Border Gateway Protocol: Concepts and Implementation

Session 1 - BGP and Research and Education Networks

University of South Carolina (USC)

Engagement and Performance Operations Center (EPOC)

Texas Advanced Computing Center (TACC)

International Networks at Indiana University (IN@IU)

Internet2 Technology Exchange

December 9, 2024

Organizers

Elie Kfoury
(USC)

Jorge Crichigno
(USC)

Jason Zurawski
(ESnet / EPOC)

Ali AlSabeh
(USC)

Corey Eichelberger
(TACC)

Hans Addleman
(IU)

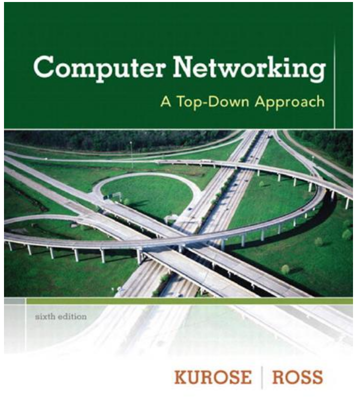
Agenda

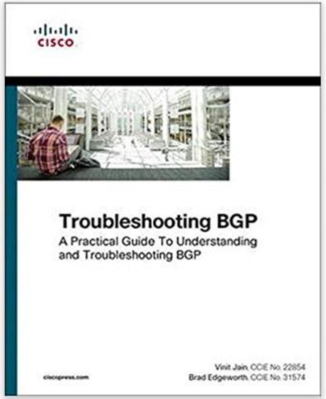
Time	Topic	Presenter
08:30 - 09:00	BGP and Research and Education Networks (RENs)	Corey Eichelberger, Hans Addleman, Jason Zurawski
09:00 - 09:45	Hands-on Session BGP 1: Essentials of BGP, EBGP, IBGP	Ali AlSabeh (USC), Jorge Crichigno (USC)
09:45 - 10:00	Break	
10:00 - 10:30	Brief overview of Local Preference and MED Attributes	Corey Eichelberger, Hans Addleman, Jason Zurawski
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 (Border Gateway Protocol) is complex and is at the heart of what makes the internet work
- BGP was created to “Just work” like TCP. It was not created for performance.
    - 35 years ago - June 1989
- 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




cisco
Troubleshooting BGP
A Practical Guide To Understanding and Troubleshooting BGP
ciscopress.com
Vinit Jain, CCIE No. 22854
Brad Edgeworth, CCIE No. 31574

What is BGP

- BGP or Border Gateway Protocol is protocol used between routers to exchange routing information and reachability information between or inside AS on the Internet.
- BGP makes the Internet work, and in most cases it just works
    - Needs to be tuned for best performance
- BGP makes routing decisions based on paths, network policies and rule-sets, communities and more. Lots of configuration choices. Also, will just work out of the box.
- Security was not a concern and not baked into the protocol
- Believes (without help) all advertisements from peers with no checks.
- It also by default can re-advertise to other peers what it learns.

AS, IGP, EGP

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

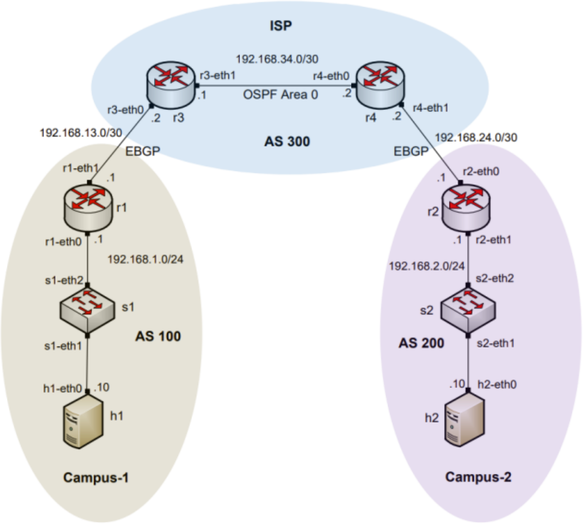
> “A set of routers under the single technical administration, using an IGP and common metrics to route packets within the AS, and using an EGP to route packets to other ASs.”

