MAX All-Hands Meeting

September 19, 2013

Executive Director's Address

Tripti Sinha

MAX All-Hands Meeting
September 19, 2013
5825 University Research Court, Suite 2500
College Park, Maryland 20740

8:15 am	**Registration and Breakfast**	
9:00 am	**Executive Director's Address**	Tripti Sinha *Executive Director, MAX*
9:40 am	**MAX Network Update** *Status of network refresh, 100G upgrade, and new PoPs.*	Tom Lehman *Director of Research, MAX*
10:15 am	**Services Update** *Summary of MAX services portfolio.*	Tom Lehman
10:30 am	**Refreshment Break**	
10:45 am	**Innovation and Advanced Services** *Update on SDNX and AWS Direct Connect.*	Jarda Flidr *Director of Services, MAX*
11:30 am	**Keynote Address**	Brian Voss *Vice President and CIO* *University of Maryland*
12 noon	**Lunch**	
1:00 pm	**Sponsored Research Projects** *Update on NSF, DOE, and DOD supported research activities.*	Tom Lehman
2:00 pm	**Participant Forum**	
3:00 pm	**Closing Remarks**	Tripti Sinha

MAX All-Hands Meeting

September 19, 2013

Introductions

Welcome to our participants and guests!

Federal Labs and Agencies

Army Research Lab
D.C. Government (through Columbia Telecommunications Corporation)
Energy Sciences Network (ESnet)/U.S. Department of Energy
Laboratory for Telecommunications Sciences
Library of Congress
NASA Goddard Space Flight Center
National Archives and Records Administration (NARA)
National Institutes of Health (NIH)
National Institute of Standards and Technology (NIST)
National Library of Medicine (NLM)
National Oceanic and Atmospheric Administration (NOAA)
National Science Foundation (NSF)
Space Telescope Science Institute (STSI)
U.S. Department of State (through GWU)
U.S. Department of Transportation (Federal Highway Administration)

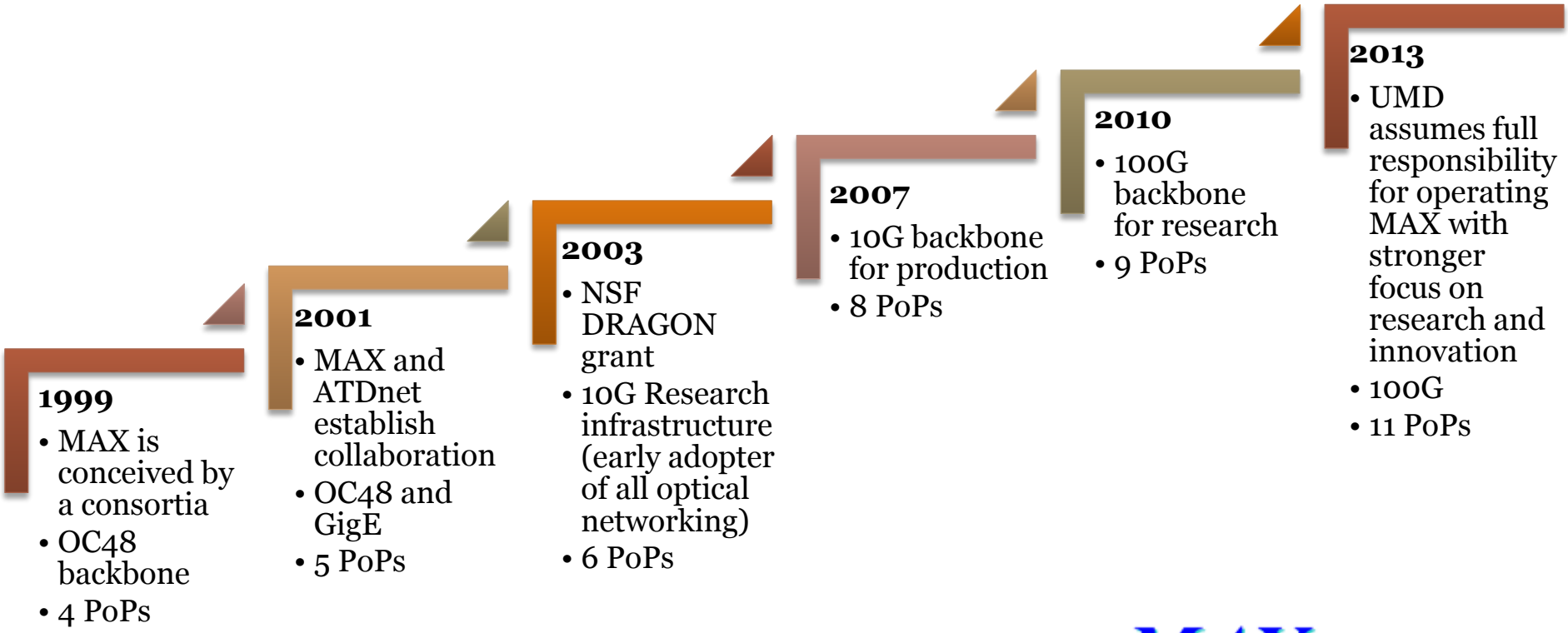
Corporate and Non-Profit

Columbia Telecommunications Corporation (CTC)
Howard Hughes Med. Institute
J. Craig Venter Institute
Northrop Grumman Corporation

Higher Education

American University
Baltimore Education & Research Network
Catholic University
Gallaudet University
GEANT / DANTE
Georgetown University
George Washington University
Internet2, D.C. Office / University Consortium for Advanced Internet Development (UCAID)
Johns Hopkins University
Johns Hopkins University – Applied Physics Laboratory (JHU-APL)
Johns Hopkins University Medical Institutions
National Center for Supercomputing Applications (NCSA) / ACCESS
Network Virginia (aggregating the state of Virginia)
Smithsonian Institution
University of Maryland, Baltimore
University of Maryland, Baltimore County
University of Maryland, College Park
University of Maryland, University College
University System of Maryland Network (aggregating 11 campuses)
University of Southern California, Information Sciences Institute – East
Washington Research Library Consortium

Mid-Atlantic Crossroads (MAX)



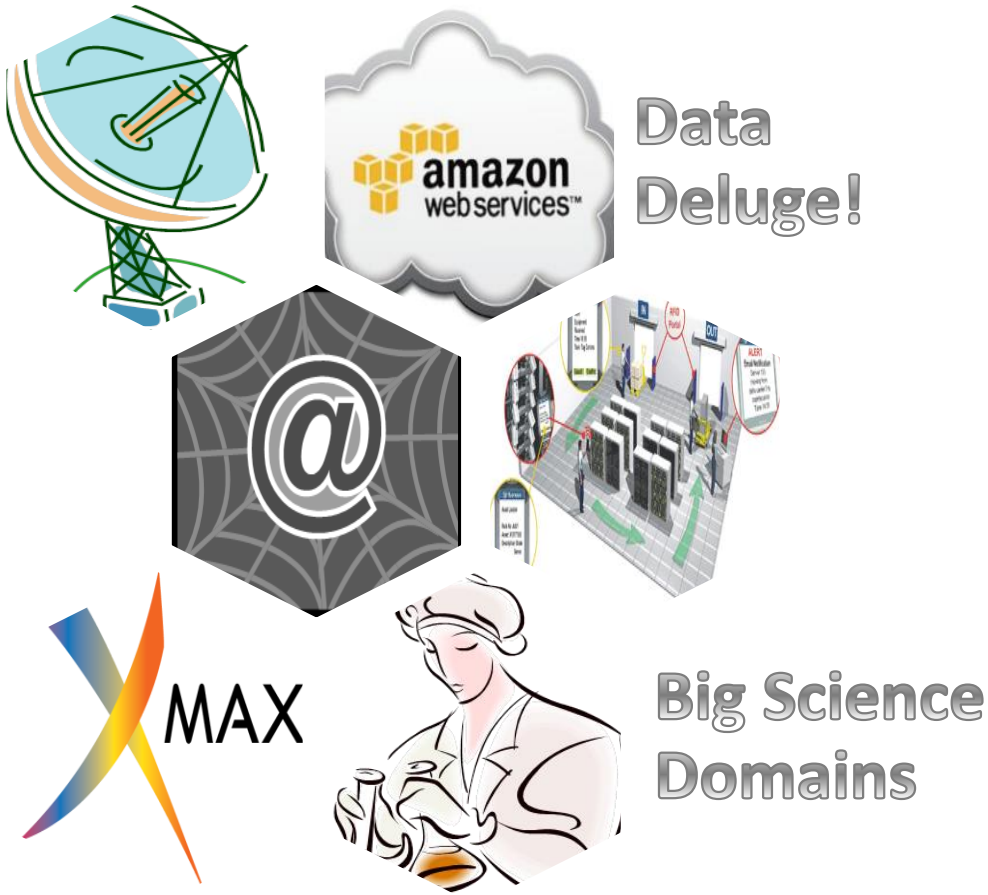
MAX IN 2013

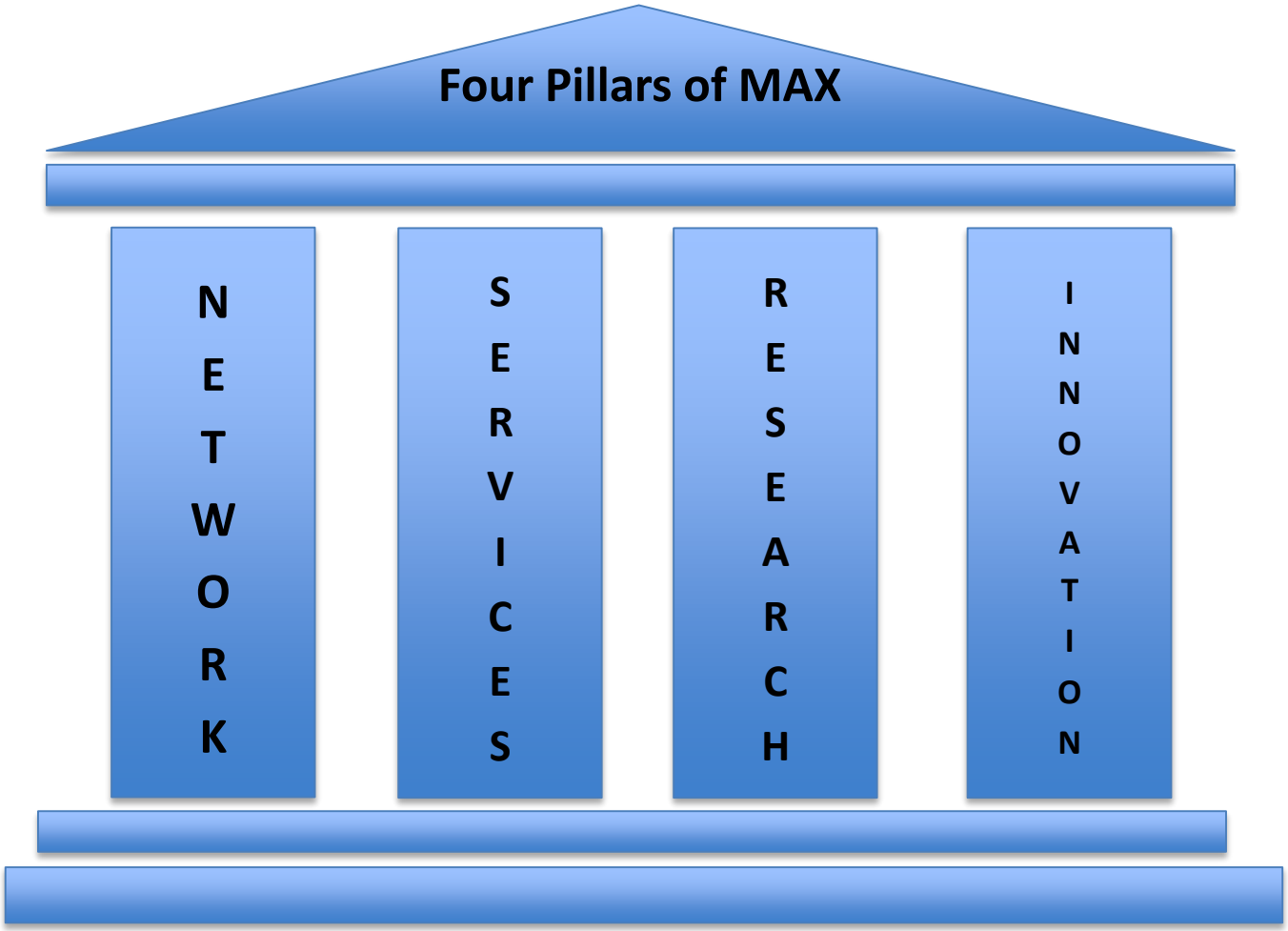
MOTIVATORS

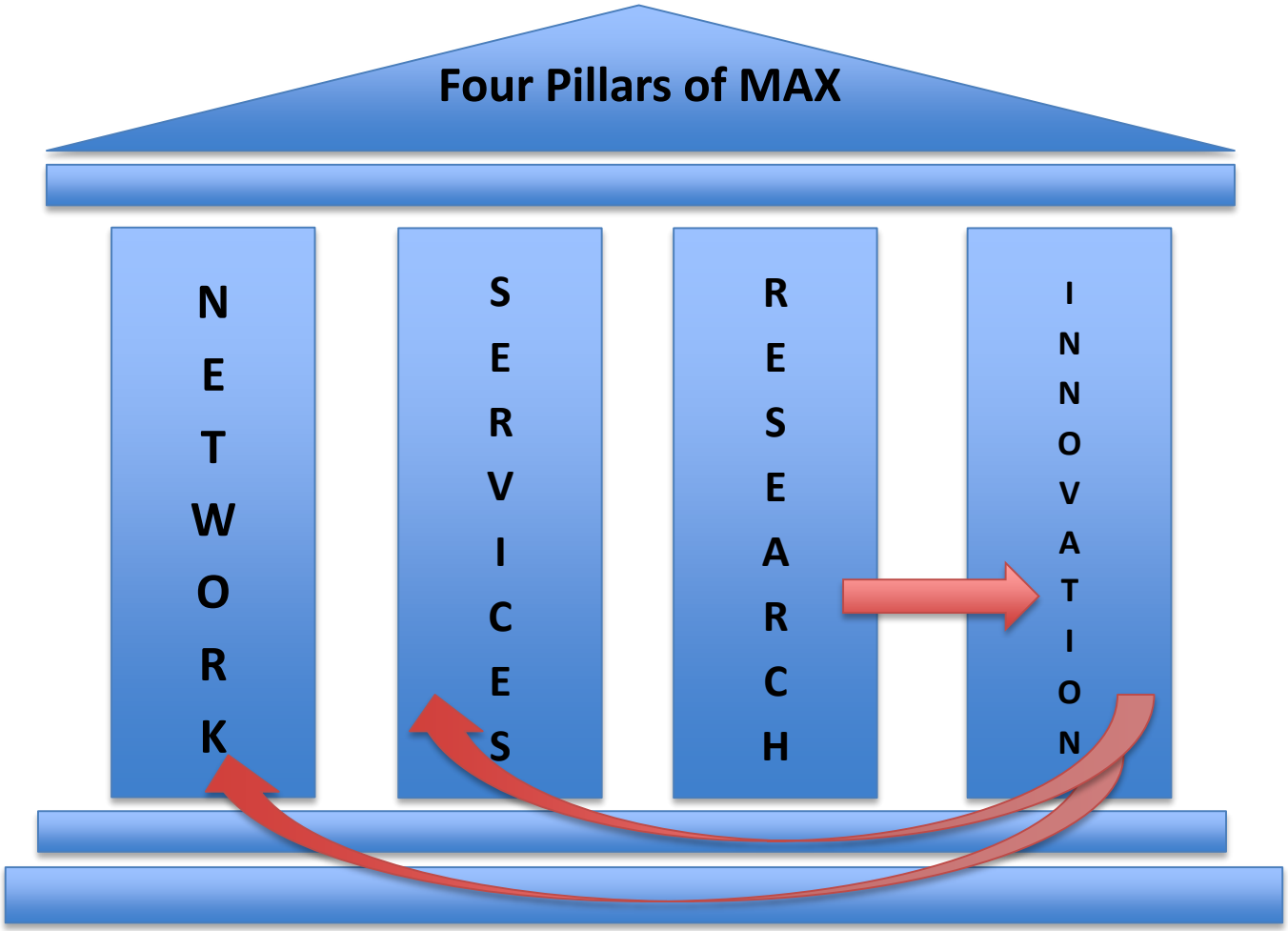
- Regional Cooperation
- Bandwidth
- R&E Networking

