A Hands-on Tutorial on BGP

University of South Carolina (USC)

Front Range GigaPop (FRGP)

Engagement and Performance Operations Center (EPOC)

Internet2 Technology Exchange

Denver, Colorado – December 5, 2022

Organizers

Jorge Crichigno
(USC)

Jason Zurawski
(ESnet / EPOC)

John Hernandez
(FRGP)

Organizers

Elie Kfoury
(USC)

Ali AlSabeh
(USC)

Workshop Website

- All material is posted on the website of the tutorial
  http://ce.sc.edu/cyberinfra/workshop_2022_ie2_bgp.html

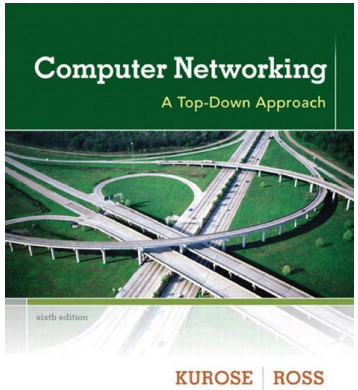
Agenda

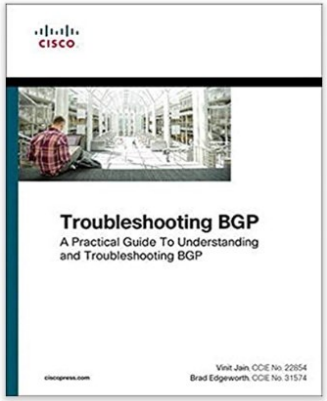
Time	Topic	Presenter
08:30 - 09:00	BGP and Research and Education Networks (RENs)	John Hernandez (FRGP), Jason Zurawski (EPOC), Jorge Crichigno (USC)
09:00 - 09:45	Hands-on Session BGP 1: Essentials of BGP, EBGP, IBGP	Ali AlSabeh (USC), Jorge Crichigno (USC)
09:45 - 10:00	Break	
10:00 - 10:30	Brief overview of Local Preference and MED Attributes	John Hernandez (FRGP), Jason Zurawski (EPOC), Jorge Crichigno (USC)
10:30 - 11:30	Hands-on Session 2: Local Preference and MED attributes	Ali AlSabeh (USC), Jorge Crichigno (USC)

BGP Fundamentals – Basic Terminology

Introduction to BGP

- BGP is complex
- Even after having read books and RFCs, students (instructors) may find it difficult to fully master BGP without having practiced it
- As critical protocol for the Internet, it is important to understand it





AS, IGP, EGP

- Routers are organized into Autonomous Systems (ASes or ASs)
- What is an AS (RFC 1771)?

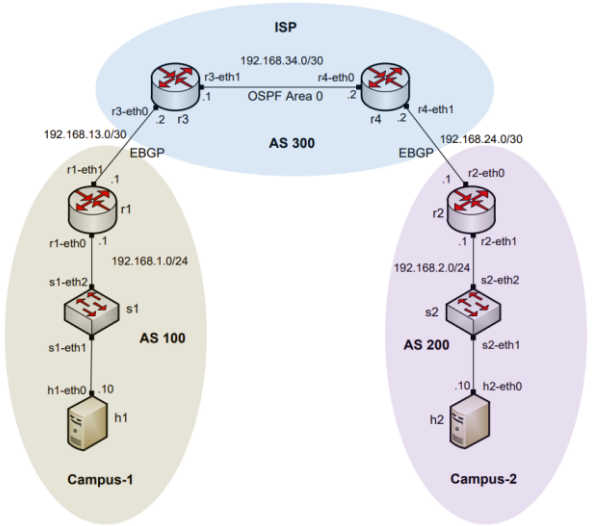
  "A set of routers under the single technical administration, using an IGP and common metrics to route packets within the AS, and using an EGP to route packets to other ASs."

- What is an Interior Gateway Protocol (IGP)?