- 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



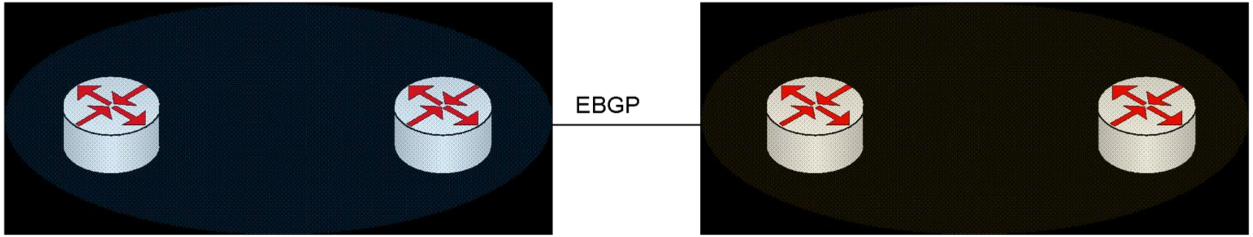
BGP In The Wild

- Over 76,000 Autonomous Systems (ASN).
- Over 1,000,000 IPv4 routes advertised.
- Over 272,000 IPv6 routes advertised.
- Each Router running BGP builds its own routing table with best path information to a subset of the internet.

Data from: https://bgp.he.net/report/prefixes#_prefixes

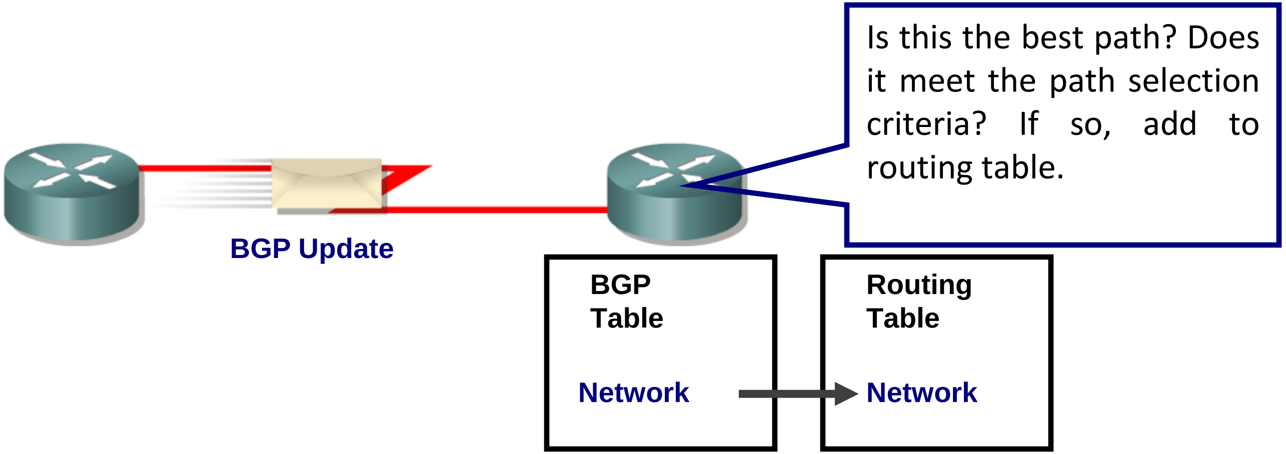
BGP Route Advertisements within an AS

- 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



BGP Update

Is this the best path? Does it meet the path selection criteria? If so, add to routing table.