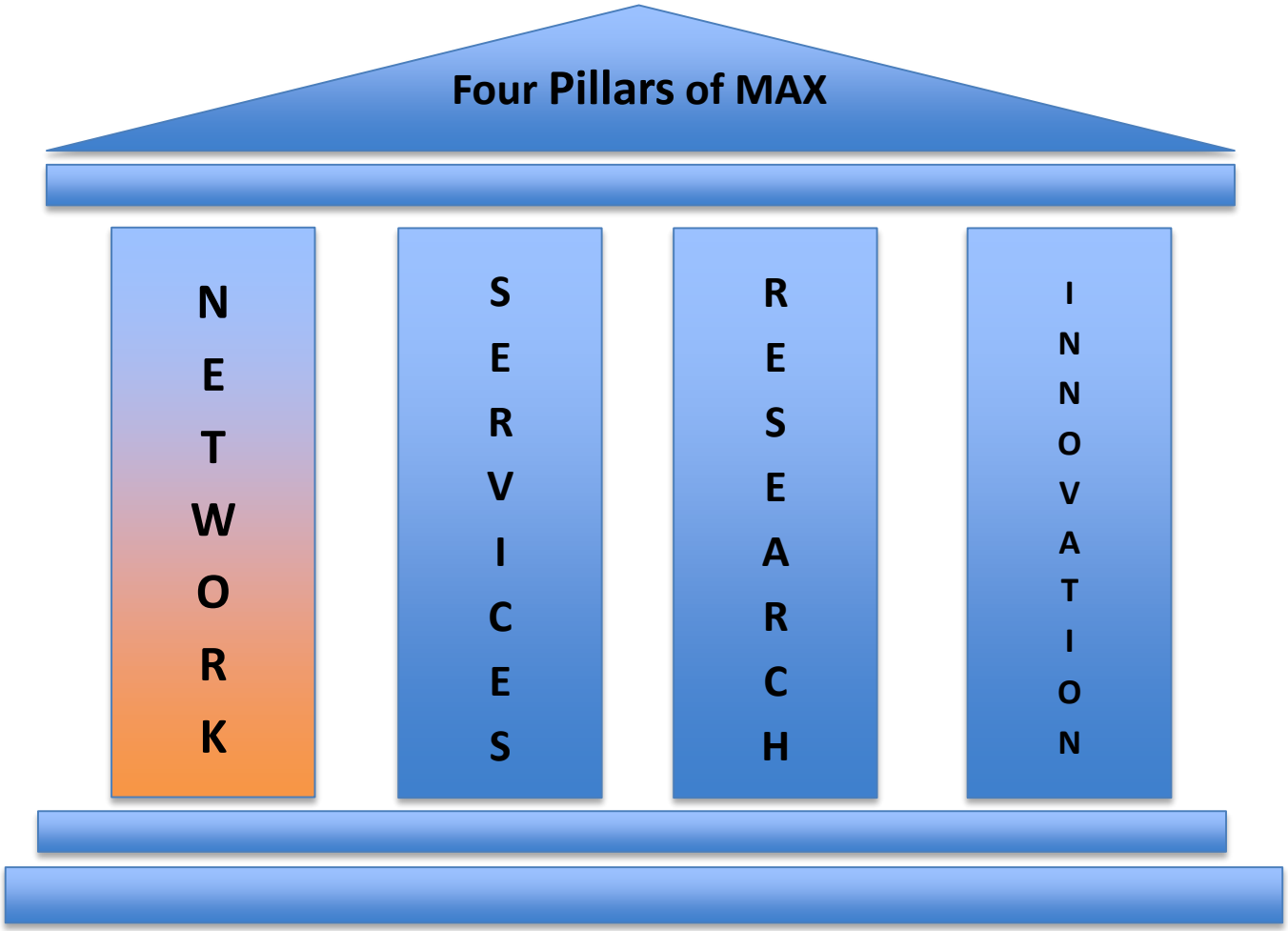
- Enable domain sciences
- Innovate
- Integrate innovations

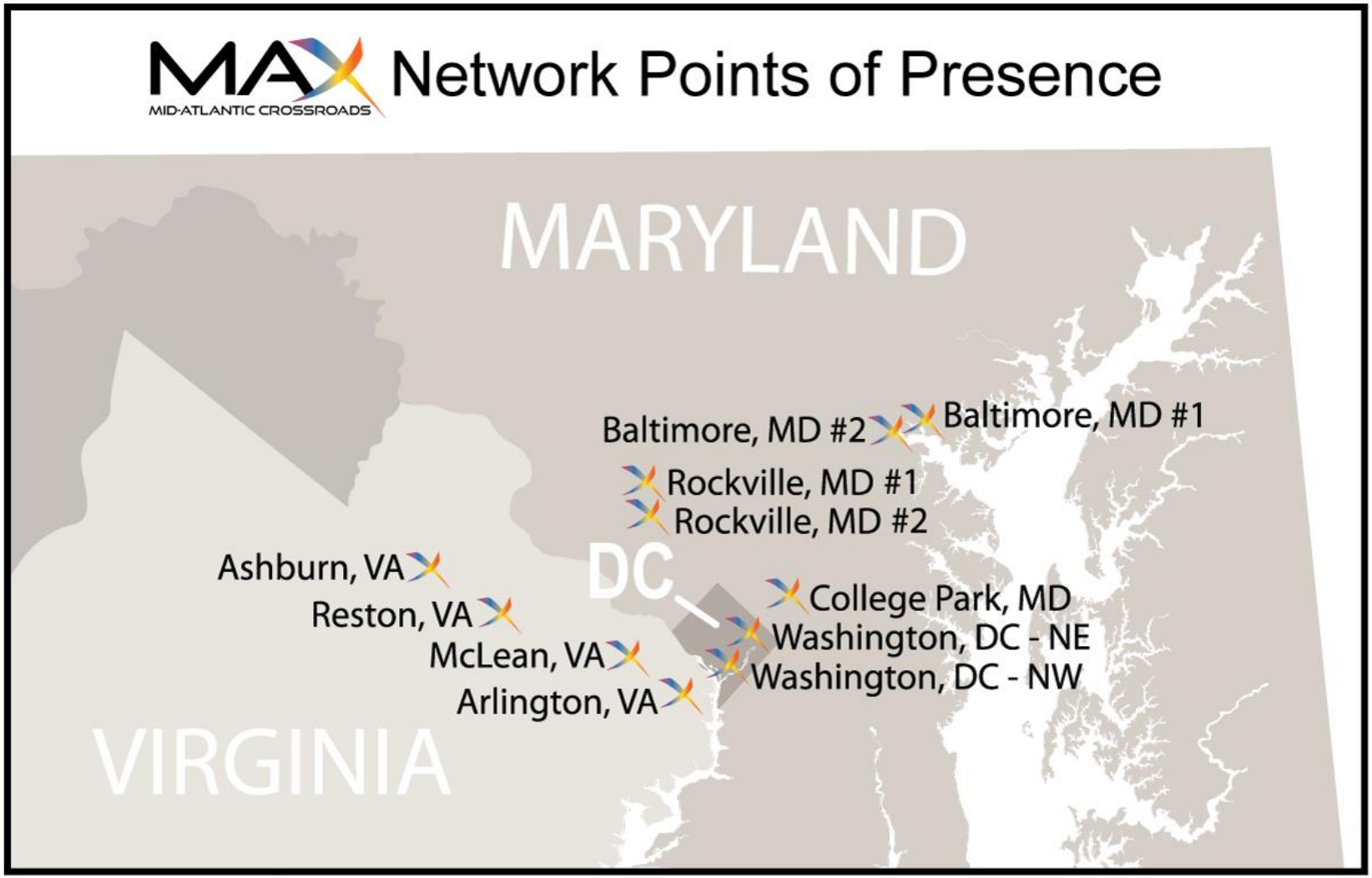
Today's world is complex!




Four Pillars of MAX
NETWORK
SERVICES
RESEARCH
INNOVATION

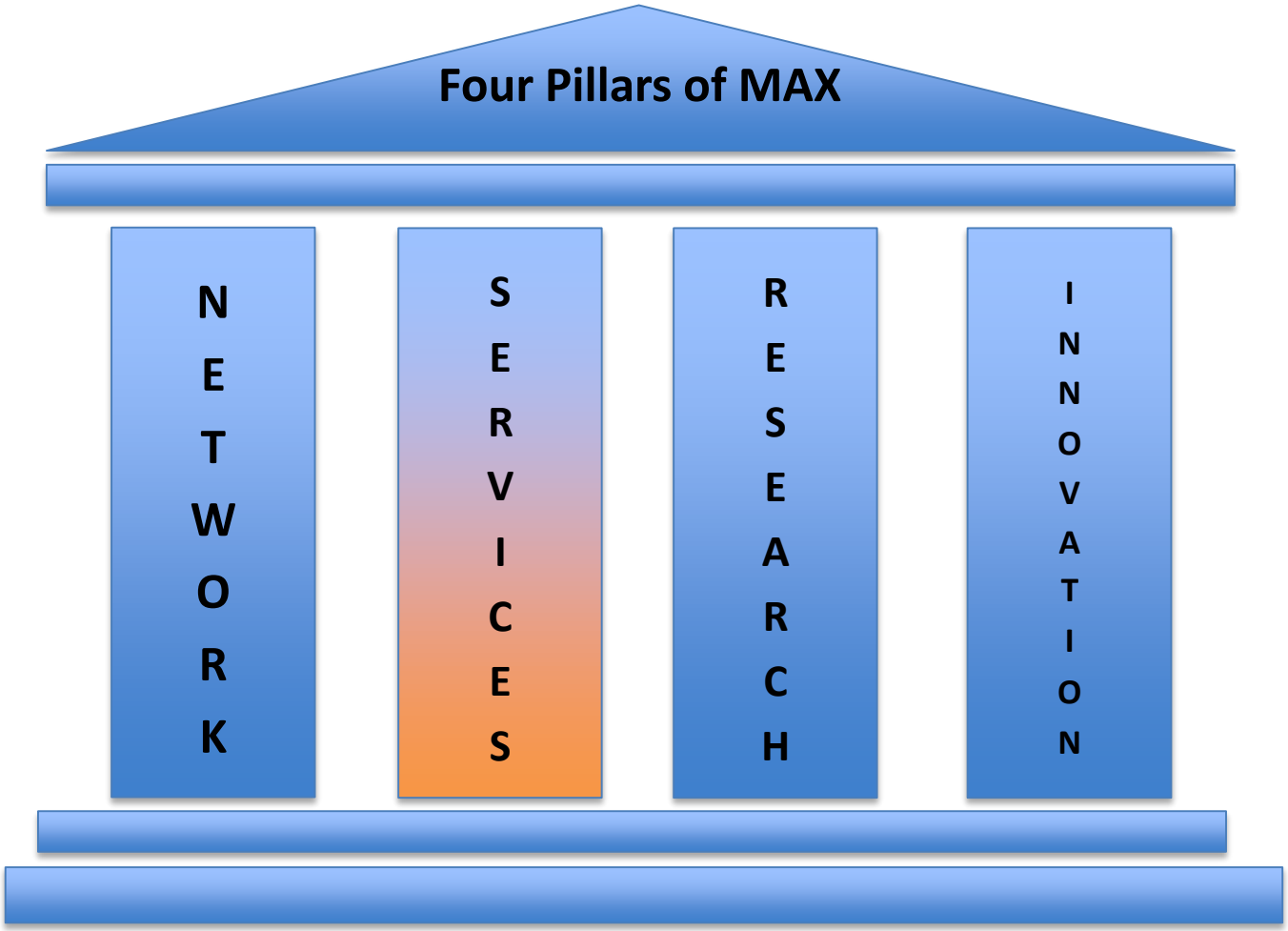

Four Pillars of MAX
NETWORK
SERVICES
RESEARCH
INNOVATION


Four Pillars of MAX
NETWORK
SERVICES
RESEARCH
INNOVATION

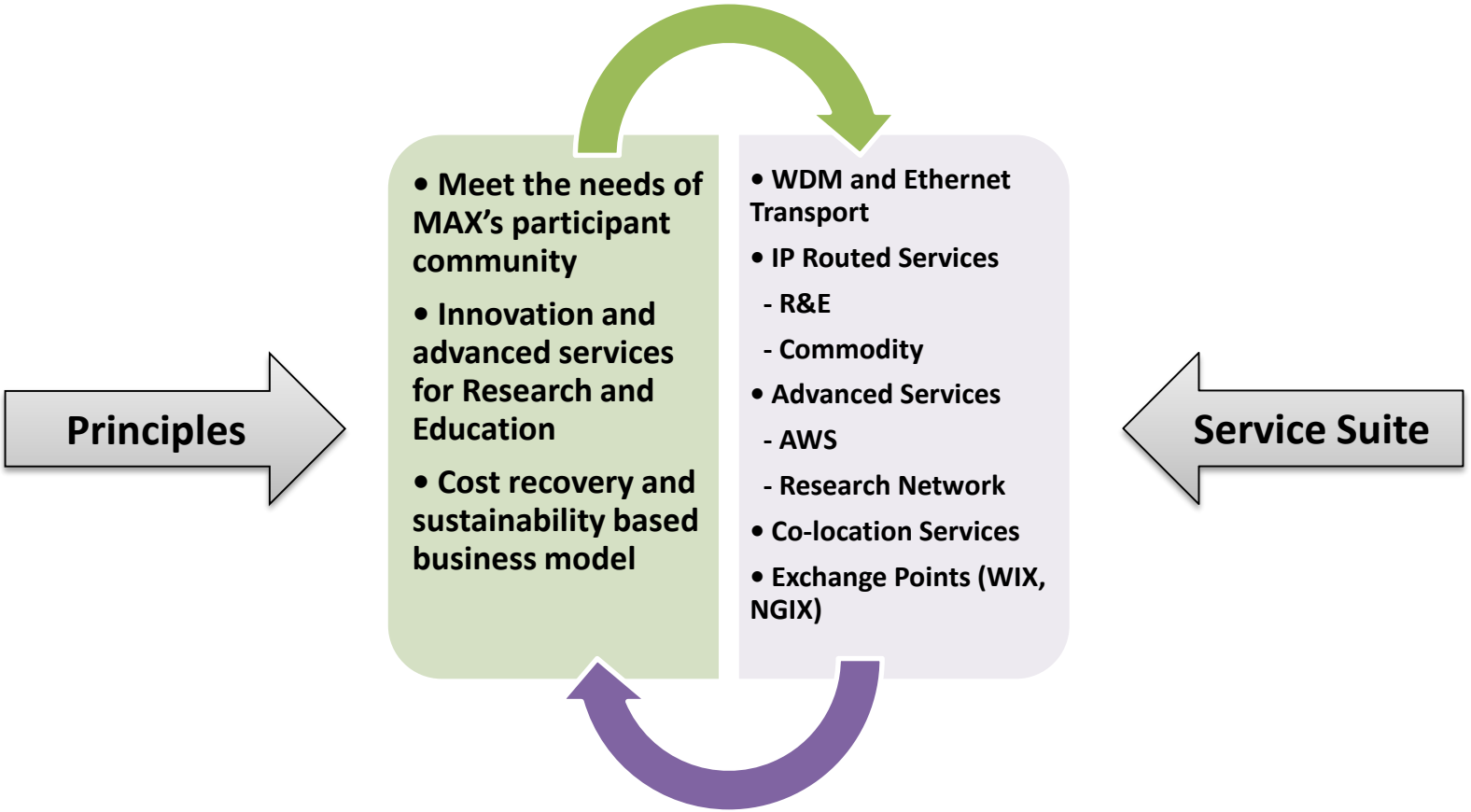

MID-ATLANTIC CROSSROADS
Network Points of Presence
MARYLAND
Baltimore, MD #2
Baltimore, MD #1
Rockville, MD #1
Rockville, MD #2
Ashburn, VA
DC
Reston, VA
College Park, MD
McLean, VA
Washington, DC - NE
Arlington, VA
Washington, DC - NW
VIRGINIA

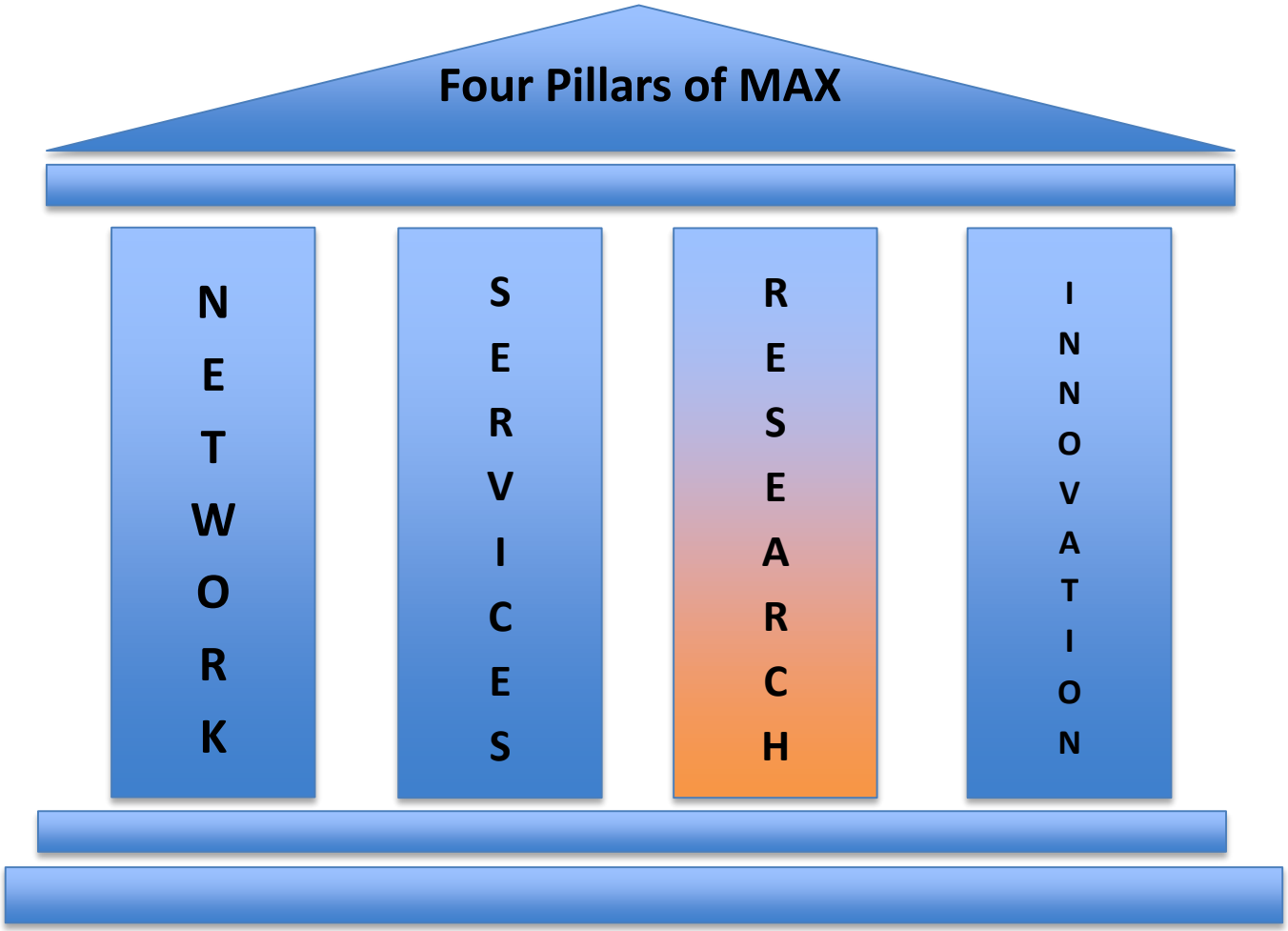
Network

- A footprint in Maryland, Virginia and Washington DC defined by 11 PoPs
- Network Refresh – consolidated production and research 100G DWDM Fujitsu infrastructure
- Professionally monitored by UMD's NOC
- Embedded infrastructure in support of advanced and innovative services

