EPOC
Engagement and Performance Operations Center

BGP Essentials

Ken Miller, Jason Zurawski
ken@es.net, zurawski@es.net
ESnet / Lawrence Berkeley National Laboratory

Modern Cyberinfrastructure for Research Data Management Workshop
University of Central Florida
February 16-17, 2023

TACC TEXAS ADVANCED COMPUTING CENTER
ENERGY SCIENCES NETWORK
INDIANA UNIVERSITY

National Science Foundation Award #1826994

Outline

- ***Preliminaries***
- R&E routing architecture
  - What it is
  - Why it matters
  - Examples of problems
  - Simplified ESnet &. FRGP Routing Architecture
- BGP Steering mechanisms and real world examples
  - Examples
  - Localpref
  - AS Path Padding
  - Communities
  - Multi Exit Discriminators (MEDs)
- Questions

EPOC
Engagement and Performance Operations Center

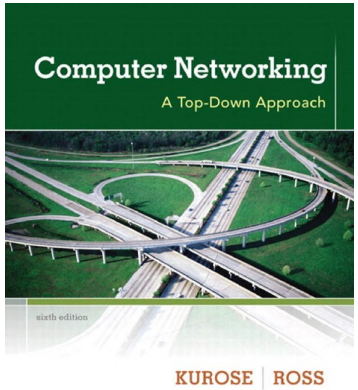
What is BGP

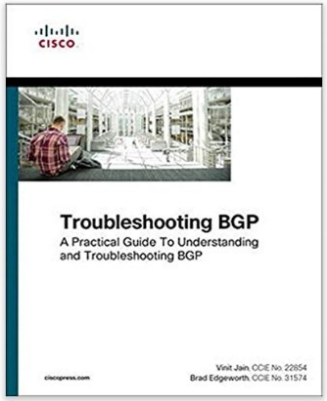
- BGP or Border Gateway Protocol is protocol used between routers to exchange routing information and reachability information between or inside AS on the Internet.
- BGP makes the Internet work, and in most cases it just works
  - Needs to be tuned for best performance
- BGP makes routing decisions based on paths, network policies and rule-sets, etc.

EPOC
Engagement and Performance Operations Center

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 a 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)?

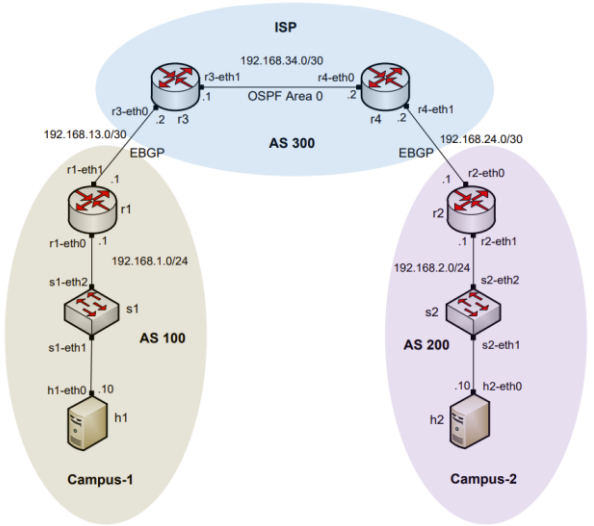
  "A set of routers under the single technical administration, using an IGP and common metrics to route packets within the AS, and using an EGP to route packets to other ASs."

- 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 Ases
  eBGP is the most prevalent example of an EGP

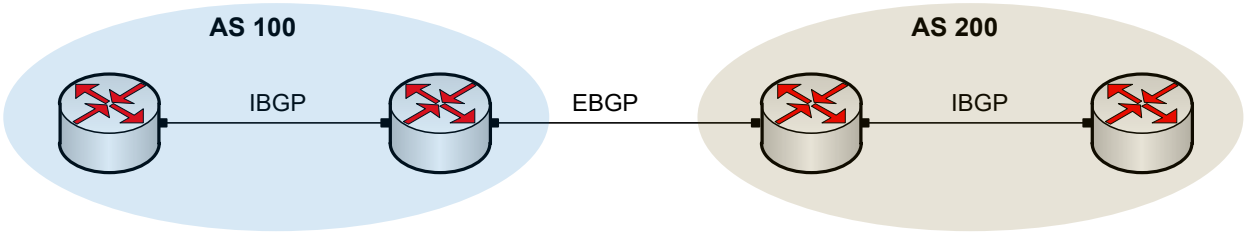


Why BGP rather than an IGP?

