It Hurts When IP

Effort to Normalize R&E Routing Policy When There are Too Many Choices

Who are We? Who are you?

- Hans Addleman - Indiana University International Networks
- Eli Dart - LBNL / ESnet
- Jon-Paul Herron - Indiana University GRNOC
- Tom Johnson - Indiana University GRNOC / Indiana Gigapop
- Jason Lee - LBNL / NERSC
- Nathaniel Mendoza - TACC
- Jason Zurawski - LBNL / ESnet

Audience Intros

Agenda

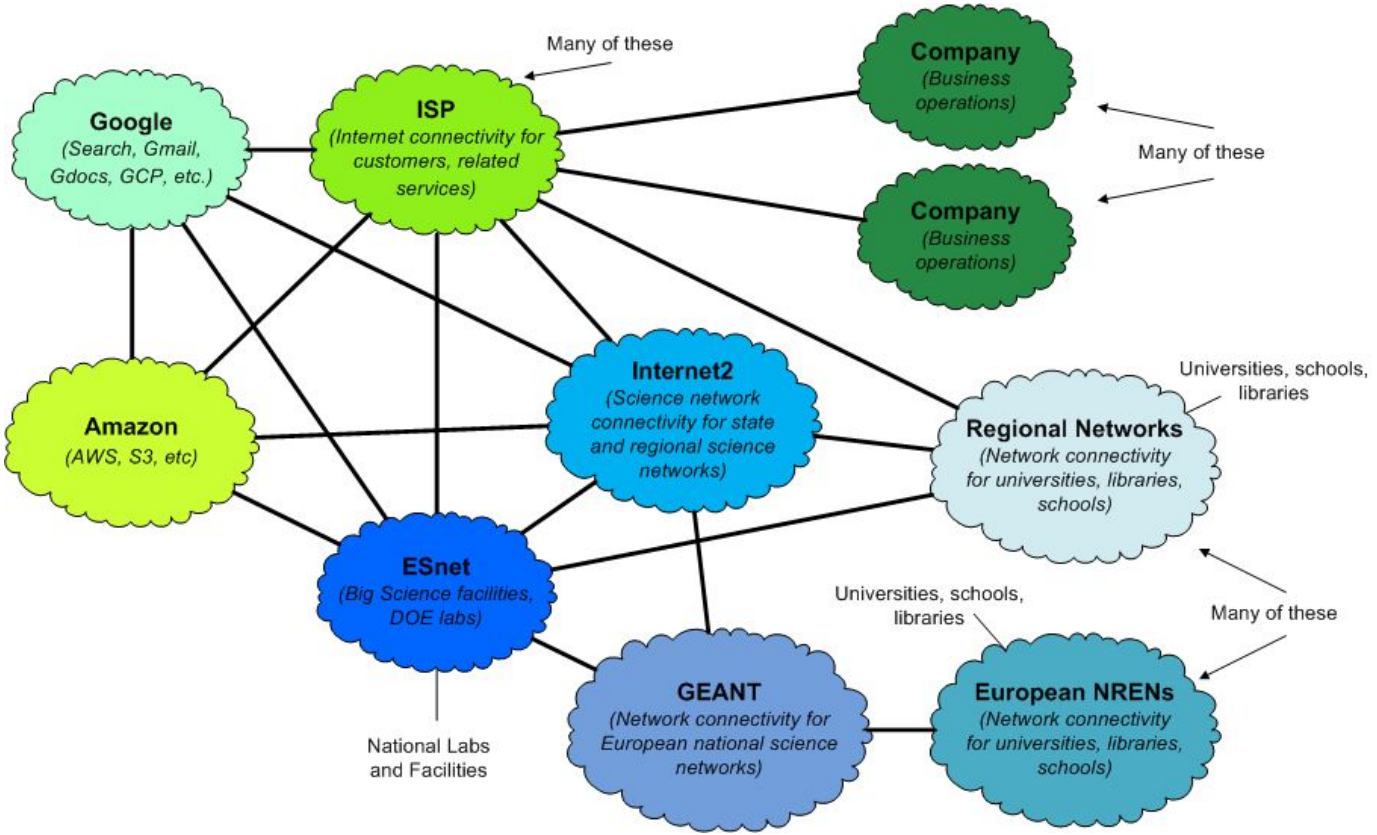
- Introduction - Jason Z
- R&E routing architecture - Eli
  - What it is
  - Why it matters
  - Examples of problems
  - Simplified ESnet Routing Architecture
- Policy Knobs on Internet2 - Hans
- Connector Usage - Tom
- Discussion - All

Why are we here?

