Research Network Challenges and Opportunities

Matt Zekauskas, matt@internet2.edu
Senior Engineer, Internet2

2006-06-29

Outline

- Internet2?
- Current network issues
  - Abilene, a 10Gbps IP backbone
- New network issues
  - Above, plus “dynamic circuits”

Internet2

- US membership non-profit organization
  - 208 University Members
  - 70 Corporate Members
  - 53 Affiliate Members
- We operate an IP backbone network
- We are not NLR, a nonprofit formed to create R&E experimental national optical network; it's moving toward production & IP, we're moving toward circuits, has been work toward merger, may happen, but dead for now (we are investors in NLR…)

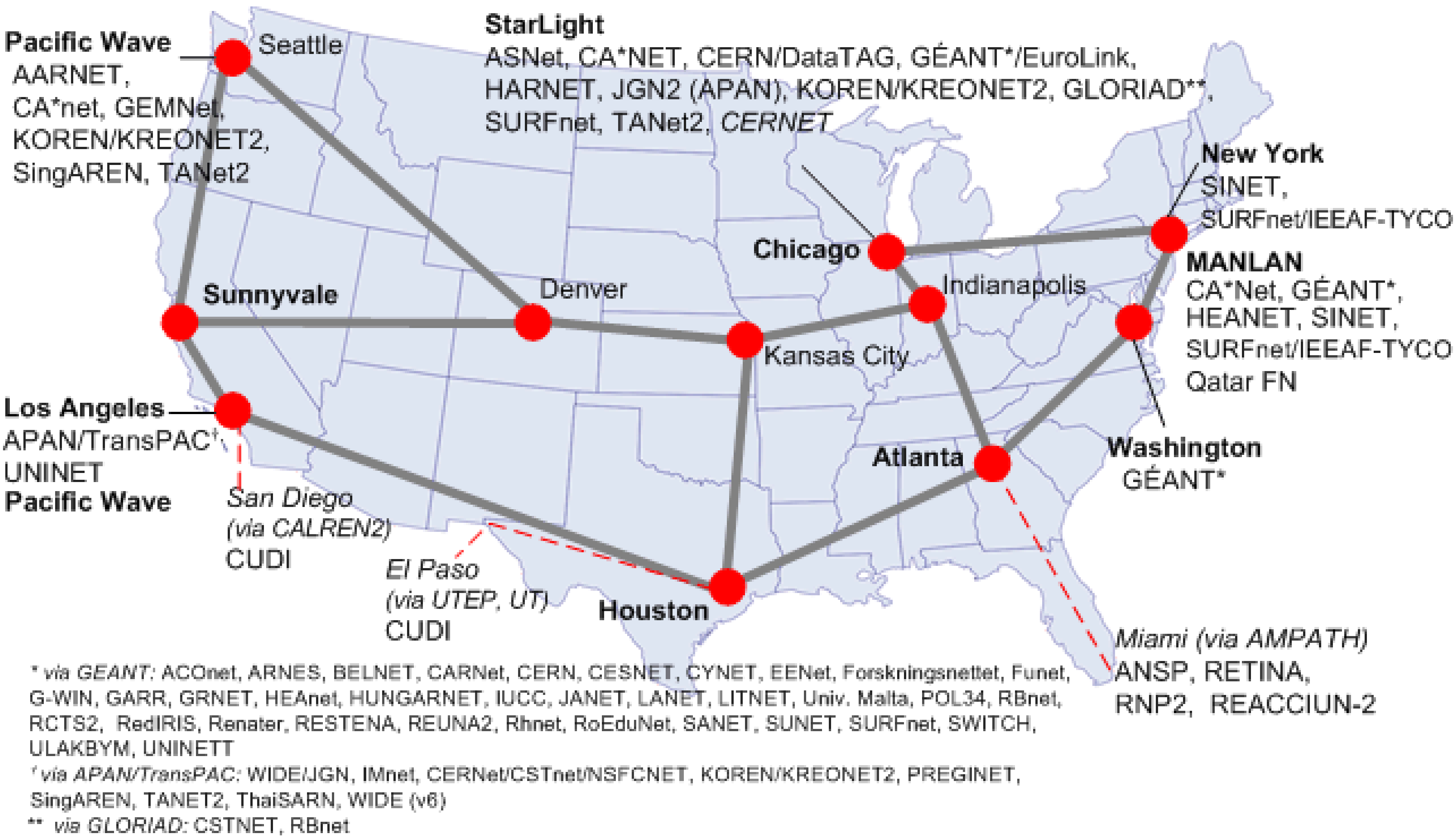
Abilene

- 10 Gbps IP Backbone
  - Carrier-provisioned OC 192 in the middle
  - Juniper T 640 routers
  - Best-effort, “overprovisioned”
  - < 2Gbps normal load
- IPv4 and IPv6, native multicast, MPLS

Abilene

- Research facilitation (data + collocation)
  - Abilene Observatory project http://abilene.internet2.edu/observatory/
- 35 Connectors
  - Mostly regional aggregators
  - Some universities
- 246 Participants
- Extensive domestic and international Research and Education (R&E) peering

Abilene, with International Peers



Measurement Capabilities

- One way latency, jitter, loss
  - IPv4 and IPv6; On-demand available
- Regular TCP/UDP throughput tests – ~1 Gbps
  - IPv4 and IPv6; On-demand available
- SNMP (Abilene NOC)
  - Octets, packets, errors; collected frequently
- Flow data (~”netflow) (ITEC Ohio)
  - Addresses anonymized by 0-ing the low order 11 bits
- Multicast beacon with historical data (NOC)
- Routing data
  - Both IGP and BGP - Measurement device participates in both
- Router data (NOC): “show” snapshots + syslog
  - See also Abilene Router Proxy

Colocation for Research

- PlanetLab
  - New future: MPLS links so can act as router w/own links and peering
- AMP: active measurement from NLANR MNA (San Diego Supercomputer Ctr.)
- PMA: passive monitoring. Currently every interface on the IPLS router is instrumented. From NLANR/MNA.