- An IGP moves packets as efficiently as possible within an AS
- A IGP does not worry about policies (limited routing policies can be enforced with IGP)
  - A corporate AS is normally 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, both within and between ASes

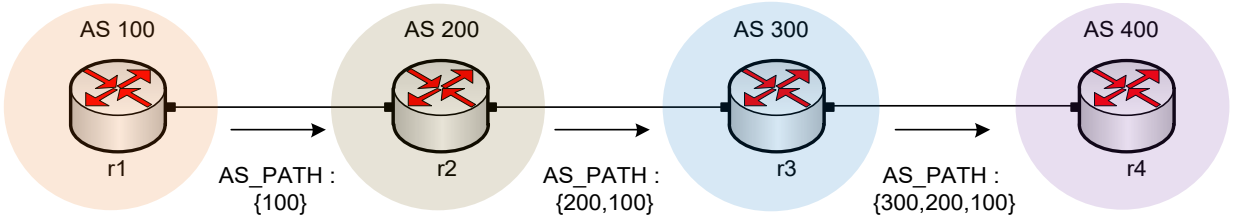
BGP Route Advertisements within an AS

- 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)



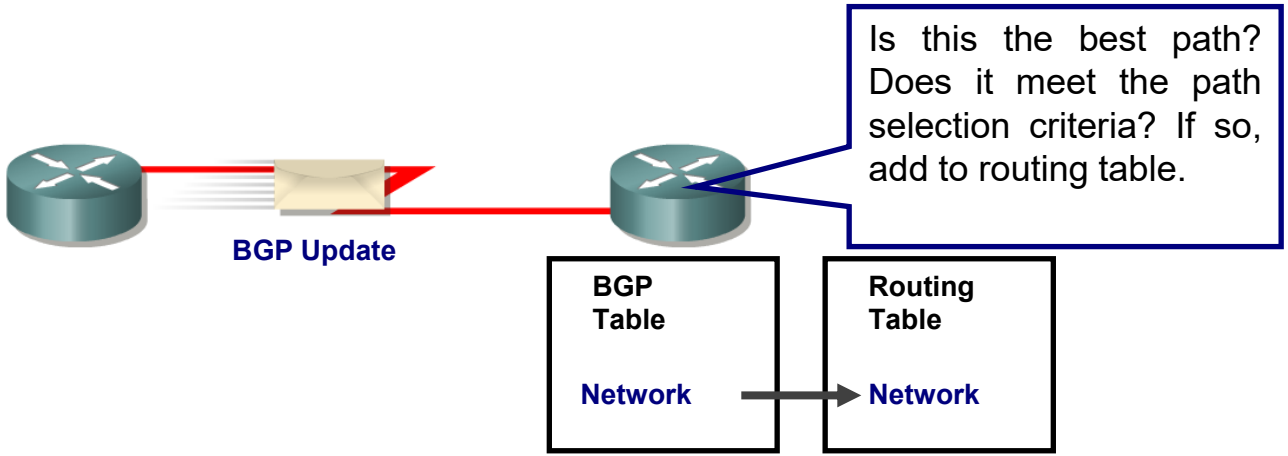
BGP Route Advertisements between ASs

- In BGP route advertisements, each border router prepends its own AS number to the route before advertising the route to the next AS



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

Agenda

- Preliminaries
- ***Commodity vs. R&E routing architecture***
  - ***What it is***
  - ***Why it matters***
  - ***Examples of problems***
  - ***Simplified ESnet & FRGP Routing Architecture***
- BGP Steering mechanisms and real world examples
  - Localpref
  - AS Path Padding
  - Communities
  - Multi Exit Discriminators (MEDs)
  - Examples (Good and Bad)
- Questions