  A routing protocol used to exchange routing information within an AS (e.g., RIP, OSPF)

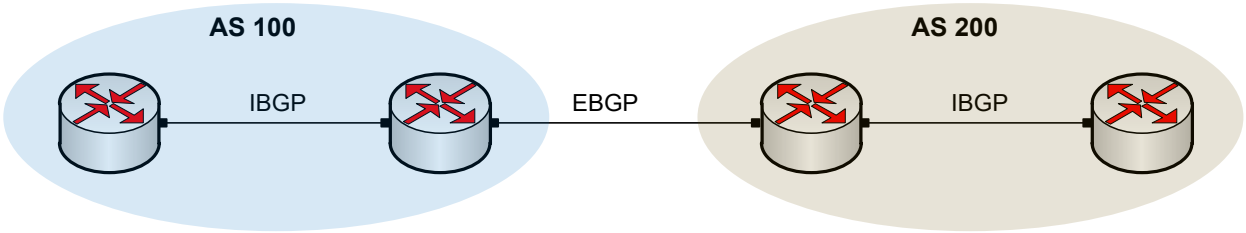
- What is an Exterior Gateway Protocol (EGP)?

  A routing protocol used to exchange routing information between AS



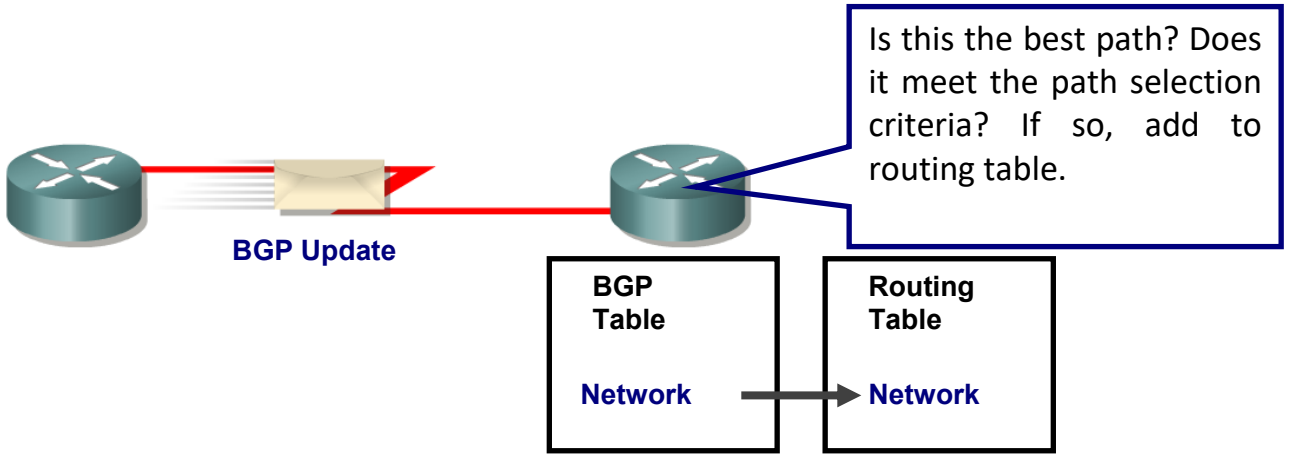
BGP Route Advertisements within an AS

- BGP advertisements from an AS to another is referred to as External BGP (EBGP)
- BGP advertisements within an AS is referred to as internal BGP (IBGP)



