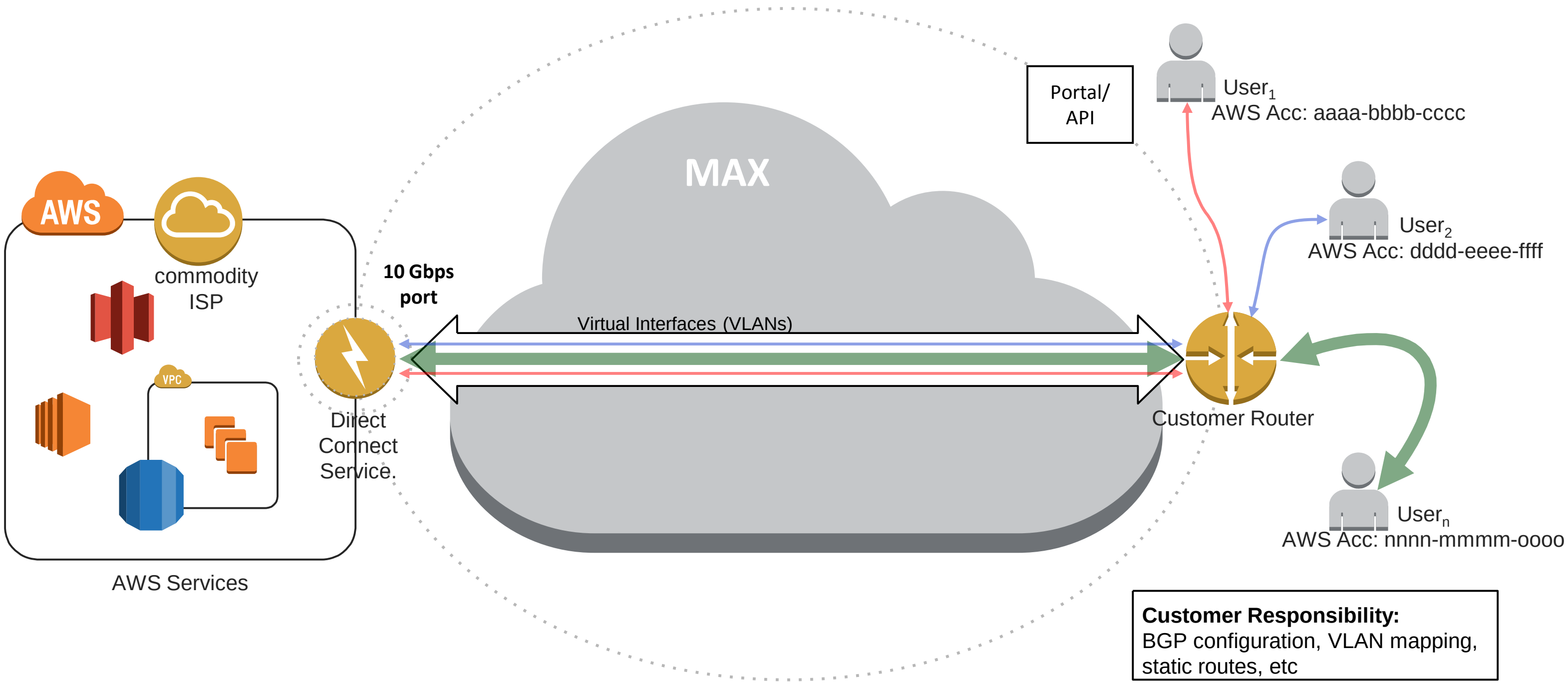
Managed Peering Summary

- What it is:
  - Special purpose, dedicated (10Gbps) connection to the services offered by AWS at Northern Virginia (*us-east-1*)
  - Dynamic: provisioned by MAX on demand
    - It Is not persistent
- What it is not:
  - Offload connection for general purpose Amazon/AWS traffic
- Intended usage:
  - Specific *big-data* transfers to/from AWS, data-intensive computation at AWS, *etc*.
- Long-term options
  - More bandwidth capacity can be provisioned

Dynamic (L3) Service

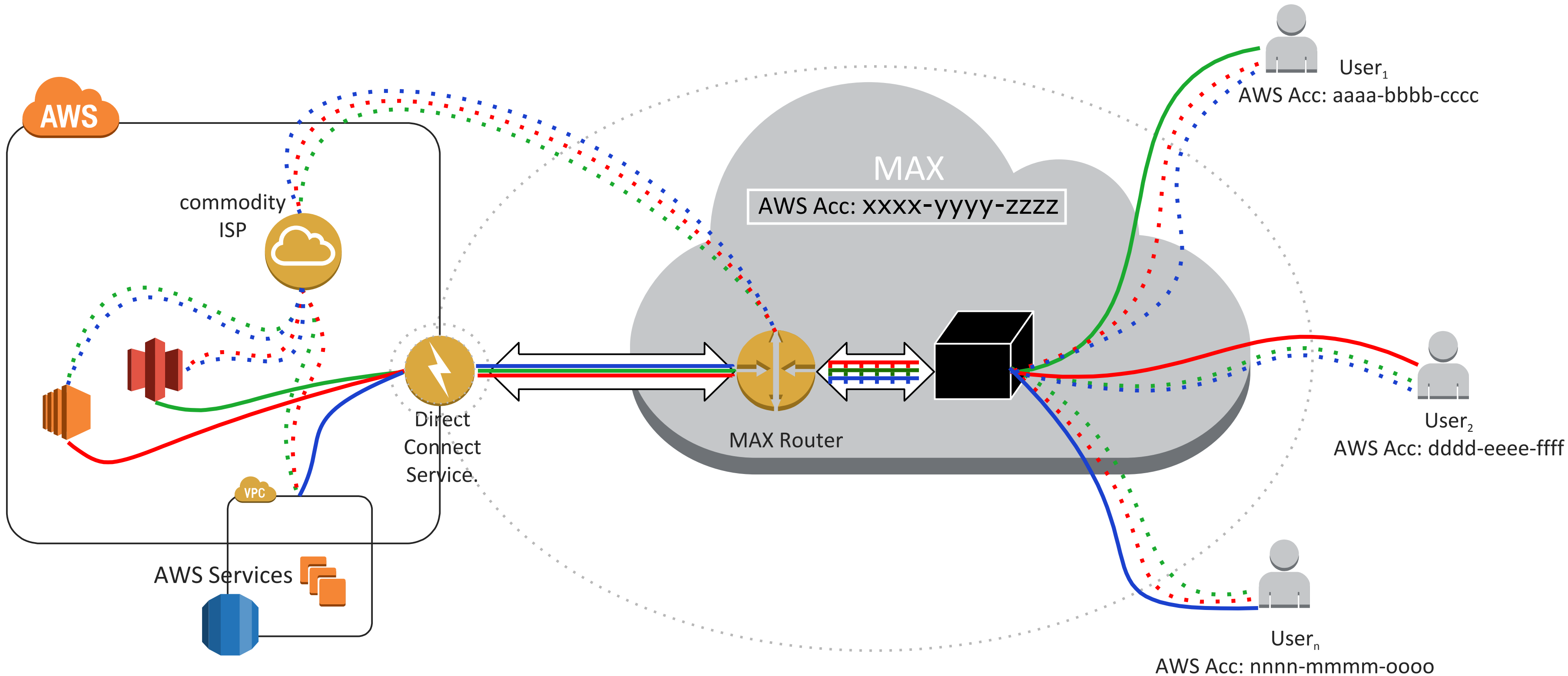


L2 Service



Customer Responsibility:
BGP configuration, VLAN mapping, static routes, etc

Dynamic Usage Example



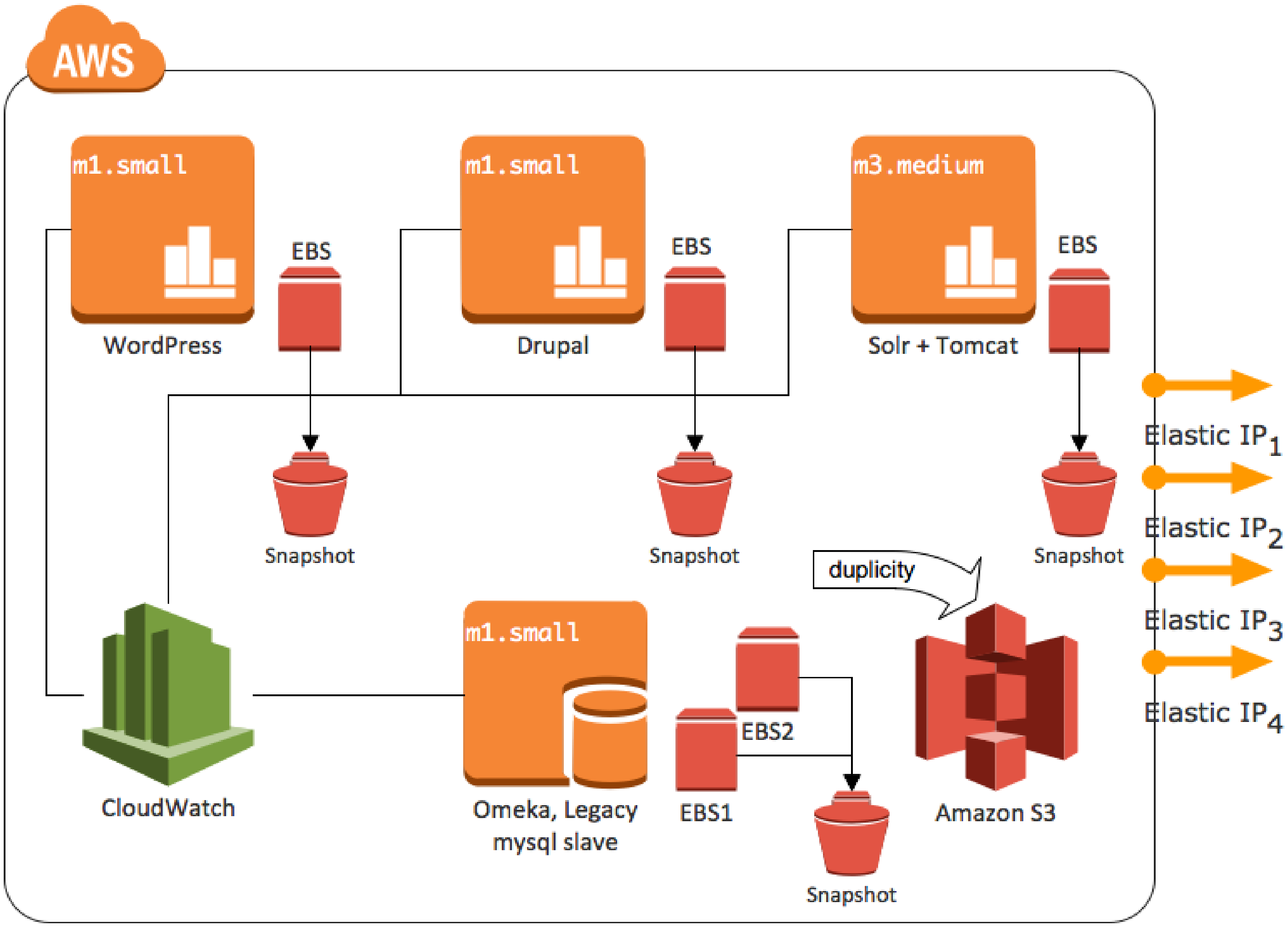
Legend:

Time Slices

- t_0:$user_1$ accesses Glacier Service over Direct Connect resold to them by MAX, the rest will see commodity routes only
- t_1:$user_2$ accesses EC2 Service over Direct Connect resold to them by MAX, the rest will see commodity routes only
- t_2:$user_3$ accesses VPC Service over Direct Connect resold to them by MAX, the rest will see commodity routes only

MAX account administrative control: only Direct Connect, resold to MAX customers dynamically

Legacy System Migration (Example)