EPOC
Engagement and Performance Operations Center

BGP in the wild

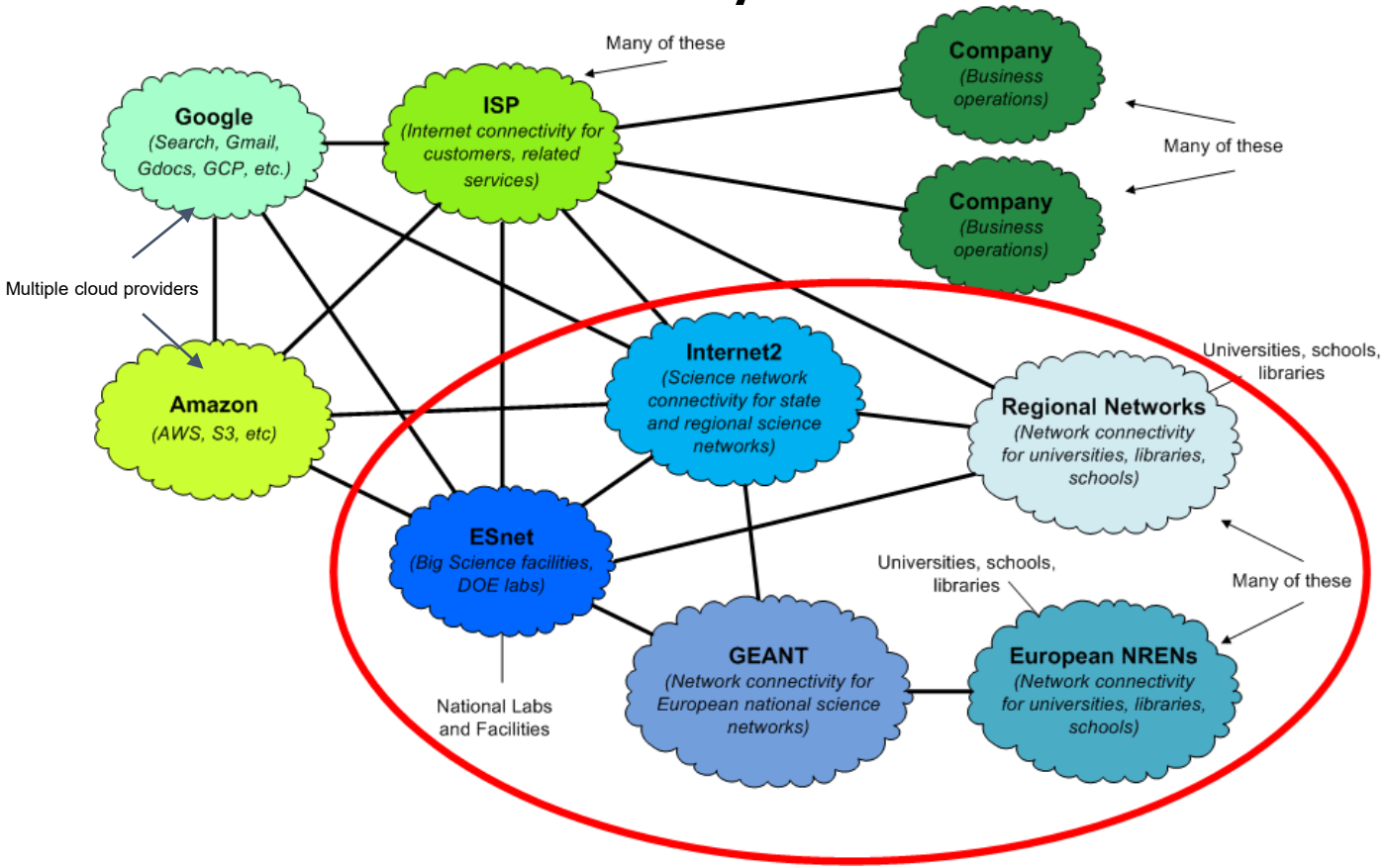
- Over 74,000 Autonomous Systems (ASN) in March 2022.
- Over 1,000,000 IPv4 routes advertised.
- Over 182,000 IPv6 routes advertised.

- Each Router running BGP builds its own routing table with best path information to a subset of the internet.
- 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

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

R&E vs. Commodity



EPOC
Engagement and Performance Operations Center

R&E Routing Architecture Vs. Commodity.

- 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!

EPOC
Engagement and Performance Operations Center

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.

Agenda

- Preliminaries
- Commodity vs. R&E routing architecture
  - What it is
  - Why it matters
  - Examples of problems
  - Simplified ESnet & FRGP Routing Architecture
- ***BGP Steering mechanisms and real world examples***
  - ***Localpref***
  - ***AS Path Padding***
  - ***Communities***
  - ***Multi Exit Discriminators (MEDs)***
  - ***Examples (Good and Bad)***
- Questions