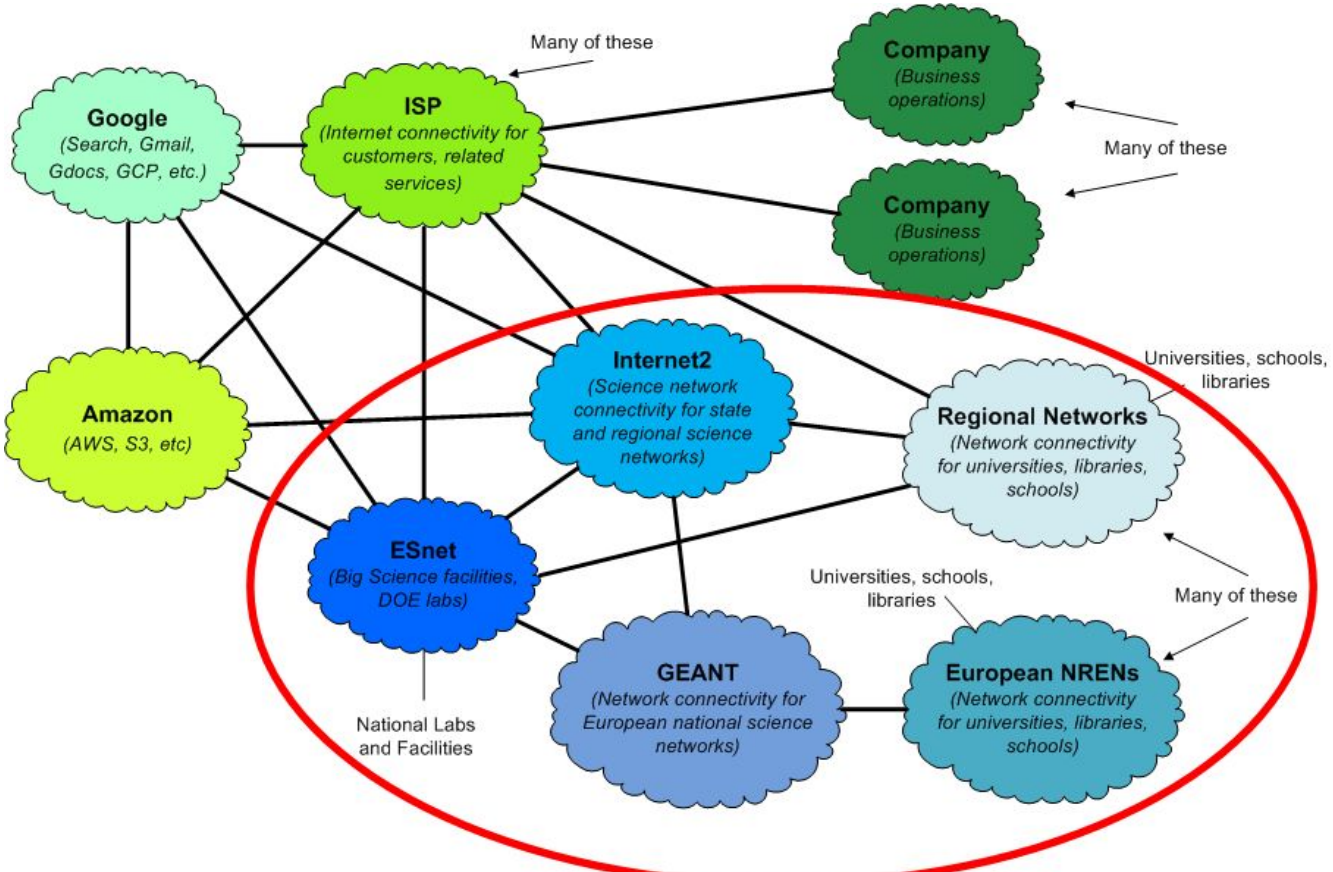
R&E Routing Architecture

- Keep R&E traffic on R&E paths if possible
  - Bandwidth
  - Performance Engineering
  - Deterministic behavior
- We all have to do our part
  - All routing decisions made locally
  - Emergent behavior is important
- Pictures to follow