BGP Table

Network

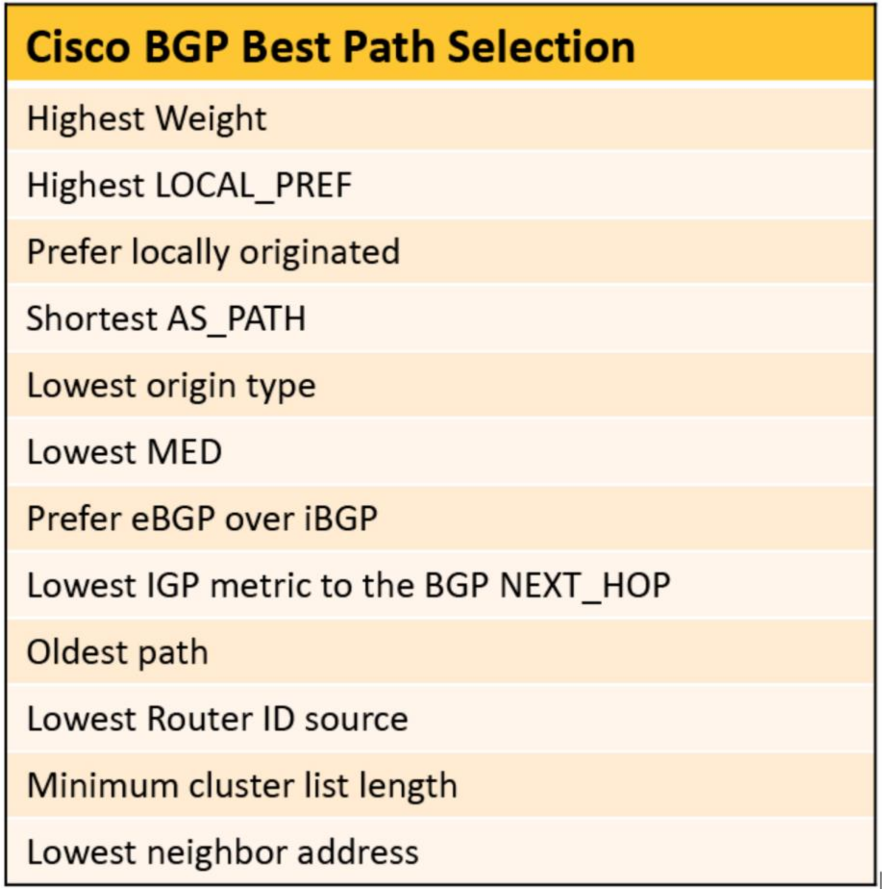
Routing Table

Network

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

BGP - Best Path Care and feeding

- BGP just works in many cases but needs tuned for performance
- Best path selection is a 10+ step process!
- Common steering mechanisms:
  - Localpref
  - Communities
  - AS Padding
  - MEDs