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

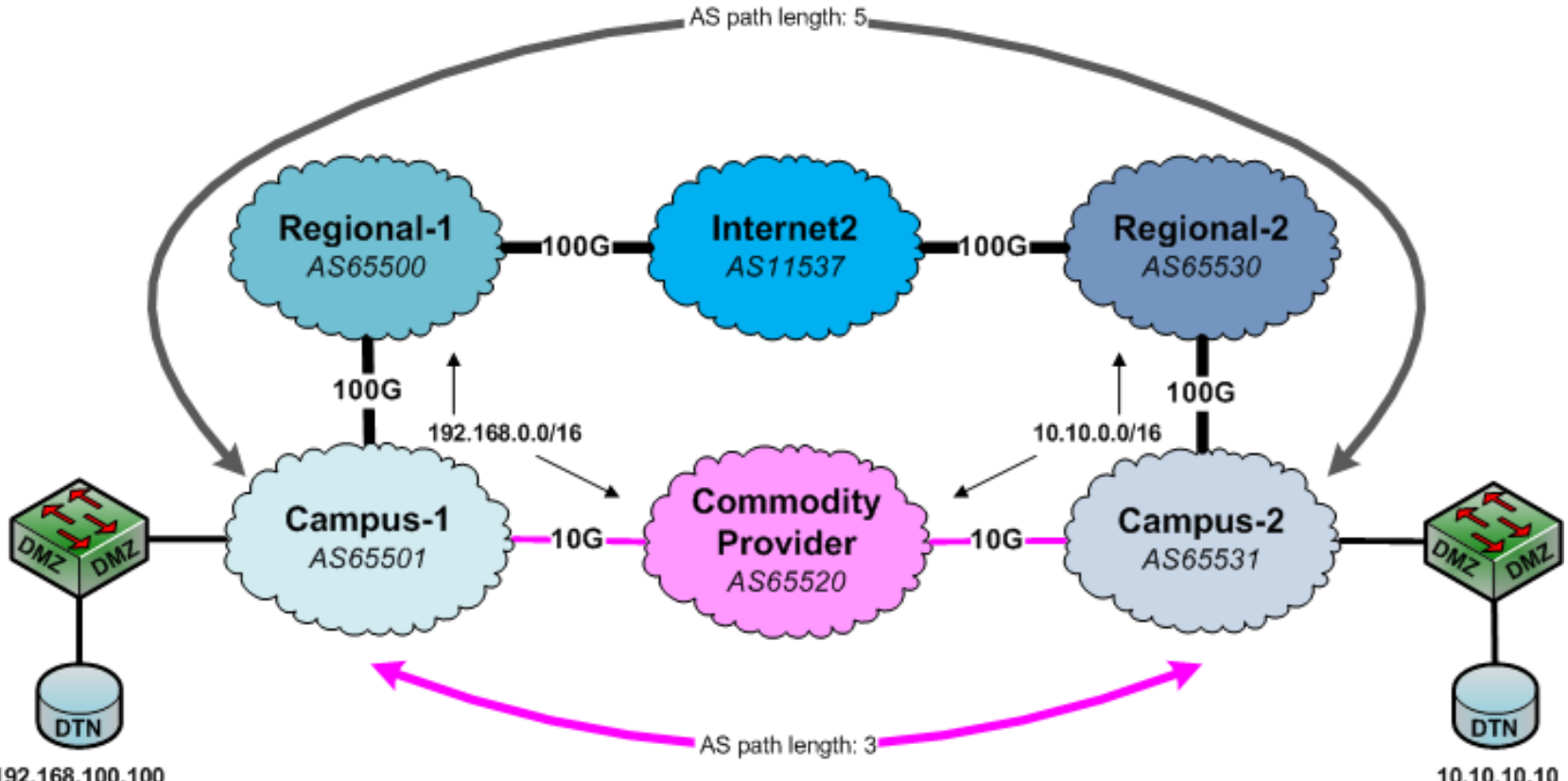
EPOC
Engagement and Performance Operations Center

BGP - 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

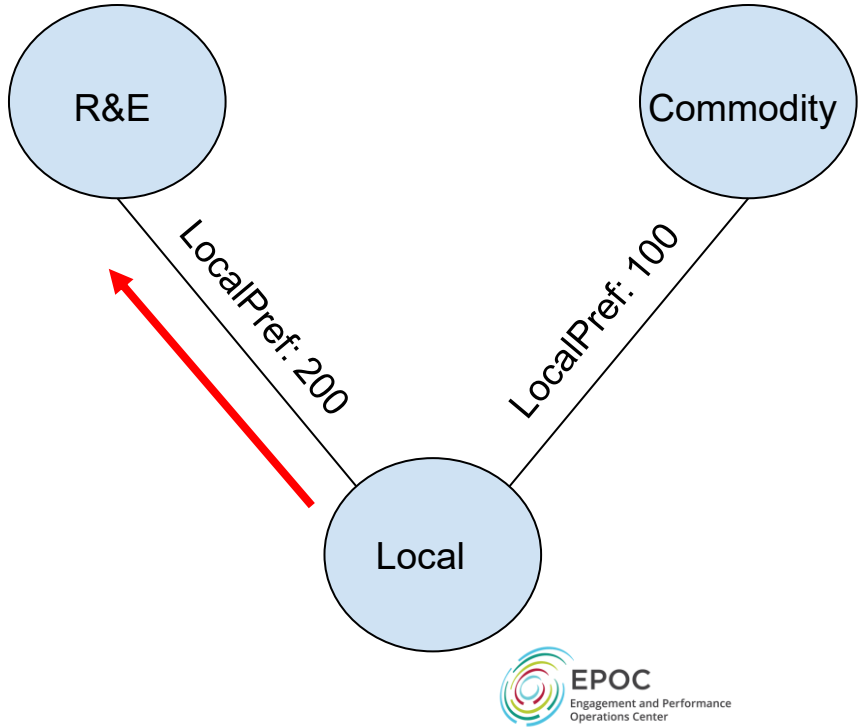
Cisco BGP Best Path Selection
Highest Weight
Highest LOCAL_PREF
Prefer locally originated
Shortest AS_PATH
Lowest origin type
Lowest MED
Prefer eBGP over iBGP
Lowest IGP metric to the BGP NEXT_HOP
Oldest path
Lowest Router ID source
Minimum cluster list length
Lowest neighbor address