R&E Routing Architecture

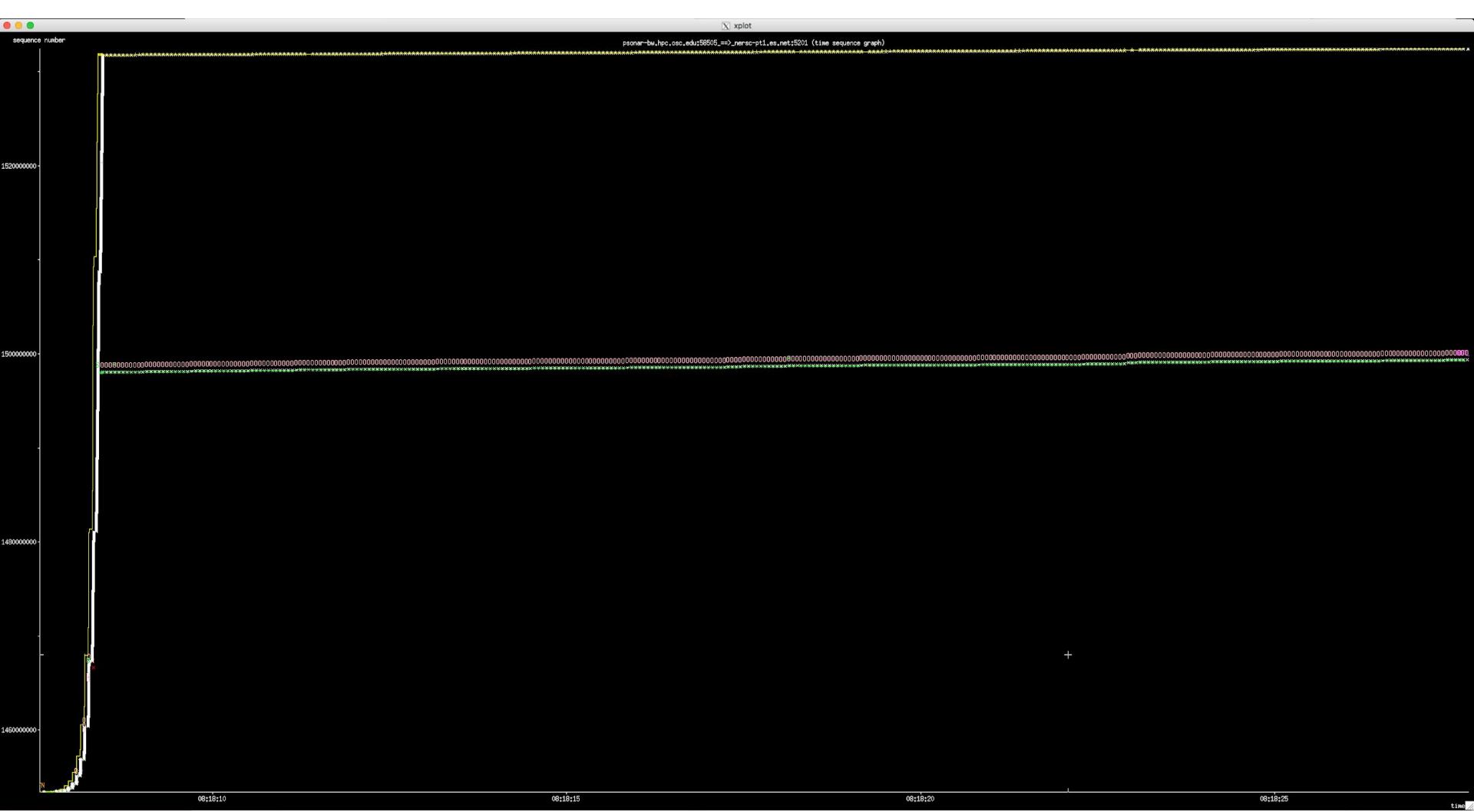


R&E Routing Architecture



Why does this matter? Example 1 - OSC

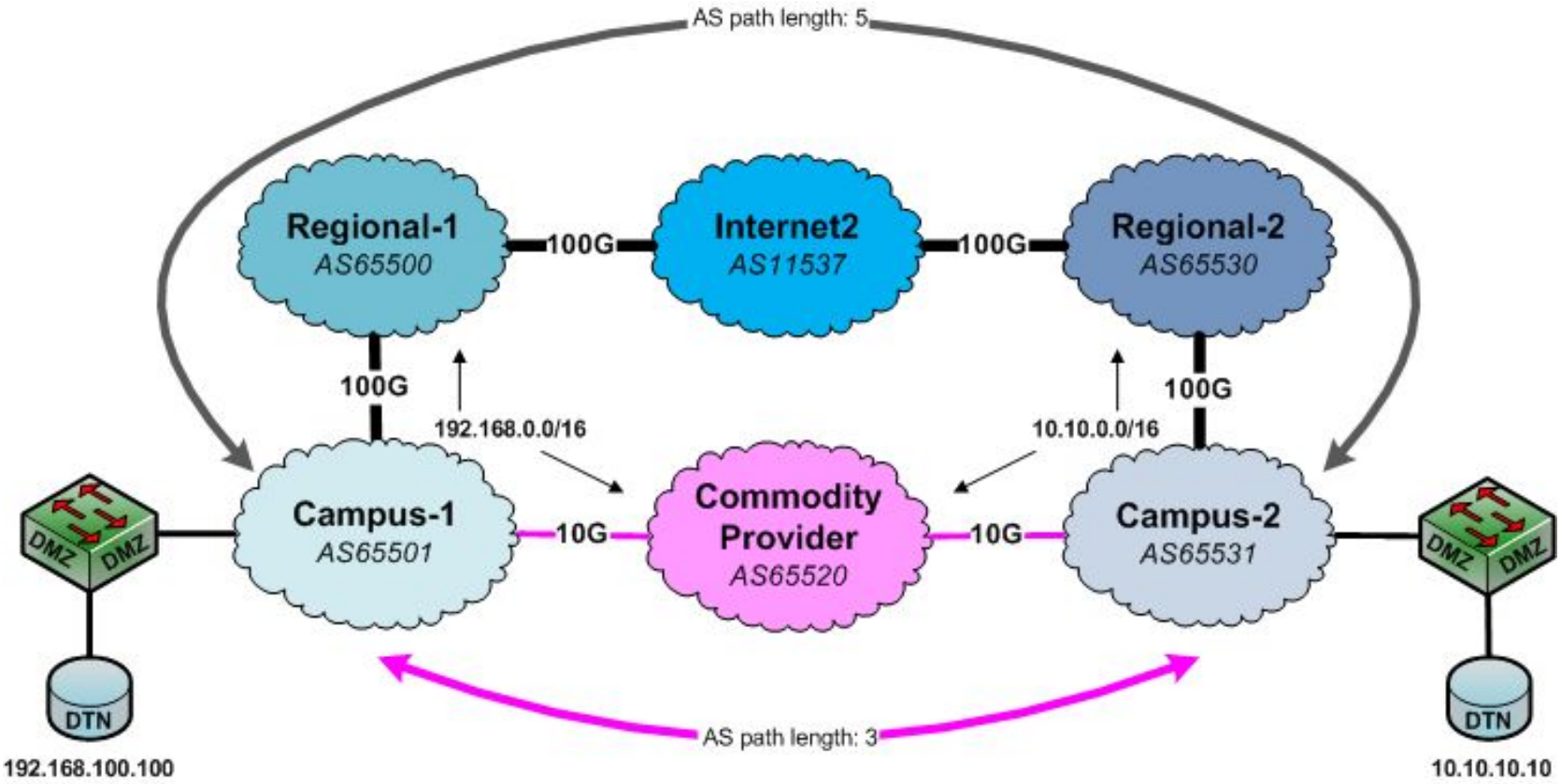
- Data transfers between Ohio Supercomputer Center and NERSC were slow
- Turns out they were going over commodity instead of R&E paths
- Commodity networks often throttle high-speed flows
  - In their world a traffic spike means a DoS attack
  - In our world it’s another scientist doing their normal thing
- 1000-word picture follows...