Other network research stuff

- Buffer sizing project (Stanford):
  - Reduce buffers available to router interfaces (software controlled)
  - Take an anonymized but correlated packet trace
  - Look for throughput and latency anomalies
- Rapid raw SNMP to test link capacity measurement programs
- Occasionally run programs on behalf of researchers on backbone machines

Similarities with Commercial Networks…

- Daily usage among universities
  - It’s IP
  - Email, web, file sharing, video conferencing, …
  - If you communicate with another university (or R&E entity) it just works

… and Differences

- Big Science datasets
  - Lots of very large transfers
  - Seen 7Gbps UDP from Caltech to CERN
- Lots of high-end video
  - 20 Mbps streams
  - 100's of Mbps HDTV
- Multicast

Security…

- Concerns similar to other larger backbones
- We have Arbor Peakflow SP
- Minimal staff… we distribute some work to the Research & Education ISAC
- Lots of small operators (with big pipes) (and small staffs) tied together
  - Operational coordination is a challenge

Routing

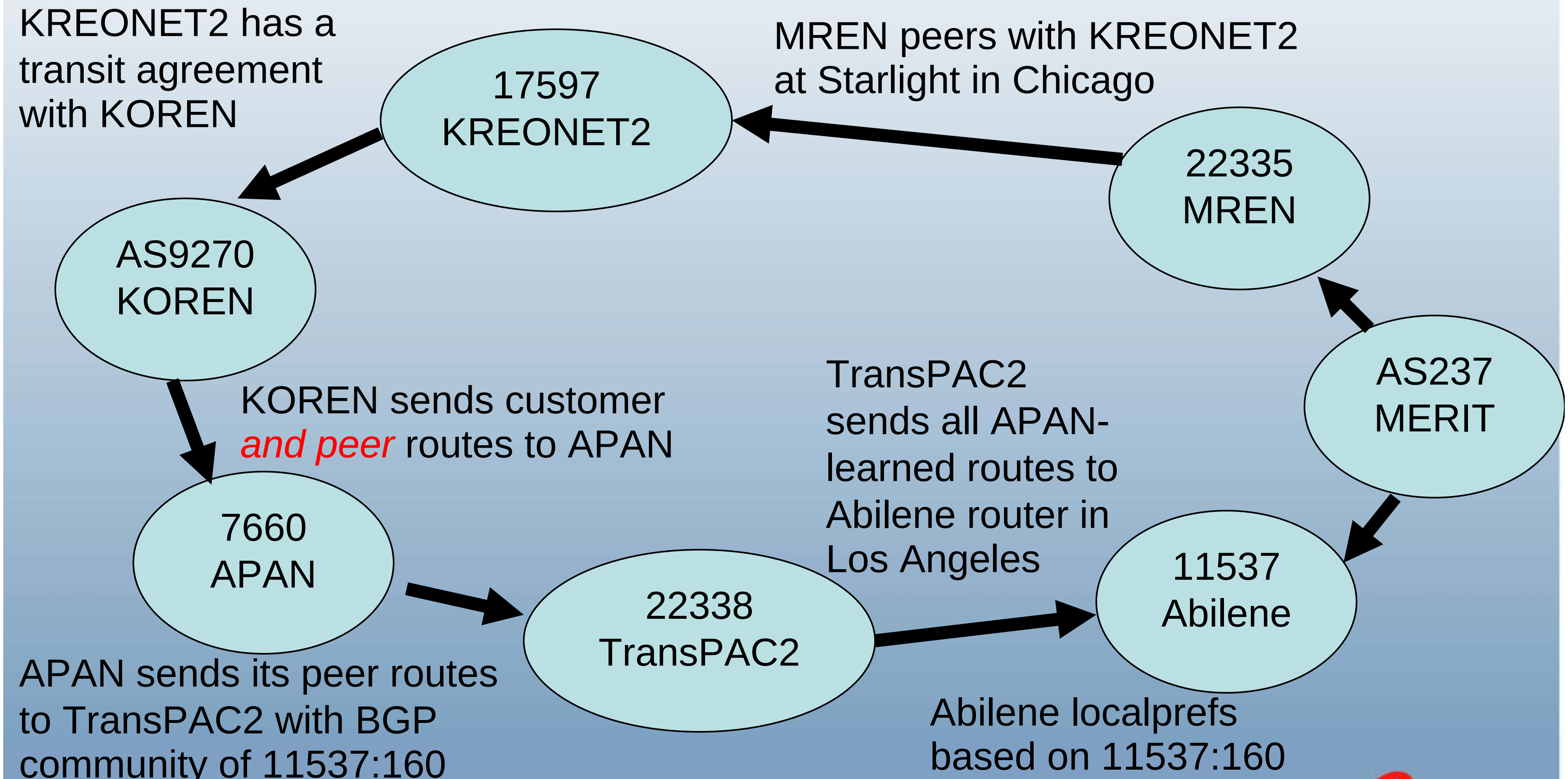
- Unlike the commercial world where business concerns drive a sane, mostly hierarchical structure, R&E networks tend to peer (and provide transit) promiscuously
- Connectivity is paramount
- Often driven by demos
- Special peerings/announcements don't necessarily get taken down, are forgotten

Michigan is in Korea?