MAX MID-ATLANTIC CROSSROADS

AWS

Amazon Web Services

Compute & Networking

Direct Connect
Dedicated Network Connection to AWS

EC2
Virtual Servers in the Cloud

Route 53
Scalable Domain Name System

VPC
Isolated Cloud Resources

Storage & Content Delivery

CloudFront
Global Content Delivery Network

Glacier
Archive Storage in the Cloud

S3
Scalable Storage in the Cloud

Storage Gateway
Integrates On-Premises IT Environments with Cloud Storage

Database

DynamoDB
Predictable and Scalable NoSQL Data Store

ElastiCache
In-Memory Cache

RDS
Managed Relational Database Service

Redshift
Managed Petabyte-Scale Data Warehouse Service

Deployment & Management

CloudFormation
Templated AWS Resource Creation

CloudTrail
User Activity and Change Tracking

CloudWatch
Resource and Application Monitoring

Elastic Beanstalk
AWS Application Container

IAM
Secure AWS Access Control

OpsWorks
DevOps Application Management Service

Trusted Advisor
AWS Cloud Optimization Expert

Analytics

Data Pipeline
Orchestration for Data-Driven Workflows

Elastic MapReduce
Managed Hadoop Framework

Kinesis
Real-time Processing of Streaming Big Data

Mobile Services

Cognito
User Identity and App Data Synchronization

Mobile Analytics
Understand App Usage Data at Scale

SNS
Push Notification Service

App Services

AppStream
Low Latency Application Streaming

CloudSearch
Managed Search Service

Elastic Transcoder
Easy-to-use Scalable Media Transcoding

SES
Email Sending Service

SQS
Message Queue Service

SWF
Workflow Service for Coordinating Application Components

Applications

WorkSpaces
Desktops in the Cloud

Zocalo
Secure Enterprise Storage and Sharing Service

MSX

Multi Service eXchange

CC-NIE Integration Award

History: Motivation and Assumptions

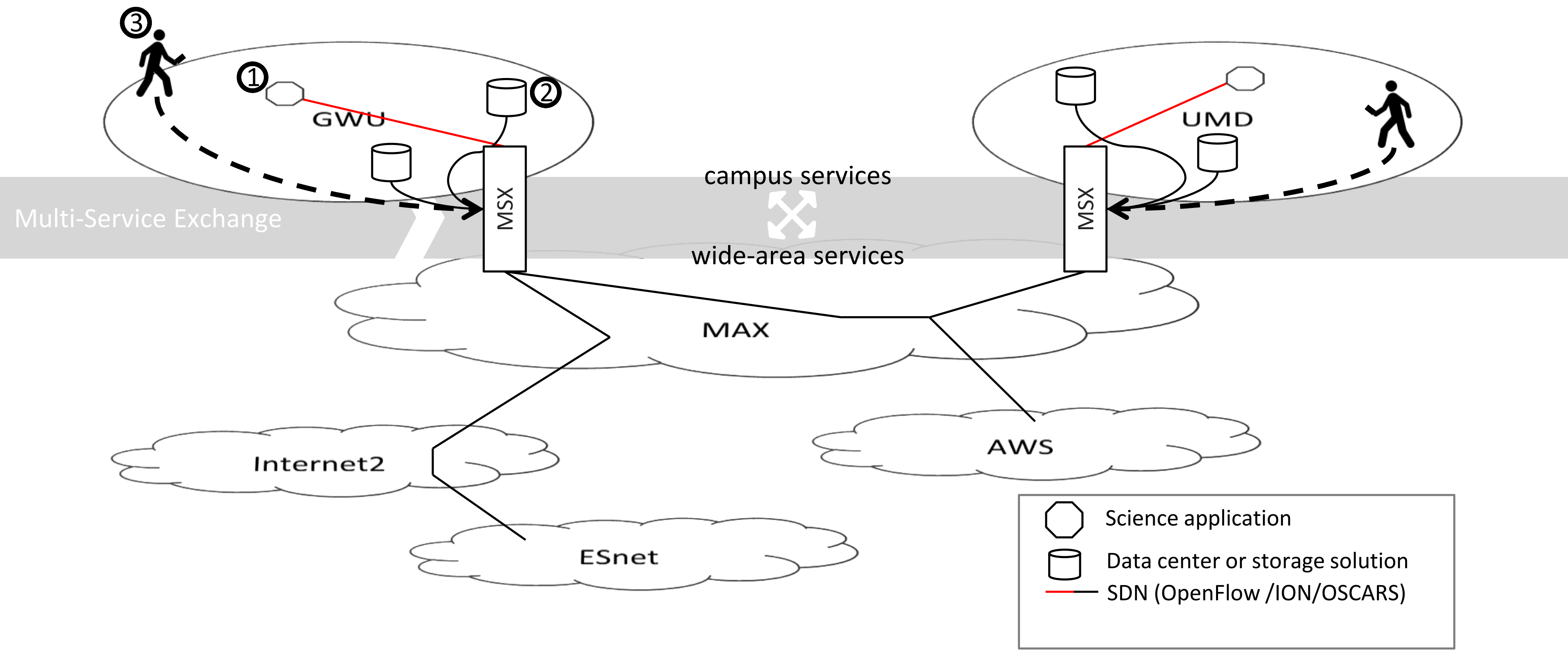
- Response to an NSF call to improve performance of data-intensive applications
  - Facilitating large data flows
- Suboptimal network performance and optimization
  - cause
    - Core vs. Edge mismatch
      - Different missions
      - Different technologies
      - Different priorities
      - Last-mile problem
    - Lack of specialized network expertise
      - “What’s Layer 2?”
      - “what’s TCP stack tuning?”
  - effect
    - Low-adoption rate: a wide spectrum of potentially beneficial and high-performance technologies are inaccessible to, or ignored by their primary users

Data-Intensive Applications

- Use cases
  - Land Cover Facility (UMD)
  - research projects encompassing the fields of remote sensing and information
  - encompasses the entire nuclear research effort at GW. Its members include the areas of Experimental, Phenomenological, and Theoretical Nuclear Physics, Astrophysics, Accelerator Physics, Reactor Physics, Nuclear Energy Research, Nuclear and Radiological Medicine, and National and International Nuclear Energy and Weapons Policy Studies, and houses the largest and most frequently accessed database of fundamental nuclear reactions in the world
  - Particle Astrophysics and High Energy Physics projects: South Pole IceCube Neutrino Observatory, The LIGO observatory, Large Hadron Collider (LHC) – tier 3 system, Open Science Grid (OSG)
  - environment at scales from individuals up to neighborhoods, cities, and metropolitan systems

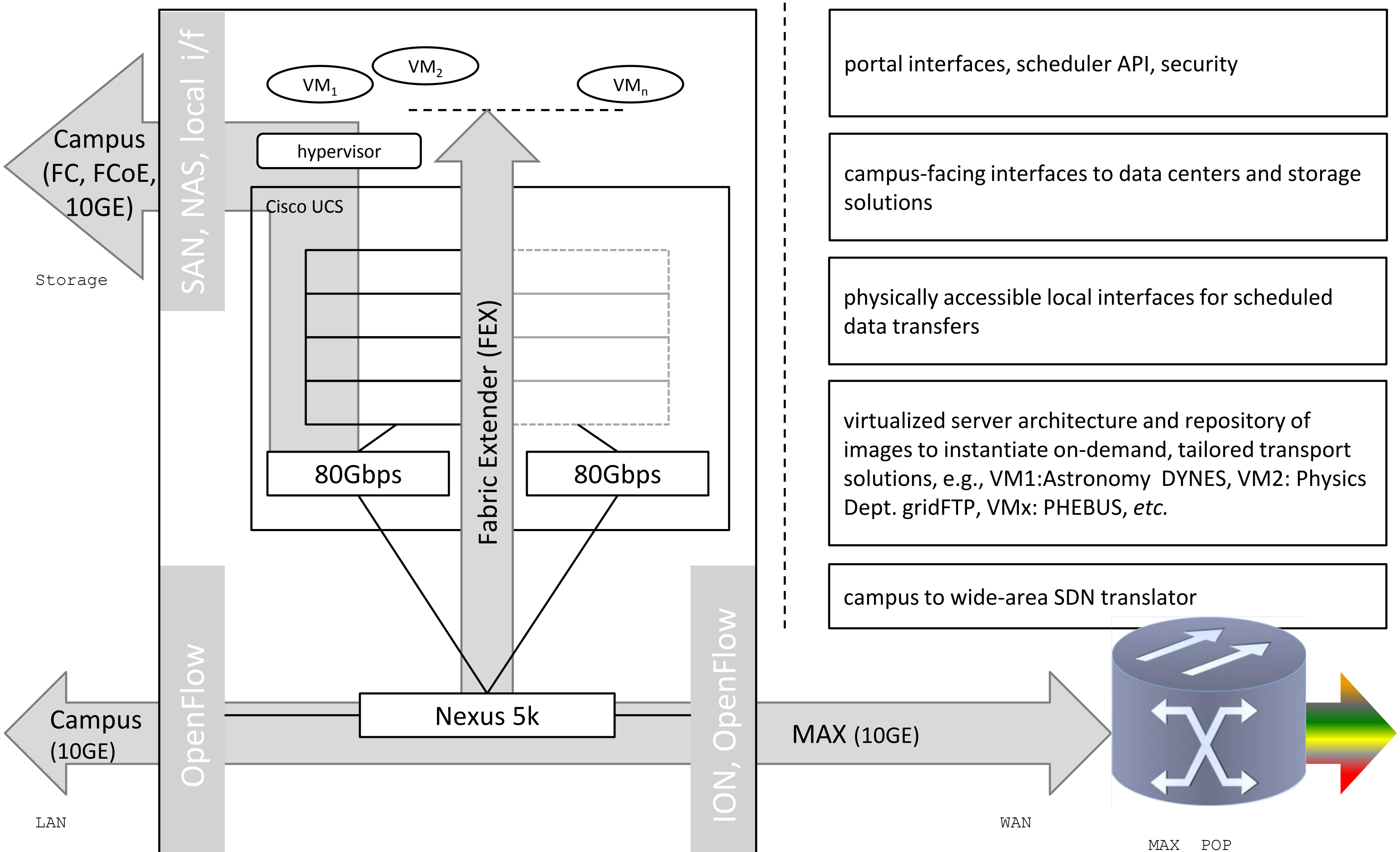
Original MSX Concept

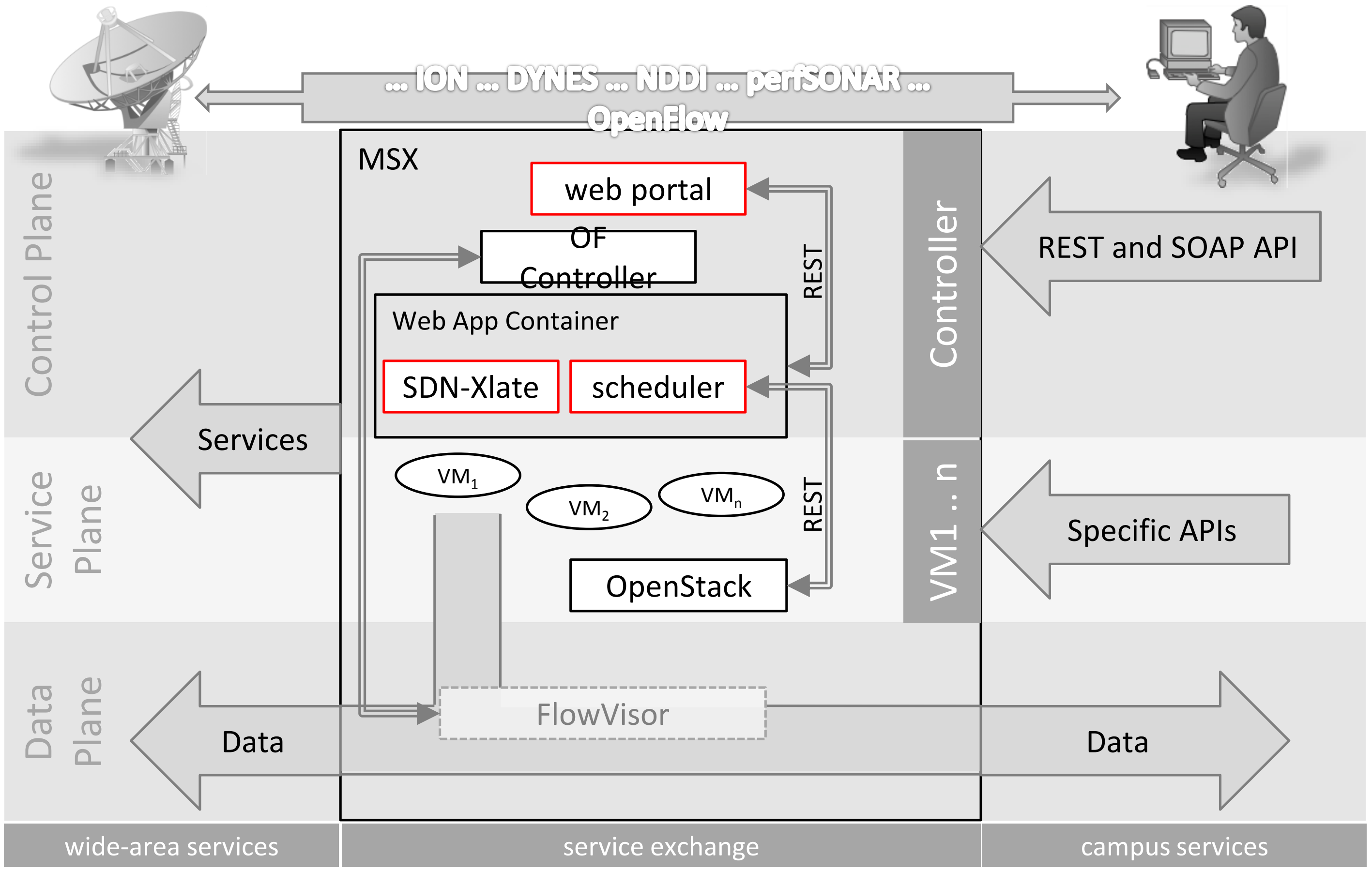
Optimized network provisioning based on SDN



hardware and virtual architecture

functionality

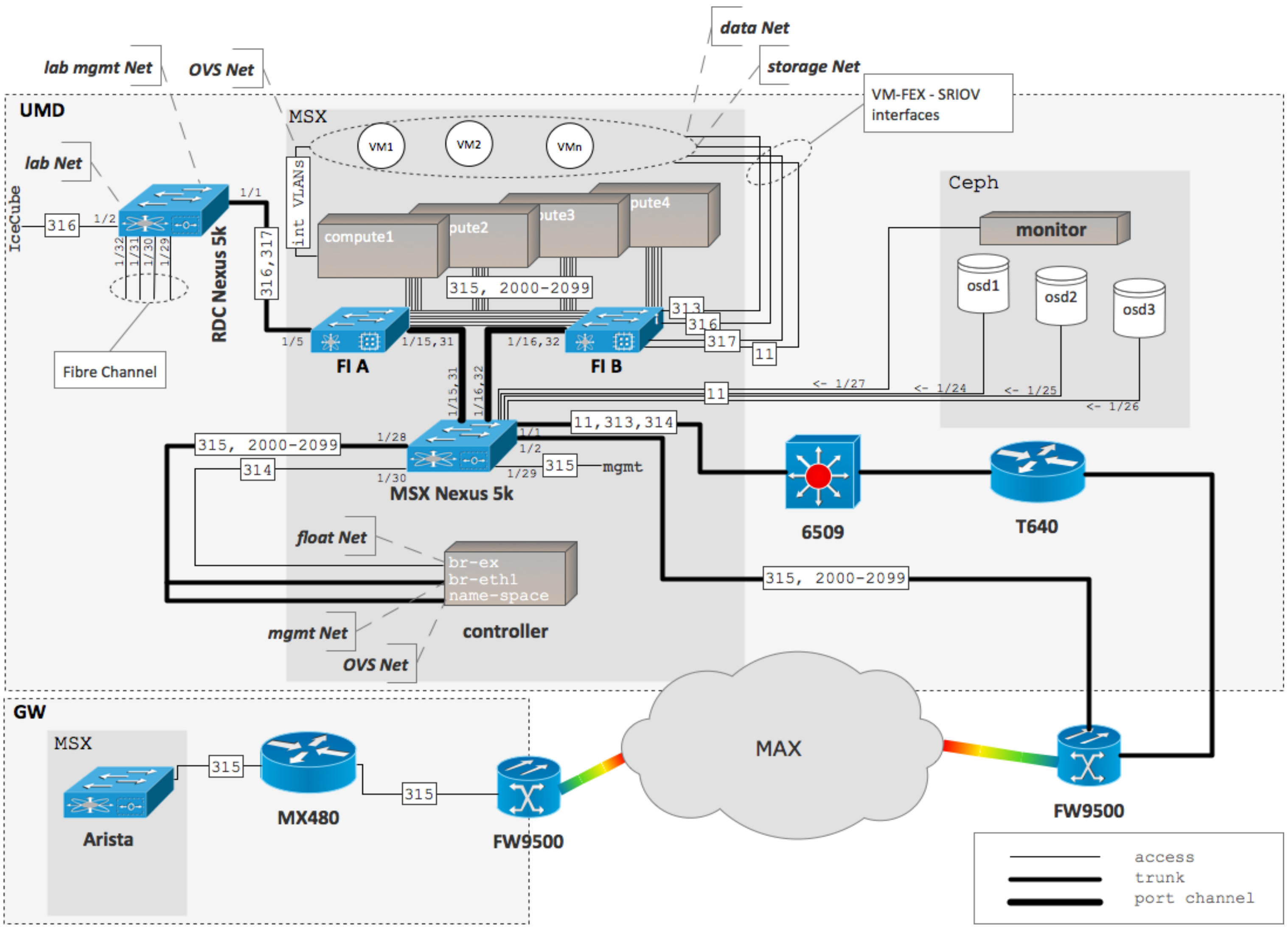



... ION ... DYNES ... NDDI ... perfSONAR ...
OpenFlow
MSX
web portal
OF
Controller
REST
Web App Container
SDN-Xlate
scheduler
VM1
VM2
VMn
REST
OpenStack
FlowVisor
Controller
VM1 .. n
REST and SOAP API
Specific APIs
Control Plane
Service Plane
Data Plane
Services
Data
Data
wide-area services
service exchange
campus services

What is MSX?

