NEXT GENERATION INFRASTRUCTURE

APAN48 July 22-26, 2019
John Hicks - Internet2
Internet2 Background: An Ecosystem View

Partial view of US R&E Ecosystems:

- Long Haul (incl Internet2)
- International Exchanges
- Regional/State Networks
- Local Campuses and Districts

Map Source: The Quilt
NGI and Internet2: Service Development + New Platform

Evolve the service portfolio of Internet2 to align with the current needs of the community.

- Focus on identified, tangible use cases
- Focus more on technical service groupings that support those use cases

Ensure the value delivered by Internet2 is clearly identifiable and fiscally sustainable

Enable rapid integration of Internet2’s technology infrastructure into the community ecosystem

- Provide flexible hardware and software building blocks to support business, technology, operations
- Modernize approaches, consider consolidation of compute, networking, storage approaches on virtualization, agile development, etc.
NGI Optical RFP: ROADM Solutions

- State-of-the-art ROADM technology seems headed for Route and Switch
  Recommended by one vendor
  More efficient design
  Lower OSNR impairment
  Wavelength filtering capabilities

- Traditional ROADM are B&S (broadcast and select)
  Recommended by all others

- All supported flexgrid and channels over 50 Ghz

- R&S technologies are still developing and are significantly more costly

- Premium DSP technology offsets the OSNR shortcoming and coherent detection can lock onto the desired wave without the need for filtering.

Subcommittee recommended against paying the premium for R&S
Optical is fundamentally an analog-based service
Type of amplifiers are used have a systemic impact on system quality
Raman amplification can improve OSNR and delivered line-side rates…
  but… at the cost of increased complexity

All finalists recommended EDFA amplification
  More cost effective
  Improvements in modem technology help defer the necessity of Raman
  A few segments might benefit from a small number of hybrid Raman

Scenarios for reuse of existing MLA3 ILA OLA sites, decreasing deployment time

Subcommittee recommended the suggested EDFA architectures
NGI Optical: Community Developed Requirements

**Current Platform:**
- Designed in 2009, installed 2011/2012
- 88 50 Ghz Channels
- Maximum 35 Gbaud, 200 G per channel
- Telecom industry oriented management kit
- Shared Platform (West Coast, ESnet)

**Planned Capabilities:**
- Improve Community Access to Optical Spectrum
  - Unrestricted “alien wave” support
  - Flexible, variable size spectrum channels
  - Extend into Cloud Interconnect Sites
- Provide drastically more bandwidth for similar investment
  - Flex-grid: “media channels” with no predefined channel spacing
  - Improve OSNR through re-installation and tuning
  - support for 400 G and longer reach
  - Lower Cost to Operate & Scale
  - Minimize Power, Space, and Maintenance Costs
- Dynamic Provisioning
  - Management kit based on open standards
- Infrastructure Sharing (Example: West Coast Shared Platform)
- Community Pricing
NGI Optical: Community Joint Recommendations

- Clear need to balance of “new technologies” vs “solid, low-risk products”

- Recommended conservative EDFA amplification strategy
  - More cost effective
  - Suggested we focus instead on modem technology / DSPs
  - Suggest future implementation of hybrid Raman

- Confirmed major requirements were met
  - Unrestricted “alien wave” support
  - 400 G + line side (on roadmap)
  - 6 - 10 OLS Year Lifespan
  - Lower Overall Cost to Operate
  - Management kit based on open standards
Production 400G coherent transponders are here
   Client side will arrive Q2 2020
   New system is optimized for 95 Gbaud; all single spans are supported

Production 400G QSFP-DD client optics are likely to come Q3-2020
   Extremely small power envelopes - low launch powers (-7 dB !!)
   Clear applications for point-to-point within metropolitan areas
   We intend to leverage them when available

Transcontinental alien wave capability is a key differentiating factor
   System should accommodate them at 95 Gbaud
   Software, system balancing, and operations protocols are key
NGI Optical: Next Steps

We are finalizing...

...the final OLS designs and BOMs...
...the contracts with our selected vendors...
...our final OLS upgrade phasing.

Expecting upgrade kick-off later this summer.

Transponder plan will evolve from packet architectures and designs.
Packet: RFI process & Discovery activities
Current Backbone and R&E Network

- MPLS with full mesh of RSVP Label Switched paths (LSPs)
  - Secondary paths are signaled and in standby
  - Shared Risk Link Groups (SRLG) are used to diversify primary and secondary paths
  - Very little traffic engineering applied
- Uses Constrained Shortest Path First (CSPF) to determine where the packet is forwarded
- Metrics are determined by fiber route kilometers
Current Layer 3 Services

Uses MPLS based L3VPNs

- DDoS
- Cloud Connect
- I2PX (old TRCPS)
- LHCONE
- XSEDE
- Science OSG/DTN
Quick Recap on Existing L2 Services at Internet2

- “Typical” OESS point-to-point circuit using shortest path (ELINE)
  - “Kompella” (BGP Signalled) L2VPN
    - BGP Signalled Control Plane (family L2VPN), MPLS Service Label
- Multipoint OESS circuit (ELAN)
  - Virtual Private Lan Service (VPLS)
    - BGP Signalled Control Plane
    - MPLS Data Plane
    - Flood and Learn
    - Multihoming support is only active/passive
- “User-defined” Point to Point OESS Circuits
  - Legacy “Circuit Cross Connect” (CCC)
    - Uses a dedicated LSP per Circuit
    - Unpopular - not supported on some platforms (Cisco/Arista/?)
- Miscellaneous hand-provisioned circuits
  - “Martini” Pseudowire (l2circuit)
Enter Ethernet VPN (EVPN)

- RFC 7432 “BGP MPLS-Based Ethernet VPN”
- RFC 7209 – Requirements for Ethernet VPN (EVPN)
- RFC 7348 – Virtual eXtensible Local Area Network (VXLAN)

Why the demand?