- Chris Robb of the NOC adds a Michigan route from MERIT to Abilene
- Router says get there via Korea(!)

```
chrobb@IPLSng-re0# run show route 198.110.96.0
inet.0: 9808 destinations, 15350 routes (9775 active, 0 holddown, 35
hidden)
+ = Active Route, - = Last Active, * = Both
198.110.96.0/20
           *[BGP/170] 00:45:42, MED 100, localpref 160, from 198.32.8.198
                AS path: 22388 7660 9270 9270 9270 17579 22335 237 I
                > to 198.32.8.81 via so-3/2/0.0
            [BGP/170] 00:29:46, MED 10, localpref 140
                AS path: 237 I
                > to 192.122.183.9 via so-2/1/2.512
```

Michigan is in Korea?



INTERNET2®

Routing

- Previous example, comparisons to commercial networks, and potential solutions in a talk by Chris Robb: http://www.internet2.edu/presentations/spring06/20060424-routingissues-robb.pdf
- Can we agree on what “bad” routing is? Could be policy driven…

Routing

- Tend to end up with interesting, non-optimal routes that are hard to understand
- Need more RouteViews, Looking Glass
- Alternate routing tends to be the most interesting and hardest to capture (unless a failure exposes a weird route)

Other: End-to-End Performance

- In our world of distributed responsibility, how find the reason why don't get performance expect
- Today, should get 100Mbps end-to-end for our users. Median lg. flow: 3-ish on Abilene.
- Additional instrumentation to help segment problem (moving toward perfSONAR, joint work with Europeans, other R&E networks)
- http://e2epi.internet2.edu/

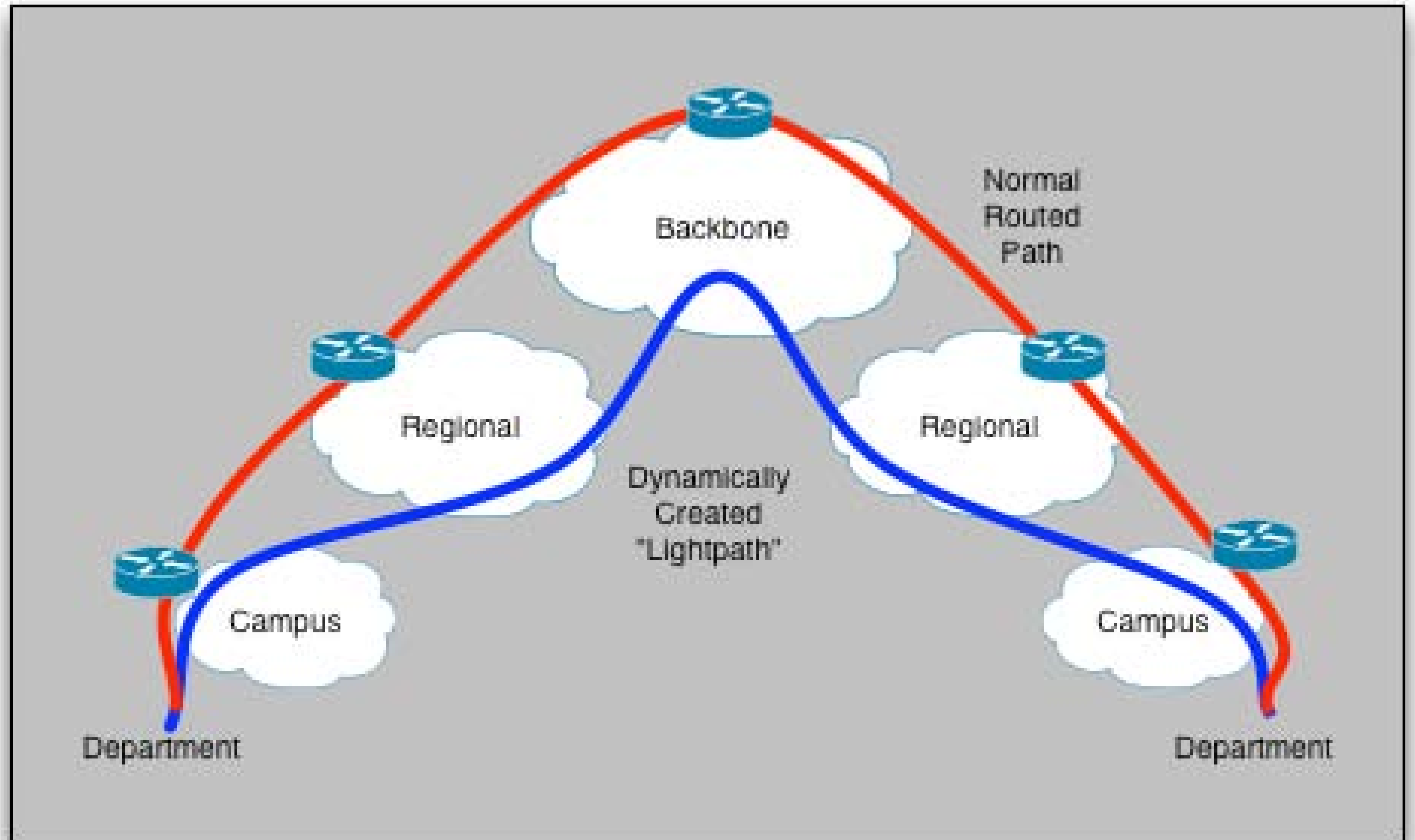
Next revision of Abilene

- October 2007 - End of current Abilene transport agreement (SONET links)
  – Replacement available by June 2007
  – Network design time frame: 2007-2012