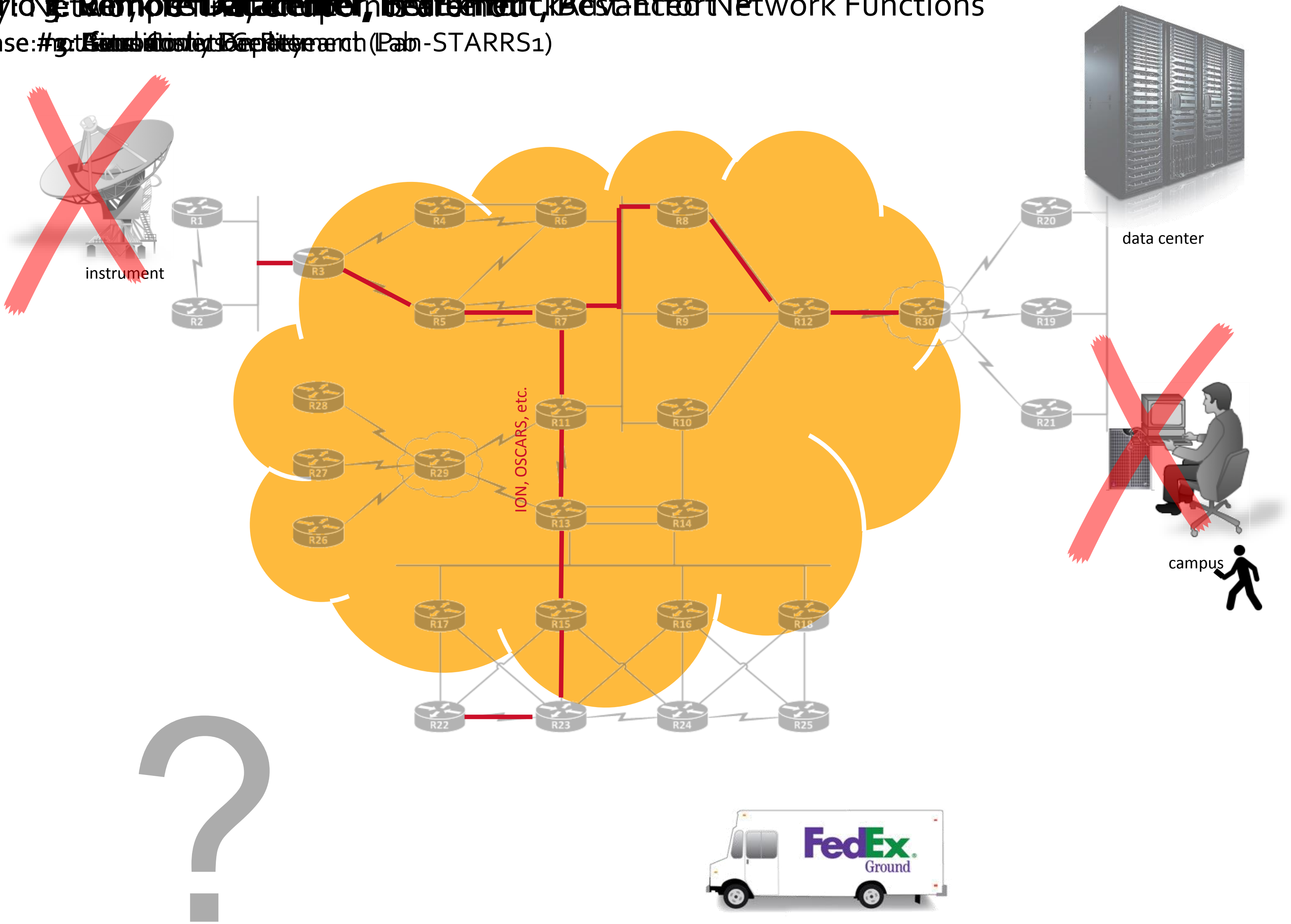
- High-performance, **serially-multitenant** platform
  - SR-IOV-enabled hosts and Virtual machines
- Integrator of the existing advanced network functions
  - best effort IP
  - AL2S
  - ION
  - DYNES
  - SDN (OpenFlow)
- Optimizer of data-intensive and compute-intensive applications
  - Fluid Edge: triangulation of the best location with respect to Storage, Compute, and Data

Actual Architecture (UMD)



Scenario 3: Remote Data Center, Instrument, Best Effort IP Network Functions

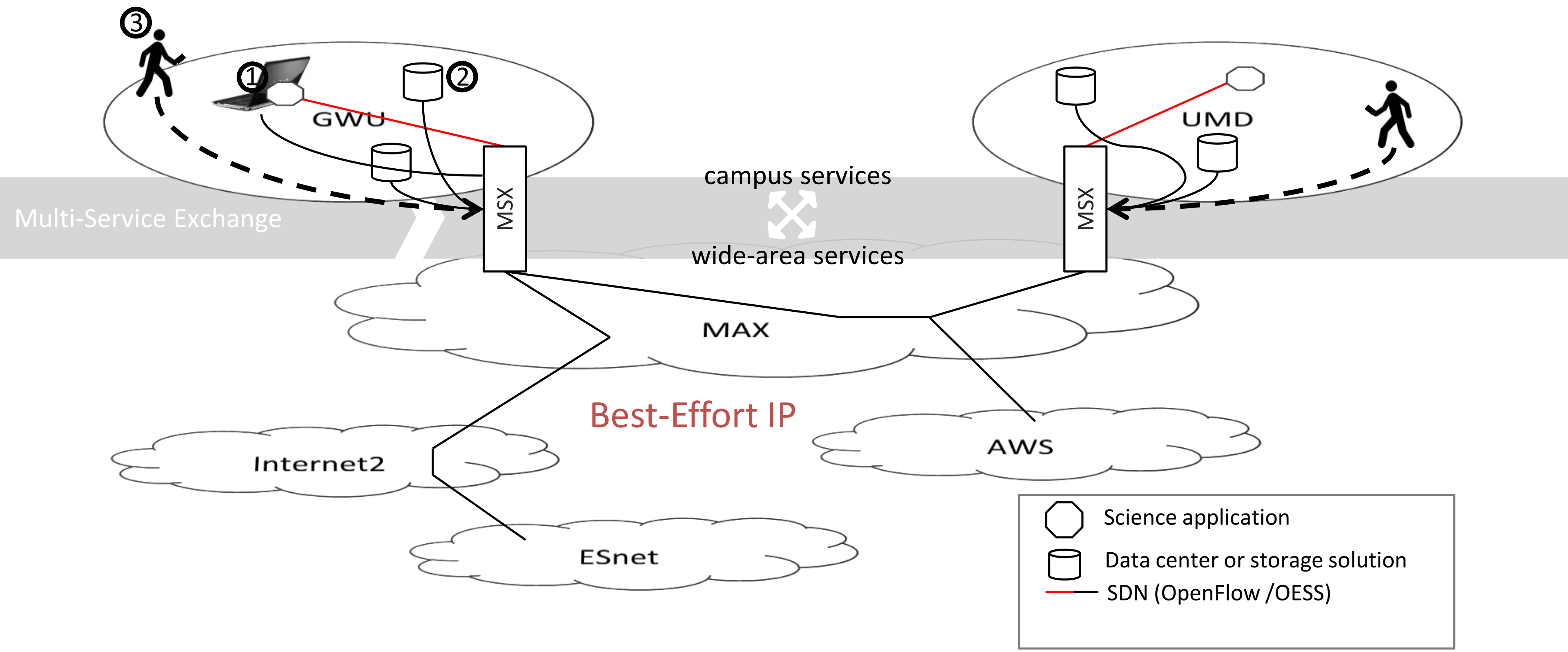
- use case: #3: ... (Pan-STARRS1)



Approach

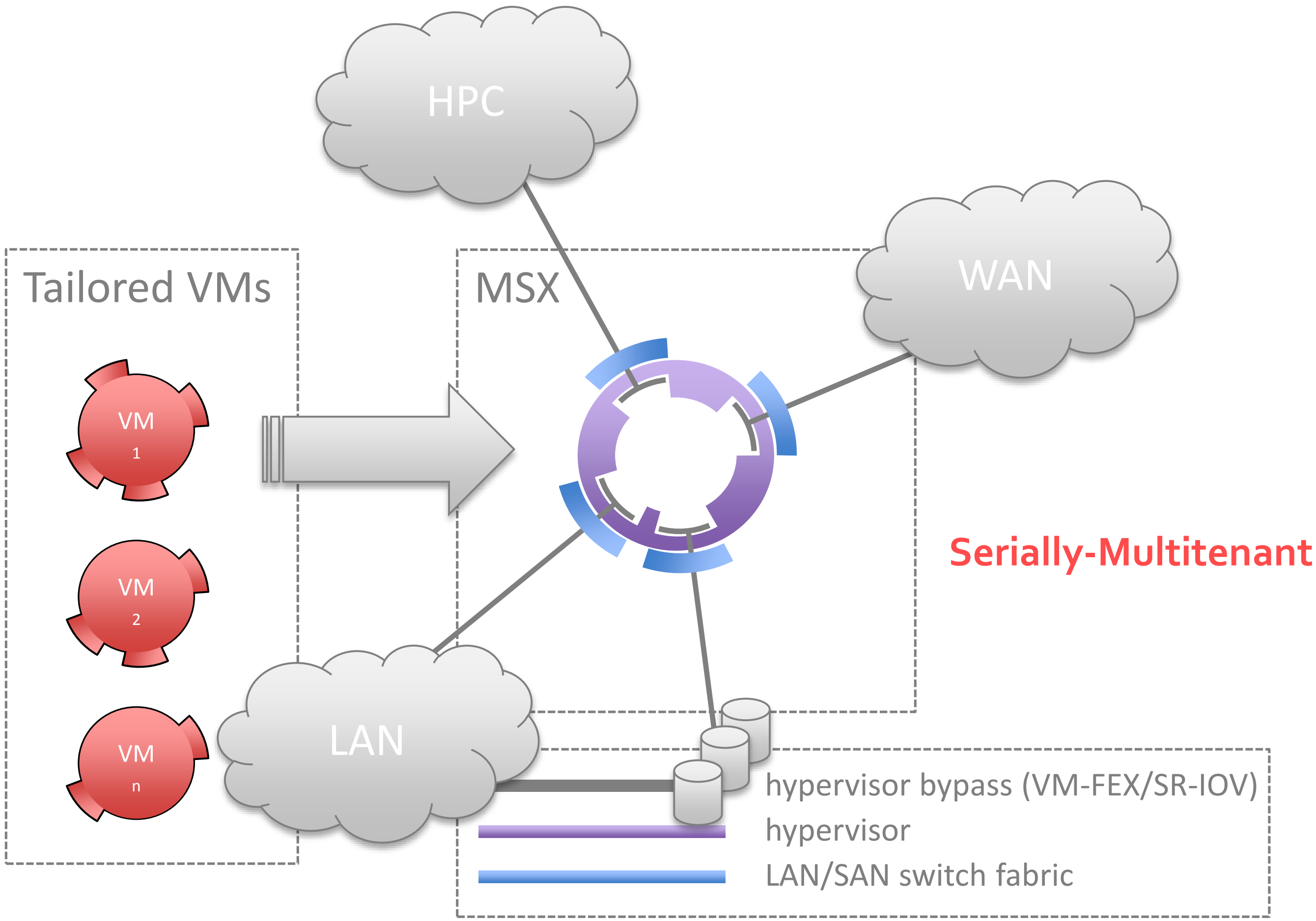
- Status Quo
  - Wide-Area Networks are well-performing. Best effort IP is sufficient
  - Endpoints' design and locations are the problem
- Solution
  - Redesign endpoints (applications)
  - Move them to the well-engineered core's edge