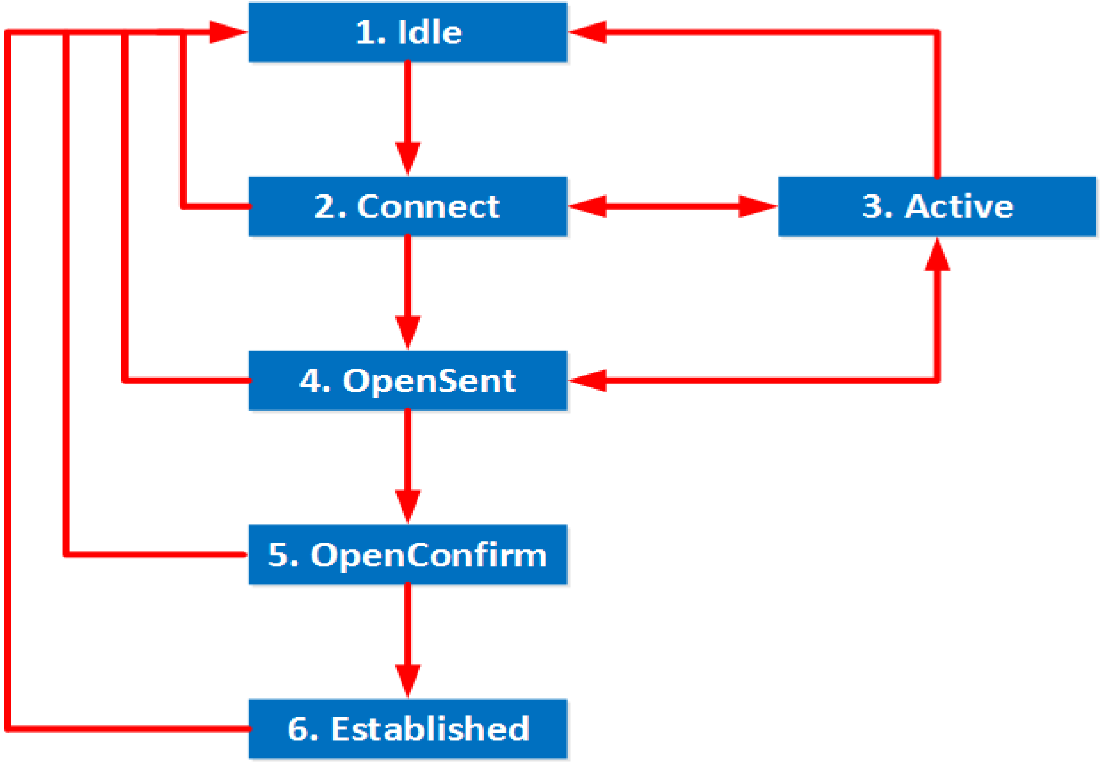


Cisco BGP Best Path Selection
Highest Weight
Highest LOCAL_PREF
Prefer locally originated
Shortest AS_PATH
Lowest origin type
Lowest MED
Prefer eBGP over iBGP
Lowest IGP metric to the BGP NEXT_HOP
Oldest path
Lowest Router ID source
Minimum cluster list length
Lowest neighbor address