New Network Requirements

- Requirements multi-dimensional, for example:
  - Provide capabilities at all network layers (layer)
  - Provide capabilities for both short term and long term applications or projects (duration)
  - Provide capabilities at a variety of different levels of robustness, from production to experimental (robustness)
- An infrastructure consisting of dark fiber, a significant number of waves, and a production quality IP network
  - Create a new architecture for the R&E community
- New features: dynamic provisioning, hybrid models (combinations of circuit and packet switching)

Next Generation Overview



Fast forward to the past?

- Gee, looks similar to ATM SVCs
- Telephony circuits
- Revenge of the “bell-heads”?
- Also seems like a continuation of the search for QoS / guarantees
  – Does Internet video really work?
  – Sometimes hard to overprovision everywhere

What’s different?

- Multiple administrative domains
- Some applications do highly desire dedicated capacity
  - Physics Large Hadron Collider data: 2*10Gbps CERN to US hot all the time
  - e-VLBI: 1Gbps + from multiple radio telescopes to a central correlator
  - “GRID” middleware wants to schedule the network like it schedules CPUs

What’s different?

- Large research networks cooperating on experimentation and implementation (Internet2, GEANT, ESnet)
- Some promising control plane technology

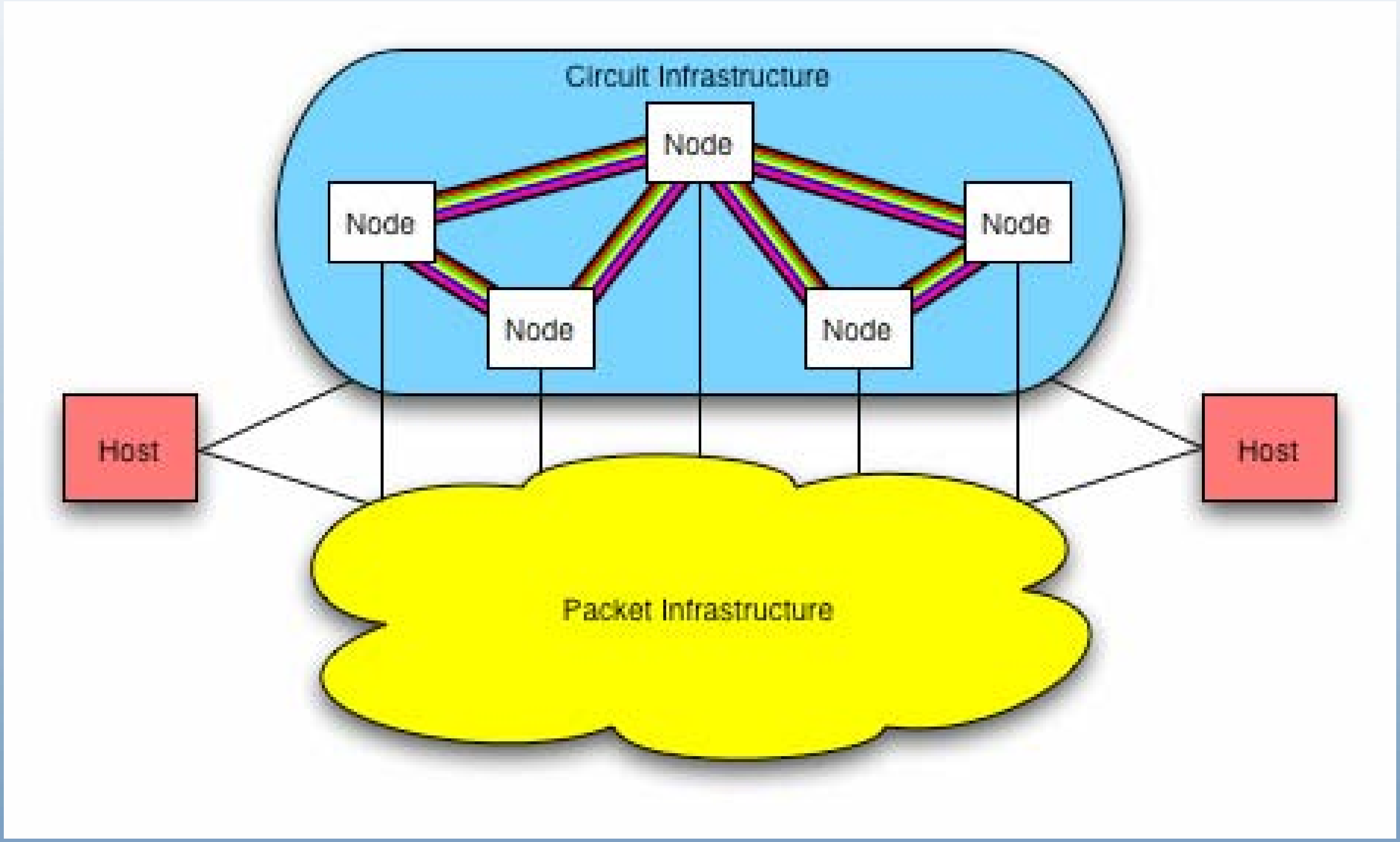
Dynamic Provisioning

- Dynamic provisioning across administrative domains
  - Setup on the order of seconds to minutes
  - Durations on the order of hours
- Switching may require unique partnerships and development of capabilities on hardware platforms
  - For example, being able to isolate user capabilities at switching nodes
  - There is interest from commercial carriers from the point of view of providing additional services
- All this should be transparent to the user
  - View as a single network
  - Hybrid aspects must be built into the architecture