Evolution of MSX

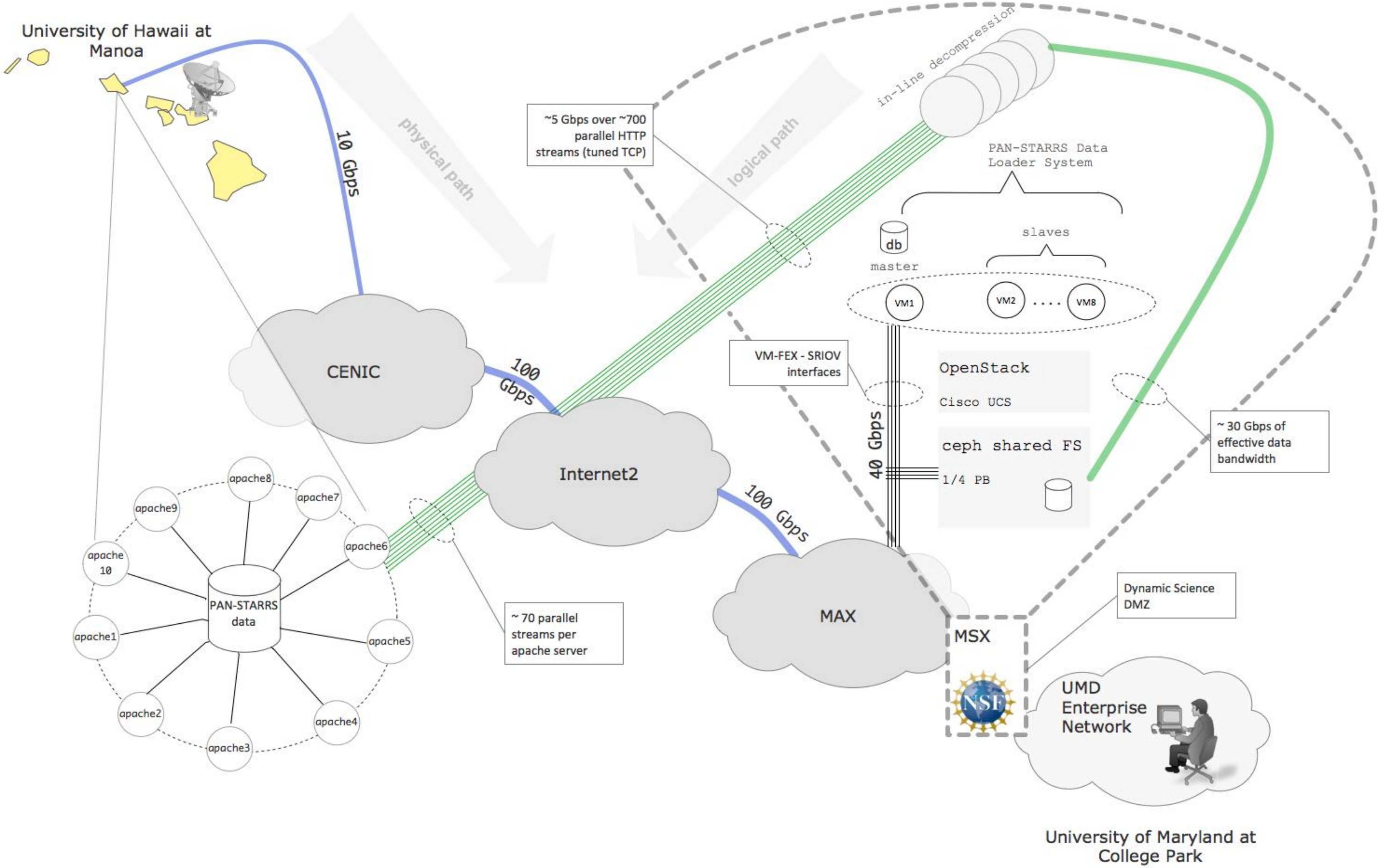


Solution

Flexible, Dynamic, Cost effective, High-Performance and – most importantly – it works
(but - labor-intensive)

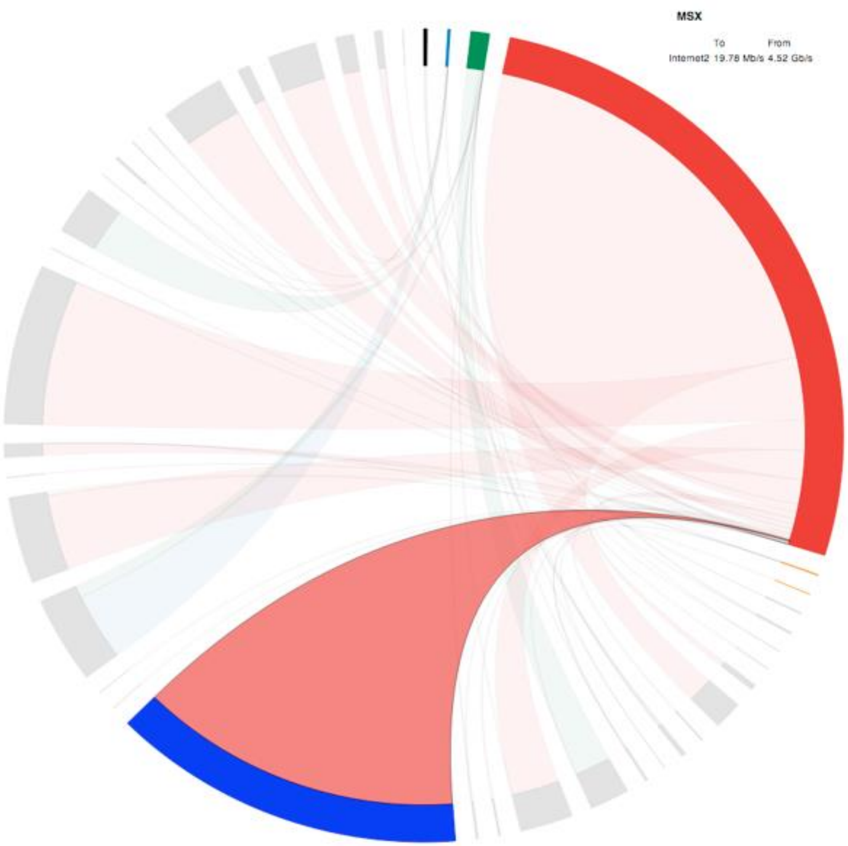


MSX ecosystem: Schematic Overview



Endpoint engineering example

Hawaii – UMD, Best-effort IP: up to 6 Gbps



Summary

- It is much more effective to move the CPU to the data than the other way
- Domain Science Application-level, endpoint integration is labor intensive
- Hardware platform
    - Open
    - Flexible
    - Modular
- Dynamic, On-Demand, Fluid Science DMZ
  - Beta service mode

Future Plans

- Additional Services Integration (HPC)
  - Tom will address those

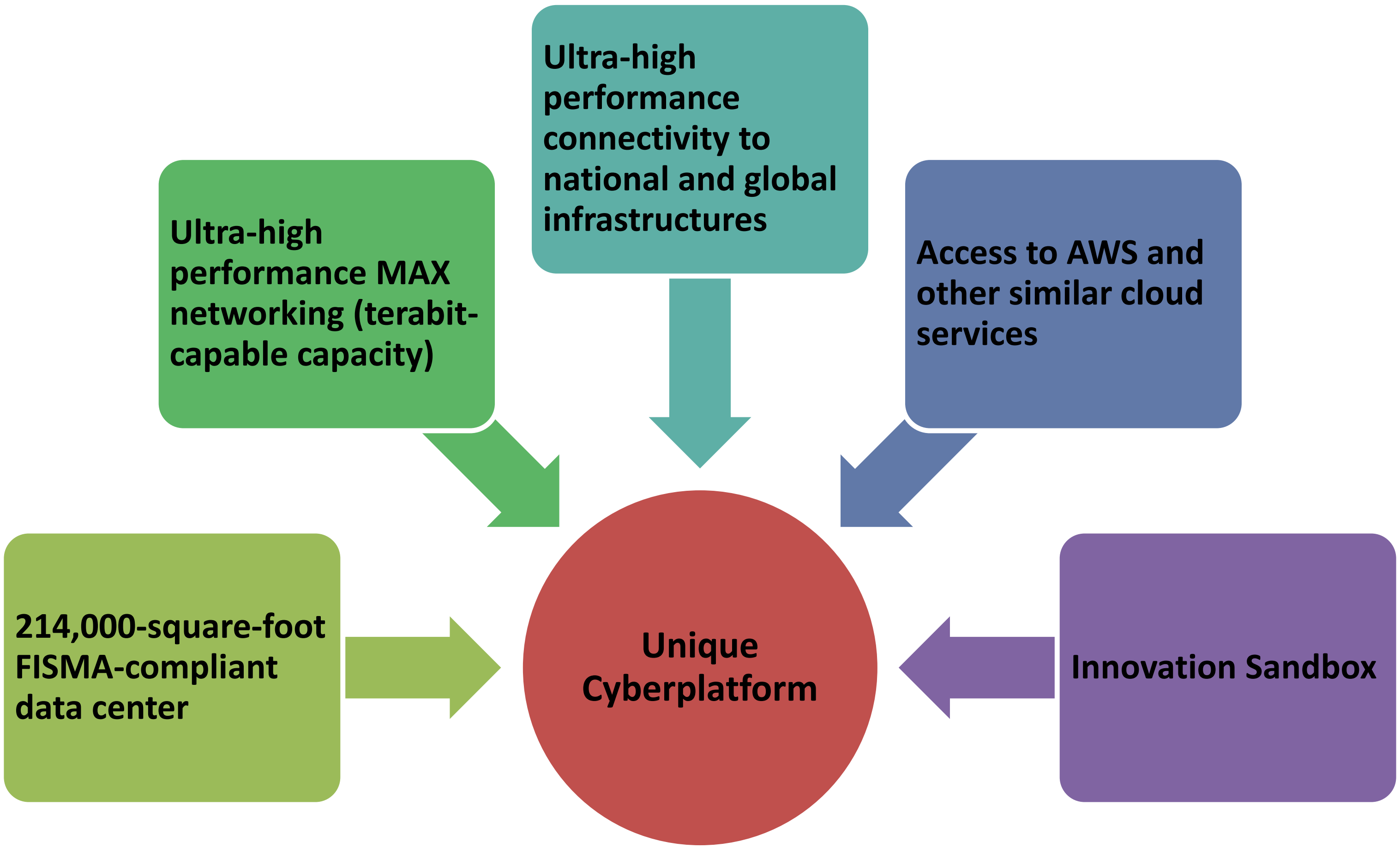
Thank You

Strategic Partnership

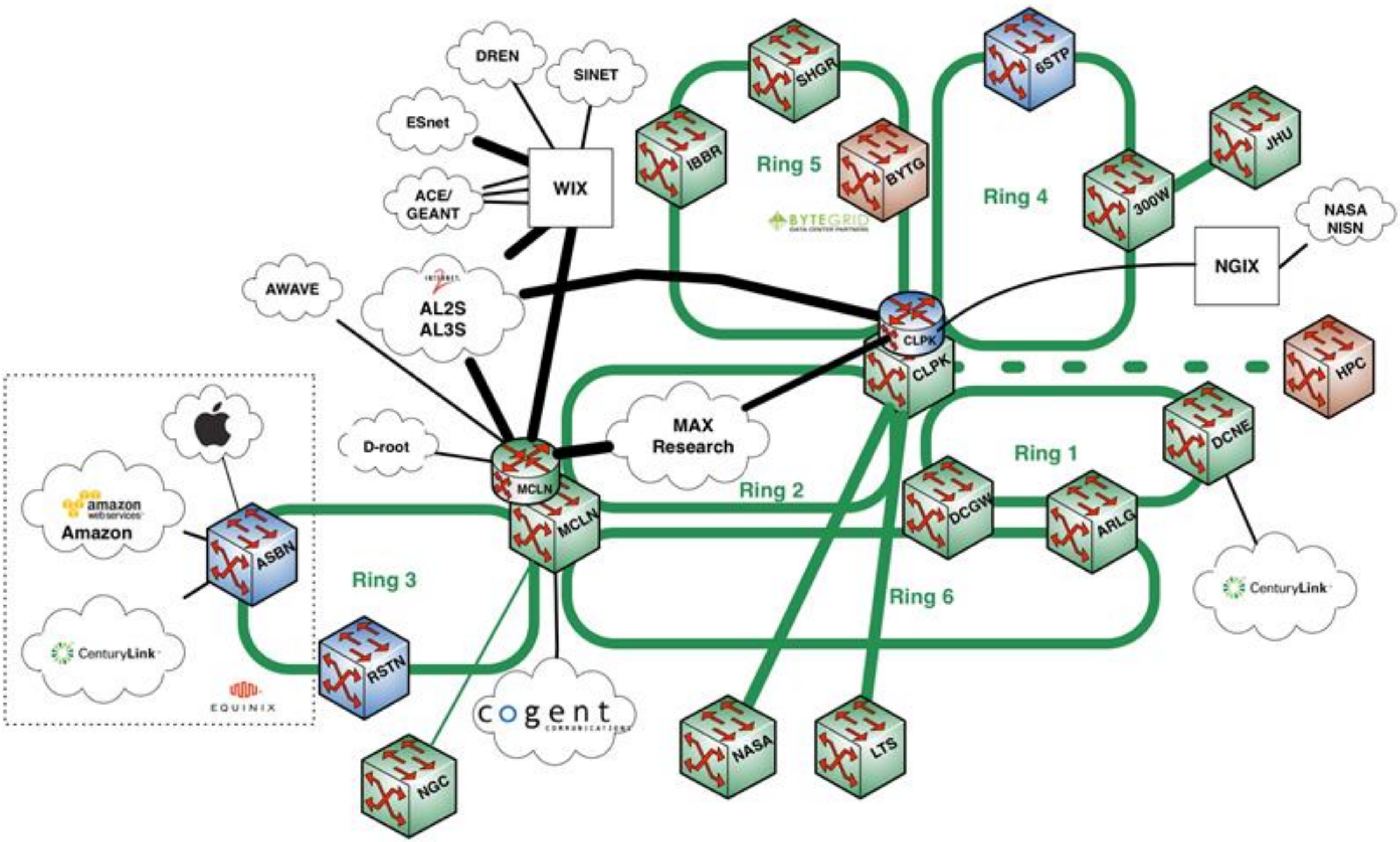
MAX
MID-ATLANTIC CROSSROADS

BYTEGRID®
DATA CENTER PARTNERS

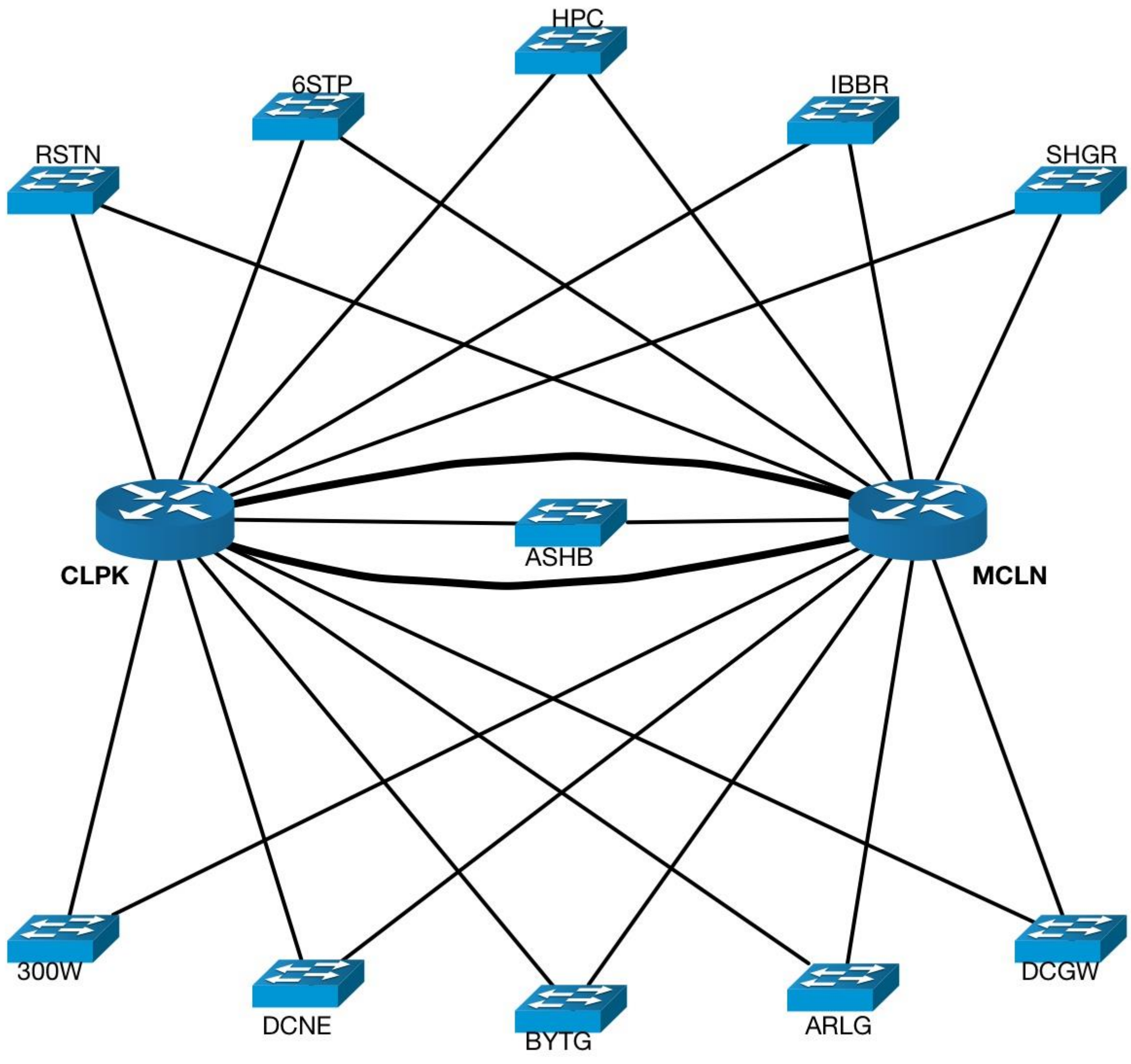
MAX-BYTEGRID Partnership – a unique platform