BGP Neighbor States

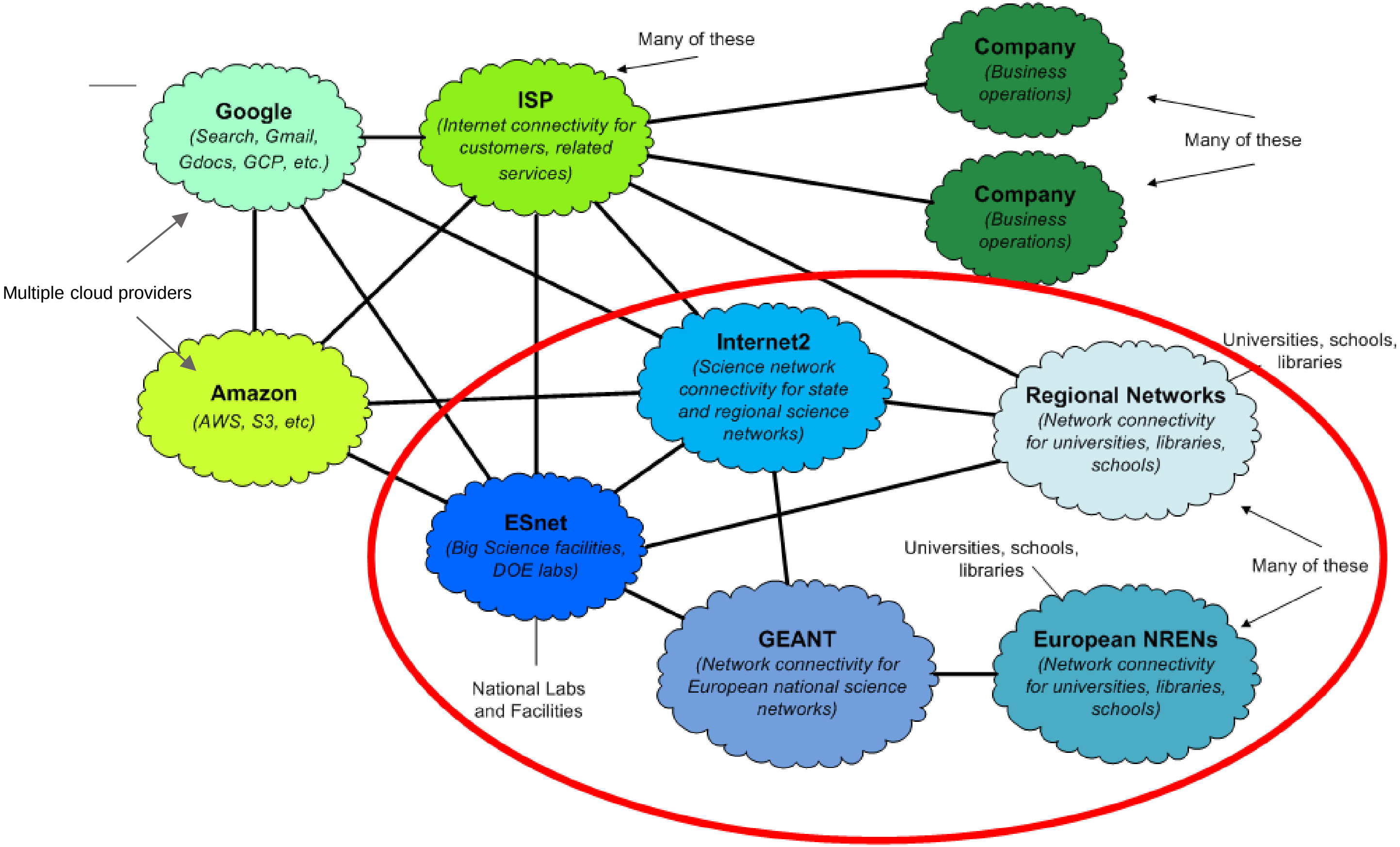
- **Idle:**
  - In this state, the router has not yet established a TCP connection with its BGP neighbor. The router is not actively attempting to establish a connection.
- **Connect:**
  - After the Idle state, the router transitions to the Connect state. In this state, the router initiates a TCP 3-way handshake connection with its neighbor by sending an initial TCP SYN packet and the other end responds to it with a TCP ACK packet and sends its own TCP SYN packet.
  - In the last initiator responds with a TCP ACK packet and TCP connections are established. Once the TCP 3-way handshake connection is established, the BGP router moves to the next state.
- **Active:**
  - If the TCP connection fails to establish within a certain timeout period, the router enters the Active state. In this state, the router repeatedly attempts to establish a TCP connection with its neighbor. This state indicates that the router is actively trying to connect but has not yet succeeded.
- **OpenSent:**
  - When the TCP connection is successfully established, the router transitions to the OpenSent state.
  - In this state, the router sends an Open message to its neighbor, which includes its BGP capabilities and other information. The router waits for an Open message from its neighbor to proceed to the next state.
- **OpenConfirm:**
  - After receiving the Open message from its neighbor, the router enters the OpenConfirm state. In this state, the router waits for a Keepalive message from its neighbor to confirm that the neighbor has also reached the OpenConfirm state. Once the Keepalive message is received, the router moves to the next state.
- **Established:**
  - The router enters the Established state when it has received a Keepalive message from its neighbor. In this state, the routers can exchange BGP updates, such as routing information and path attributes. In the BGP established state, the router is now fully operational and actively exchanging routing information with its neighbor.
  - If there is a disruption or issue in the BGP session, the router may transition back to the Idle state and repeat the process of establishing the neighbor relationship.