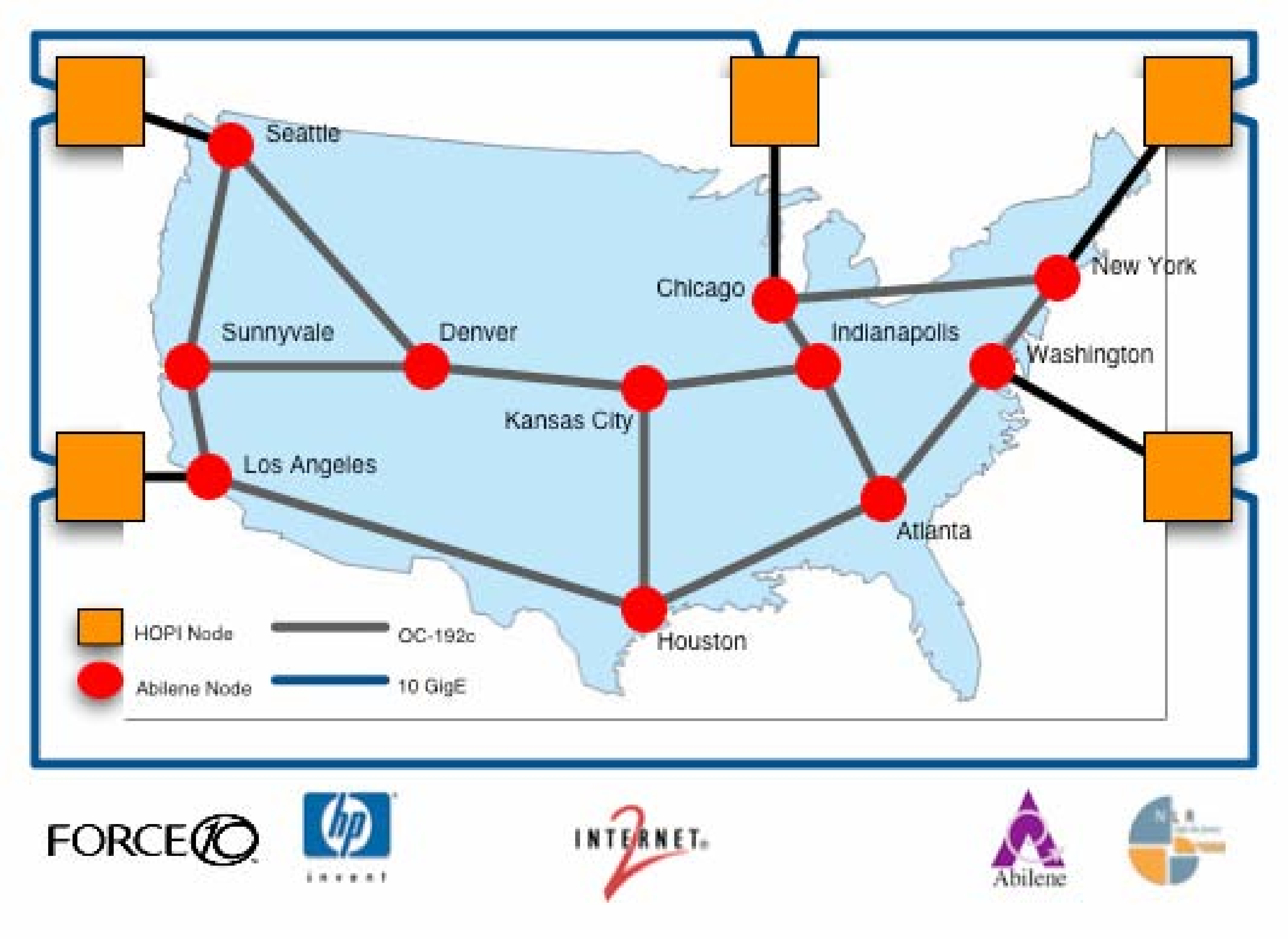
HOPI Project - Overview

- We expect to see a rich set of capabilities available to network designers and end users
  - Core IP packet switched networks
  - A set of optically switched waves available for dynamic provisioning
- Examine a **hybrid** of shared IP packet switching and dynamically provisioned optical lambdas
- HOPI Project – Hybrid Optical and Packet Infrastructure - how does one put it all together?
  - Dynamic Provisioning - setup and teardown of optical paths
  - Hybrid Question - how do end hosts use the combined packet and circuit switched infrastructures?
  - HOPI is a testbed for experiments, not a production network
  - We are using experiment results to guide the next generation network