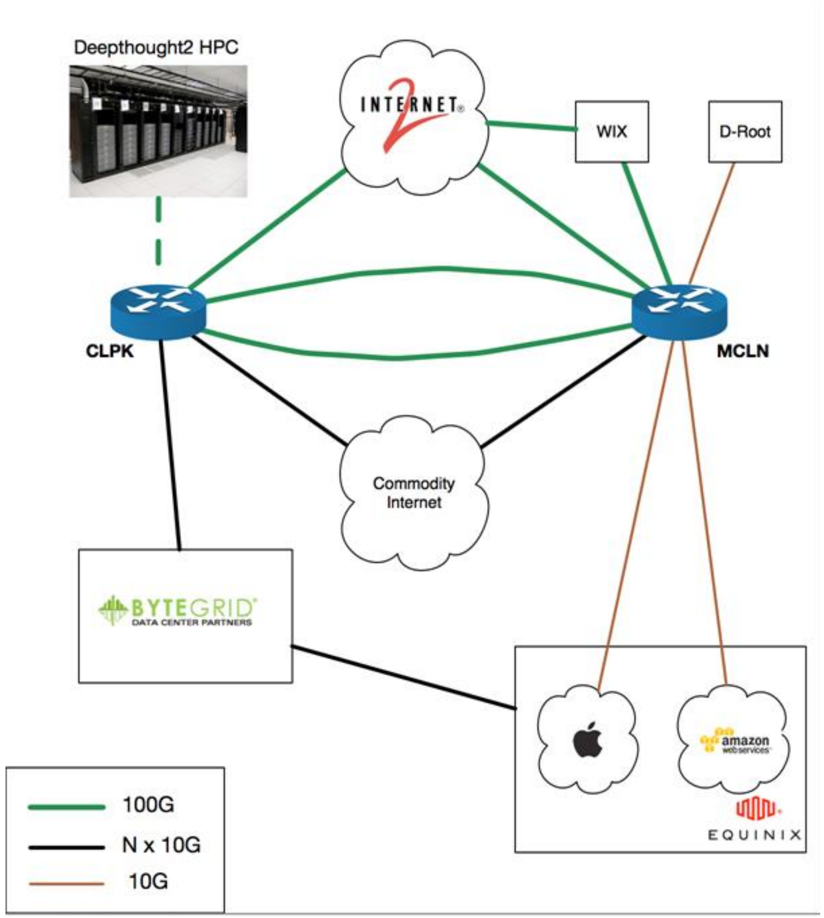
MAX-BYTEGRID Partnership – a unique platform



MAX-BYTEGRID Partnership – a unique platform



MAX-BYTEGRID Partnership – a unique platform



MID-ATLANTIC CROSSROADS

Don Goodwin

Executive Vice President

BYTEGRID®
DATA CENTER PARTNERS

MID-ATLANTIC CROSSROADS

MAX Sponsored Research Projects

Update on NSF, DOE, and DOD supported research activities

Research Program - Objectives

Contribute to the evolution of advanced networking within the Research and Education Community

Deploy systems and develop technologies which facilitate domain science researchers use of cyberinfrastructure

- Make these available to MAX Participants

Assist in the development of the vision, design, and deployments for the next generation MAX network

Past MAX Research Areas

WSS/ROADM based All-Optical Regional Networks

- Very early adopter of this technology; collaboration with MOVAZ (now ADVA) to deploy prototype ROADMs on MAX Research Network as part of DRAGON project
- Expanded on this work with Fujitsu based 100G DWDM equipment

Dynamic Networking

- Early developers of below IP (layer 2 and 1) dynamic services
- Developed and deployed solutions based on control plane and data plane separation as we see now in Software Defined Networking (SDN) architectures
- DRAGON software suite used as basis for initial deployments of Internet2 dynamic networking (HOPI, Dynamic Circuit Network

Past MAX Research Projects

- DRAGON (Dynamic Resource Allocation via GMPLS Optical Networks), 2003-2008, NSF
- Hybrid Optical Packet Infrastructure (HOPI), 2005-2006, Internet2
- Advanced Technology Demonstration Network (ATDnet) Collaboration, 2007-2009, NRL
- Dynamic Circuit Network (DCN), 2007-2008, Internet2
- GENI Spiral 1, 2008-2012, BBN/NSF
- MAX 100G, 2010-2012, NSF ARI

Current MAX Research Topics

MAX Research and Advanced Services

- Research activities are shifting from lower level optical topics to higher level cyber-infrastructure integration and advanced services development

High Level Objectives

- Integration between applications/workflows, compute, storage, instruments, and networks
- Develop advanced services which facilitate flexible use of cyber-infrastructure by domain scientist
- Accomplish this in the emerging world of big data, a variety of compute options (HPC, Clouds, local compute), and increasingly distributed environments

Current MAX Research Topics

We have identified the key technical areas where we think R&D focus is needed to accomplish these objectives:

- End-to-End Flow Management
  - where E2E now includes the storage and compute end-systems
- Science Workflow Integration
- Network Virtualization (NV)
- SDN (Software Defined Networks)
  - Layer 3, Layer 2, Layer 1
- SDI (Software Defined Infrastructure)
- Multi-domain service provisioning, federation

Have active research projects addressing these issues now

MAX Research Infrastructure

MAX operates a Research Network in parallel with the Production Network to allow these types of research and development activities to occur

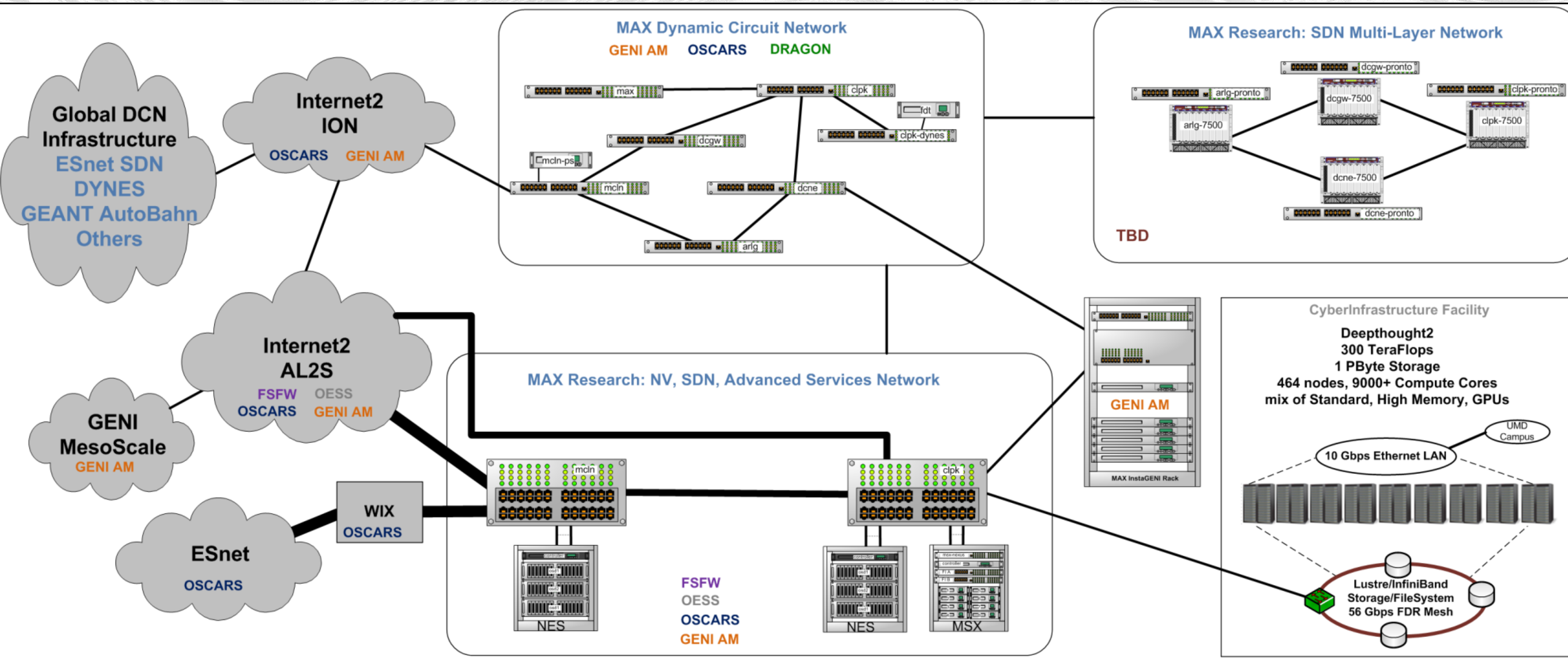
First Generation: NSF DRAGON project all-optical MOVAZ ROADM dark fiber based footprint; mirroring the production network

Second Generation: NSF ARI initiated all-optical Fujitsu 9500 WSS based dark fiber based footprint; mirroring the production network.

- Early deployment of 100G DWDM Technologies (Fujitsu 100G Transponder Serial numbers 1 and 2 deployed on MAX)
- Most of this infrastructure is being absorbed as part of the MAX Production Refresh for 100G networking

Third Generation: Building that now, "Research and Advanced Services Development Network", more details on the following slides

MAX Research Infrastructure



Control Plane Software:

GENI AM - GENI Aggregate Manager

OSCARS - InterDomain Dynamic Circuits

DRAGON - GMPLS Dynamic Circuits

FSFW - FlowSpace Firewall (Internet2 Network Virtualization)

OESS - Internet2 Circuit and Multipoint Provisioning

TBD - To Be Developed

- This in the process of being built
- Constructed using a mix of lambdas and vlan partitioning on the underlying MAX network

ControlPlane Software Overview

FlowSpace Firewall (FSFW) (Internet2)

- Network Virtualization
- Runs on Internet2 Production Network

Open Exchange Software Suite (OESS) (Internet2)

- Point-to-Point and Multi-Point VLAN provisioning
- Runs on Internet2 Production Network

OSCARS (ESnet)

- Multi-Domain Point-to-Point VLAN provisioning

DRAGON (MAX)

- GMPLS based multi-vendor, multi-technology provisioning

GENI Aggregate Manager (GENI Project)

- GENI Resource and Federation interaction software

Research Projects

- **High Performance Computing with Data and Networking Acceleration (HPCDNA)**
  - NSF CC-NIE
- **Resource Aware Intelligent Network Services (RAINS)**
  - DOE Office of Science
- **100G Connectivity for Data-Intensive Computing at JHU**
  - NSF STCI
- **GENI Stitching and Computation Enhancements (GENIStitch)**
  - NSF GPO (GENI Project Office)
- **Network Survivability via Failure Identification and Rapid Network Restructure (NetSurvive)**
  - DOD DTRA

High Performance Computing with Data and Networking Acceleration (HPCDNA)

- **NSF CC-NIE Project**
  - UMD Principal Investigators (PIs) Tripti Sinha, Tom Lehman, and Xi Yang from MAX and Saurabh Channan from the Global Land Cover Facility (GLCF) and Paul Torrens from the Geosimulation Research Laboratory
- **Motivation:**
  - Domain sciences are facing **big data** challenges. They need HPC!
  - However, there was a missing link between high performance computing and big scientific data.
  - Many groups with big scientific data considered HPC center a "walled garden" in which they could not easily get data in or out.
  - The simplified answer is: We need integration of data processing (**compute**), data **storage** and data movement (**network**). But how?
- **Solution:**
  - Extending the internal high performance data storage and access system in the core of the HPC system to high performance external storage systems embedded within high performance networks.