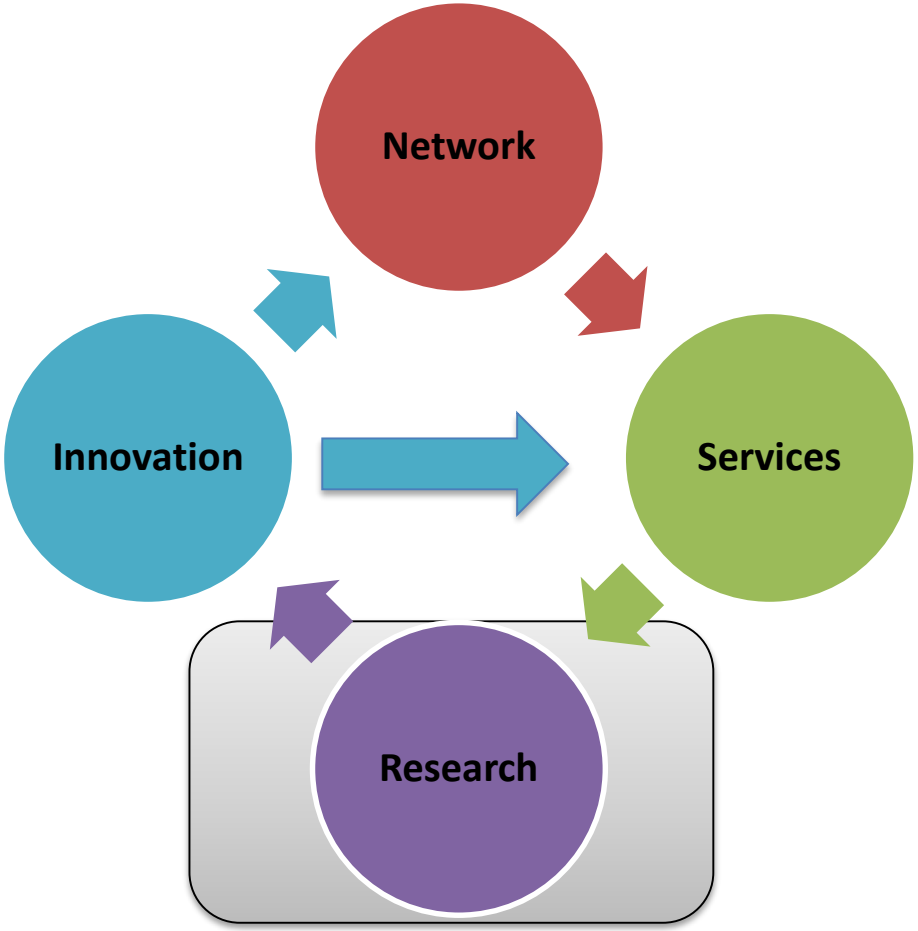

Four Pillars of MAX
NETWORK
SERVICES
RESEARCH
INNOVATION

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


Four Pillars of MAX
NETWORK
SERVICES
RESEARCH
INNOVATION

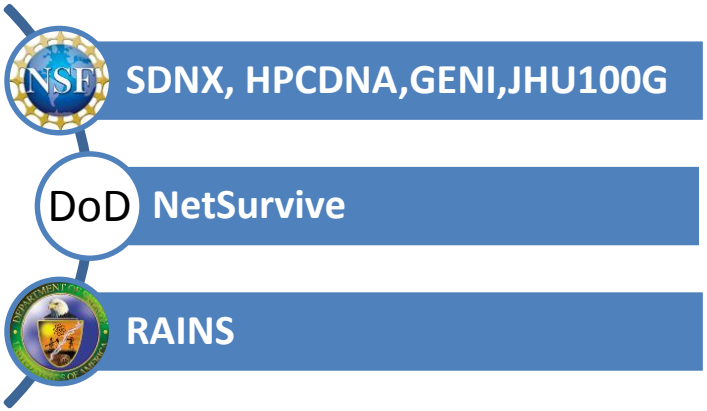
The cycle of innovation and advanced services

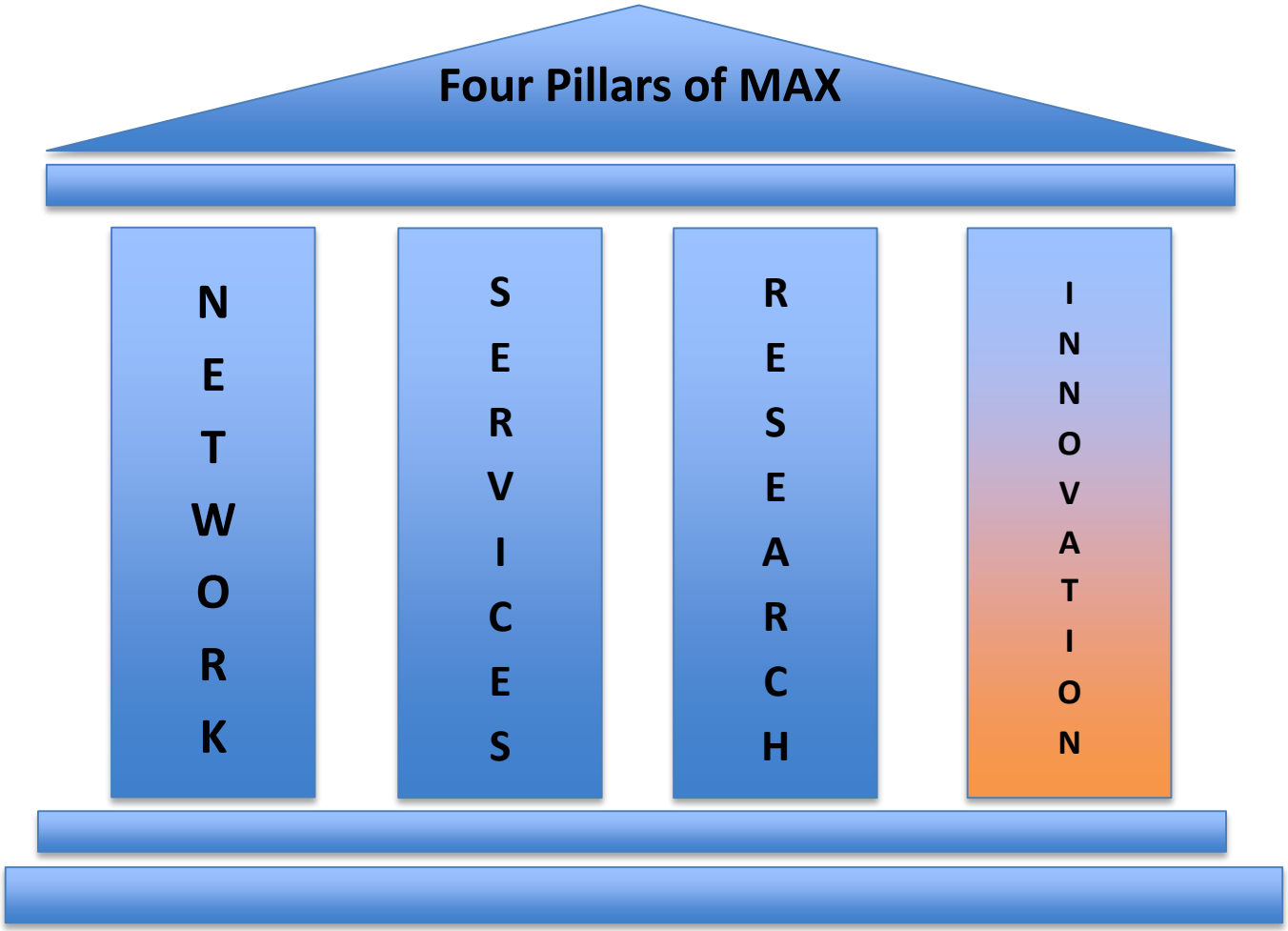


Research

The state of MAX's sponsored research activities is vibrant!

Tom Lehman is at the helm of research

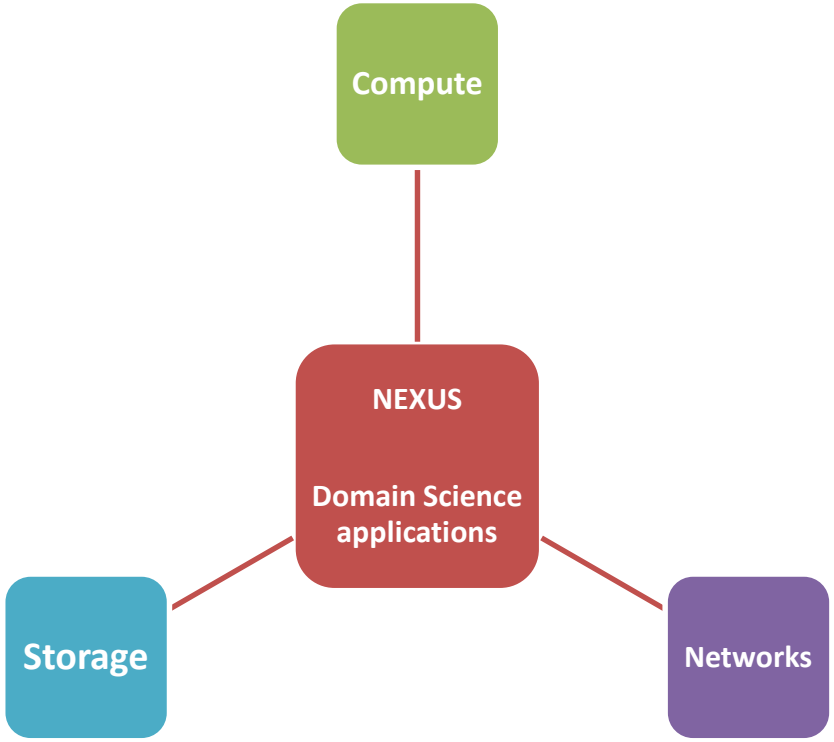



Four Pillars of MAX
NETWORK
SERVICES
RESEARCH
INNOVATION

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

MID-ATLANTIC CROSSROADS

Questions?

MAX Network Update

Status of network refresh, 100G upgrade, and new Points of Presence (PoPs)

Tom Lehman
Director of Research

Network Refresh

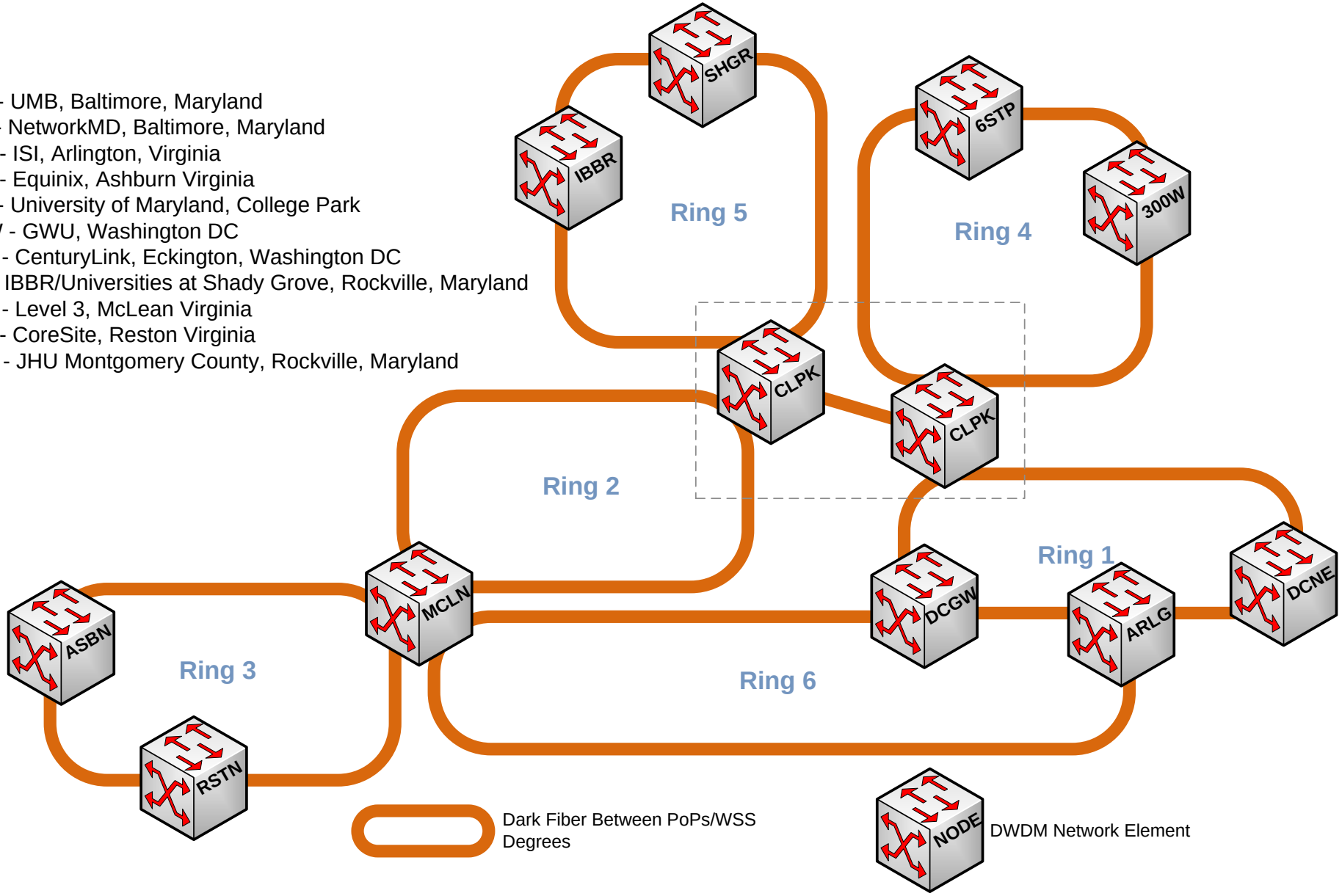
- Upgrading Layer 1, Layer 2, and Layer 3 to 100 Gbps capable
  - Initial Layer 1 100G DWDM Refresh funded by NSF Infrastructure Grant (Aug 2010- Jan 2013)
  - MAX is now building on that infrastructure
- Connecting to Internet2 at 2 x 100G
  - Layer 3/Layer 2 Combo Port
  - AL2S and R&E IP Routed Infrastructure

MAX Network Overview

- Dark Fiber based network
- 11 Points of Presence (PoPs) across Maryland, DC, Virginia
- Dense Wave Division Multiplexing (DWDM) based
- Mix of 10G and 100G Network Equipment currently, evolving to all 100G capable soon for core components (DWDM, switches, routers)
- Interconnects to wide area infrastructures
- Network Services: Layer 1, Layer 2, and Layer 3
- Advanced services (AWS, Research Network, SDN)

MAX Fiber/DWDM Footprint

300W - UMB, Baltimore, Maryland
6STP - NetworkMD, Baltimore, Maryland
ARLG - ISI, Arlington, Virginia
ASBN - Equinix, Ashburn Virginia
CLPK - University of Maryland, College Park
DCGW - GWU, Washington DC
DCNE - CenturyLink, Eckington, Washington DC
IBBR - IBBR/Universities at Shady Grove, Rockville, Maryland
MCLN - Level 3, McLean Virginia
RSTN - CoreSite, Reston Virginia
SHGR - JHU Montgomery County, Rockville, Maryland



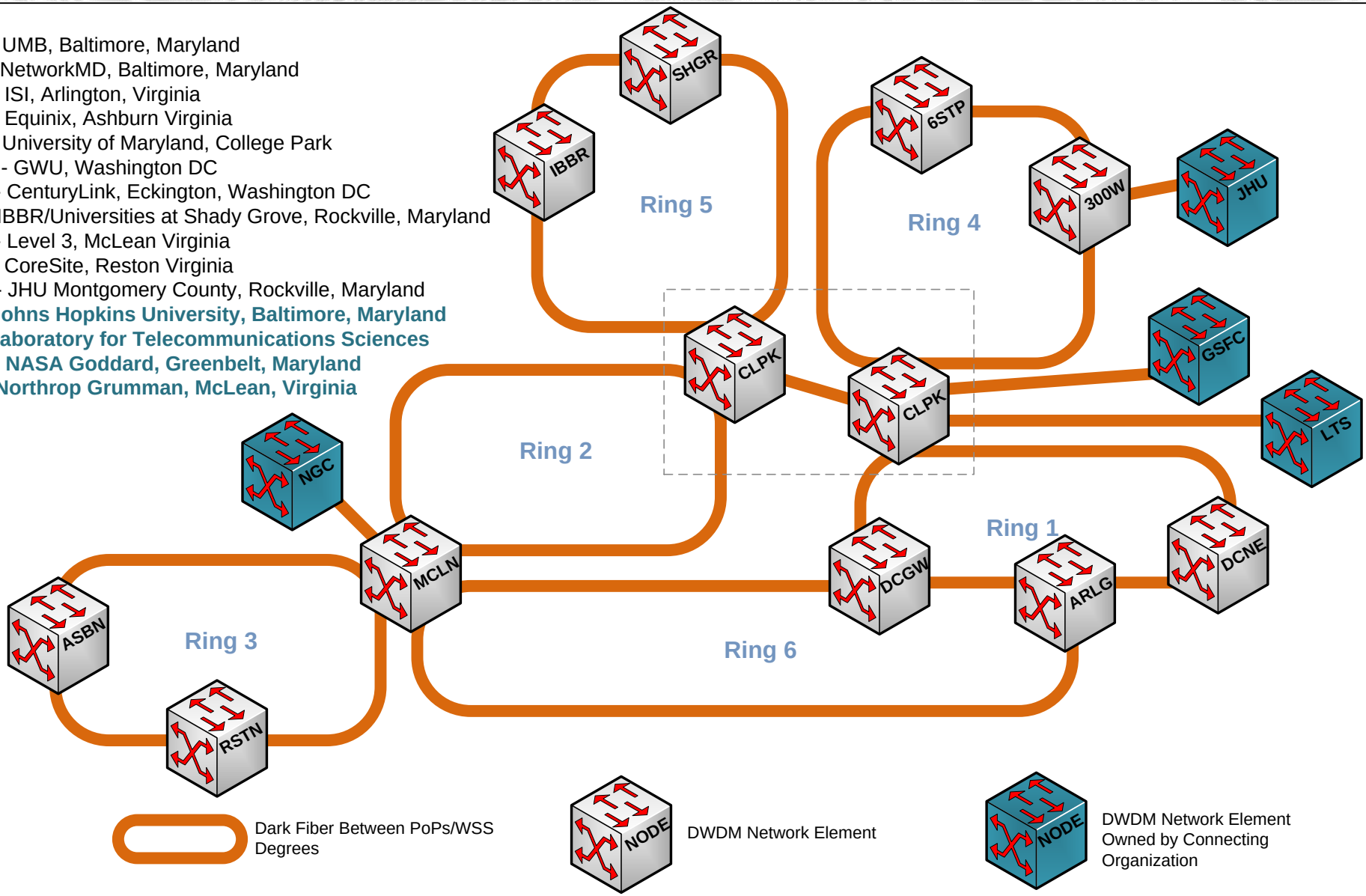
Recently Added PoPs

- Shady Grove Life Sciences Center Area
- IBBR (Institute for Bioscience and Biotechnology Research) in Rockville, Maryland
  - Deployed and in service
- Johns Hopkins University Montgomery County Campus
  - partnership with JHU
  - equipment deployed, scheduled to be online in October 2013