BGP AS Path Length Illustrated



LocalPref (e.g. “outbound route”)

- Indicates to routers in the AS which path is preferred to exit the AS (higher is better)
- Can be done per-prefix
- Modifies path for outbound traffic
- Exchanged only among routers within the same AS (passed only via IBGP, not via EBGP)
- very frequently used in provider networks



EPOC
Engagement and Performance Operations Center

Public BGP Community Strings offered by Internet2

- https://noc.net.internet2.edu/i2network/maps-documentation/documentation/bgp-communities.html
- Set LocalPref on your advertised prefixes
  - 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

EPOC
Engagement and Performance Operations Center

BGP Community Strings

- A community string is a number value that the peer uses like a tag.
- Tagging prefixes with communities tells the peer to handle the prefixes in a special way.
- Can make changes to routing policy based on per prefix strings
- Prefixes can have multiple community strings
- Can provide useful information about the prefix
- Communities that might be useful to external networks should be made public
  - Provides a mechanism for peers to affect a network's internal behavior
  - Common uses: change local preference, DDoS mitigation
- Look for upstream networks published communities
  - Regional?
  - National?

EPOC
Engagement and Performance Operations Center

AS Path Padding

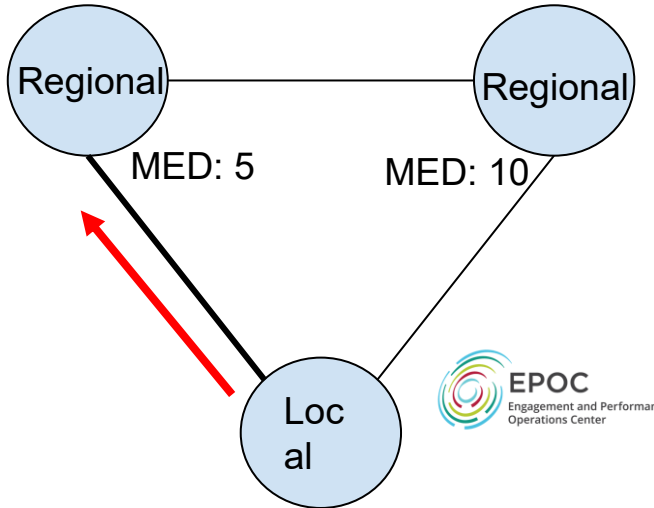
- BGP will choose shortest AS Path
- Add one or more copies of your AS# to prefixes advertised to specific neighbors.

* 180.208.59.0/24 202.112.61.57 - - - 4538 4538 24364 **133465 133465 133465** 65300 i

EPOC
Engagement and Performance Operations Center

(MED) Multi Exit Discriminator (e.g. “inbound route”)

- Indicates to external neighbors the preferred path into an AS
- Useful when you have N+1 connections to a network
- Lowest number preferred
- MED can be sent to EBGP peers:
  - Routers propagate a MED within their AS
  - But do not pass it on to the next AS
- This may, or may not, be honored by the neighboring AS
  - Back to first principles – forwarding decisions are made locally



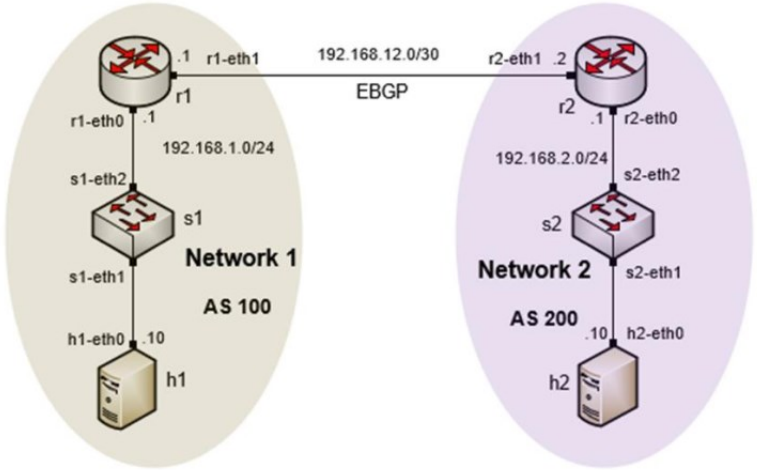
Why does this matter? Example 1 - OSC

- 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

Next-hop Attribute

- Unlike IGPs, BGP routes AS by AS, not router by router
- The next-hop address for a network from another AS is an IP address of the entry point of the next AS along the path to that destination network
- This default behavior is sometimes overridden through an iBGP export/outbound policy known as “next-hop self”