HPCDNA Key Features

Network Embedded Storage (NES)

- Ceph distributed storage system:
  - Parallel file system for high performance
  - Distributed locations for replication
  - 250 TB x 2~3 sites
- Well engineered and attached to MAX 100G infrastructure:
  - 100G MAX regional network at L2 and L3.
  - 100G to I2 AL2S
  - OpenFlow capable Ethernet layer
  - 10G AWS Direct Connect

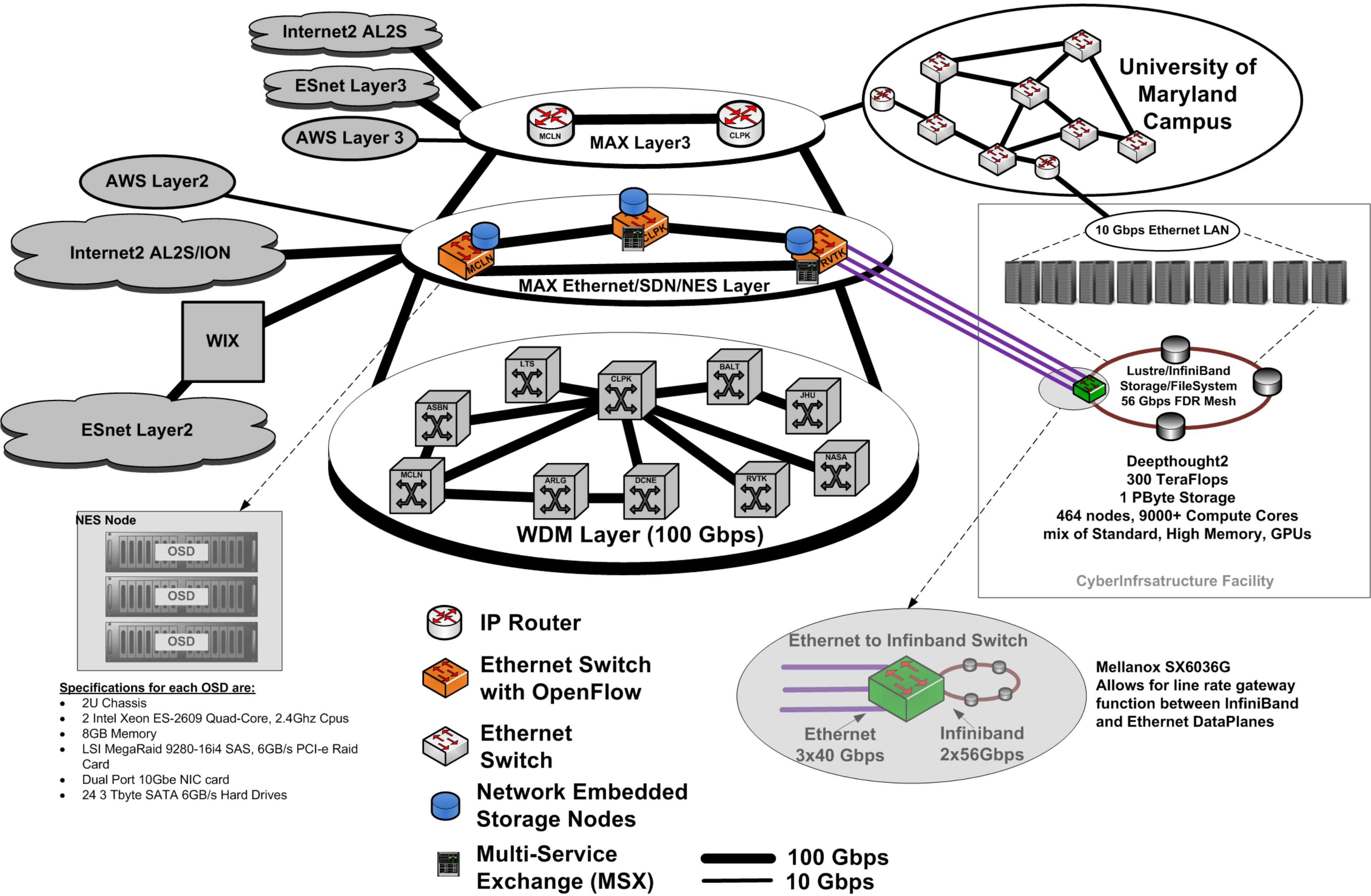
HPC and HPN Integration

- HPC file system to NES system integration
  - DeepThought2 Lustre file system bridge to HPCDNA NES
  - 56Gbps InfiniBand facing Deepthought2
  - 3 x 40Gbps Ethernet facing NES
- File system access options
  - Block storage mount (POSIX)
  - S3/Swift object storage API
  - Network storage (NFS/SMB)

HPCDNA Enabled Application Workflow

- Providing a rich tool set to integrate into big scientific data application workflows.
- NES for high-performance data launching, landing and staging.
- A high-performance storage and cache gateway.
- Ability to access the high-performance NES from inside Deepthought2 HPC cluster.
- Federated with UMD authentication system (potentially also with InCommon etc.)
- External AWS integration for hybrid cloud workflows.

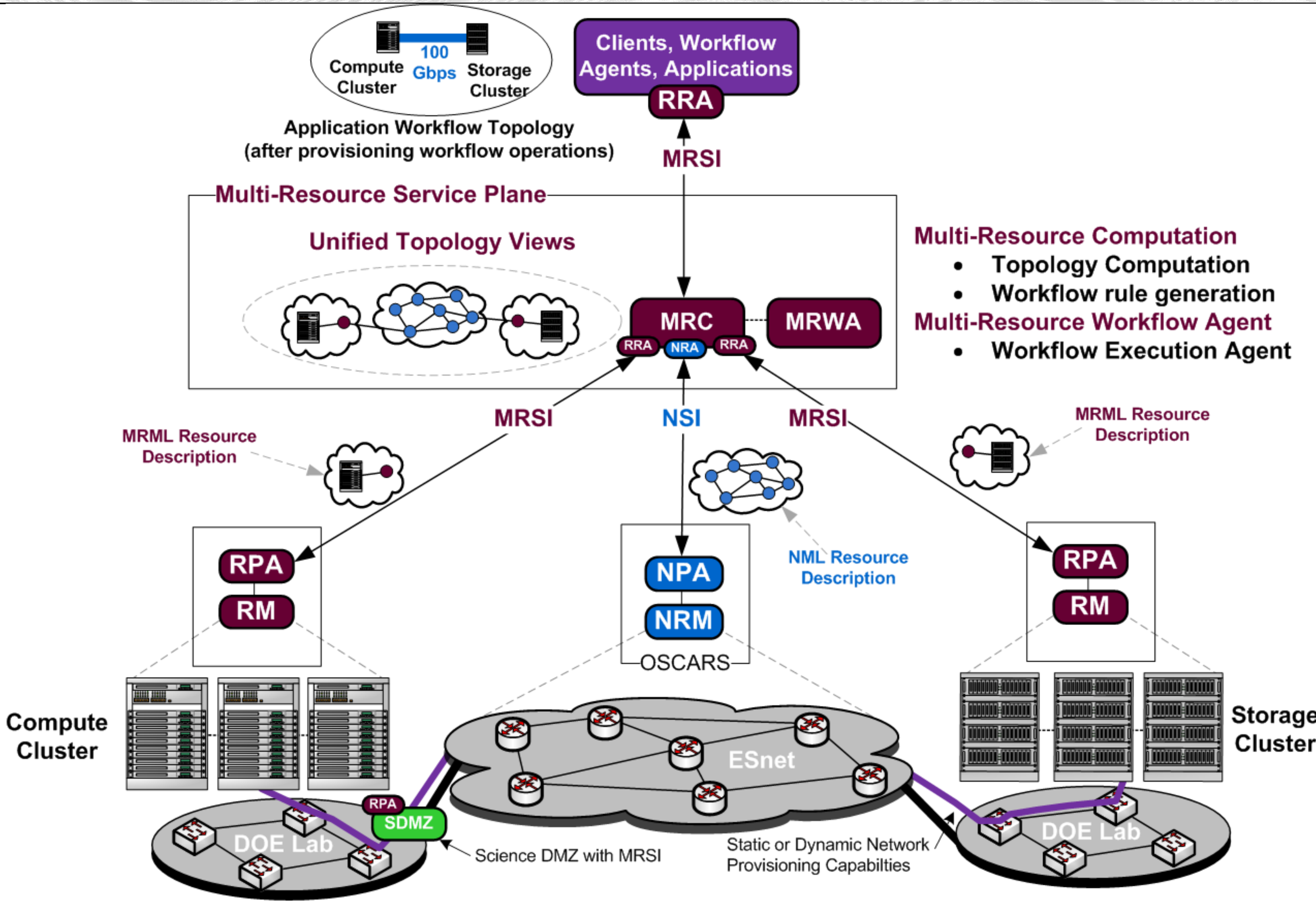
HPCDNA Architecture



Resource Aware Intelligent Network Services (RAINS)

- Funded by DOE
- UMD/MAX (lead organization), PIs Tom Lehman, Xi Yang
- Argonne National Laboratory (ANL), PIs Raj Kettimuthu, Linda Winkler
- Motivation:
  - A wide range of science applications need for flexible and seamless integration across multiple resources to support workflows.
  - Advanced networking infrastructures and capabilities are the cornerstone technology to enable this integration.
  - Today's dynamic network service development is focused exclusively on network topologies and resources.
  - Challenge remains to determine how their domain specific compute and storage resources are connected to the dynamic network infrastructure.
- Solution:
  - Developing technologies that enable the integration of domain specific (compute and storage) resources with the Network Service Plane (**NSP**) and the Intelligent Network Services (**INS**).

RAINS Architecture

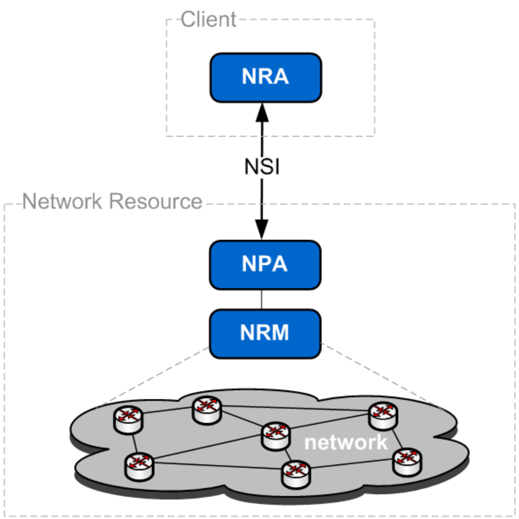


<u>DOE Lab and ESnet Network Resource Approach</u>

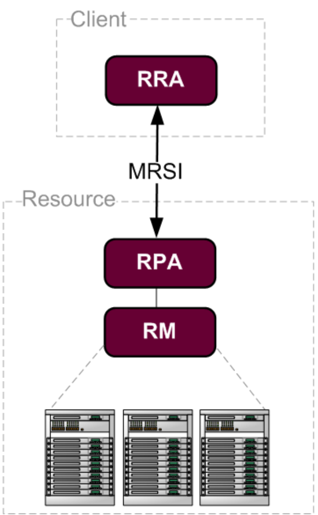
- ESnet services such as OSCARS dynamic provisioning will be incorporated into the MRSP ecosystem
- DOE Lab networks may not have a dynamic provisioning capability. Planning to work to extend the lab ScienceDMZ connections and features to support the MRSP. This may include placement of a MRSI interface agent to "cover" the Science DMZ.

Multi-Resource Service Interface (MRSI)

- **Purpose:**
  - Every MRSP participant acts as a service provider or a service requestor or both and is connected to the ecosystem through a set of services.
  - Through the MRSI, MRSP can provide common and open-standard mechanisms for requesting, querying and monitoring diverse types of resources
  - It also provides common mechanisms for security and policy management.
- **Model:**



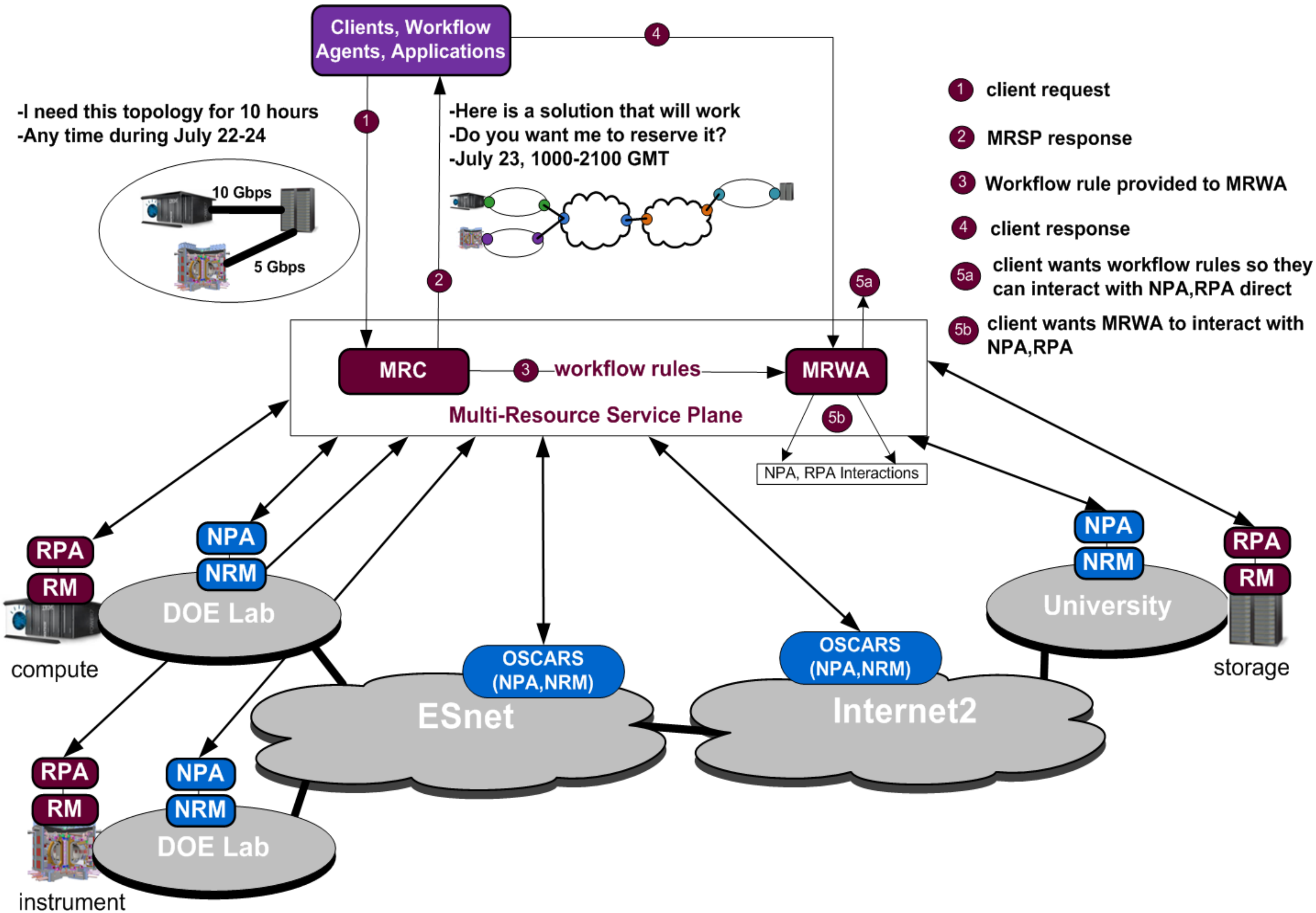
Network Service Interface Model
NRA - Network Requester Agent
NSI - Network Service Interface
NPA - Network Provider Agent
NRM - Network Resource Manager



Multi-Resource Service Interface Model
RRA - Multi-Resource Requester Agent
MRSI - Multi-Resource Service Interface
RPA - Multi-Resource Provider Agent
RM - Resource Manager

- **Focus:**
  - Make OSCARS an MRSI compatible resource provider agent (RPA).
  - Develop a new MRSI RPA to cover Magellan OpenStack clouds.
  - Wrap KBase or make Shock and AWE resource managers MRSI compatible.