MAX Fiber/DWDM Footprint

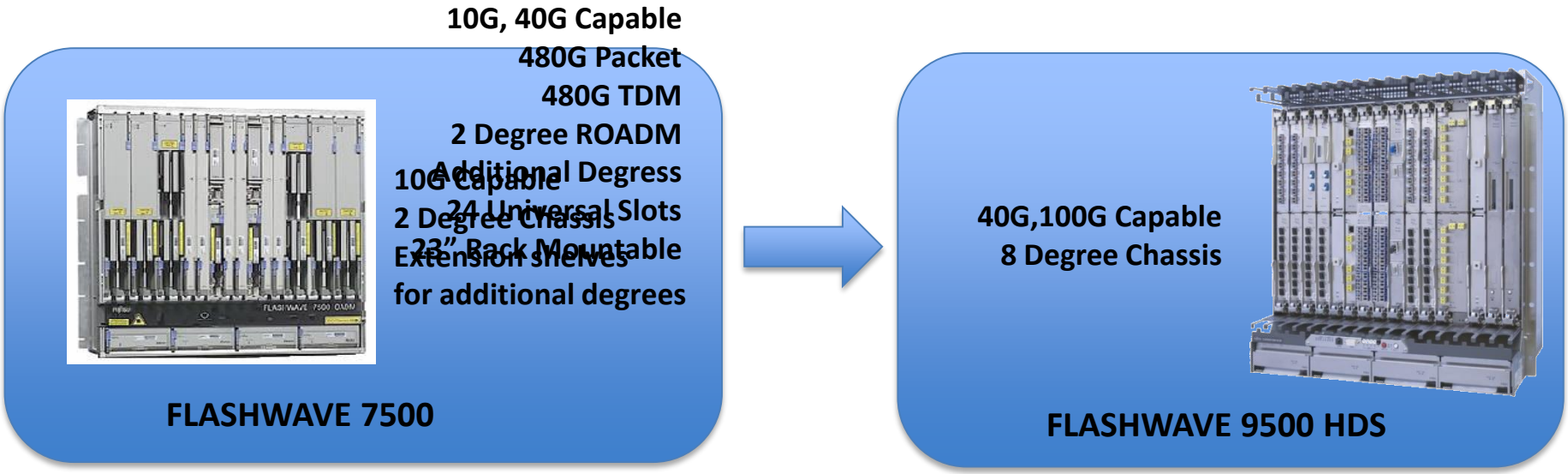
MID-ATLANTIC CROSSROADS

300W - UMB, Baltimore, Maryland
6STP - NetworkMD, Baltimore, Maryland
ARLG - ISI, Arlington, Virginia
ASBN - Equinix, Ashburn Virginia
CLPK - University of Maryland, College Park
DCGW - GWU, Washington DC
DCNE - CenturyLink, Eckington, Washington DC
IBBR - IBBR/Universities at Shady Grove, Rockville, Maryland
MCLN - Level 3, McLean Virginia
RSTN - CoreSite, Reston Virginia
SHGR - JHU Montgomery County, Rockville, Maryland
JHU - Johns Hopkins University, Baltimore, Maryland
LTS - Laboratory for Telecommunications Sciences
GSFC - NASA Goddard, Greenbelt, Maryland
NGC - Northrop Grumman, McLean, Virginia



DWDM Elements

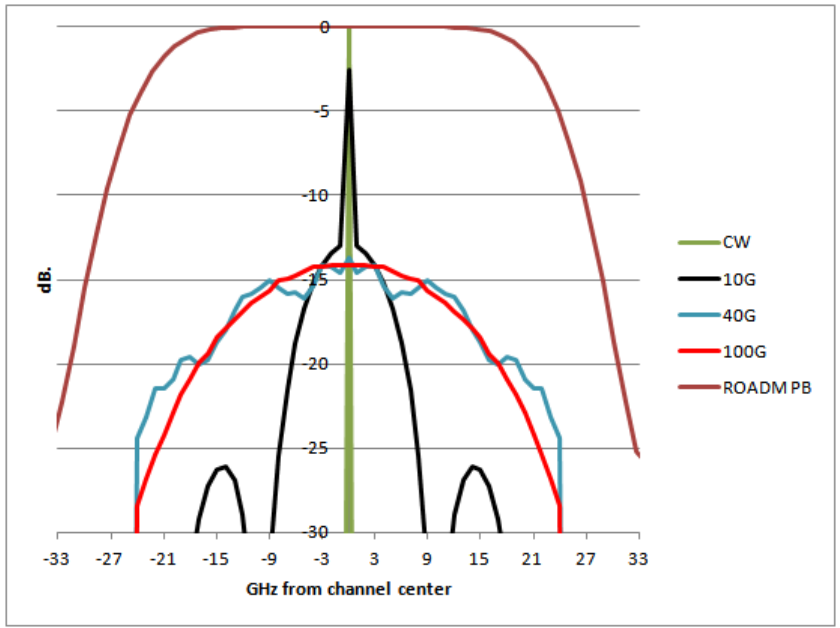
- Currently a mix of Fujitsu 7500s and 9500s
- 10 9500s in the network today, expect to have around 14 when complete with refresh
- Evolving to all 9500s by early 2014



- Fujitsu 9500 supports up to 88 lambdas per degree
- 100G Transponders used on MAX
- 10x10G Muxponders also to be used

FLASHWAVE® 9500 100G Transponder

- Client interface
  - 100GBASE-LR4, 100GBASE-SR10, or 10x10MSA
- Network interface
  - Full-Band Tunable DWDM single-carrier DP-QPSK
  - Spectral width compatible with 50 GHz spaced ROADM.
    - Allows switching/sharing of 100G wavelengths by ROADM
    - Operates with or without Dispersion Compensation
  - Can mix 10G and 100G wavelengths on same fiber



So what do we do with this advanced multiple 100G DWDM Network?

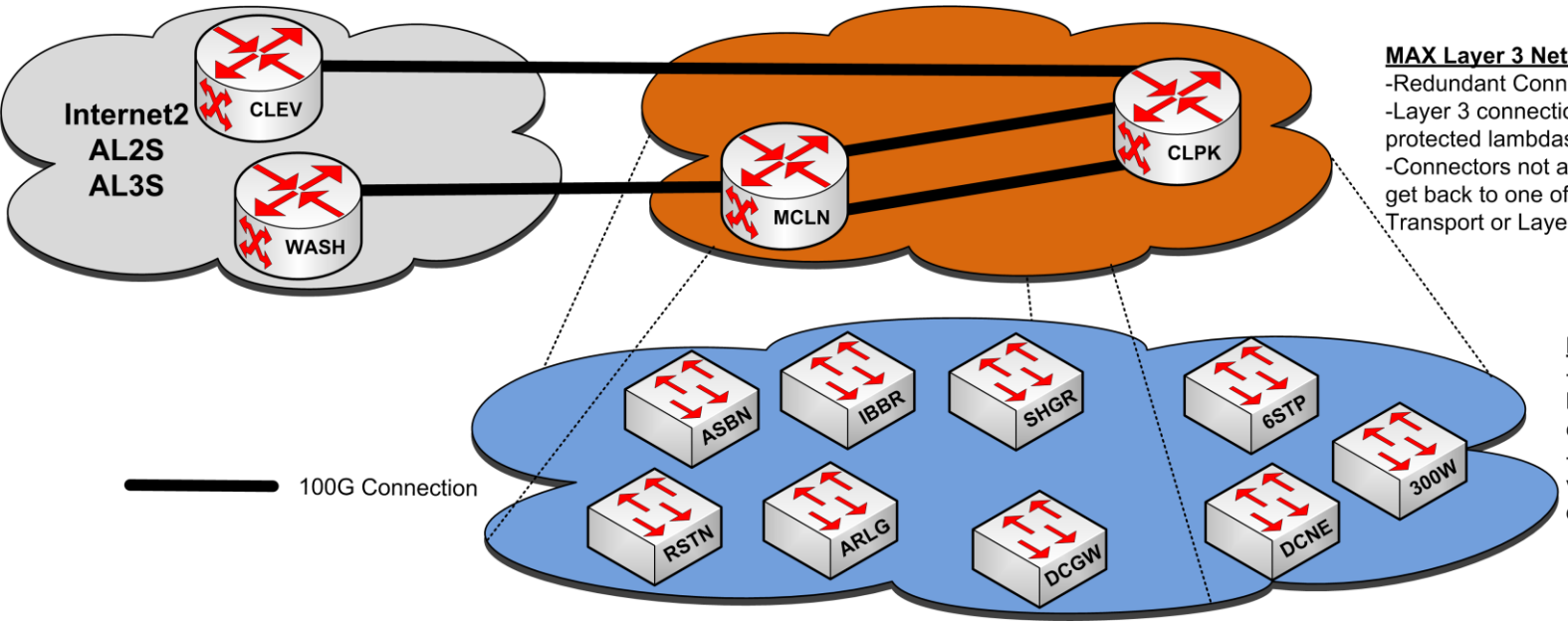
- Production Network
  - Layer 3, Layer 2, and Layer 1 Services
  - Advanced Services (AWS Direct Connect, MSX)
- Research Network
  - OpenFlow, Software Defined Networking, Dynamic Circuit Network/ION Service/AL2S services
  - MAX Network as a GENI Aggregate, MAX InstaGENI Rack

Production Network

- Layer 2 Network Elements
  - Juniper EX4550 at most PoPs
  - Juniper MX960 at College Park and McLean
- Layer 3 Network Elements
  - Juniper MX960 at College Park and McLean
- Juniper MX960
  - 100 Gbps capable
  - replacing the current Juniper T640s
- Juniper Equipment ordered, deliveries ongoing
  - installs to begin this month

Production Network

Layer 3 / Layer 2 View

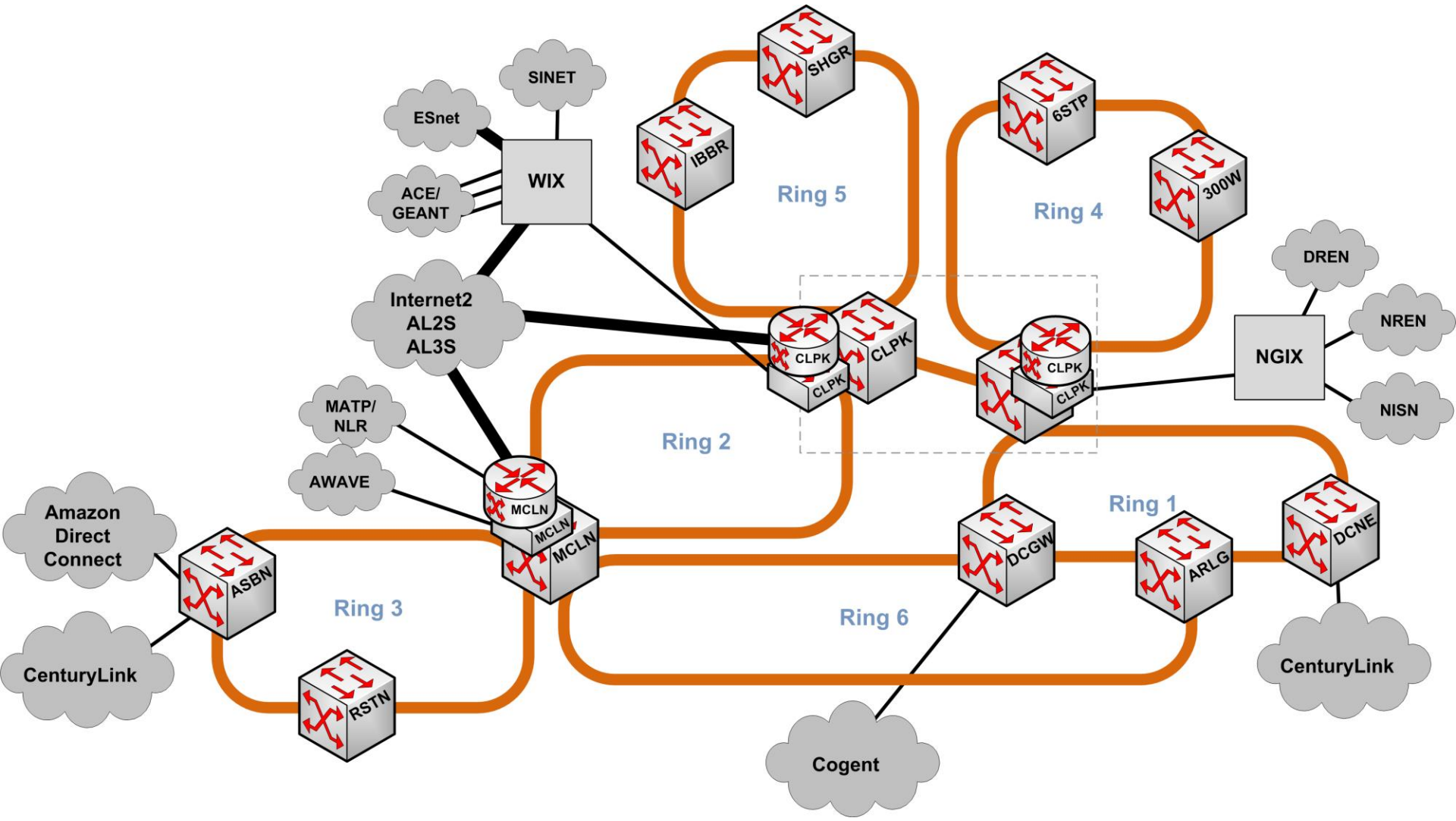


MAX Layer 3 Network
-Redundant Connections to different I2 Routers
-Layer 3 connections provisioned via optically protected lambdas over diverse fiber
-Connectors not at the CLPK or MCLN PoPs can get back to one of the core routers via a Layer 2 Transport or Layer 1 Transport service

MAX Layer 2 Network
-Redundant Layer 2 Connections back to two core Routers from each switch
-Layer 2 connections provisioned via optically protected lambdas over diverse fiber

Production Network

External Network Connection View



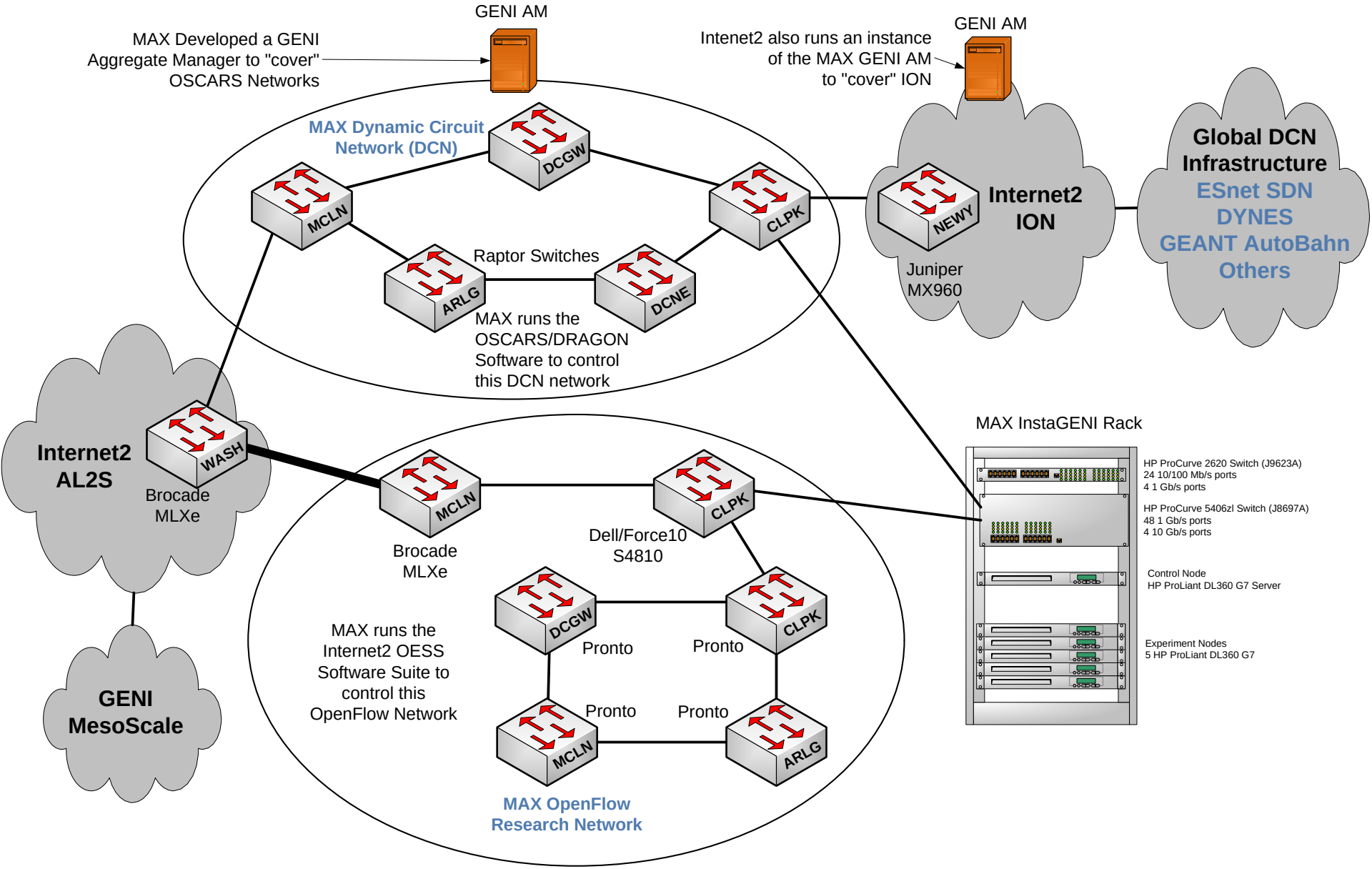
WIX – Washington International Exchange

- WIX is a facility jointly managed by MAX and Internet2
- Consists of a 100G capable Brocade MLXe located at Level3 Facility in McLean, Virginia
- Provides standard exchange point services
- Connectors include Internet2 AL2S, MAX, ESnet, ACE/GEANT, SINET.
- Connections to NGIX for peering between exchange points (DREN,NISN)