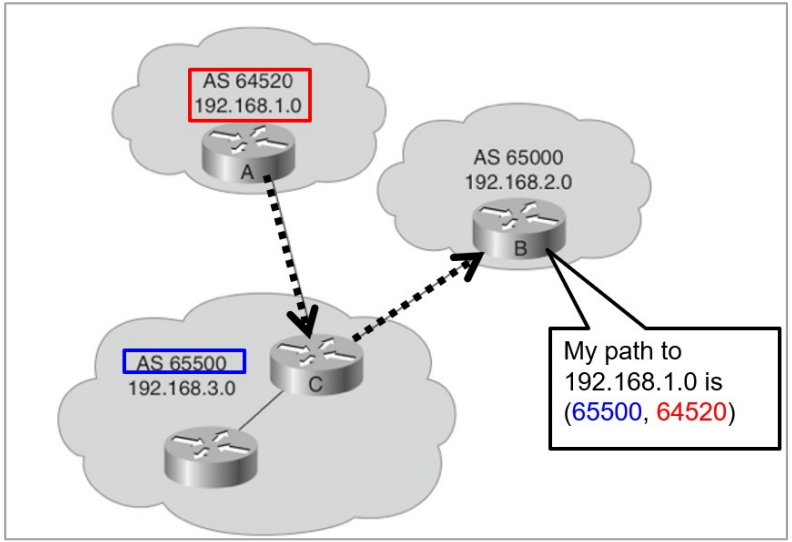
BGP table router r1

```
frr-pc# show ip bgp
BGP table version is 2, local router ID is 192.168.12.1, vrf id 0
Default local pref 100, local AS 100
Status codes:  s suppressed, d damped, h history, * valid, > best, = multipath,
               i internal, r RIB-failure, S Stale, R Removed
Nexthop codes: @NNN nexthop's vrf id, < announce-nh-self
Origin codes:  i - IGP, e - EGP, ? - incomplete

   Network          Next Hop            Metric LocPrf Weight Path
*> 192.168.1.0/24   0.0.0.0                  0         32768 i
*> 192.168.2.0/24   192.168.12.2             0             0 200 i
```

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

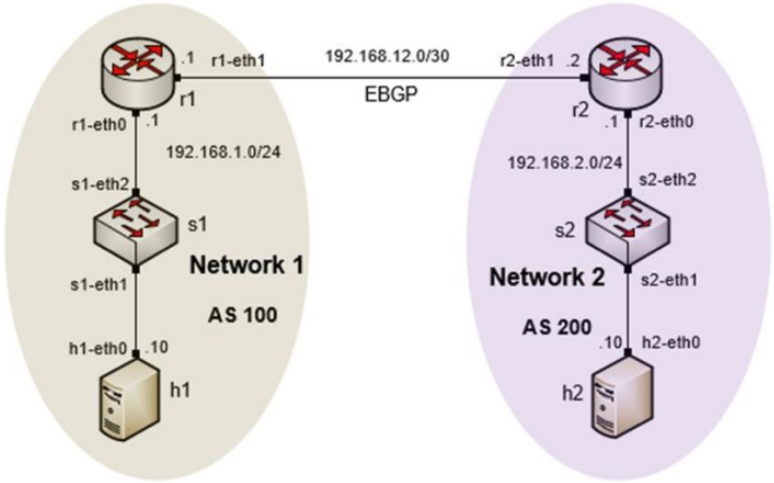
The AS-Path Attribute

- Whenever a route update passes through an AS, the AS number is prepended to that update
- Router A: advertises network 192.168.1.0 in AS 64520
- Router C: prepends its own AS number to it and advertises the route to Router B
- Router B: the path to reach 192.168.1.0 is:
  - 65500, 64520



Origin Attribute

- Defines the origin of the path information
- The origin attribute can be one of three values:
- **IGP (“i”)**
  - The route is interior to the originating AS
  - Normally when the **network command** is used
- **EGP (“e”)**
  - The route is learned via **EGP**
  - EGP is legacy and no longer supported
- **Incomplete (“?”)**
  - The route’s origin is unknown / some other means
  - It usually occurs when a route is **redistributed into** BGP