BGP State Diagram



BGP Use in R&E Networks (generalities)

R&E vs. Commodity

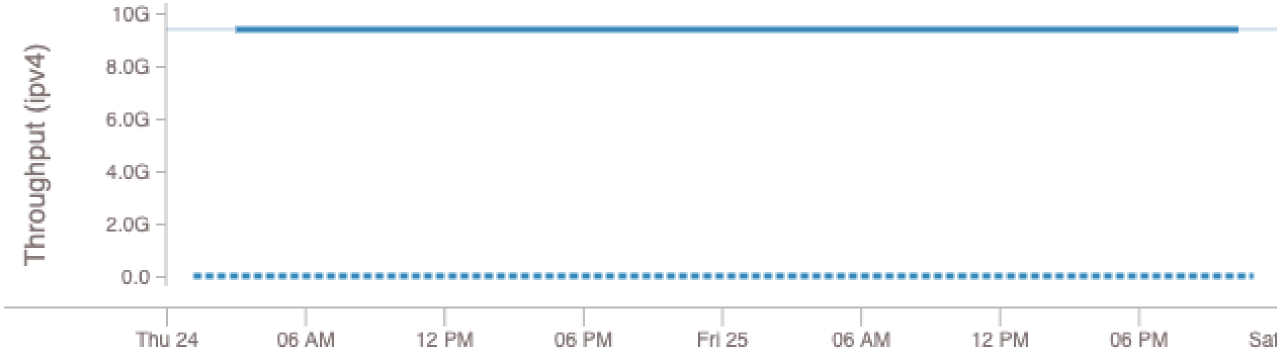

Many of these
Google
(Search, Gmail, Gdocs, GCP, etc.)
ISP
(Internet connectivity for customers, related services)
Company
(Business operations)
Company
(Business operations)
Many of these
Multiple cloud providers
Amazon
(AWS, S3, etc)
Internet2
(Science network connectivity for state and regional science networks)
Universities, schools, libraries
Regional Networks
(Network connectivity for universities, libraries, schools)
ESnet
(Big Science facilities, DOE labs)
Universities, schools, libraries
Many of these
National Labs and Facilities
GEANT
(Network connectivity for European national science networks)
European NRENs
(Network connectivity for universities, libraries, schools)

R&E Routing Architecture

- Research and Education Networks
  - Bandwidth
  - Performance Engineering
  - Deterministic behavior
  - Community
- Commodity Networks
  - Traffic shaping
  - DoS protections
  - Unknown architecture
- R&E networks are engineered to support science while commodity networks are not
  - Keep the science traffic on the science networks!

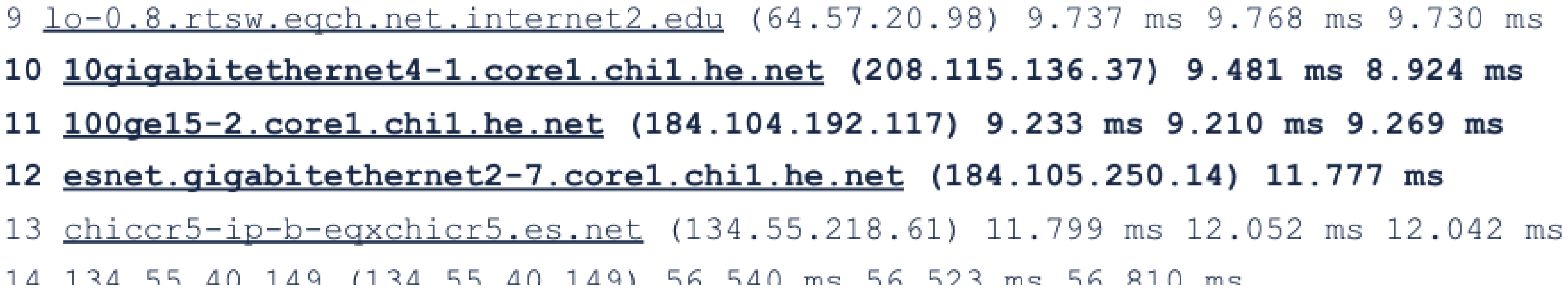
Why does this matter? Example 1 - OSC

- 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 - <u>this should be stopped!</u>
  - **R&E:** another scientist doing normal things - <u>this is core mission!</u>



