EPOC
Engagement and Performance Operations Center

Network Strategy to Enable Data Intensive Science

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

Virtual Training Workshop for Educators and Network Engineers on High Speed Network Protocols and Security, May 4-6, 2020

Eli Dart, ESnet / LBNL

Hans Addleman, Indiana University

ESnet
ENERGY SCIENCES NETWORK

INDIANA UNIVERSITY

Agenda

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

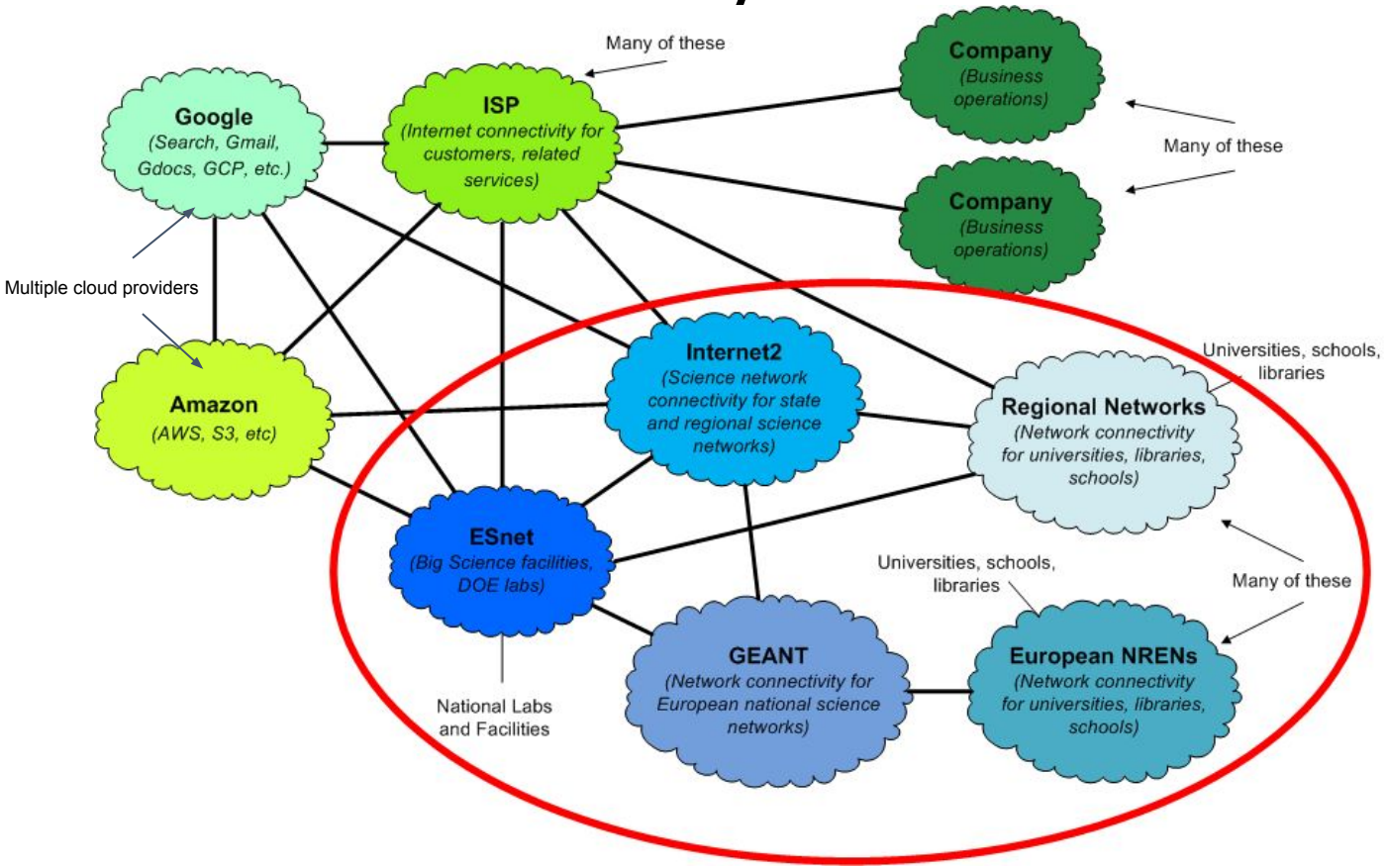
EPOC
Engagement and Performance Operations Center

Engagement and Performance Operations Center (EPOC)