BGP table router r1

```
frr-pc# show ip bgp
BGP table version is 2, local router ID is 192.168.12.1, vrf id 0
Default local pref 100, local AS 100
Status codes:  s suppressed, d damped, h history, * valid, > best, = multipath,
               i internal, r RIB-failure, S Stale, R Removed
Nexthop codes: @NNN nexthop's vrf id, < announce-nh-self
Origin codes:  i - IGP, e - EGP, ? - incomplete

   Network          Next Hop            Metric LocPrf Weight Path
*> 192.168.1.0/24   0.0.0.0                  0         32768 i
*> 192.168.2.0/24   192.168.12.2             0             0 200 i
```

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

Administrative Distance

- A router may run multiple routing protocols / static routes
- If BGP and OSPF are configured on a router, both protocols may provide different best paths (analogous to map software)
- How does the router know which protocol to choose?
  - The route with lower Administrative Distance is installed in the routing table

Route Source	Administrative Distance
Connected	0
Static	1
EIGRP summary route	5
External BGP	20
Internal EIGRP	90
IGRP	100
OSPF	110
IS-IS	115
RIP	120
External EIGRP	170
Internal BGP	200

```
frr-pc# show ip route
Codes: K - kernel route, C - connected, S - static, R - RIP,
       O - OSPF, I - IS-IS, B - BGP, E - EIGRP, N - NHRP,
       T - Table, v - VNC, V - VNC-Direct, A - Babel, D - SHARP,
       F - PBR, f - OpenFabric,
       > - selected route, * - FIB route, q - queued route, r - rejected route

B   192.168.2.0/24 [200/0] via 192.168.23.1, r3-eth1, 00:34:38
O>* 192.168.2.0/24 [110/20] via 192.168.23.1, r3-eth1, 00:49:22
C>* 192.168.3.0/24 is directly connected, r3-eth0, 00:52:03
O   192.168.23.0/30 [110/10] is directly connected, r3-eth1, 00:49:32
C>* 192.168.23.0/30 is directly connected, r3-eth1, 00:52:03
```

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

The Weight Attribute - For “Outbound Route”

- vendor-specific attribute (e.g. Cisco)
  - Juniper has a different mechanism to achieve a similar result
- Configured locally and not propagated to any other routers
- Higher weight is preferred
- Weight takes precedence over Local Preference
- Value from 0 to 65535
- Default is 32768
- Default is 0 for routes not originated by this router

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

BGP Table

- Internal version number of the table
- This number is incremented whenever the table changes

```
frr-pc# show ip bgp
BGP table version is 3, local router ID is 192.168.23.2, vrf id 0
Default local pref 100, local AS 200
Status codes:  s suppressed, d damped, h history, * valid, > best, = multipath,
               i internal, r RIB-failure, S Stale, R Removed
Nexthop codes: @NNN nexthop's vrf id, < announce-nh-self
Origin codes:  i - IGP, e - EGP, ? - incomplete

   Network          Next Hop            Metric LocPrf Weight Path
  i192.168.1.0/24   192.168.12.1              0    100      0 100 i
*>                  192.168.13.1             0             0 100 i
*>i192.168.2.0/24   192.168.23.1             0    100      0 i
*> 192.168.3.0/24   0.0.0.0                  0         32768 i
```

Status Code

- Displayed at the beginning of each line in the table
- This example is for Cisco