Commodity vs R&E OSC Troubleshooting 2

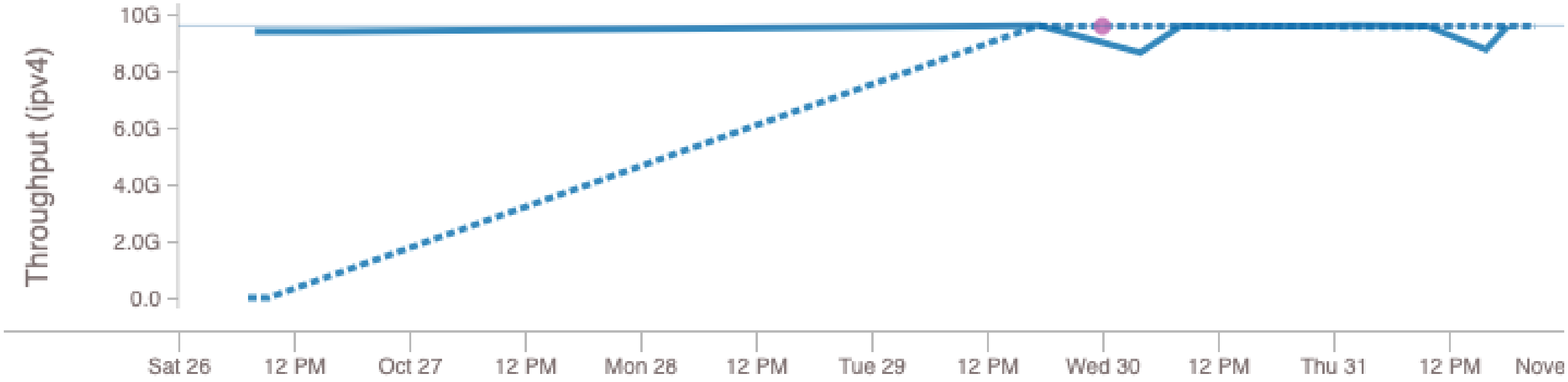
- OSC Engineer found a memory allocation issue on border router causing the routing table to not fully populate.
  - This kept the best path to ESnet out of the table
- ESnet engineer found an out of date routing configuration as well



```
9 lo-0.8.rtsw.eqch.net.internet2.edu (64.57.20.98) 9.737 ms 9.768 ms 9.730 ms
10 10gigabitethernet4-1.core1.chi1.he.net (208.115.136.37) 9.481 ms 8.924 ms
11 100ge15-2.core1.chi1.he.net (184.104.192.117) 9.233 ms 9.210 ms 9.269 ms
12 esnet.gigabitethernet2-7.core1.chi1.he.net (184.105.250.14) 11.777 ms
13 chiccr5-ip-b-eqxchicr5.es.net (134.55.218.61) 11.799 ms 12.052 ms 12.042 ms
14 134.55.40.149 (134.55.40.149) 56.540 ms 56.523 ms 56.810 ms
```

Commodity vs R&E: OSC Results

- Performance improved substantially
- Another example of the need for a Routing Working Group