xplot
sequence number
psonar-bw.hpc.csc.edu:58505_==>_nersc-pt1.es.net:5201 (time sequence graph)
1520000000
1500000000
1480000000
1460000000
08:18:10
08:18:15
08:18:20
08:18:25
time

Other examples

- https://connect.geant.org/2017/05/15/taking-it-to-the-limit-testing-the-performance-of-re-networking
- https://indico.geant.org/event/1/contributions/11/attachments/47/207/190521_-_PT_TNC2019_v8.pdf
- Common theme: R&E networks are engineered to support science while commodity networks are not
  - This shouldn't surprise us - high speed science is what we've been doing for years
  - But this means we have to keep the science traffic on the science networks!
- To first order, this means we need to override BGP's use of AS path length when choosing between R&E and commodity paths
  - R&E path will be longer in the general case (more organizations involved)
  - R&E bandwidth capabilities (provisioned capacity and per-flow data rate) much higher
  - Use normal BGP route selection between R&E routes, and between commodity routes

BGP AS Path Length Illustrated



ESnet Routing Architecture (High-Level, Simplified)

- Routing policy applied at ingress (import policy on peerings)
  - Routing policy sets communities based on peering type
  - Routing policy sets localpref set based on peering type - simplified version:
    - ESnet site - high
    - R&E peering - medium
    - Commercial Peering - low
    - Transit - very low
  - Communities control route announcement behavior to sites and peers
  - Localpref controls forwarding behavior within ESnet network
- This allows us to group routes based on connectivity capability and type of peer organization, and use normal BGP route selection within those groups
  - Forwarding is sane and high performance
  - This is more complex than a campus needs (we’re a national backbone), but ideas still hold

Handoff from Eli to Hans

Example 2 - Institution Redacted

- 2 peerings to Regional provider.
  - 1x100G, 1x10G
- Asymmetry caused traffic to come back into campus via the congested 10G
- Asymmetry in Campus network in addition

Before

Interval	Throughput
0.0 - 10.0	27.97 Mbps

After

Interval	Throughput
0.0 - 10.0	717.75 Mbps

Example 3 - Institution Redacted

- Routing Asymmetry
  - Preferring comercial path out
  - R&E path in

1 University 1 1.103 ms mtu 9000 bytes
2 Regional 2.163 ms mtu 1500 bytes
3 Regional to ISP link 5.425 ms mtu 1500 bytes
4 Hurricane Electric (206.223.118.37) 13.309 ms mtu 1500 bytes
5 Hurricane Electric (184.105.81.205) AS6939 17.328 ms mtu 1500 bytes
6 Hurricane Electric (184.105.65.166) AS6939 21.361 ms mtu 1500 bytes
7 Hurricane Electric to University 2(184.105.48.246) AS6939 24.856 ms mtu 1500 bytes
8 University 2 mtu 1500 bytes
9 University 2 mtu 1500 bytes
10 University 2 perfSONAR node mtu 1500 bytes

Example 3 Part 2 - Redacted

University 2 Route *[BGP/170] 9w6d 05:38:46, MED 0, localpref 150
University 2 Route *[BGP/170] 1w2d 09:49:01, MED 0, localpref 100

- Multiple Routing tables advertised from Regional to Campus
- Does the University blend those tables and just go with BGP’s best choice?