Research Network

- Network Equipment
  - Brocade MLXe (100G, OpenFlow), Dell/Force10 S4810 (10G, OpenFlow), Raptor (10G), Pronto (10G, OpenFlow)
- Software and Services
  - Dynamic Circuit Network (DCN) (OSCARS/DRAGON Software)
  - OpenFlow (Internet2 OESS Software)
- GENI Services and Equipment
  - DCN as GENI Aggregate via MAX developed Aggregate Manager

Research Network



MID-ATLANTIC CROSSROADS

MAX Services Update

Summary of MAX services portfolio

MAX Services

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

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

- For information on specific services and pricing contact Tom Lehman (tlehman@maxgigapop.net) or Jarda Flidr (jflidr@maxgigapop.net)
- Or email services@maxgigapop.net

MID-ATLANTIC CROSSROADS

Innovation and Advanced Services

Jarda Flidr

Director of Services

- AWS Direct Connect
- SDNX – MSX: MultiService eXchange

Definition

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

Motivation

- Enabling researchers and their applications
  - Well-engineered paths to major destinations
    - AWS
  - Network, Compute, and Data optimization
    - Domain Science Application integration
    - Science instrument integration

AWS DIRECT CONNECT SERVICE

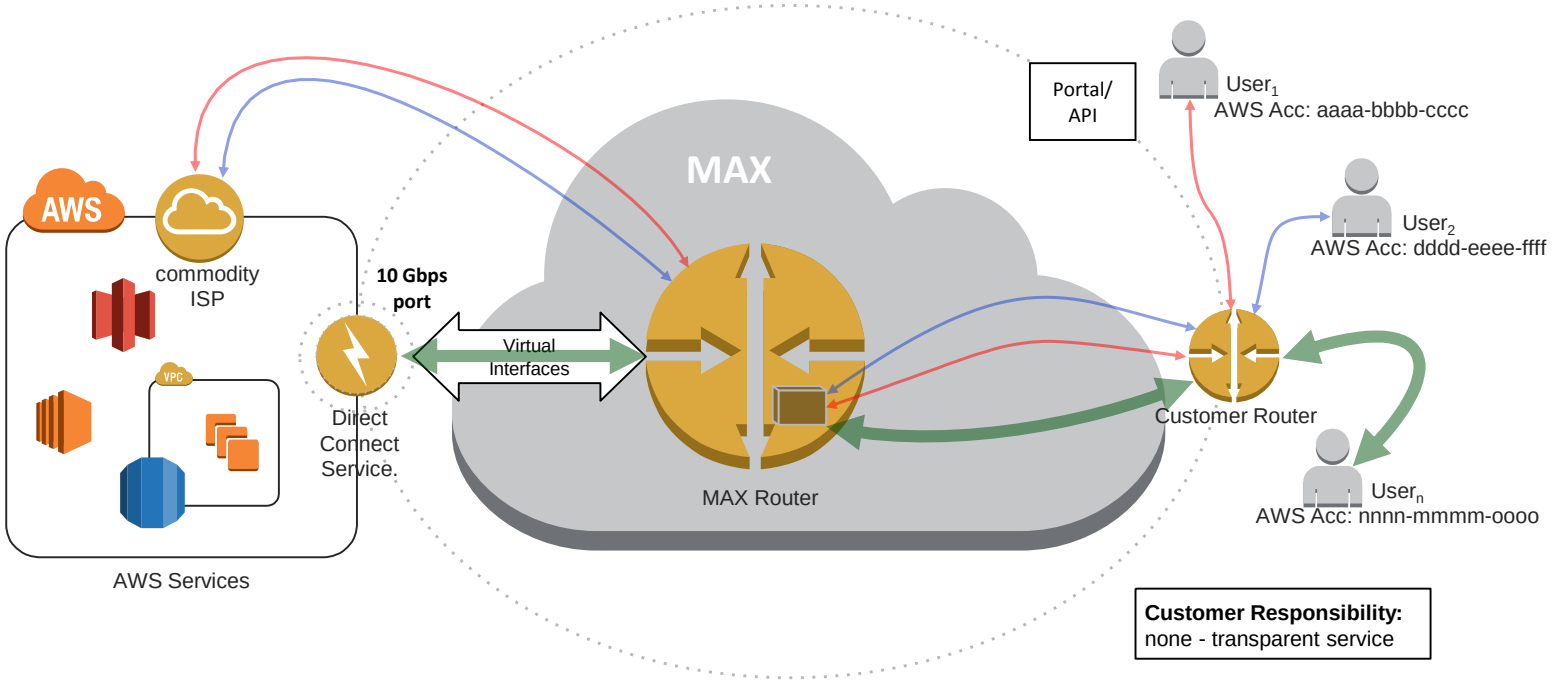
Direct Connect Overview

- 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
- Benefits
  - Discounted data pricing
  - Dedicated path
  - Private BGP peering for VPC integration

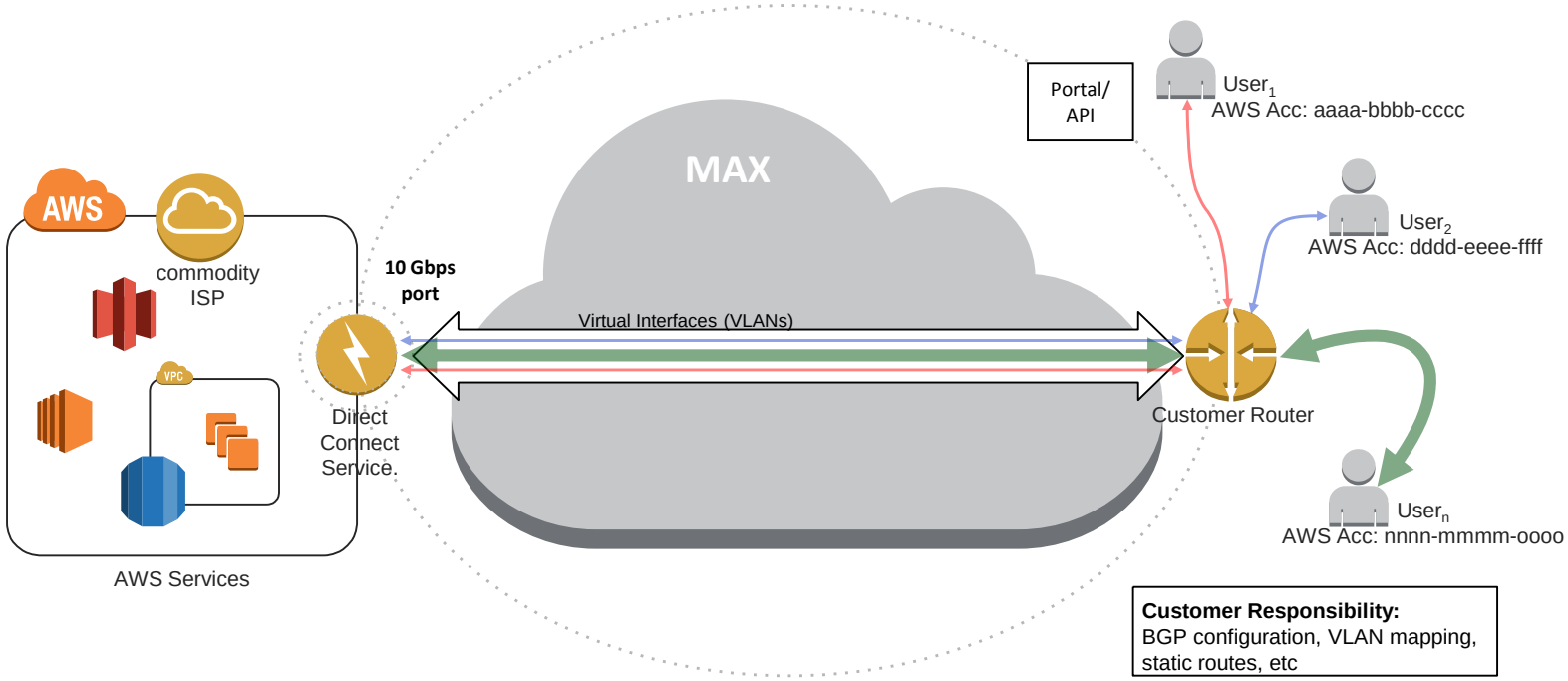
MAX Direct Connect Overview

- 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*.
- Persistent options
  - More physical ports can be provisioned

Dynamic (L3) Service

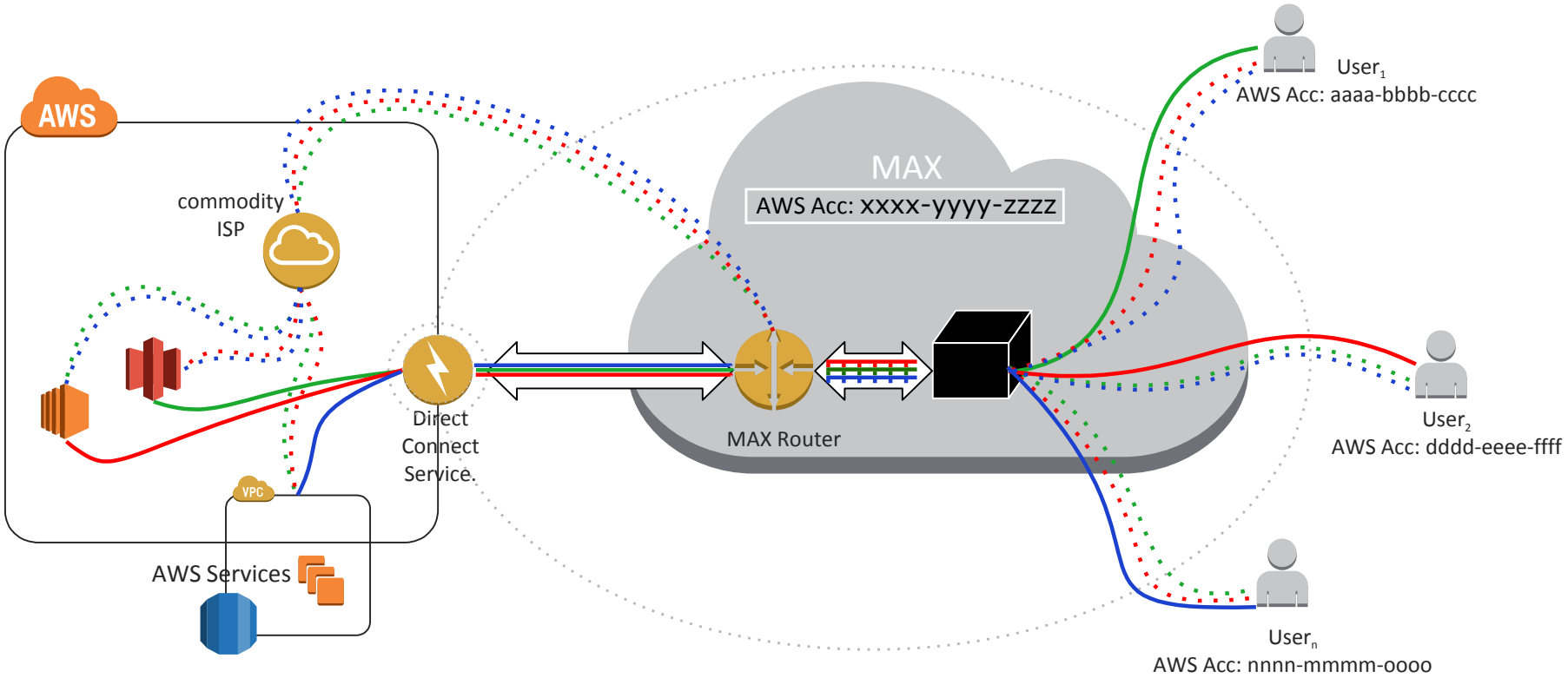


L2 Service


MAX
Portal/
API
User1
AWS Acc: aaaa-bbbb-cccc
User2
AWS Acc: dddd-eeee-ffff
Usern
AWS Acc: nnnn-mmmm-oooo
AWS
commodity
ISP
VPC
AWS Services
10 Gbps
port
Direct
Connect
Service.
Virtual Interfaces (VLANs)
Customer Router
Customer Responsibility:
BGP configuration, VLAN mapping,
static routes, etc

MID-ATLANTIC CROSSROADS

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

If anybody is interested …

Amazon Web Services

Compute & Networking

- **Direct Connect** — Dedicated Network Connection to AWS
- **EC2** — Virtual Servers in the Cloud
- **Elastic MapReduce** — Managed Hadoop Framework
- **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** NEW — Managed Petabyte-Scale Data Warehouse Service

Deployment & Management

- **CloudFormation** — Templated AWS Resource Creation
- **CloudWatch** — Resource and Application Monitoring
- **Data Pipeline** — Orchestration for Data-Driven Workflows
- **Elastic Beanstalk** — AWS Application Container
- **IAM** — Secure AWS Access Control
- **OpsWorks** NEW — DevOps Application Management Service

App Services

- **CloudSearch** — Managed Search Service
- **Elastic Transcoder** NEW — Easy-to-use Scalable Media Transcoding
- **SES** — Email Sending Service
- **SNS** — Push Notification Service
- **SQS** — Message Queue Service
- **SWF** — Workflow Service for Coordinating Application Components

SDNX

Software Defined Network eXchange: End-to-End, Dynamic Science DMZ

CC-NIE Integration

Award Number: 1246386

- Integration project funded by NSF
- In transition to become a service

Motivation

- Utilizing the existing advanced network functions
  - best effort IP
  - AL2S
  - ION
  - DYNES
  - SDN (OpenFlow)
- Optimizing Domain Science (campus research) applications' performance
  - Fluid Edge: triangulation of the best location with respect to Storage, Compute, and Data

Current Situation

- Suboptimal network performance and optimization
  - The 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?”
  - Result
    - Low-adoption rate: a wide spectrum of potentially beneficial and high-performance technologies are inaccessible to, or ignored by their primary users

Goal

- Integration of Advanced Network Functions with Domain Science Applications
  - Assumption
    - All the necessary technology pieces exist
  - Method
    - Focus on a set of particular Domain Science Applications (7 use cases): application workflows, algorithms and code
    - Integration at the application level
    - Hardware requirements: Multi-Service eXchange platform (MSX)
  - Intended outcome
    - Optimized data-related workflows of arbitrary Domain Science Applications
      - “classic” Science DMZ?: in some cases yes, in some not.

Data-Intensive Applications Use Cases

- Land Cover Facility (UMD)
  - research projects encompassing the fields of remote sensing and information systems; leader in the field of land cover and land use mapping, which has a long history in the development of algorithms and data products
- Biostatistics Center (GWU)
  - serves as the coordinating center for large-scale multi-center clinical trials and epidemiologic studies and processes large amounts of data from exome sequencing.
- Department of Astronomy (UMD)
  - Pan-STARRS1 and Large Synoptic Survey Telescope projects: two data-intensive research programs in the area of time-domain astrophysics
- Institute for Massively Parallel Applications and Computing Technology (GWU)
  - establishing an interdisciplinary GWU academic excellence program in High-Performance Computing (HPC) that spans research, education, and outreach, bringing large bio-informatics data sets from Biology and from the medical school and the Children's National Medical Center for phylogenetic analyses, as well as Medical Images with multiple modalities for processing in Ashburn

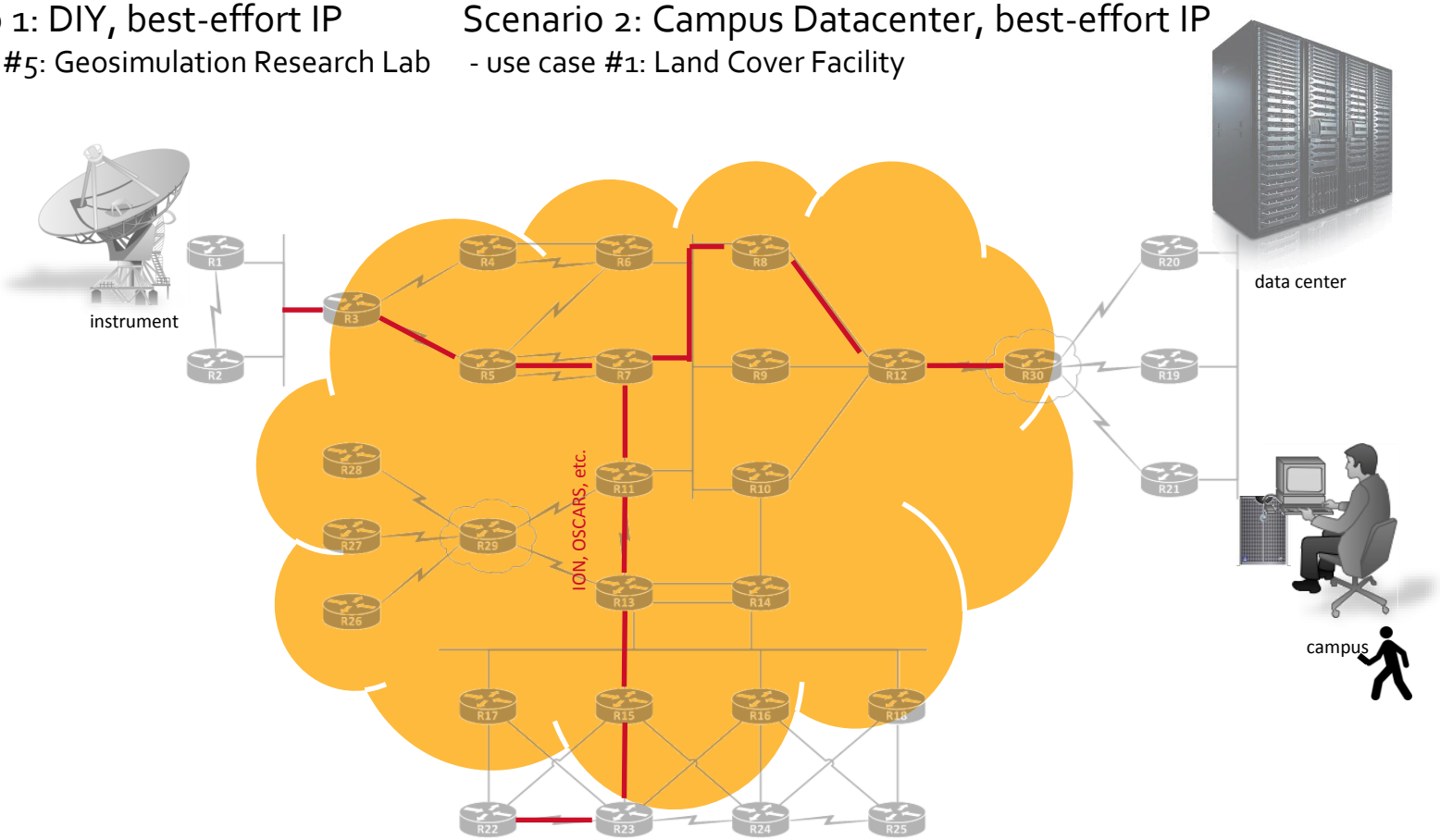
Data-Intensive Applications Use Cases

- Geosimulation Research Laboratory (UMD)
  - building very detailed and behaviorally realistic models of how householders and developers interact with each other and the built, social, economic, policy environment at scales from individuals up to neighborhoods, cities, and metropolitan systems
- The Institute for Nuclear Studies (GWU)
  - 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. The high users of bandwidth in GWINS are the Experimental Nuclear and Astrophysics groups.
- Physics Department (UMD)
  - 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)