- Joint project between Indiana University and ESnet
  - PI Dr. Jennifer Schopf, co-PI Jent (IU GlobalNOC) and Zurawski (ESnet)
- Partnerships with regional, infrastructure, and science communities that span the NSF and DOE continuum of funding.
- 5 Focus Areas: Roadside Assistance and Consultation, Application Deep Dives, Network Analysis (Netsage), Services “in a box” (DMZ, perfSONAR, etc), Training

EPOC
Engagement and Performance Operations Center

R&E vs. Commodity



EPOC
Engagement and Performance Operations Center

R&E Routing Architecture

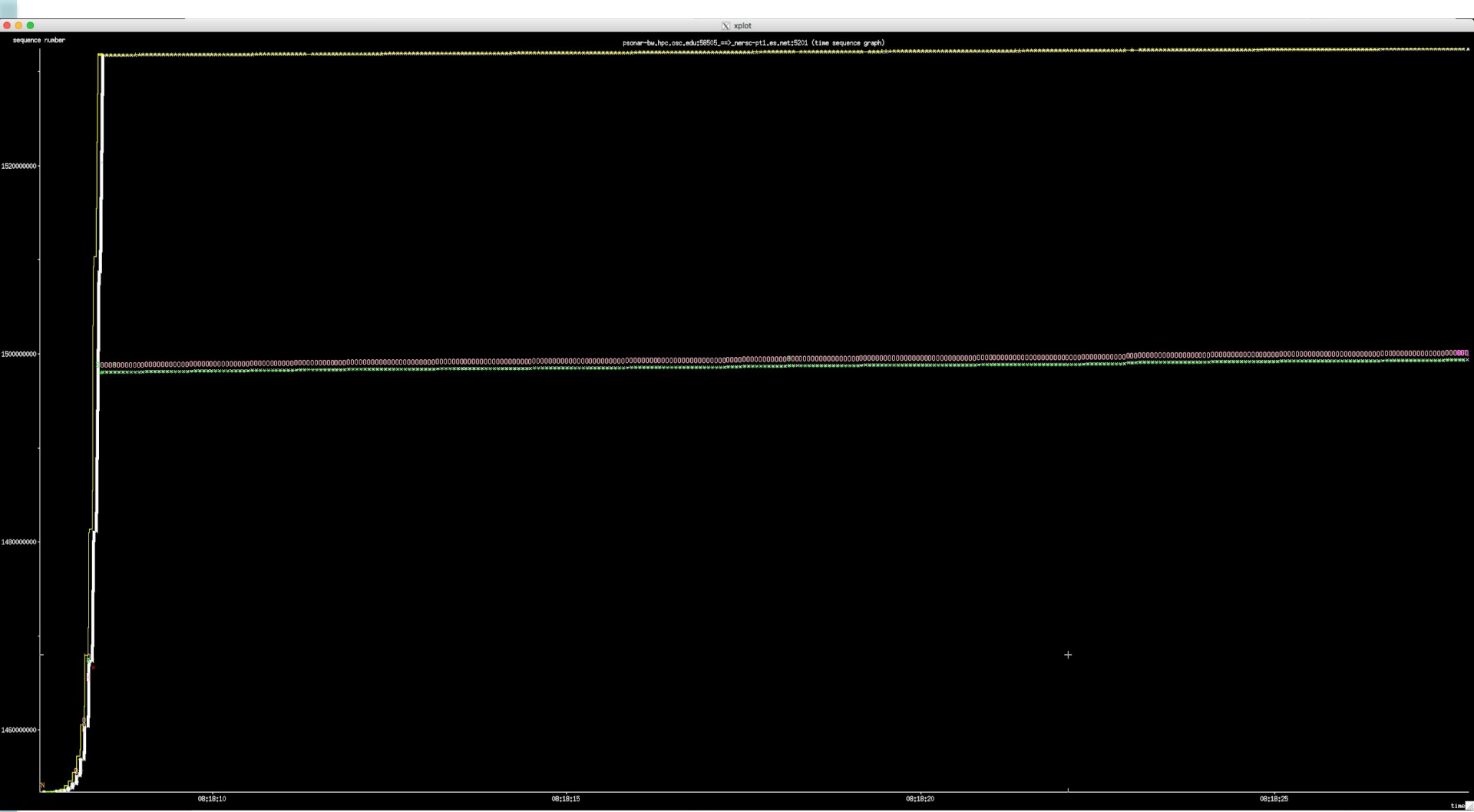
- R&E networks engineered for science traffic
- Keep R&E traffic on R&E paths if possible
  - Bandwidth
  - Performance Engineering
  - Deterministic behavior
- We all have to do our part
  - All routing decisions made locally
  - Emergent behavior is important
- Motivation examples to follow

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.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
  - 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!