BGP – Best Path

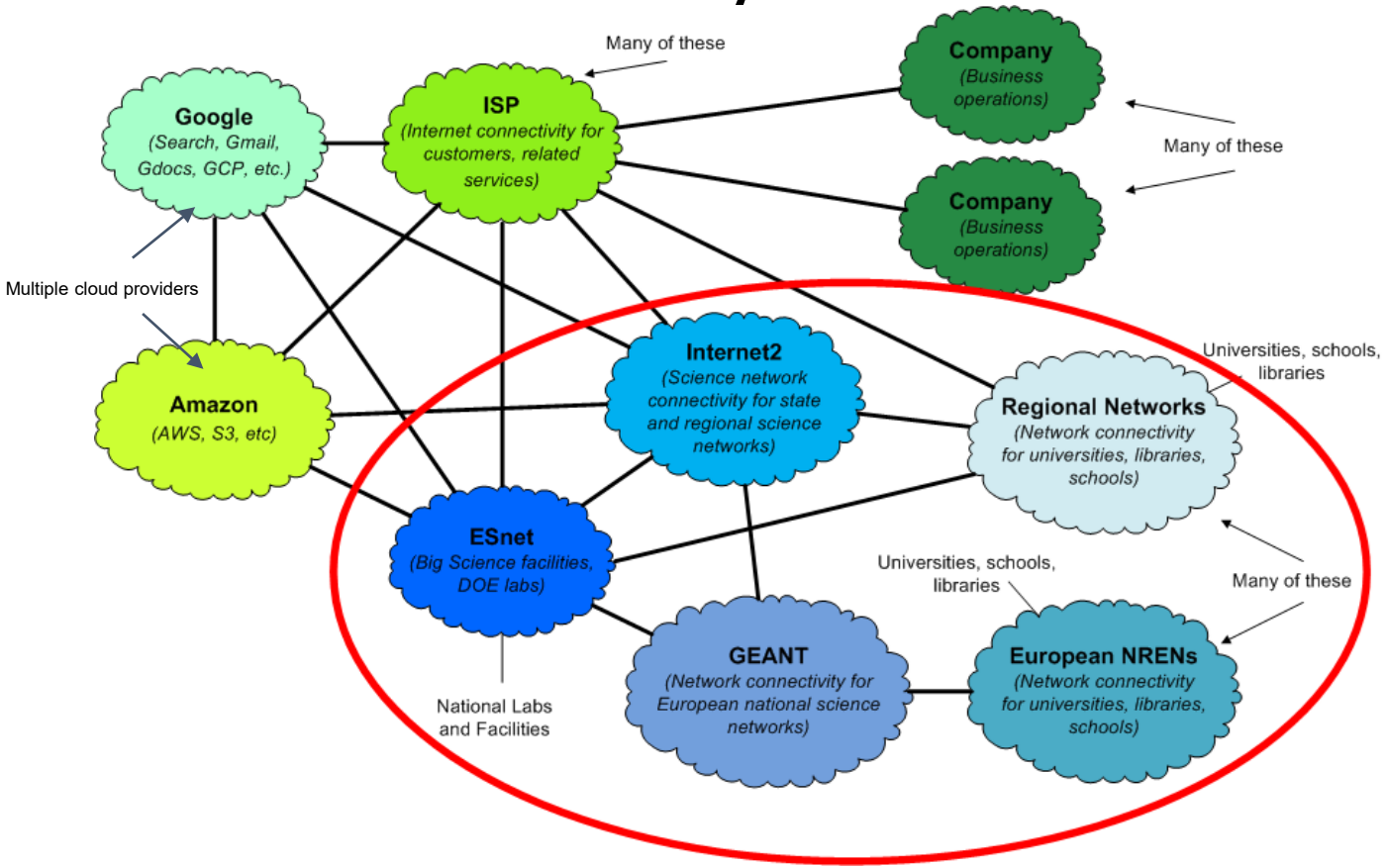
- The main goal is to provide interdomain routing
- BGP selects one path as the best path
- It places the selected path in its routing table and propagates the path to its neighbors



Rick Graziani, “Implementing Cisco IP Routing,” Cisco Press, 2015

BGP Use in R&E Networks (generalities)