MID-ATLANTIC CROSSROADS

Scenario 1: DIY, best-effort IP
- use case #5: Geosimulation Research Lab

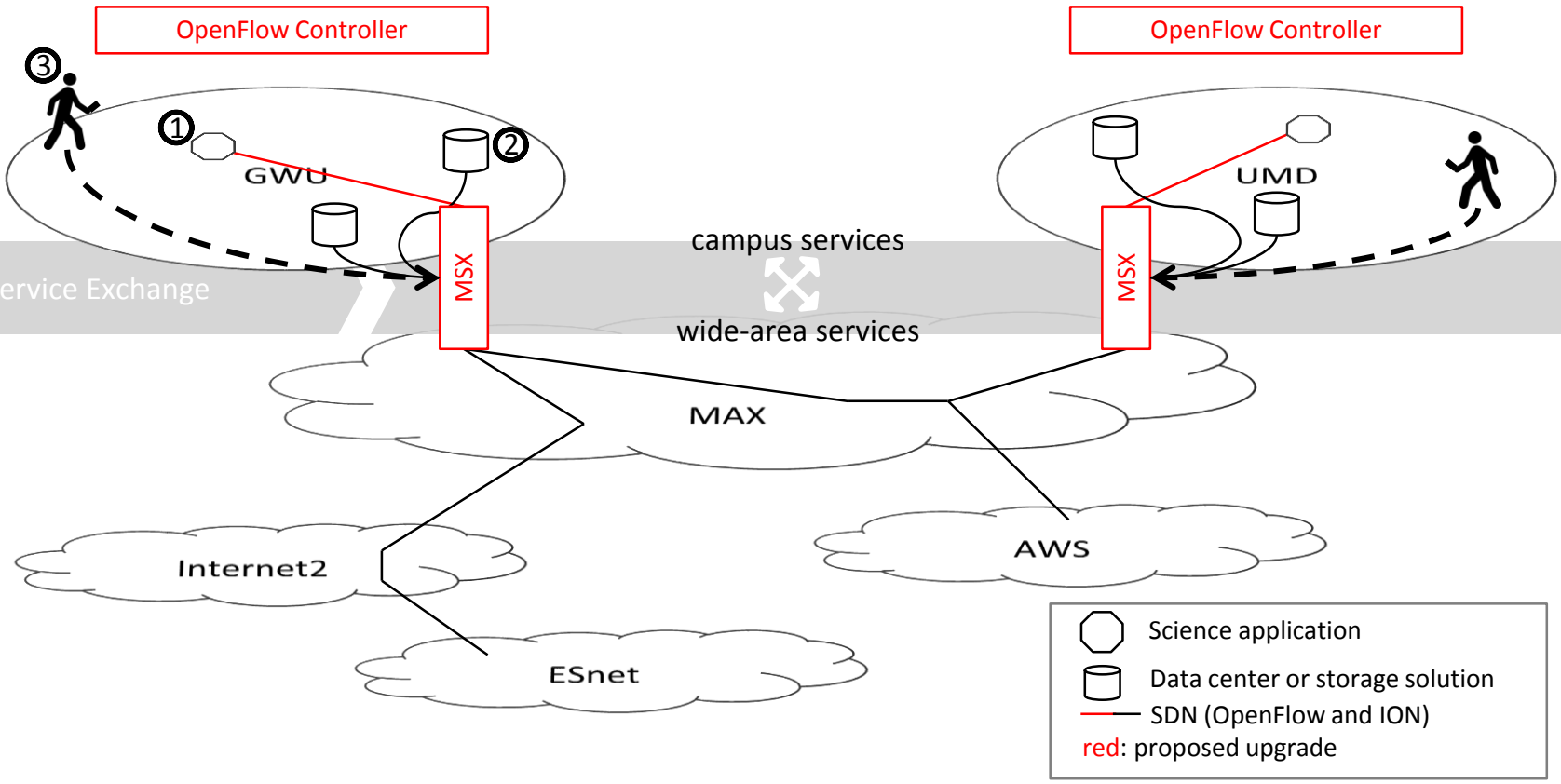
Scenario 2: Campus Datacenter, best-effort IP
- use case #1: Land Cover Facility



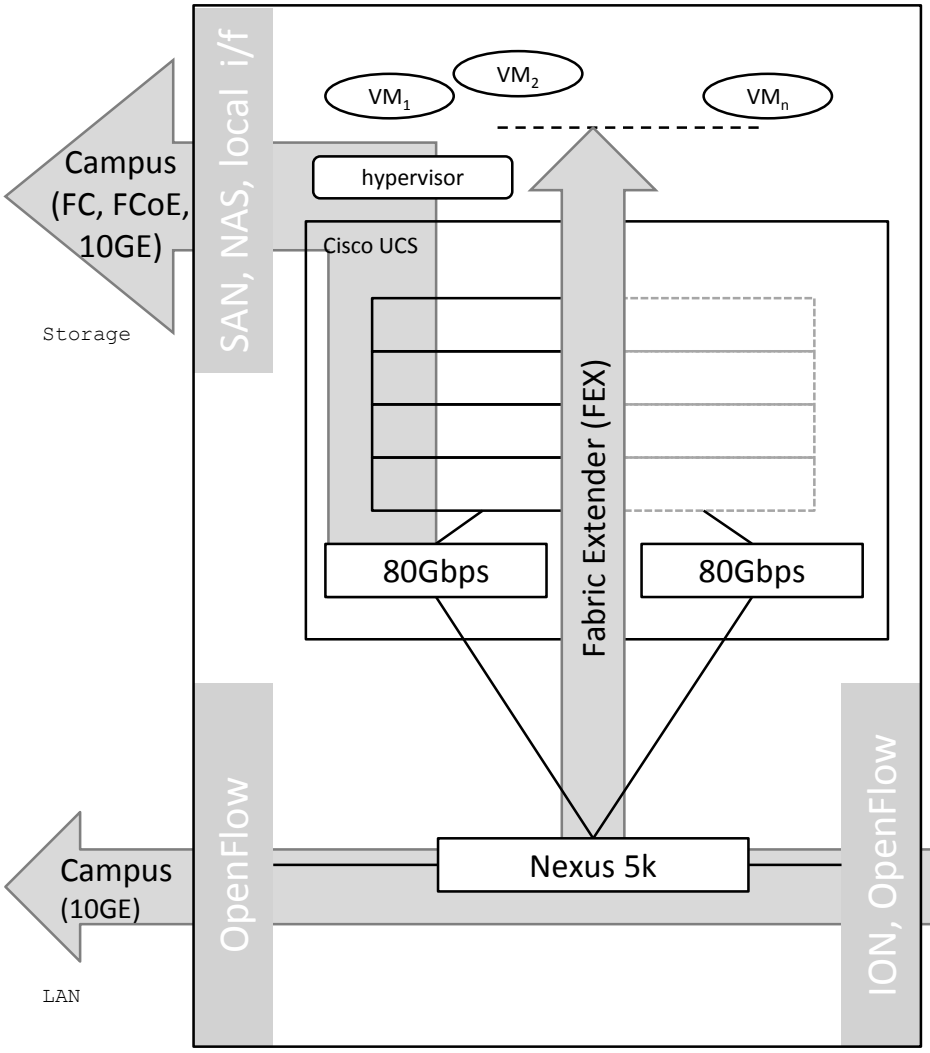
Scenario 3: Remote Datacenter, Instrument, Best-Effort IP
- use case #3: Astronomy Department (Pan-STARRS1)

Scenario 4: Remote Datacenter, FedEx Truck
- use case #2: Biostatistics Center

Scenario 5: Remote Datacenter, Instrument, Advanced Network Functions
- use case: not found

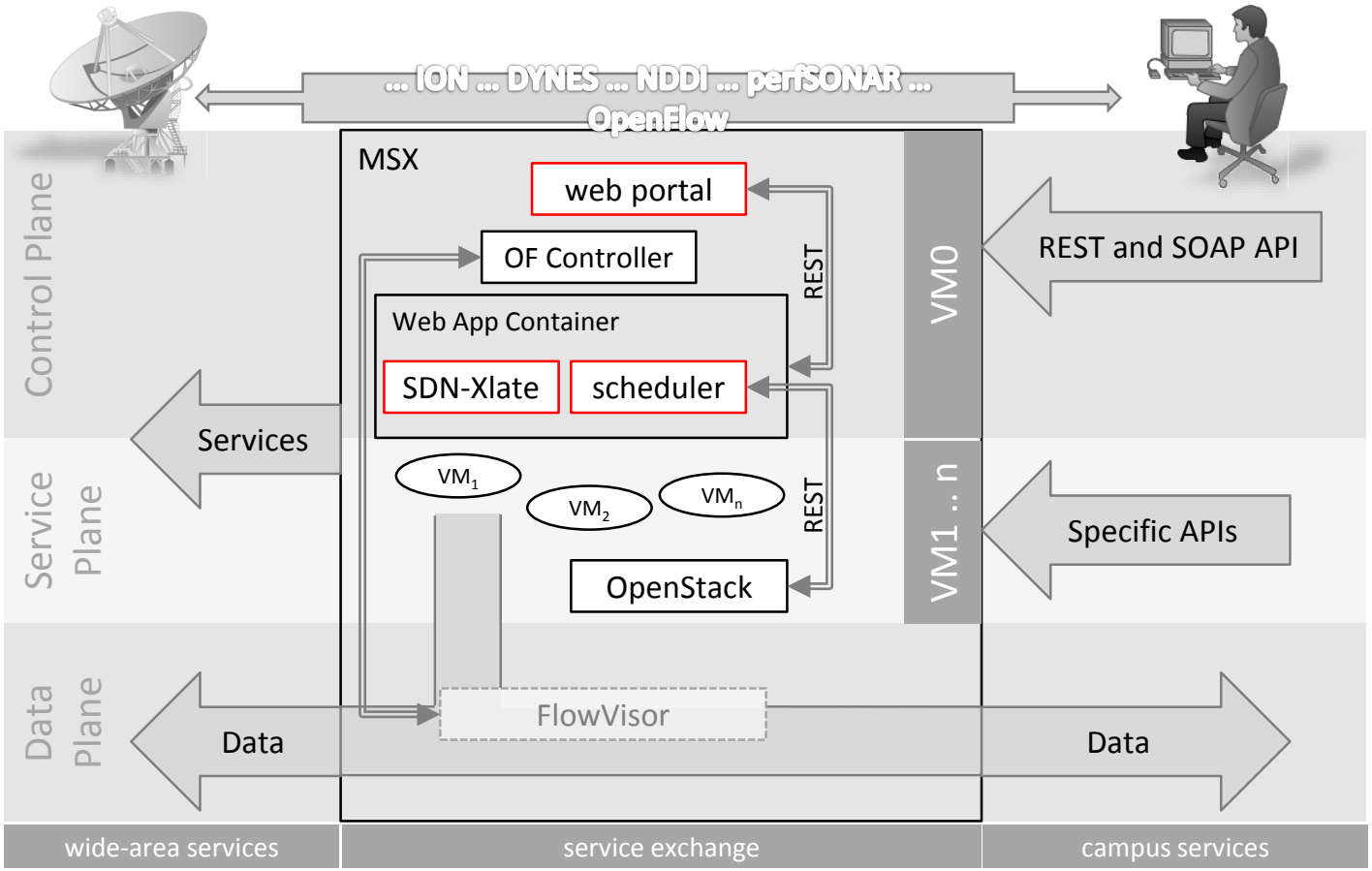

OpenFlow Controller
OpenFlow Controller
3
1
2
GWU
UMD
MSX
MSX
campus services
wide-area services
Multi-Service Exchange
MAX
Internet2
AWS
ESnet
Science application
Data center or storage solution
SDN (OpenFlow and ION)
red: proposed upgrade

hardware and virtual architecture



functionality

- portal interfaces, scheduler API, security
- campus-facing interfaces to data centers and storage solutions
- physically accessible local interfaces for scheduled data transfers
- virtualized server architecture and repository of images to instantiate on-demand, tailored transport solutions, e.g., VM1:Astronomy DYNES, VM2: Physics Dept. gridFTP, VMx: PHEBUS, *etc.*
- campus to wide-area SDN translator


... ION ... DYNES ... NDDI ... perfSONAR ...
OpenFlow
MSX
web portal
OF Controller
Web App Container
SDN-Xlate
scheduler
REST
REST
VM0
VM1 .. n
REST and SOAP API
Specific APIs
Services
VM1
VM2
VMn
OpenStack
FlowVisor
Data
Data
Control Plane
Service Plane
Data Plane
wide-area services
service exchange
campus services

Summary

- Domain Science Application-level integration (90 % of effort)
- Hardware platform (10% of effort)
  - Open
  - Flexible
  - Modular
- Dynamic, On-Demand, Fluid Science DMZ

Future Plans

- Additional Services Integration
  - Tom will address those

MID-ATLANTIC CROSSROADS

Q&A

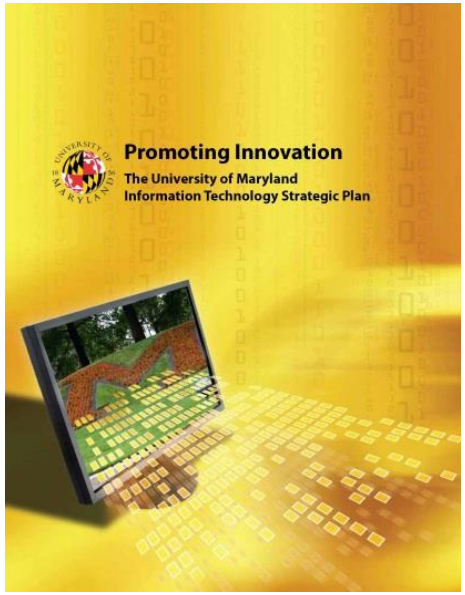
BRIAN D. VOSS

VICE PRESIDENT AND CIO
UNIVERSITY OF MARYLAND

MAX Organizational Structure

- What we did on [y]our summer vacation
  - MAX and the University of Maryland
- MAX's mission and that of IT@UMD
  - UMD's IT Strategic Plan supports MAX

    …and Vice Versa



MAX is still MAX – *and MAX is More*

- MAX continues serving R&E community
- MAX even more relevant to research and innovation
  - The Four Pillars – Network, Services, Research & Innovation
  - *Tripti's 'The Holy Triad' … and more in Cyberinfrastructure*
- MAX and High Performance Computing