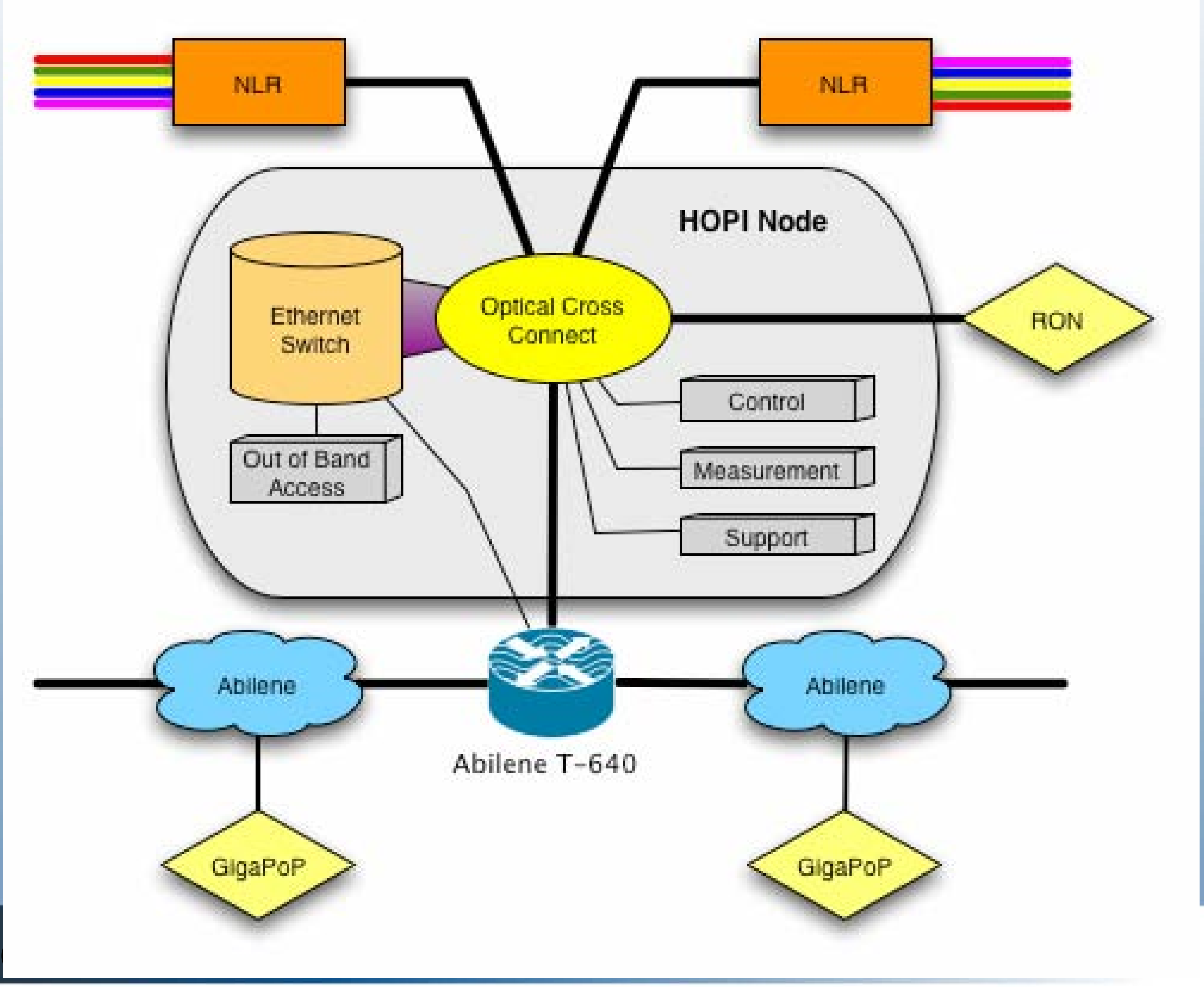
HOPI General Problem



HOPI Topology



HOPI Node


NLR
NLR
HOPI Node
Ethernet Switch
Optical Cross Connect
RON
Control
Out of Band Access
Measurement
Support
Abilene
Abilene
Abilene T-640
GigaPoP
GigaPoP

HOPI Deployment

- Connections to other US testbeds:
  - UltraLight (High-energy Physics)
  - UltraScienceNet (Department of Energy)
  - CHEETAH (National Science Foundation funded project)
  - DRAGON (another NSF funded project)
- Anticipate a circuit from NY to London (through MANLAN) to attach to GEANT2 testbeds (~July 2006)
- First experiments: cross-domain control plane

INTERNET2®

Next Generation Design

- Use dedicated fiber from Level3
  - They maintain fiber, optical platform
  - We have full control over provisioning
- Built on Infinera platform providing innovative optical technology
  - Simple and convenient add/drop technology
  - Simple and convenient wave setup
  - Demonstrated high reliability in initial period of operation on the Level3 network
  - Economics of Infinera system are disruptive in the market place

Next Generation Design

- Control Plane
  - Initial: manual, or “semi-manual”
  - Near term: carry over DRAGON control plane from HOPI testbed
  - Long term: ?