- One service type to handle ELINE and ELAN services (and also Layer3).
- Service providers like BGP for standardization. Central Control Plane. Ability to use routing policies.
- Decouples control plane from Data Plane. We no longer require MPLS.
- Active-Active Multihoming a requirement.
- Optimizations in regards to handling Broadcast / Unknown Unicast / Multicast (BUM) Traffic
EVPN Advantages

- Multiple types of multihoming including Active-Active multihoming, with loop prevention/mitigation
- Multiple service delivery methods including bundling multiple vlans per EVPN instance or dedicating an instance per vlan. (vlan-aware bundle, vlan-based, EVPN-VPWS)
- Support for “mass-withdraw” of MAC addresses - Faster convergence during attachment circuit failures.
- Integrated broadcast suppression - Proxy ARP / ND ARP
- Support for IGMP to increase multicast efficiency
- EVPN leverages MP-BGP to distribute MAC address information (amongst other things).
Our Plans for EVPN

Short Timeframe
- Migrate multipoint/ELAN OESS Circuits to EVPN - No more VPLS.
  - Clear advantages for EVPN vs VPLS.
  - Multipoint solution doesn’t require traffic engineering support, so the choice of encapsulation (VXLAN/MPLS) is not critical.
- New point-to-point OESS Circuits to be provisioned using EVPN (Non-Traffic Engineered)

Long Term Vision
- Migrate all P2P Layer2 Circuits to EVPN, including the strict paths.
- We need to follow the EVPN TE situation closely before making this move.
Current Data Transport on Internet2

- IS-IS + RSVP + MPLS
- The overwhelming majority of traffic on our network (R&E, I2PX, AL2S), takes shortest path routing using a full mesh of core LSPs
  - 2,000+ core LSPs
- OESS provisioned circuits
  - 752 Layer 2 circuits total
  - 25 manually provisioned paths
- Tactical traffic engineering is rare, our backbone can handle most failures without needing to readjust traffic flows
- Do not use bandwidth reservation or autobandwidth
Segment Routing (SR) satisfies our goals

- **Simplicity**
  - Minimizes signaling - labels are signaled via IGP, no need for LDP or RSVP
  - Minimizes state within the network/transit LSRs - state is implemented as a per-packet label stack and imposed by ingress router

- **Support for standard LSP full mesh** - does not require a controller

- **Equal-Cost Multi-Path (ECMP)**

- **Fast reroute**
  - SR uses Topology Independent Loop Free Alternate (TI-LFA)
    - computed repair path is same as post convergence path for SPF LSPs
  - 100% coverage
  - Protection does not impose additional signaling requirements

- **Supports Traffic Engineering**
  - Programmed static path on ingress router
  - Controller, e.g., OESS
  - BGP-SRTE signalling
  - Path Computation Element Protocol (PCEP)
We drive via cities, landmarks, points of interest.
NGI Packet: SR Advanced Use Cases

Security gateway
- Science flows unimpeded
- Higher risk, unknown flows inspected

Diagram:
- DTN to GridFTP
- GridFTP to DTN
- HTTP, Telnet, ???
NRP (National Research Platform)
PRP/NRP Nautilus k8s cluster

- The “NRP” (National Research Platform) is an evolution of the PRP (Pacific Research Platform) - (Larry Smarr, Tom DeFonti)

- PRP/NRP runs a kubernetes cluster (John Graham and Dima Mishin – UCSD) and CEPH/EdgeFS and provides an environment for GPU compute, storage, and cache.

- Member are encouraged to bring resources to the platform (e.g. DTN, GPU, storage, …, etc.)

- Can also request namespace to run scientific workflow

- https://nautilus.optiputer.net/
3rd NRP Work Shop

- September 24-25, 2019 - Minneapolis, MN
- We want to look at use cases that need (want) an environment not served by current available resources (EXSEDE, OSG, …, etc.)
Internet2 – OSG collaboration
OSG Internet2 collaboration

- Data Federation built on tools from LHC community.
  - Exabyte scalable global CI
- Distributed data sources within discoverable namespace.
  - Supports highly dynamic tree of data origins and caches.
- Caching policies are local
Implications

• An organization can join the federation with their own “data origin” and their own partition of the global namespace.
• A cache owner can decide on caching policies for different parts of namespace.
• This allows the owner to selectively serve only a subset of the community that use the federation.

We can build community specific “storage overlays” with this technology.
Caching Pilot

- Objective: Experiment with extending current campus-based StashCache into Internet2 backbone
- Choose initial locations based on proximity to existing caches
- Experimental results will drive changing/extending configuration – potential for commercial cloud services, etc.
- Interconnection – experiment with virtual network service options optimize for manageability, resiliency, security, etc.
Current NRP/OSG deployment

[Diagram showing the current NRP/OSG deployment across the United States, with cities and connections marked.]
Future NRP/OSG deployment
NEXT GENERATION INFRASTRUCTURE