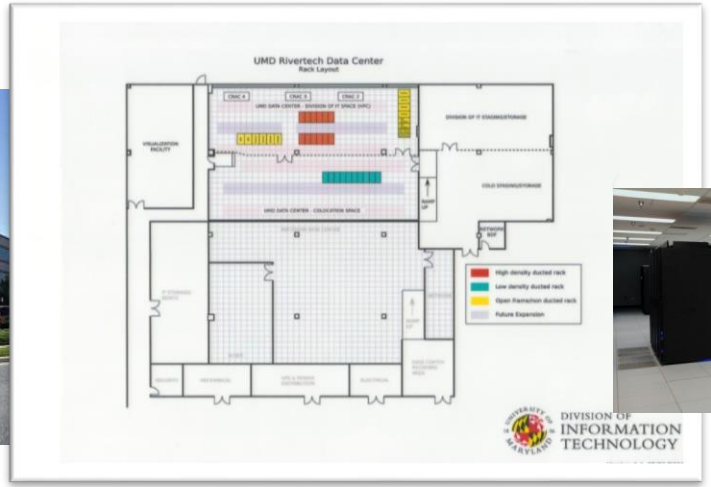
MID-ATLANTIC CROSSROADS

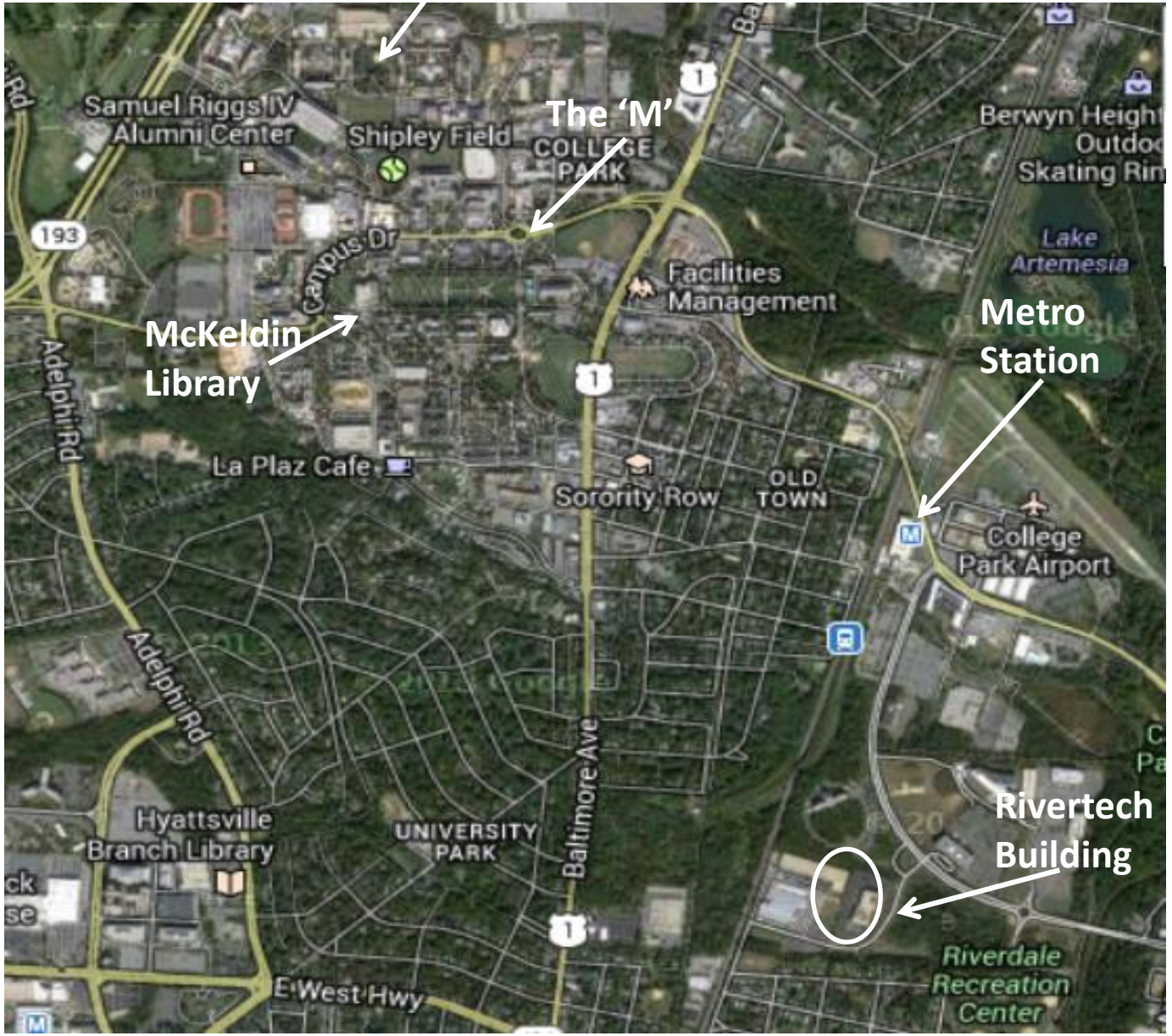
Rivertech Data Center

HPC/Cyberinfrastructure @ UMD

Implementing Recommendations in the ITSP

- Recommendation 1, Action Item 1 [phase-1]
- Recommendation 4, Action Item 1 [phase-1]

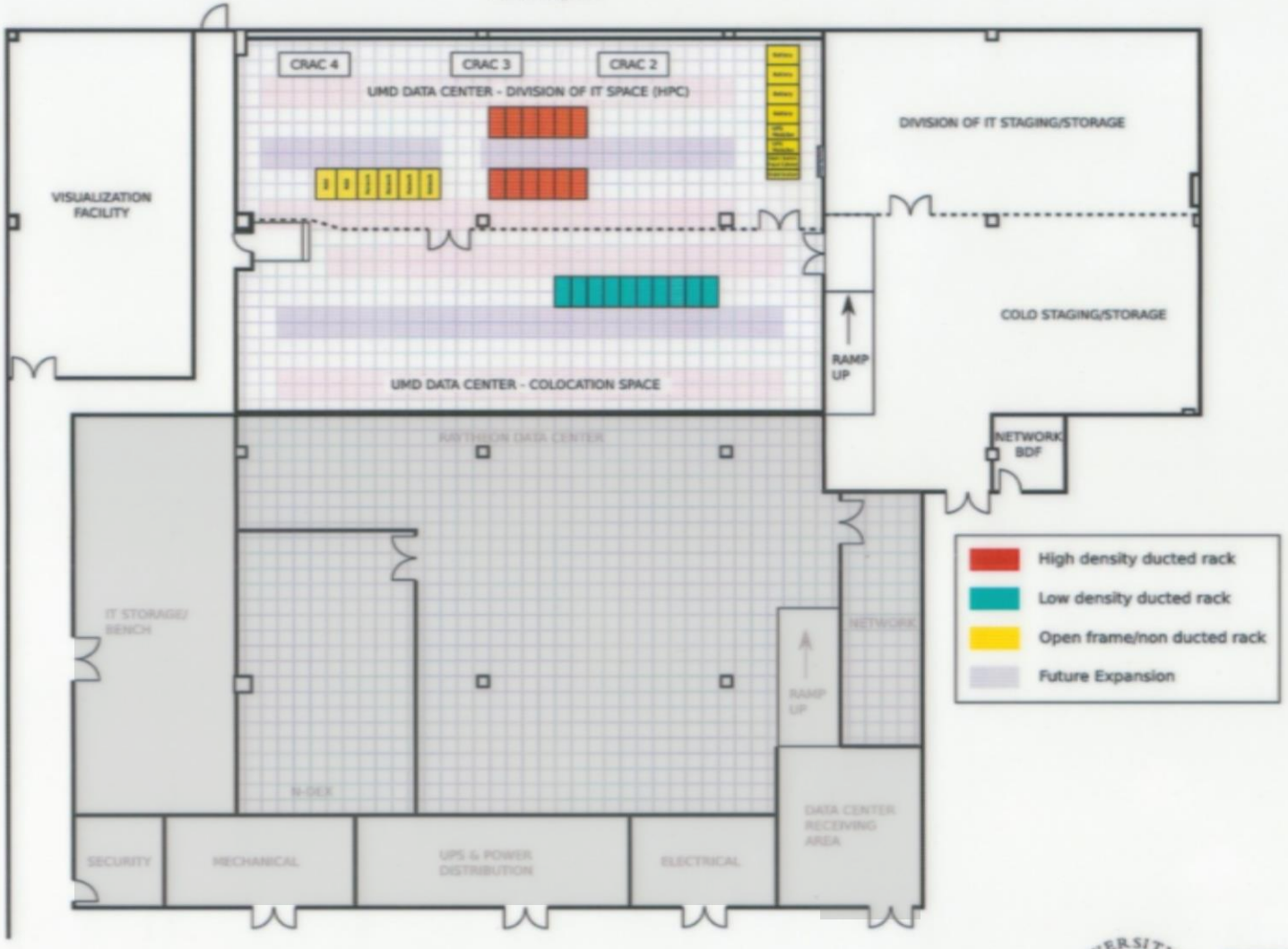



Patuxent
Samuel Riggs IV Alumni Center
Shipley Field
The 'M'
COLLEGE PARK
Berwyn Height
Outdo
Skating Rin
193
Campus Dr
Lake Artemesia
Facilities Management
McKeldin Library
Metro Station
Adelphi Rd
La Plaz Cafe
Sorority Row
OLD TOWN
College Park Airport
Adelphi Rd
Hyattsville Branch Library
UNIVERSITY PARK
Baltimore Ave
Rivertech Building
E West Hwy
Riverdale Recreation Center

MID-ATLANTIC CROSSROADS

UMD Rivertech Data Center

Rack Layout



DIVISION OF INFORMATION TECHNOLOGY

Version 4.4 05/31/2013

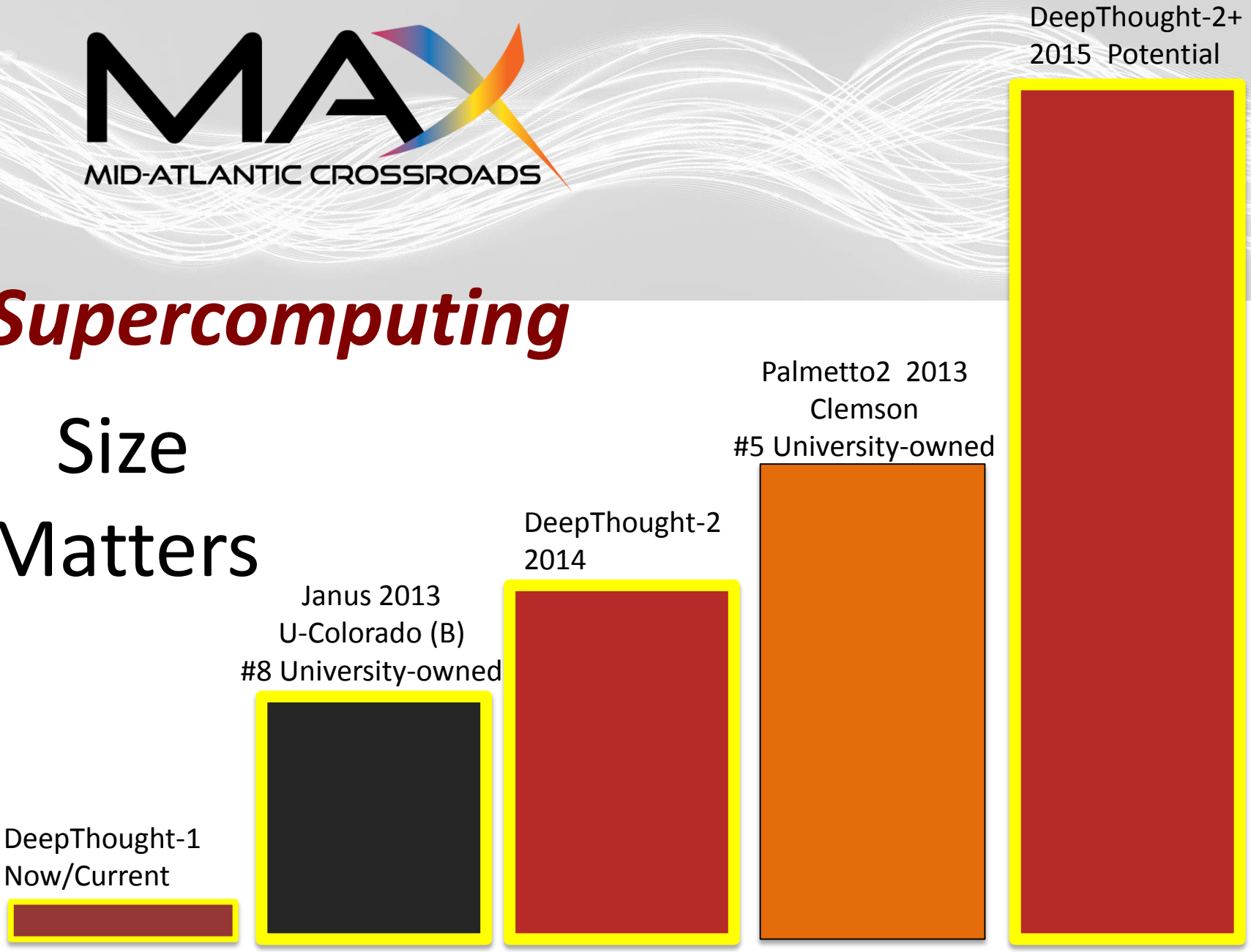
DeepThought2 … and more

- **New Parallel Compute Cluster & Storage**
  - ~5000-6000 Computational Cores
  - Mixture of node-types (“standard,” high-memory, GPUs)
  - Order of magnitude increase over current Deepthought
  - >1PB Storage – Lustre; FDR Infiniband
- **Development Compute Cluster**
  - Repurposed DeepThought elements used to allow ‘prep’ work for moving jobs to production resources (Deepthought2, National Centers, JHU, etc.)
- **Visualization Resources**
  - Showpiece

MID-ATLANTIC CROSSROADS

Supercomputing

Size Matters



MID-ATLANTIC CROSSROADS

JHU-UMD HPC Facilities

- **$30-million investment by State**
  - Sponsored by Senate President Miller
  - Passed in 2012
- **MOU worked out between UMD & JHU**
  - Location will be in Baltimore
  - Modular 'POD/Container' Design
  - Massive number of 'plain vanilla' processors
- **Ongoing design and governance discussions**
- **Expected online by early 2015**

MID-ATLANTIC CROSSROADS

TACC-Stampede
NSF Center
U-Texas Austin

Supercomputing

Size

REALLY

Matters

DeepThought-2
2014

DeepThought-2+
2015 Potential

BigRed2 2013
Indiana U.
#2 University-owned

Joint JHU-UMD
HPC Assets 2015

Carter 2013
Purdue U..
#1 University-owned

Q&A

MID-ATLANTIC CROSSROADS

MAX Sponsored Research Projects

Update on NSF, DOE, and DOD supported research activities

Sponsored 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 ($645K for two years, start Oct. 1st 2013)
  - 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 LA2S
  - OpenFlow capable Ethernet layer
  - 10G AWS Direct Connect

HPC and HPN Integration

- HPC file system to NES system integration
  - DeepThough Lustre file system bridge to HPCDNA NES
  - 2 x 56Gbps InfiniBand facing DeepThought
  - 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 DeepThought HPC cluster.
- Federated with UMD authentication system (potentially also with InCommon etc.)
- External AWS integration for hybrid cloud workflows.

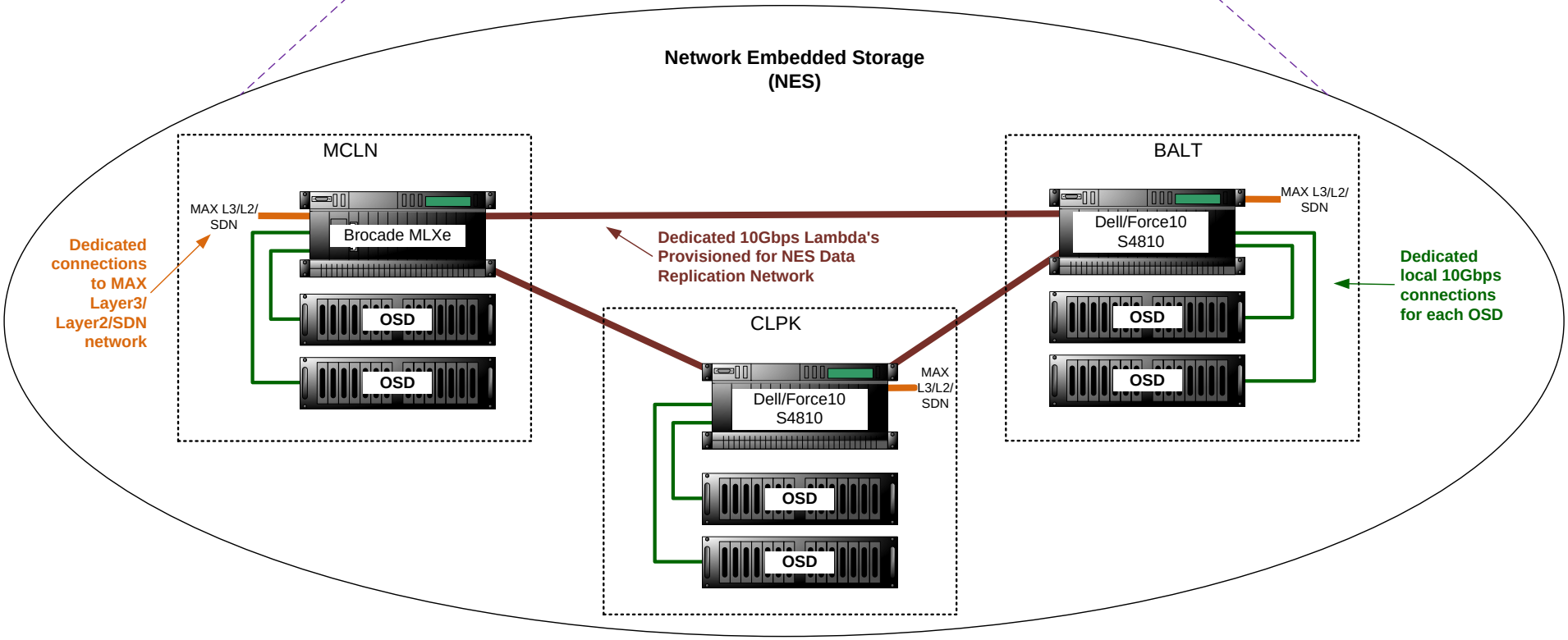
MAX Network Embedded Storage

Clients

- Users inside UMD HPC System (DeepThought)
- SDNx Multi-Service Exchange (MSX) Multi-Tenant Virtual Machines
- Any user on UMD Campus using the local lab systems
- Users inside Amazon Cloud (on EC3 compute instances)