DRAGON

- Dynamic Resource Allocation over GMPLS Optical Networks
- NSF-funded project
- Network Aware Resource Broker
- Virtual Label Switched Router
- Application Specific Topologies
- http://dragon.east.isi.edu/

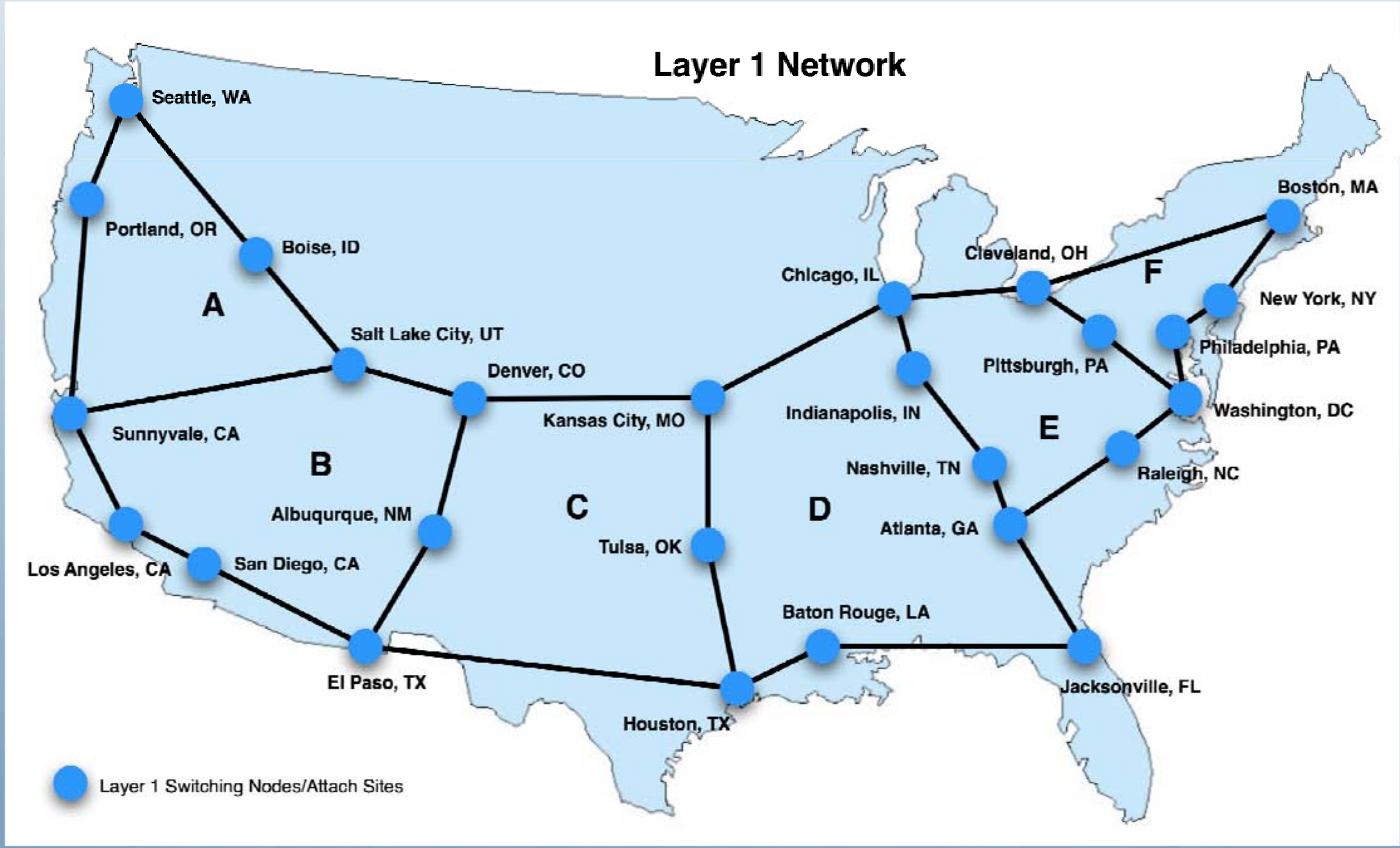
Next Generation Design

- Architecture has maximum flexibility. Every connector can access every wave on the system if needed
- System includes grooming capabilities - lightpaths can be built over Ethernet or SONET
  – Can take advantage of advanced SONET cpabilities like GFP, VCAT, and LCAS
  – Capable of lightpath provisioning to the campus

Next Generation Design

- Attachment expected to evolve to 2 x 10 Gbps connections
  - 10 Gbps IP connection
  - 10 Gbps point-to-point connection (capable of STS-1 granularity lightpaths provisioned in seconds), most likely provision using Ethernet (GFP based)
  - Hybrid capabilities
- Expect 20 - 24 connectors
  - Simple and consistent connection scheme
  - Promoting aggregation
  - Need input and discussion on exceptional cases