R&E vs. Commodity



EPOC
Engagement and Performance Operations Center

R&E Routing Architecture

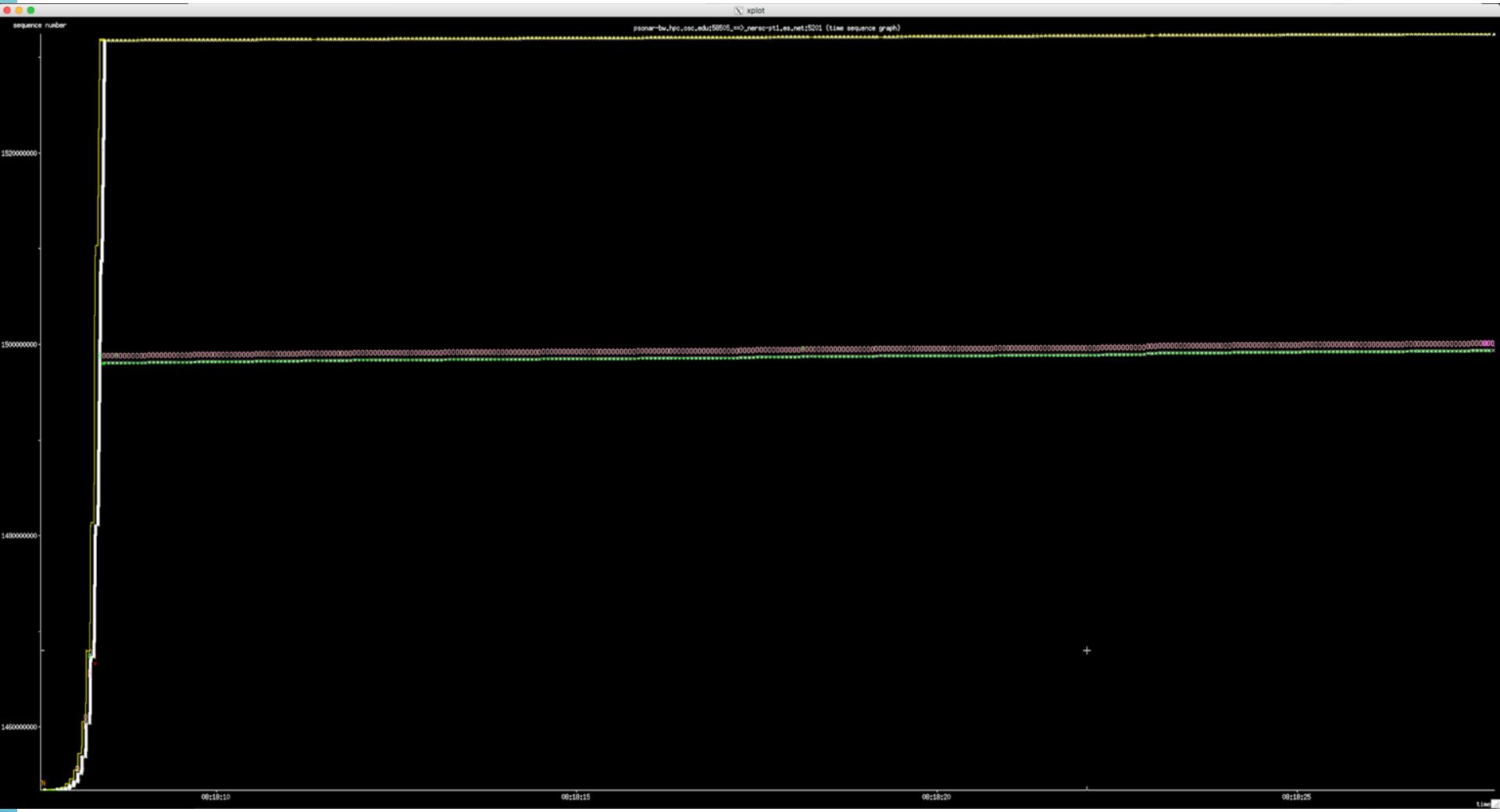
- R&E networks engineered for science traffic
- Keep R&E traffic on R&E paths if possible
  - Bandwidth
  - Performance Engineering
  - Deterministic behavior
- We all have to do our part
  - All routing decisions made locally
  - Emergent behavior is important
- Motivation examples to follow

EPOC
Engagement and Performance Operations Center

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
  - What does a multi-gigabit traffic spike mean?
  - **Commodity:** another DoS attack - this should be stopped!
  - **R&E:** another scientist doing normal things - this is core mission!
- What does it look like on the wire?

EPOC
Engagement and Performance Operations Center


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

Example 2

- 2 peerings to Regional provider.
  - 1x100G, 1x10G
- Asymmetrical traffic to coming back into campus via the congested 10G

Before

Interval	Throughput
0.0 - 10.0	27.97 Mbps

After

Interval	Throughput
0.0 - 10.0	717.75 Mbps

EPOC
Engagement and Performance Operations Center

Example 3

- 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 perfSONAR node mtu 1500 bytes