RAINS Ecosystem Vision



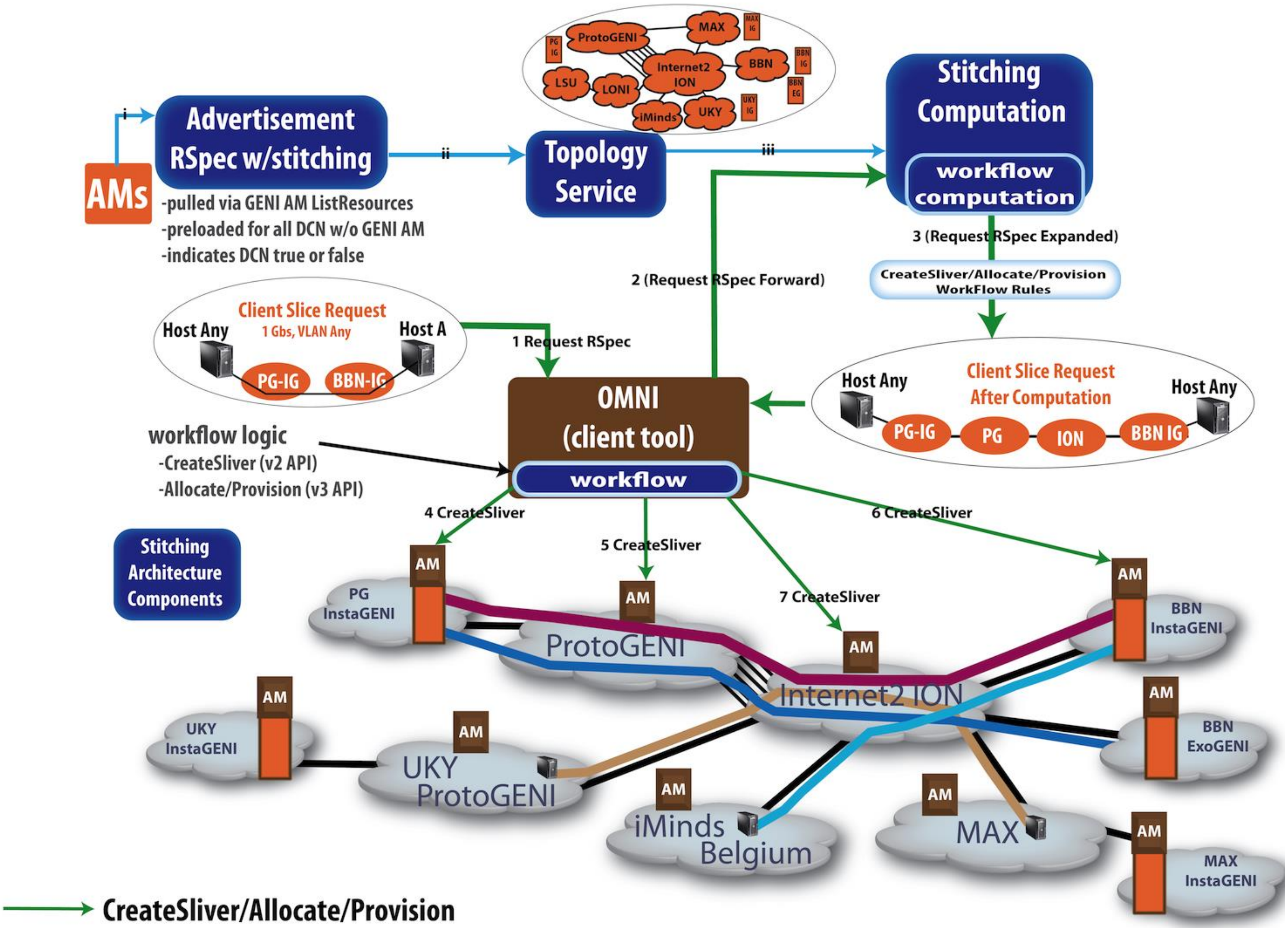
100G Connectivity for Data-Intensive Computing at JHU

- **NSF STCI Project (three years, started January 2012)**
  - Lead Organization Johns Hopkins University, Alex Szalay (PI)
  - MAX providing networking support, Tom Lehman (co-PI)
  - Utilizing the 100G infrastructure between JHU and MAX
- **Objectives:**
  - Support efforts to move Big Data to/from JHU Data-Scope to national scale computation facilities
  - JHU Data-Scope is a novel instrument to observe and visualize large data sets in real-time
- **Current Activities:**
  - Working on facilitating data transfers and both Layer2 and Layer3 to/from several sites including SDSS (Sloan Digital Sky Survey), LANL (Los Alamos National Lab), and Fermilab

Global Environment for Network Innovations (GENI) , GENI Stitching

- **Background:**
  - "GENI is a virtual laboratory for exploring future internets at scale, creates major opportunities to understand, innovate and transform global networks and their interactions with society."
  - GENI consists of interconnected and federated "aggregates" the provide virtualized compute and network resources, a.k.a. "slices", to experimenters.
  - Each GENI aggregate joins their resources to the community by implementing a set of well defined APIs.
- **MAX:**
  - Is a GENI aggregate providing DRAGON network to use by the GENI community
  - Has deployed a GENI InstaGENI Rack that provides additional resources including openflow networking
  - Is a key contributor for architecting and developing GENI infrastructures and technologies.
  - MAX focus is on the GENI Stitching Architecture, Development, Deployment
  - Current Activities; AL2S support for GENI, multi-point topologies, workflow negotiation techniques, topology computations to support user tools

GENI Stitching Architecture



NetSurvive

Network Survivability via Failure Identification and Rapid Network Restructure (NetSurvive)

- **Funded by DTRA**
- **University of New Mexico is prime contractor. UMD/MAX and University of South Florida (USF) are subcontractors.**
- **Overview:**
  - Focus on protection and restoration against backbone disruptions and large-scale failures that involve many network elements and multiple network administrative domains.
  - Design Survivability Aware Intelligent Network Service Plane Architecture to address both pre-emptive protection and post-failure restoration services.
  - Use MAX and other infrastructures to create multi-domain multi-layer testbed for prototyping and evaluating the service plane architecture and survivability algorithms and workflows
  - Apply Software Defined Networking (SDN) to multi-domain transport networks

Thanks

OpenStack Cloud: Architecture, Development, and Deployment

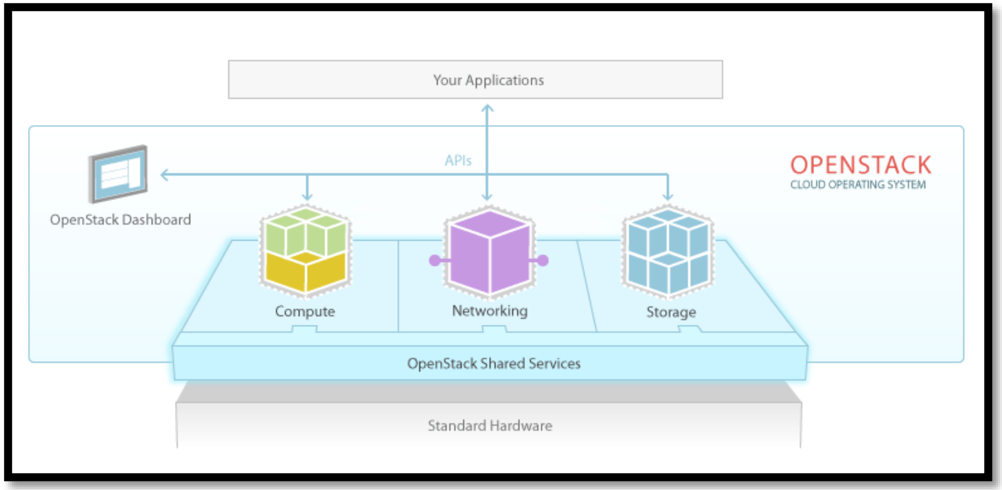
Christian Johnson
Advanced Systems Developer

OpenStack

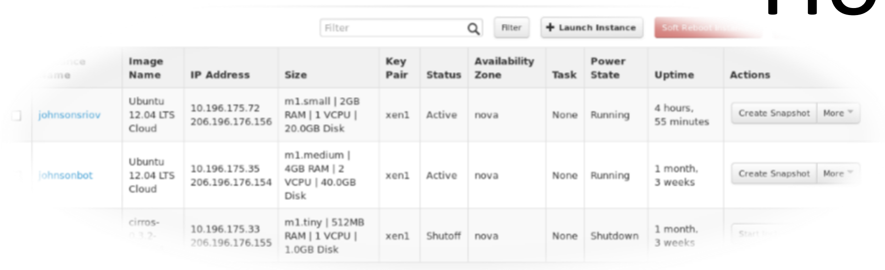
- Founded by Chris Kemp, NASA's first CTO, who was tasked with leading NASA's cloud initiatives in 2010. In collaboration with Rackspace, Kemp released OpenStack as one of the first open source solutions to "democratize web –scale computing" and provide a cloud solution via an Infrastructure as a Service (IaaS) architecture
- The OpenStack community is now comprised of over 18,000 developers, researchers, and corporations contributing to the source repository.

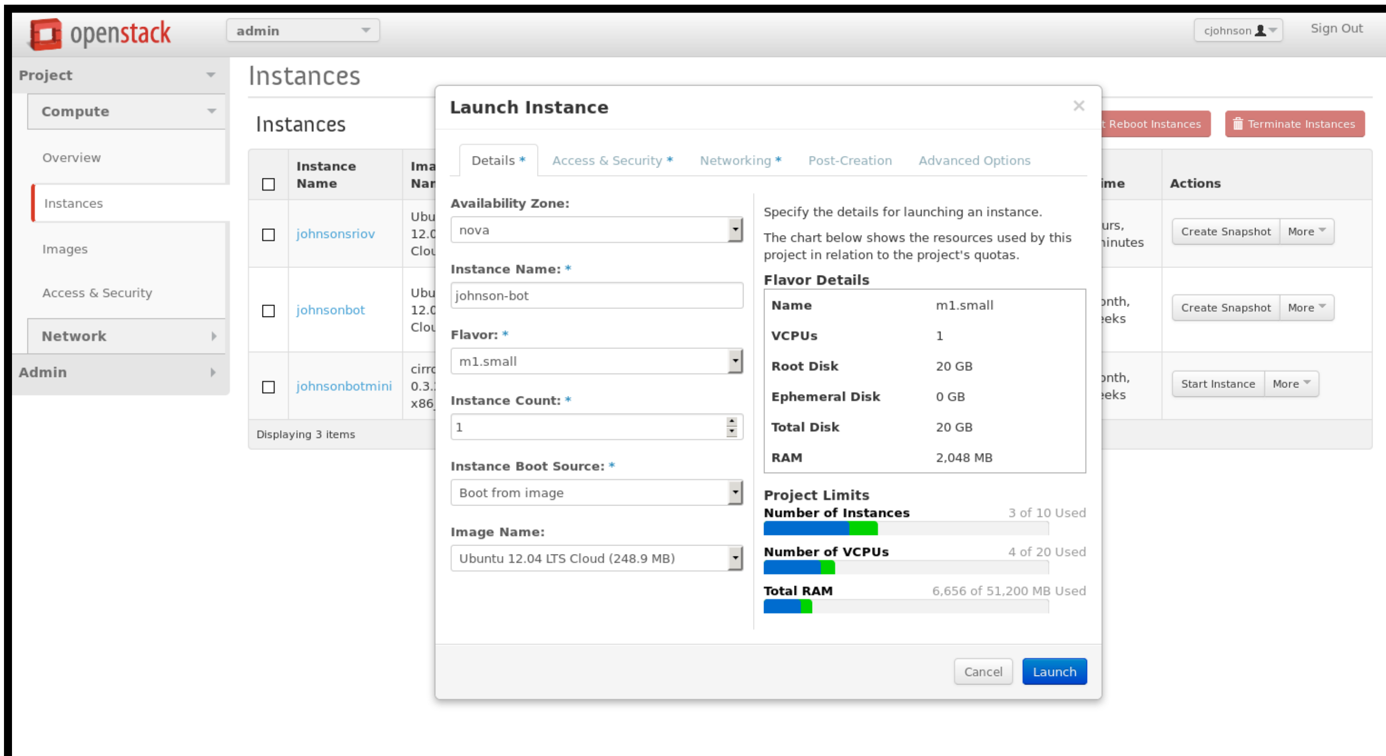
Software Configuration

- All machines using Ubuntu 14.04 LTS 5 year extended support.
- OpenStack Icehouse natively supported on 14.04
  - Python 2.7, Libvirt 1.2.2, Kernel 3.13
- Installed infrastructure services: neutron, nova, horizon, keystone, glance
- MySQL/apache/PHP on management controller