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

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

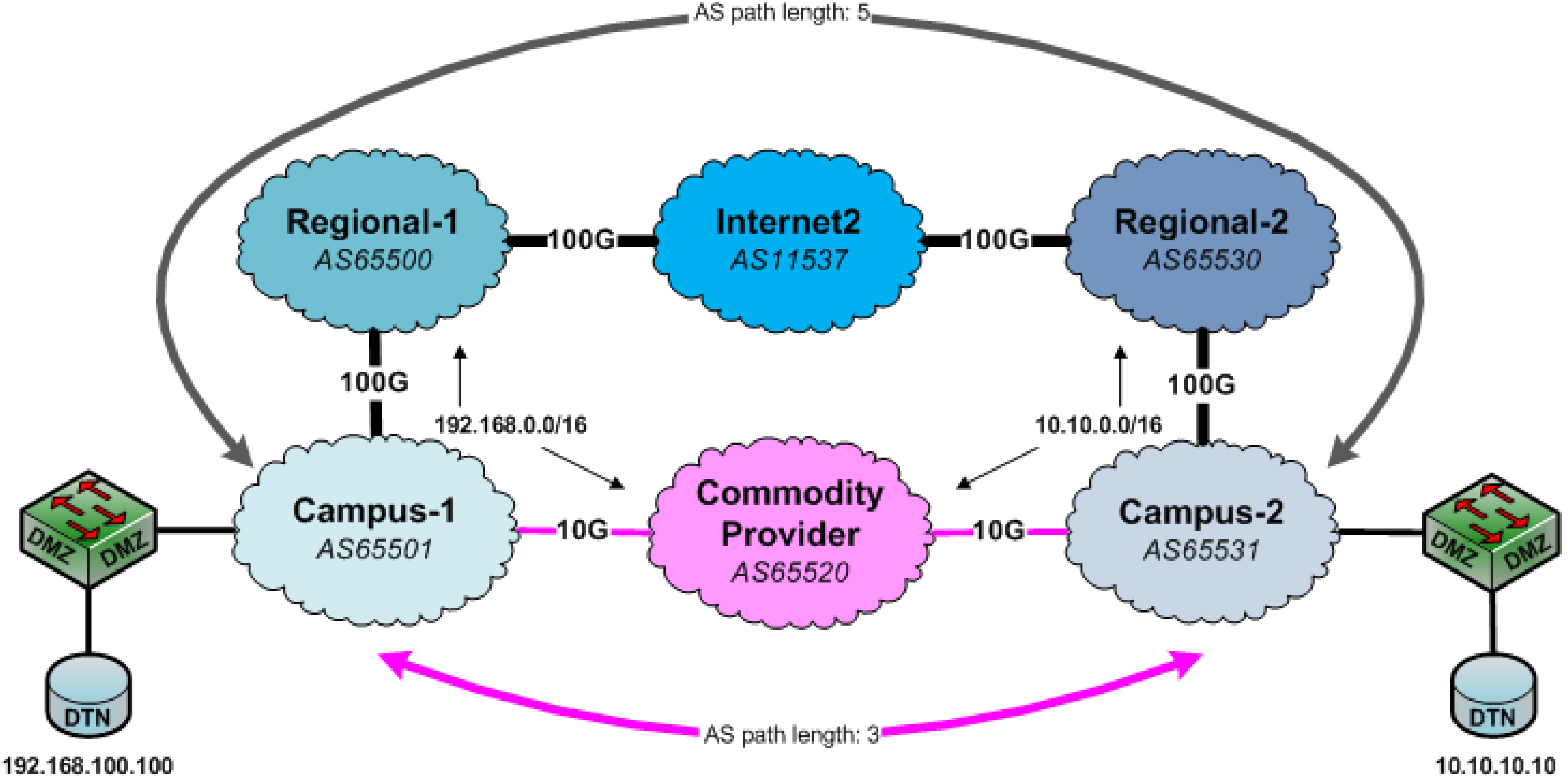
Other examples

- 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!

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

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

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

BGP Use at TACC

TACC Routing Architecture

- 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:
    - Direct Peering - High
    - R&E peering - Medium
    - Commercial Peering - Low
  - Routes tagged with direct or R&E communities re-announced into iBGP mesh
    - Allows us to better steer traffic to the correct exit router and peerings
  - Localpref steers traffic to preferred peerings

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

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 = readadvertise 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

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