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

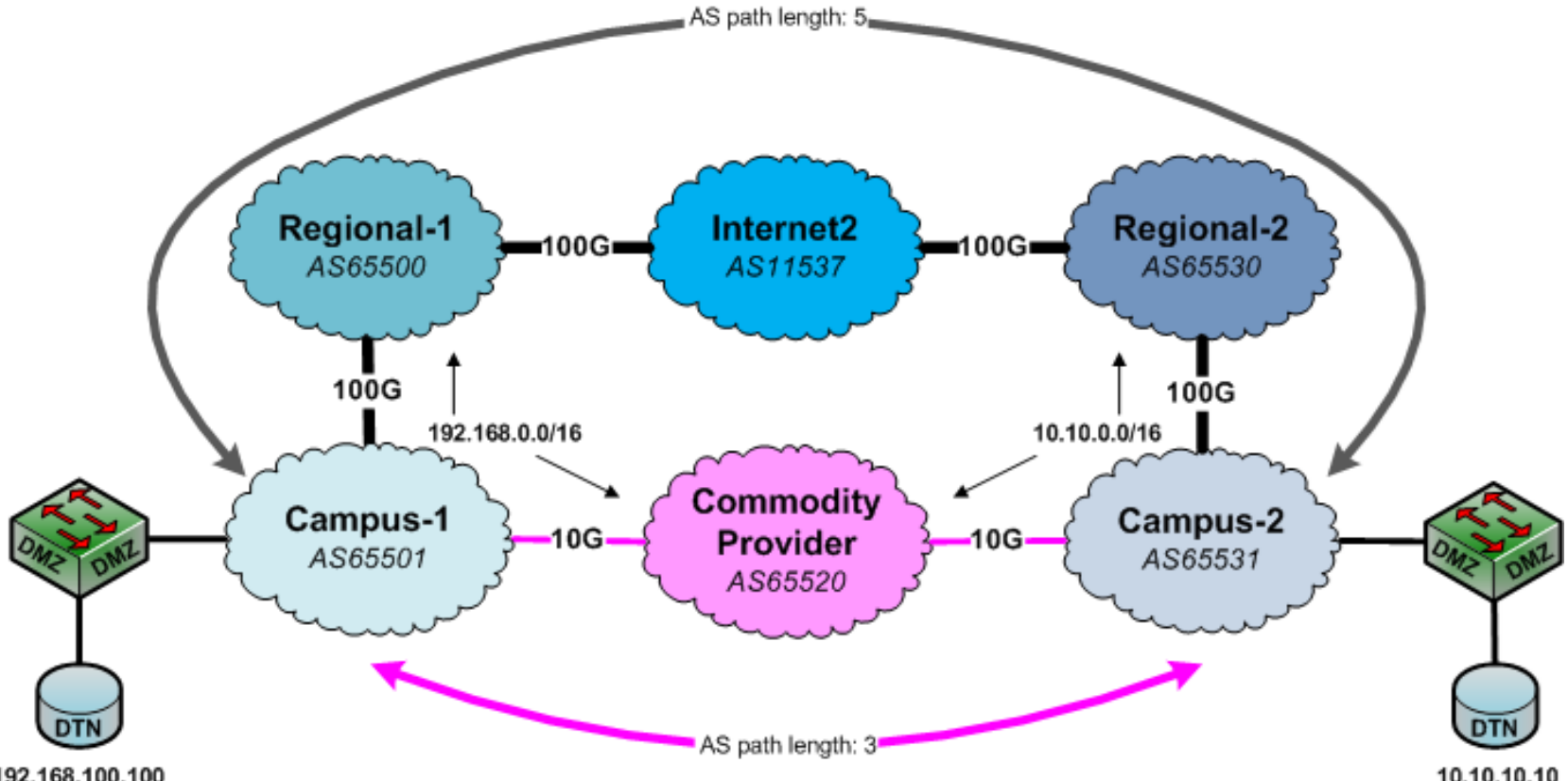
EPOC
Engagement and Performance Operations Center

So what do we do?