Layer 1 Topology

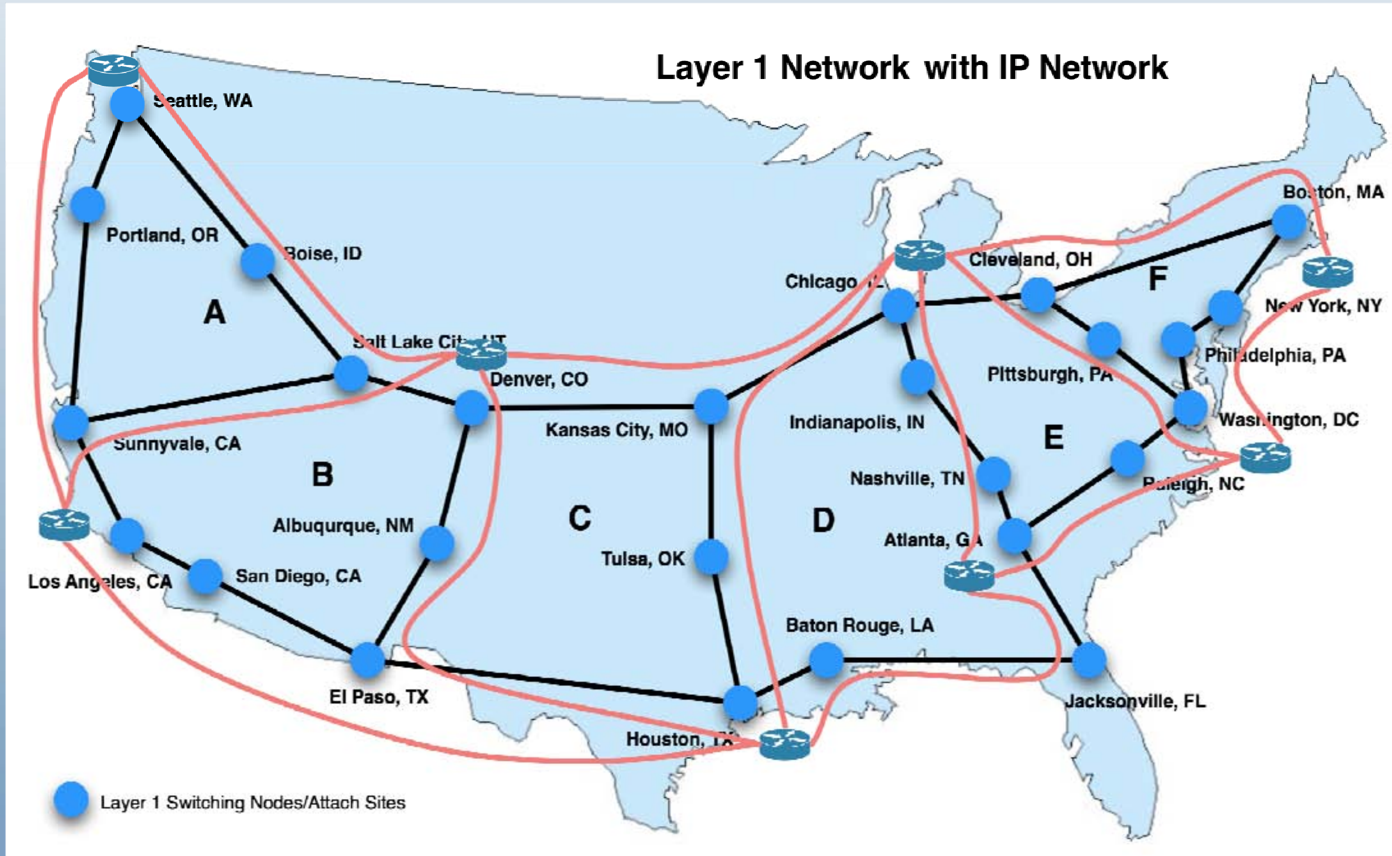


PROVISIONAL TOPOLOGY – SUBJECT TO DISCUSSION

IP Network

- IP network built on top of optical system
  – High reliability - architecture provides a variety of protection options
  – Possible commercial service offering - standard connection may include commodity services
  – Current plan is to continue to use highly reliable Juniper routers, but open to new technologies
  – May use fewer routers, emphasizing point-to-point capabilities and hybrid networking
  – Potential near term option of 40 Gbps

Layer 1 Topology with IP Network



PROVISIONAL TOPOLOGY – SUBJECT TO DISCUSSION

Observatory

- Intend to continue current IP layer observatory
- Add circuit information (control plane, errors)
- There isn't much compared with IP (utilization?)
- We will likely form a working group to do requirements and initial design

Challenges: Control Plane

- Interoperation among multiple administrative domains still a prototype
- What about settlements/economics
  - Will there be any?
  - If so, what's the least overhead required
  - "land grabs"?
- There is now a new set of potential control plane attacks

Challenges: Control Plane

- Ensuring easy-to-understand picture of allocations and unused capacity
- Verifying you deliver what was asked for

Challenges: Debugging

- If end-to-end errors with concatentated segments, possibly using different technologies (SONET, Ethernet, MPLS), find the source…
- What do you need to verify the entire end-to-end path works as a system?

Challenges: Service Definition

- Want “Gigabit Ethernet” link
  - VLAN tags? 9000 byte MTUs? Spanning Tree?
  - What if cross traffic introduces jitter?
  - What if served by bonded smaller channels in middle, and that introduces some reordering?
- Want a service that is predictable, verifiable, repeatable… and end-to-end
- http://dragon.maxgigapop.net/twiki/bin/view/DRAGON/CommonServiceDefinition

Challenges: Circuits + Routing

- People will use them as an “end-around” other security-motivated restrictions (sometimes with cause -- pieces of older campus infrastructure)
- People will end up routing IP over it, creating new channels by mistake

References

- http://abilene.internet2.edu/observatory/
- http://www.internet2.edu/presentations/spring06/20060424-routingissues-robb.pdf
- http://networks.internet2.edu/hopi/
- http://dragon.maxgigapop.net/twiki/bin/view/DRAGON/CommonServiceDefinition
- http://dragon.east.isi.edu/
- http://cans2005.cstnet.cn/down/1102/A/afternoon/DRAGON_at_CANS2005%20v2_2A.pdf

References

- http://networks.internet2.edu (next gen)
- http://events.internet2.edu/2006/designws/
  (background material, presentations on the new network)

www.internet2.edu

INTERNET 2®