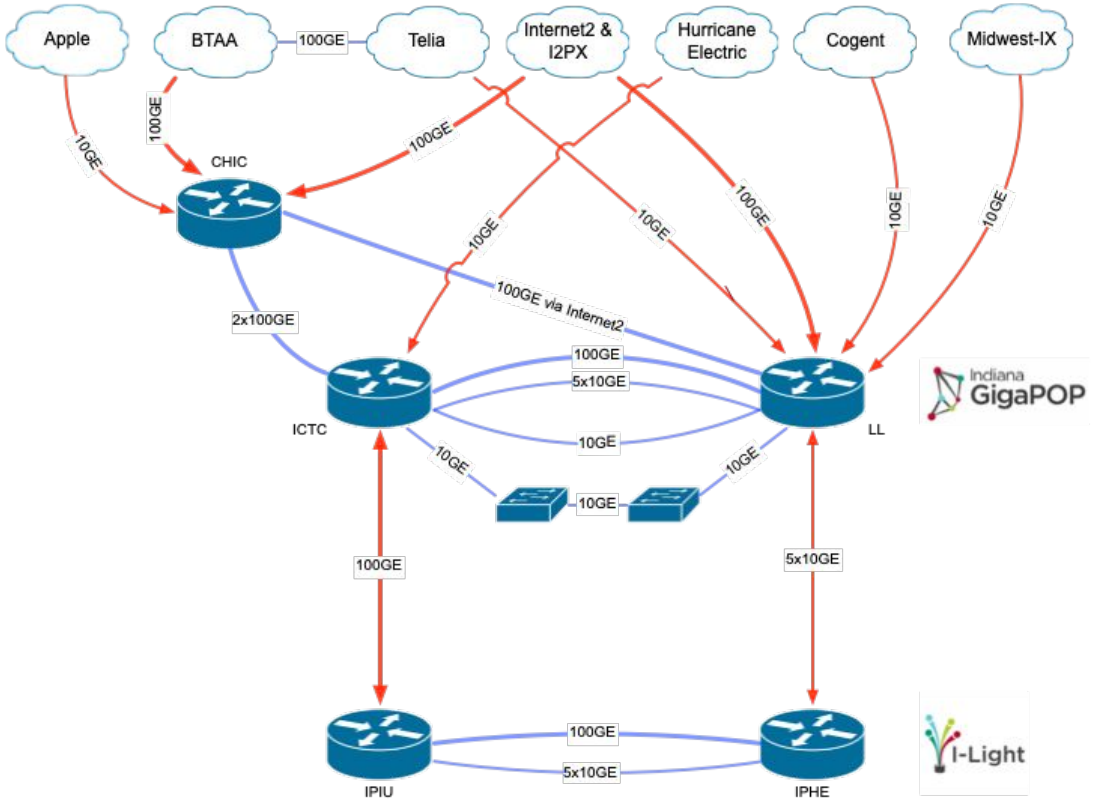
Common route steering tools

- LocalPref (BIG HAMMER)
  - Higher Better
- AS Path Padding
  - Shorter AS path better
  - Don’t overpad!
- MED
  - Lower Better
- BGP Communities
  - Can make changes to routing policy based on community tags on each routing prefix
  - Must be setup by each network and published publicly
  - Also can provide clues about the prefix

Traffic Steering mechanisms offered by Internet2

- BGP Communities!
  - https://noc.net.internet2.edu/i2network/maps-documentation/documentation/bgp-communities.html
- LocalPref?
  - Default - 100
  - 11537:40 - Low
  - 11537:160 - High
- Prefix identification?
  - 11537:5004 - Amazon
- Where does the prefix enter the network?
  - 11537:242 New York
- Emergency!
  - 11537:911 - Discard all traffic destined to these prefixes!
- AS Path Padding?
  - 65001:65000 - prepend x1

Indiana GigaPOP (ING) & I-Light (ILN)



Traffic Steering offered by ILN & ING

How does Indiana Gigapop allow traffic to be steered by Customers?

- upto /24 & /48
  - Accept upto /24 & /48 from members
    - Upto /32 & /128 for RTBH via BGP Community
- BGP Communities (set LP)
  - https://indiana.gigapop.net/ingigapop/maps_documentation/documentation/indiana-gigapop-noc-support-bgp-communities.html
  - https://noc.ilight.net/ilight/maps_documentation/documentation/i-light-bgp-communities-information.html
- AS Path Prepend

Traffic Steering offered by ING & ILN

How does Indiana Gigapop stear traffic upstream to national and other providers?

- Two VRF’s (Commodity, R&E)
- BGP Communities (set adjacent network LP)
  - Example: Internet2 (three paths)
    - two primary (11537:160)
    - one backup (11537:40)
  - Example: Telia (Commodity Transit)
    - One primary (dedicated connection in Indianapolis)
    - One backup (via shared BTAA connection in Chicago)
- AS Path Prepend
  - Commodity Transit, ‘19782 19782 19782’
  - Commodity Peering, ‘’
  - R&E, ‘’

Where do we go from here?