- To first order, this means we need to use BGP policy to keep R&E traffic on R&E networks
  - Announcements attract traffic
  - Routing determines the path the traffic takes through the network - BGP gives us the tools
- BGP is a path vector protocol
  - For a given prefix, the shorter AS path is preferred
  - If AS path length is the same, then other criteria are used, in order ("BGP path selection algorithm")
- 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)
  - Use normal BGP route selection between R&E routes, and between commodity routes
  - Remember - hop count is a legacy metric

EPOC
Engagement and Performance Operations Center

BGP AS Path Length Illustrated



BGP Use in ESnet

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

EPOC
Engagement and Performance Operations Center

Site Or Campus Routing Isn't Backbone Routing

- Many of the tools are the same (e.g. BGP policy)
- Goals are sometimes different
  - Backbone: multiple peers, resilience to route leaks, BCP38 filters, etc.
  - Campus: support security policy, keep transit costs down, etc.
  - High performance for science: common goal
  - Cost reduction: common goal (flat rate vs. charge by the bit)
- Don't try to replicate ESnet's policy on your campus perimeter
  - Not necessarily a good fit
  - **Know Your Network**
- Make sure you understand the tools you have, and use them to get as much as you can out of the infrastructure you've got
- Keep science traffic on science networks - every site has to do this unless your provider is explicitly doing it for you

EPOC
Engagement and Performance Operations Center

BGP Use in FRGP

FRGP Routing Architecture

- Front Range GigaPoP is a regional R&E network in Colorado / Wyoming / New Mexico
- Compact routing core consisting of 4 Juniper MX routers in Denver-area exchanges and strategic carrier locations.
- Dark fiber, DWDM and Metro Ethernet technologies to aggregate customer access
- We use a very similar BGP policy to Esnet
  - **Localpref** groups – same idea (**Customer > Research > Commercial Peer > Transit**)
  - On a Campus, you may only be concerned with Research and Transit
- We also use BGP Communities to tag groups of routes as they are learned
  - This helps us with announcements (export policy)
- Two explicit routing instances : research (vrf) and commercial Internet (global routing table)
  - These are implemented as MPLS/BGP Layer3 VPNs
  - Most customers using BGP have two VLAN tags and two BGP sessions, for each table
- We have a blended VRF for customers that prefer the simplicity of default routing only
  - This consists of the research VRF plus a default route that points to the commercial table

Internet DFZ routing vs Campus / LAN

- Routing asymmetry is commonly observed, expected feature of the multihomed AS
- This is because each AS makes independent routing decisions
- A step further- if you announce the same route to two external peer ASes, you should not **assume** a specific distribution of inbound traffic across those two connections. It will probably be unbalanced.
- In this scenario, we can **try** to influence what happens by making suggestions to the neighboring ASes. We will discuss some of these techniques later.
- When thinking BGP, it is useful to think about unidirectional concepts:
  - Received / Learned routes are used to send (transmit) packets
  - Advertised routes will influence where you receive traffic. – “Announcements attract traffic”

BGP hygiene - Preventing routing leaks and hijacks

- Default policy = readvertise all routes among external peers (disallowing AS loops)
- This is a reasonable policy for Internet backbone routers. The rest of us must have policy in place for proper routing behavior!
- If you are an end-site, at a minimum, you should have policy in place to ensure that you're only advertising your own route(s) to all neighbors.
- For many networks, BGP policy can be nuanced and complex. This can lead to unintended advertisements.

EXTRA STUFF

Other examples

- https://connect.geant.org/2017/05/15/taking-it-to-the-limit-testing-the-performance-of-re-networking
  - Commodity path showed two problems
    - Packet loss
    - DoS mitigation killed high-speed flows
  - Configure-before-use or test-before-use model impedes science
- https://indico.geant.org/event/1/contributions/11/attachments/47/207/190521_-_PT_TNC2019_v8.pdf
  - Multi-nation testing of R&E vs. commodity
  - Results indicate R&E paths perform better, even with more hops
    - Key point - hop count is a legacy metric because modern routers are ASIC-based
- 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!

EPOC
Engagement and Performance Operations Center

Why BGP rather than an IGP?

- An IGP moves packets as efficiently as possible within an AS
- A IGP does not worry about politics (not many routing policies can be enforced)
  - A corporate AS is not willing to carry (transit) traffic originating from a foreign AS
  - A Research and Education Network (REN) may not want to carry commercial traffic
  - Traffic starting or ending at Apple should not transit Google, etc.
- BGP is designed to handle all these cases and enforce routing policies between ASes