Easy VM Provisioning for Users with Horizon



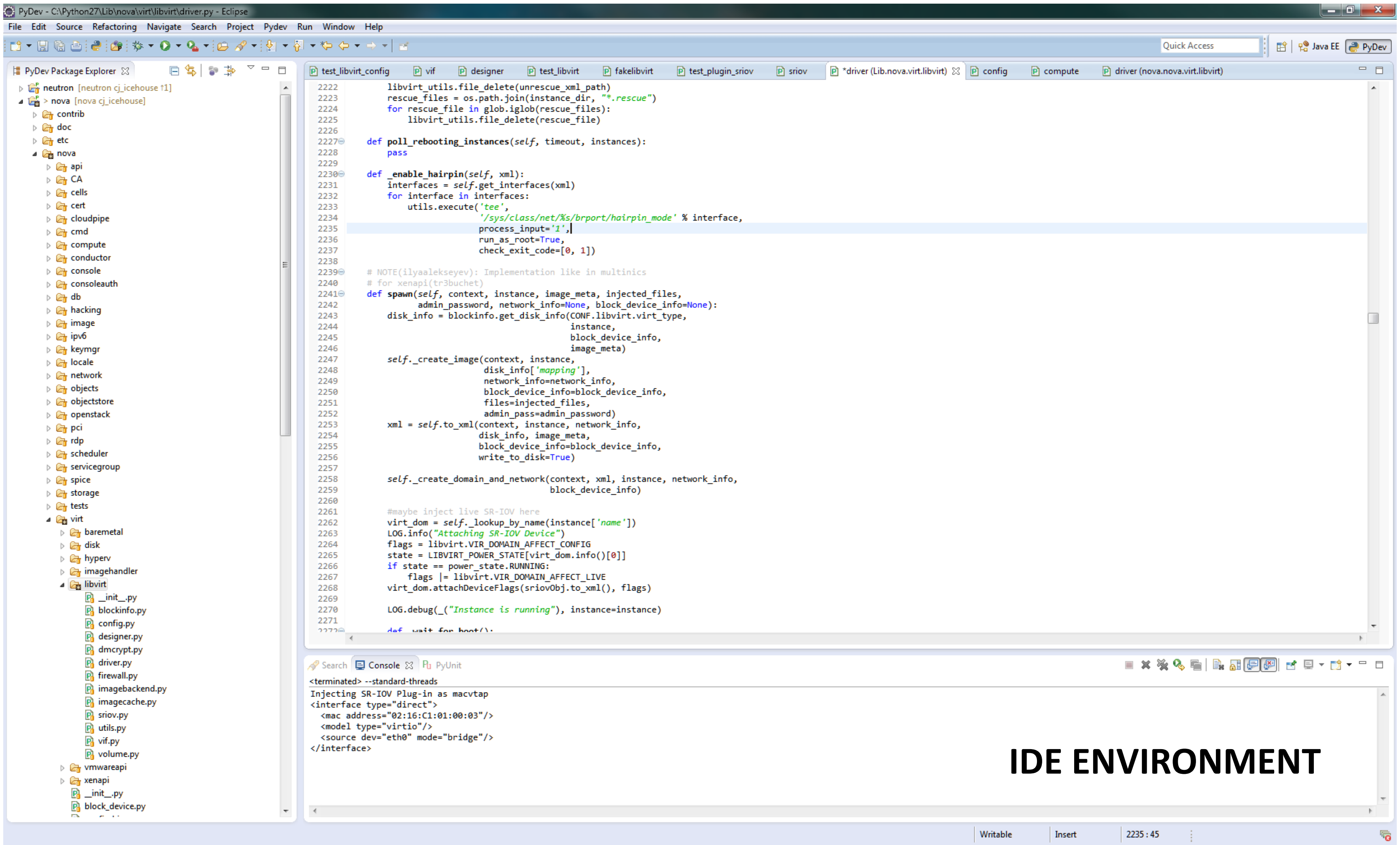


OpenStack Development

- Open source platform allows for code contributions, bug patching, modifications to core files, and plug-in development
- Source code can be cloned from git and modified directly from terminal / SSH, or with an IDE like Eclipse.
- Several pros and cons of OpenStack development

PROS	CONS
Interpreted language easy to error trace and debug. OpenStack has multiple logging objects for interactive logging and error tracing.	Core functionality is not always modular, customized development often requires direct modification to the core execution path to load plugins.
Configuration options and management easy to extend and implement via Oslo.	Code contributors develop source to meet every possible use case, often times overcomplicating core functionality and blueprints
Strong REST API endpoints for interaction with third party services and applications	Unit testing not always consistent, and not always easy to manage with dox software. (easy ways to avoid this)
New patches and code blue prints generally easy to follow (documentation for overall platform very inconsistent)	

Code management via Eclipse allows for simplification in automating code versioning via git, runtime configurations via PyDev extensions, and unit testing with PyUnit



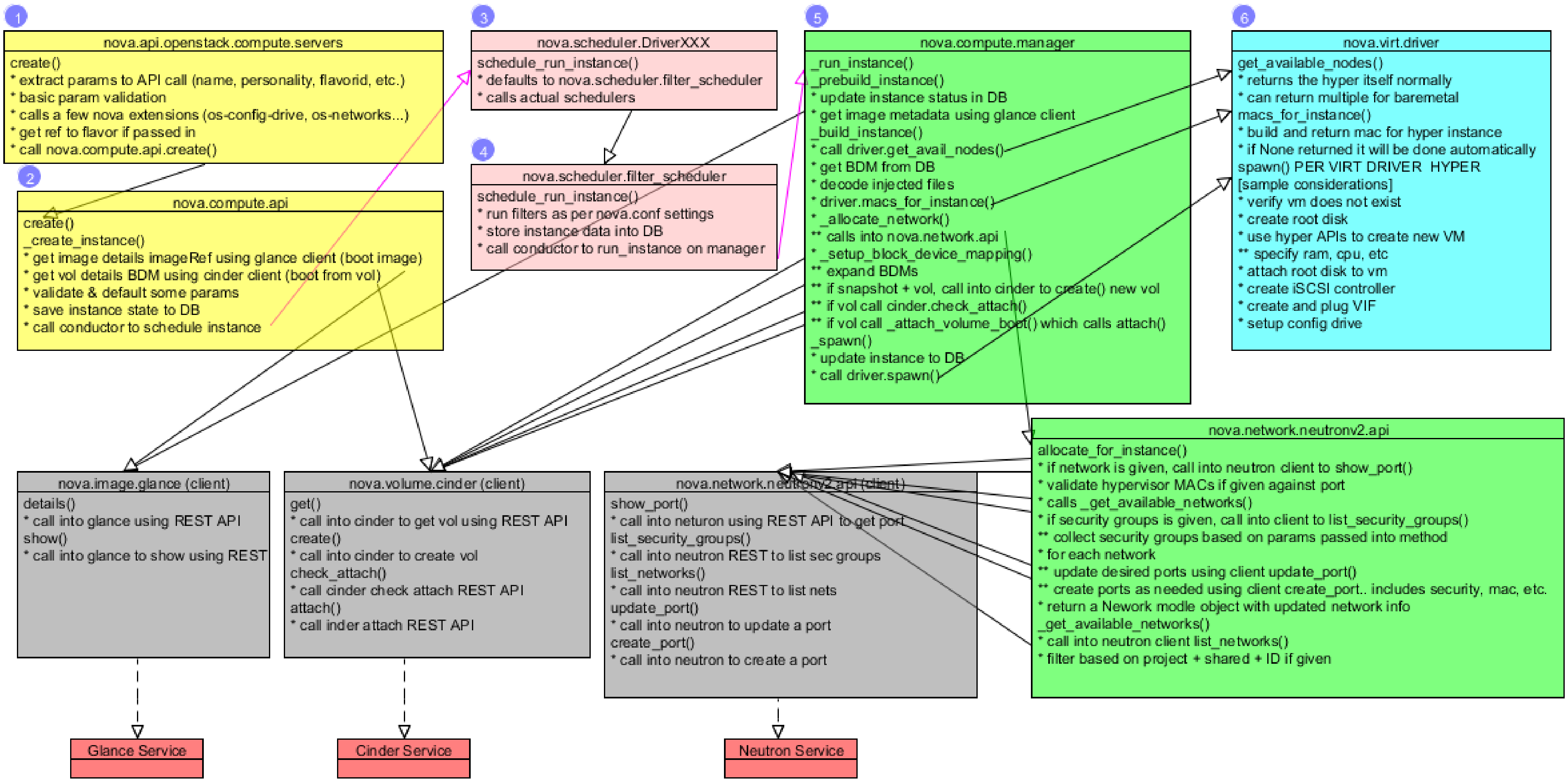
Development Goals

- As of the official Icehouse release, there is no official support via the OpenStack orchestration to support SR-IOV (Single Root I/O Virtualization)
- Goal: Develop easy to install plugin to provide libvirt functionality to instantly provision SR-IOV devices to tenant VMs
- Provide MACVTAP capability if SR-IOV hardware not available or PCI pass-through not supported by the hypervisor

Manipulating Libvirt

- The hundreds of thousands of lines that comprise the OpenStack source code only manage one file!
- Libvirt.xml used to describe and provide hardware virtualization parameters to libvirt/KVM virtualization library
- Libvirt supports SR-IOV and PCI pass-through, we only need to be concerned with the XML orchestration from OpenStack.
  - Assuming kernel, drivers, and modules are correctly supported and loaded in host to support VM capabilities

NOVA BOOT EXECUTION PATH



nova.virt.libvirt.driver

- Entry point for nova to provision and describe the VM instance in terms libvirt can understand.
  - Nova API -> scheduler -> manager -> driver -> config
- Plug-in object is instantiated via an import, and appropriate driver hooks to plugin objects
  - nova.virt.libvirt.sriov (plugin namespace)
- Plugin initialized with two lines of code:

```
from nova.virt.libvirt import sriov

sriovObj = sriov.SRIOV_Plugin()
```

nova.virt.libvirt.config

- Responsible for building libvirt.XML, and providing metadata/attributes needed for the driver to attach/detach devices during VM creation
- LibvirtConfigGuestInterface is the default class that manages adding/removing network devices

```
LibvirtConfigObject
  |
  + LibvirtConfigGuest
  + LibvirtConfigGuestDevice
      |
      +- LibvirtConfigGuestDisk
      +- LibvirtConfigGuestFilesys
      +- LibvirtConfigGuestInterface
      +- LibvirtConfigGuestInput
      +- LibvirtConfigGuestGraphics
      +- LibvirtConfigGuestChar
          |
          +- LibvirtConfigGuestSerial
          +- LibvirtConfigGuestConsole
```

MID-ATLANTIC CROSSROADS

LIBVIRT.XML

```
root@compute1: /var/lib/nova/instances/15307f1e-03d1-4f63-ab4a-0178a3121211
  GNU nano 2.2.6                         File: libvirt.xml

<domain type="kvm">
  <uuid>15307f1e-03d1-4f63-ab4a-0178a3121211</uuid>
  <name>instance-0000004a</name>
  <memory>2097152</memory>
  <vcpu>1</vcpu>
  <sysinfo type="smbios">
    <system>
      <entry name="manufacturer">OpenStack Foundation</entry>
      <entry name="product">OpenStack Nova</entry>
      <entry name="version">2014.1</entry>
      <entry name="serial">53d19f64-d663-a017-8922-003048da6e8a</entry>
      <entry name="uuid">15307f1e-03d1-4f63-ab4a-0178a3121211</entry>
    </system>
  </sysinfo>
  <os>
    <type>hvm</type>
    <boot dev="hd"/>
    <smbios mode="sysinfo"/>
  </os>
  <features>
    <acpi/>
    <apic/>
  </features>
  <clock offset="utc">
    <timer name="pit" tickpolicy="delay"/>
    <timer name="rtc" tickpolicy="catchup"/>
    <timer name="hpet" present="no"/>
  </clock>
  <cpu mode="host-model" match="exact"/>
  <devices>
    <disk type="file" device="disk">
      <driver name="qemu" type="qcow2" cache="none"/>
      <source file="/var/lib/nova/instances/15307f1e-03d1-4f63-ab4a-0178a3121211/disk"/>
      <target bus="virtio" dev="vda"/>
    </disk>
    <interface type="bridge">
      <mac address="fa:16:3e:54:ce:ad"/>
      <model type="virtio"/>
      <source bridge="qbr1889fbd6-1a"/>
      <target dev="tap1889fbd6-1a"/>
    </interface>
    <serial type="file">
      <source path="/var/lib/nova/instances/15307f1e-03d1-4f63-ab4a-0178a3121211/console.log"/>
    </serial>
    <serial type="pty"/>
    <input type="tablet" bus="usb"/>
    <graphics type="vnc" autoport="yes" keymap="en-us" listen="0.0.0.0"/>
    <video>
      <model type="cirrus"/>
    </video>
  </devices>
</domain>

                                        [ Read 52 lines ]
^G Get Help     ^O WriteOut     ^R Read File    ^Y Prev Page    ^K Cut Text     ^C Cur Pos
^X Exit         ^J Justify      ^W Where Is     ^V Next Page    ^U UnCut Text   ^T To Spell
```

Injection Routine

- Once plug-in hook is loaded, the initialization functions (spawn, resume, suspend, etc.) functions need to call the plug-in to attach SR-IOV device or MACVTAP interface

```
def attach(self, dom, flags, state, power_state, libvirt):
    LOG.info("Attaching " + self.SRIOVInterface.network_type + " Device")
    if state[dom.info()[0]] == power_state.RUNNING:
        flags |= libvirt.VIR_DOMAIN_AFFECT_LIVE
    dom.attachDeviceFlags(self.to_xml(), flags)
```

MACVTAP

- The current development platform does not have SR-IOV hardware as PCI pass through modules installed.
- To achieve similar capability, use MACVTAP
- Direct interface, uses virtio driver interface and bridges to network device
- Supports live migration of VMs

Results

- VM now has second device that has its provided route and increased performance capabilities (as it directly writes to hardware registers)

```
00:03.0 Ethernet controller: Red Hat, Inc Virtio network device
00:06.0 Ethernet controller: Red Hat, Inc Virtio network device
eth0      Link encap:Ethernet  HWaddr fa:16:3e:54:ce:ad
          inet addr:10.196.175.72  Bcast:10.196.175.255  Mask:255.255.255.0
          inet6 addr: fe80::f816:3eff:fe54:cead/64 Scope:Link
          UP BROADCAST RUNNING MULTICAST  MTU:1500  Metric:1
          RX packets:106419 errors:0 dropped:0 overruns:0 frame:0
          TX packets:107599 errors:0 dropped:0 overruns:0 carrier:0
          collisions:0 txqueuelen:1000
          RX bytes:11526199 (11.5 MB)  TX bytes:22722968 (22.7 MB)

eth1      Link encap:Ethernet  HWaddr 02:16:c1:01:00:03
          inet addr:206.196.176.170  Bcast:206.196.176.191  Mask:255.255.255.192
          inet6 addr: fe80::16:c1ff:fe01:3/64 Scope:Link
          UP BROADCAST RUNNING MULTICAST  MTU:1500  Metric:1
          RX packets:77658 errors:0 dropped:0 overruns:0 frame:0
          TX packets:299 errors:0 dropped:0 overruns:0 carrier:0
          collisions:0 txqueuelen:1000
          RX bytes:5649728 (5.6 MB)  TX bytes:24966 (24.9 KB)
```

Benefits of SR-IOV for HPC Networking

- Lower CPU utilization (by up to 50%)
- Lower network latency (by up to 50%)
- Higher network throughput (by up to 30%)
- Best suited for specialized workloads where high volume traffic is generated for HPC workloads.

Estimates based on performance reports by Microsoft Hyper-V team