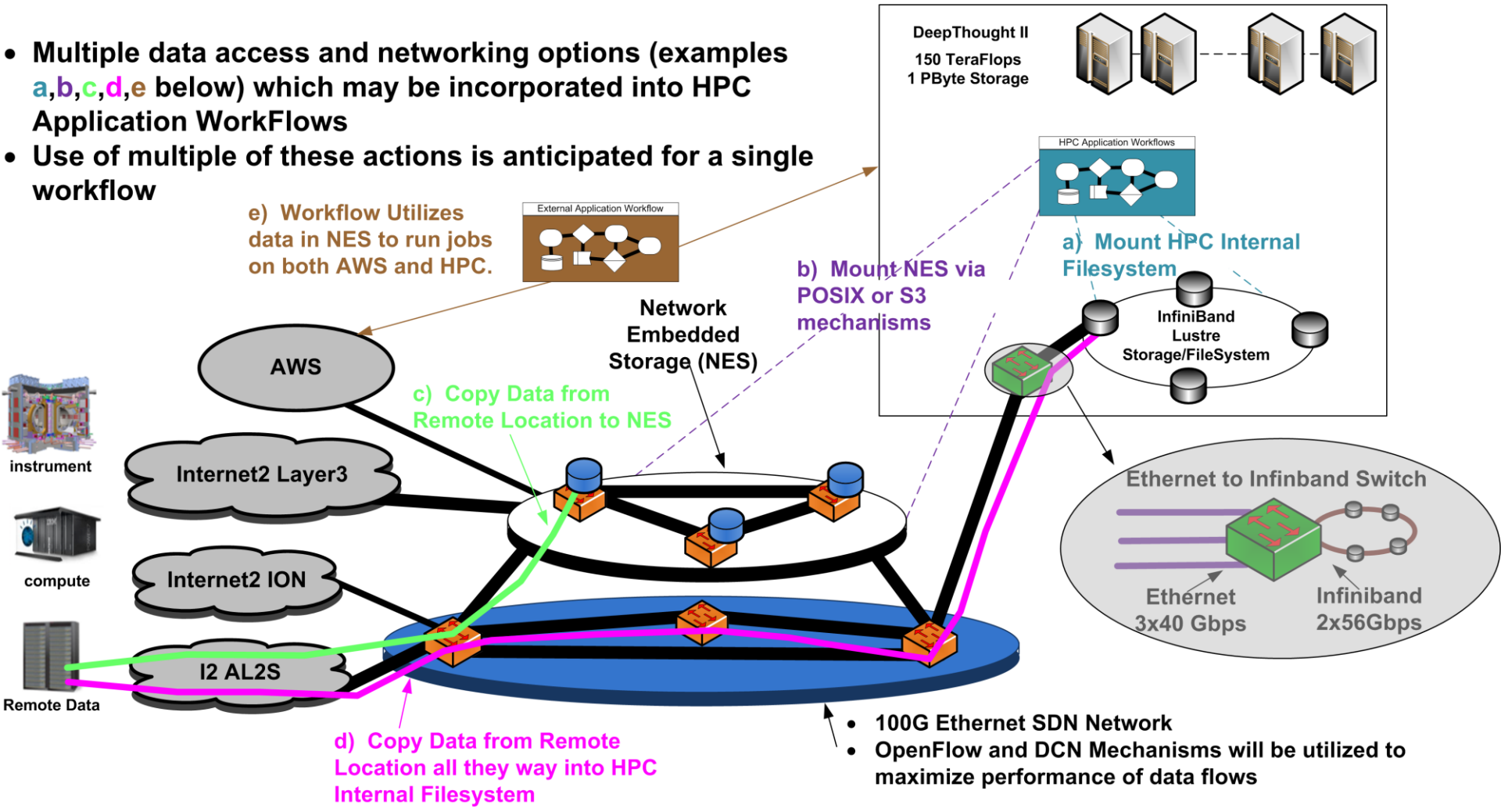
NES API will offer multiple Client Access Options and Features

- Block Storage Mount (Posix Compliant)
- S3 Object Data Storage
- NSF/SMB
- SDN control of network paths (local and wide area) which synchronize with data movement operations



HPCDNA Architecture

- Multiple data access and networking options (examples a,b,c,d,e below) which may be incorporated into HPC Application WorkFlows
- Use of multiple of these actions is anticipated for a single workflow



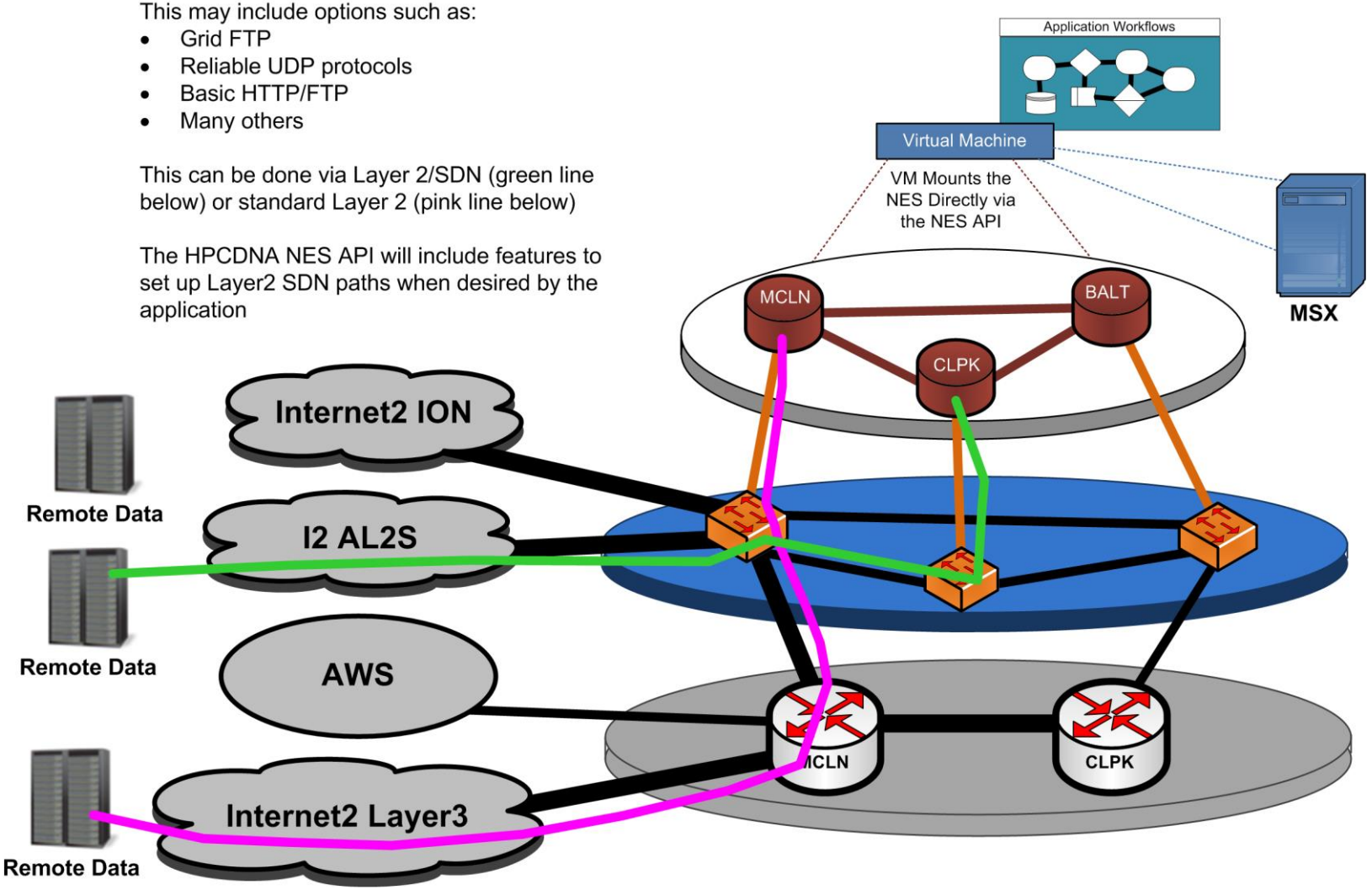
HPCDNA MSX Integration

Once the VM has mounted the NES, the VM can then use whatever mechanism is most efficient (and possible based on remote data system) to move data directly to the NES storage system. This may include options such as:

- Grid FTP
- Reliable UDP protocols
- Basic HTTP/FTP
- Many others

This can be done via Layer 2/SDN (green line below) or standard Layer 2 (pink line below)

The HPCDNA NES API will include features to set up Layer2 SDN paths when desired by the application



MID-ATLANTIC CROSSROADS

Resource Aware Intelligent Network Services (RAINS)

- Funded by DOE ($1.8M Total / $900K MAX, 4 years, start Sep. 1st 2013)
- UMD/MAX (lead organization), Tom Lehman, Xi Yang
- Argonne National Laboratory (ANL), Narayan Desai, 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 Highlights

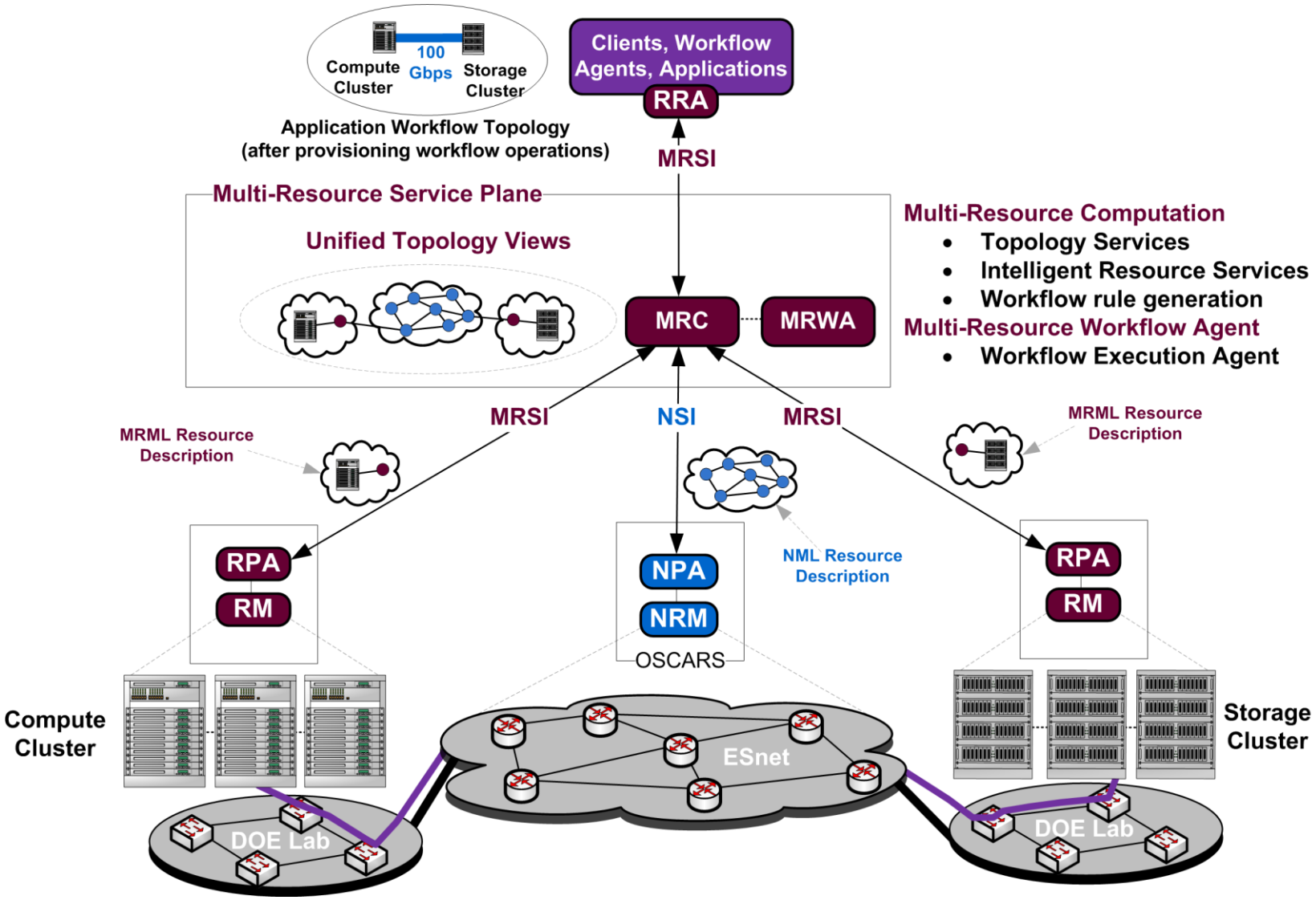
Building RAINS Multi-Resource Service Plane (MRSP)

- Providing a mechanism to describe a unified view across compute, storage, and network resources in terms of their interconnection and their services.
  - Multi-Resource Markup Language (MRML)
  - Multi-Resource Service Interface (MRSI)
- Prototyping key intelligent Resource Services (IRS) to facilitate the integration
  - Topology and workflow computation
  - Inter-domain, inter-layer, inter-resource stitching
  - Workflow assistance

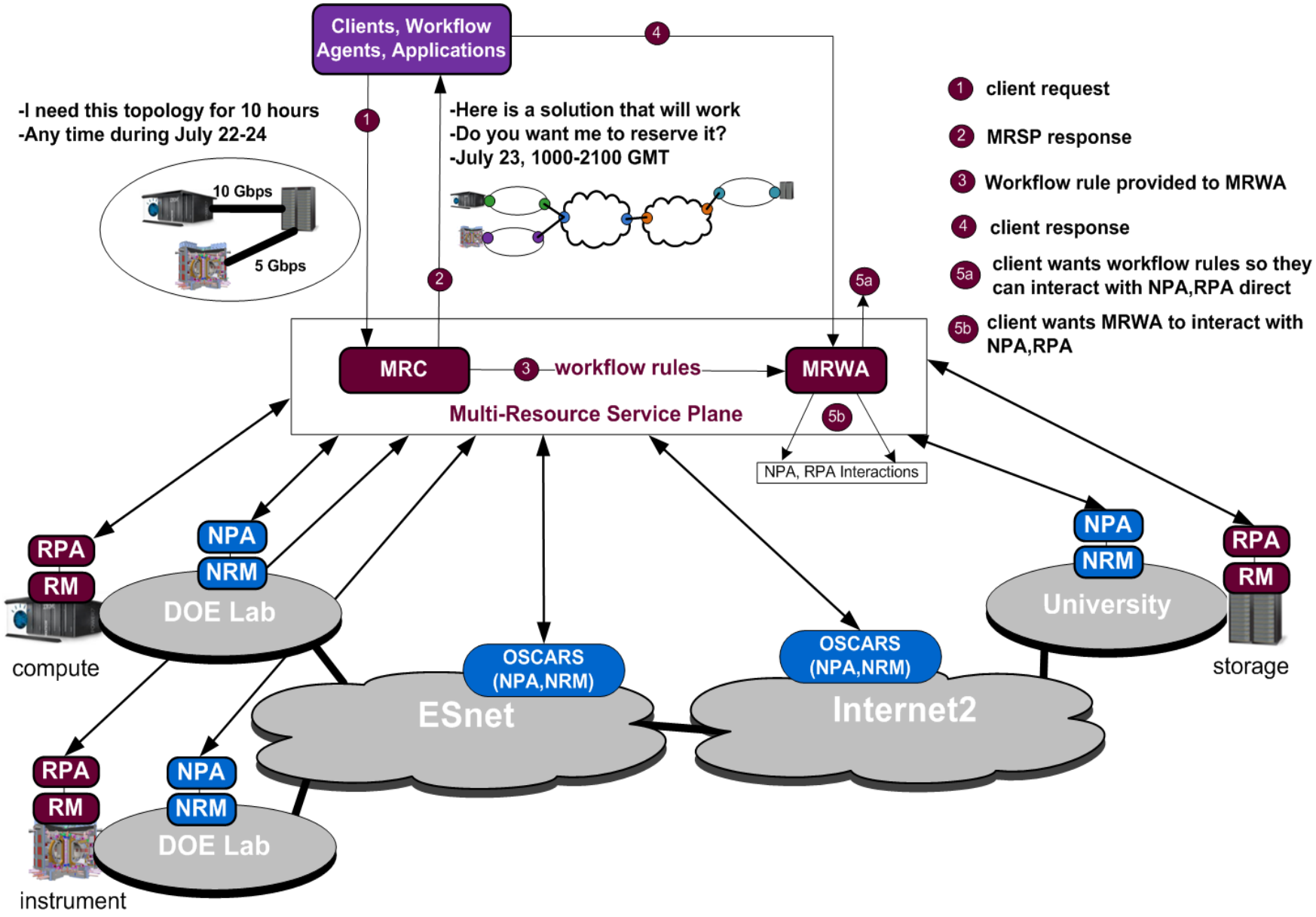
Design Driven by A Focused "Data Aware Networking" Use Case - KBase

- A DOE multi-institutional effort led by LBL to build a unified set of services enabling biologists and bioinformaticians to generate and test new hypotheses and models relating to microbes, plants and microbial communities.
- Real scientific applications using:
  - OpenStack backed Magellan cloud computing systems at ANL and NSERC.
  - ESnet OSCARS and other advanced networking technologies.
- RAINS uses KBase as a vehicle to determine the best architecture for general use and standardization for MRSP.

RAINS Architecture



RAINS Ecosystem Vision



MID-ATLANTIC CROSSROADS

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:
  - A Supercomputing 2013 demonstration is be planned around moving Sloan Digital Sky Survey (SDSS) data from JHU to Oakridge National Laboratory Titan Computation Facility

Global Environment for Network Innovations (GENI)

- Funded by NSF since 2008: 5 years of continuous funding, 4 spirals, 3 projects.
- 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.

Achievements

MAX MANFRED Substrate/Aggregate

- Rich set of resources for GENI community
  - DRAGON optical network and GMPLS control plane
  - PlanetLab nodes
  - Ability to reach DCNs in rest of world via IDC / OSCARS
- Recent additions
  - InstaGENI rack
  - MAX OpenFlow research network
  - 100G network and connection to ION AL2S

MAX Aggregate Manager

- Developed AM software for controlling DCN through IDC/OSCARS
  - Advertise DCN resources in GENI RSpec v3 format
  - Support GENI AM API v2
  - Federate with other GENI aggregates and clearinghouses
- Adopted by Internet2 as production ION Aggregate Manager.
  - Support OSCARS API v5 and v6.
  - Can be used to translate any other DCN into a GENI aggregate.

GENI Stitching

- Designed the GENI Stitching Architecture
  - Helped GENI community to build infrastructures that allow experimenters to create slices seamlessly across multiple aggregates.
  - Initially focused on Layer-2 VLAN stitching.
- Developed GENI Stitching Computation Service (SCS) Software
  - Facilitate multi-aggregate topology computation and workflow assistance.
  - Being used in routine stitching operations.

Upcoming Effort

- Newly funded project "GENI Stitching and Computation Enhancements (GENIStitch)" to start on Oct. 1st 2013.
- Goals:
  - Enhancing GENI stitching for operation / production use.
  - Adding advanced stitching features, including
    - multi-point VLAN bridging stitching,
    - support for IP/MPLS tunnel, IP GRE tunnel, native OpenFlow, VxLAN, etc.
    - support for negotiation workflow.
  - Utilizing the stitching features to assist researchers and tools in multiple phases of the experiment lifecycle. For example:

    It is possible for experiment tools autonomously determine what is the best type of stitching capability to use when crossing two aggregates. Stitching service will provide computation features to facilitate such flexible-type stitching workflow.

MID-ATLANTIC CROSSROADS

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

- Funded by DTRA: three years, started in May 2013.
- University of New Mexico is prime contractor. UMD/MAX is subcontractor.
- Background:
  - Under the context of WMD attack and other massive disasters, network survivability research needs to address scenarios different than those previously studied in the literature.
  - Focus on protection and restoration against backbone disruptions and large-scale failures that involve many network elements and multiple network administrative domains.
- Project Highlights:
  - 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 network survivability scenarios.