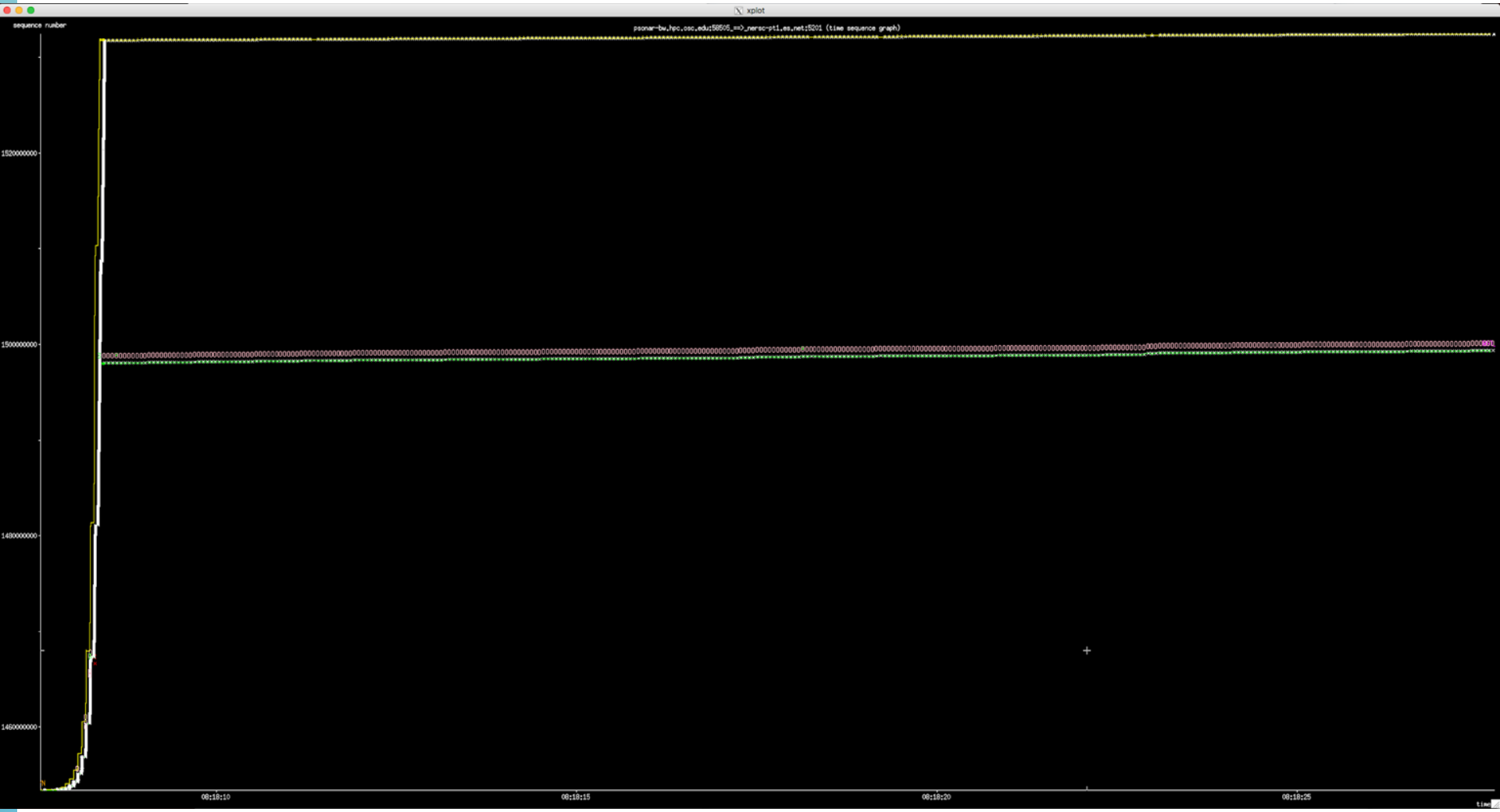
```
frr-pc# show ip bgp
BGP table version is 3, local router ID is 192.168.23.2, vrf id 0
Default local pref 100, local AS 200
Status codes:  s suppressed, d damped, h history, * valid, > best, = multipath,
               i internal, r RIB-failure, S Stale, R Removed
Nexthop codes: @NNN nexthop's vrf id, < announce-nh-self
Origin codes:  i - IGP, e - EGP, ? - incomplete

   Network          Next Hop            Metric LocPrf Weight Path
  i192.168.1.0/24   192.168.12.1             0    100      0 100 i
*>                  192.168.13.1             0             0 100 i
*>i192.168.2.0/24   192.168.23.1             0    100      0 i
*> 192.168.3.0/24   0.0.0.0                  0         32768 i
```

Code	Meaning
s	Table entry is suppressed
d	Table entry is dampened
h	Table entry history
*	Table entry is valid
>	Table entry is the best entry to use for this network
i	Table entry was learned via an internal BGP session
r	Table entry is a RIB-failure
S	Table entry is stale
=	Table entry has multipath to use for this network
b	Table entry has a backup path to use for this network
x	The table entry has a best external route to

EXAMPLES

- OSC to NERSC Routing
- Regional Network with 2 paths of different capacity
- Campus with commodity and R&E paths

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

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

Agenda

- Preliminaries
- Commodity vs. R&E routing architecture
  - What it is
  - Why it matters
  - Examples of problems
  - Simplified ESnet & FRGP Routing Architecture
- BGP Steering mechanisms and real world examples
  - Localpref
  - AS Path Padding
  - Communities
  - Multi Exit Discriminators (MEDs)
  - Examples (Good and Bad)
- ***Questions***

EPOC
Engagement and Performance Operations Center

Questions?

Transfer Performance problems? EPOC is here to help!

- epoc@tacc.utexas.edu
- https://epoc.global/

NSF Award: 1826994

EPOC
Engagement and Performance Operations Center

EPOC
Engagement and Performance Operations Center

BGP Essentials

Ken Miller, Jason Zurawski
ken@es.net, zurawski@es.net
ESnet / Lawrence Berkeley National Laboratory

Modern Cyberinfrastructure for Research Data Management Workshop
University of Central Florida
February 16-17, 2023

TACC TEXAS ADVANCED COMPUTING CENTER
ENERGY SCIENCES NETWORK
INDIANA UNIVERSITY

National Science Foundation Award #1826994