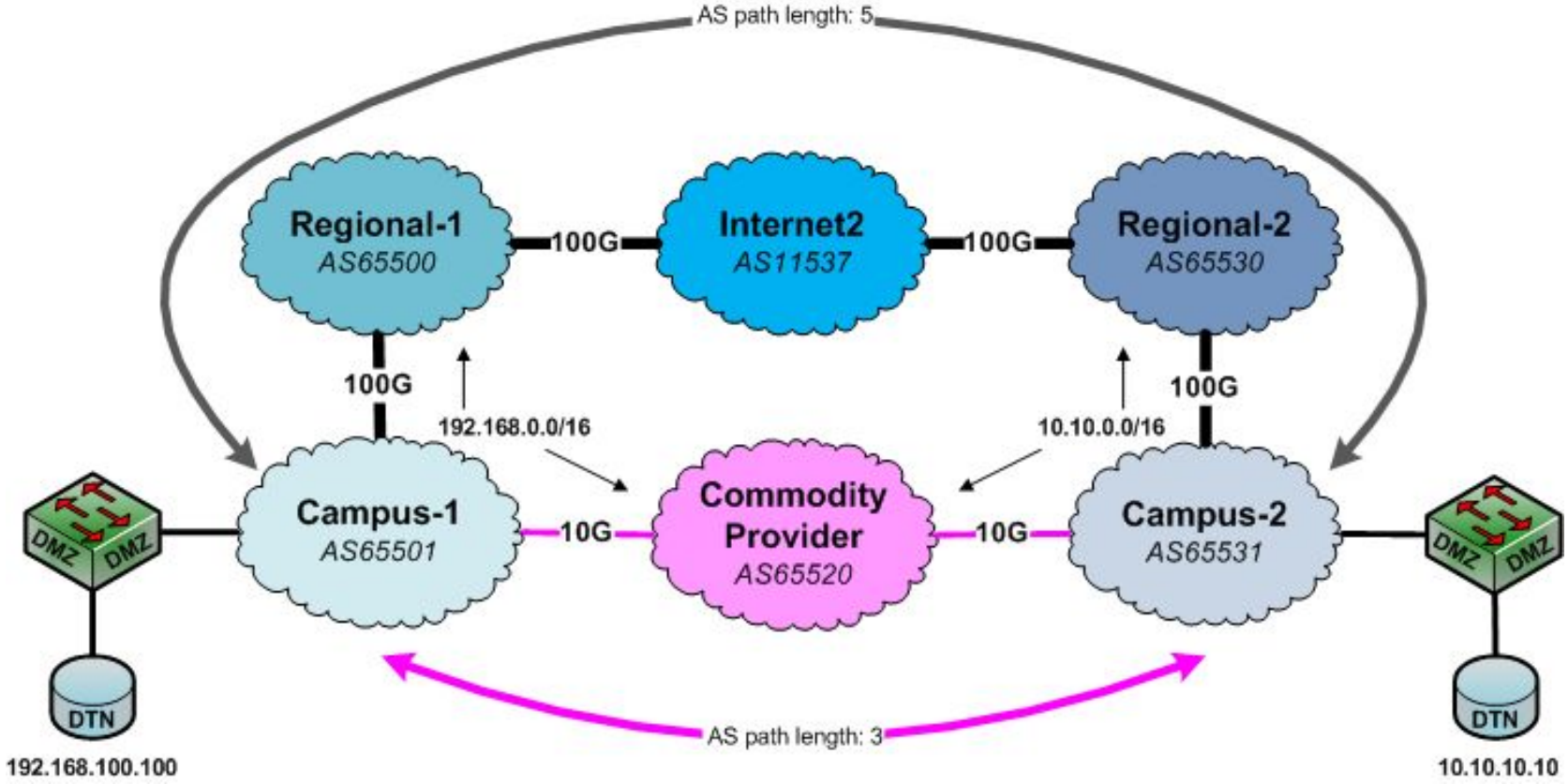
EPOC
Engagement and Performance Operations Center

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



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

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

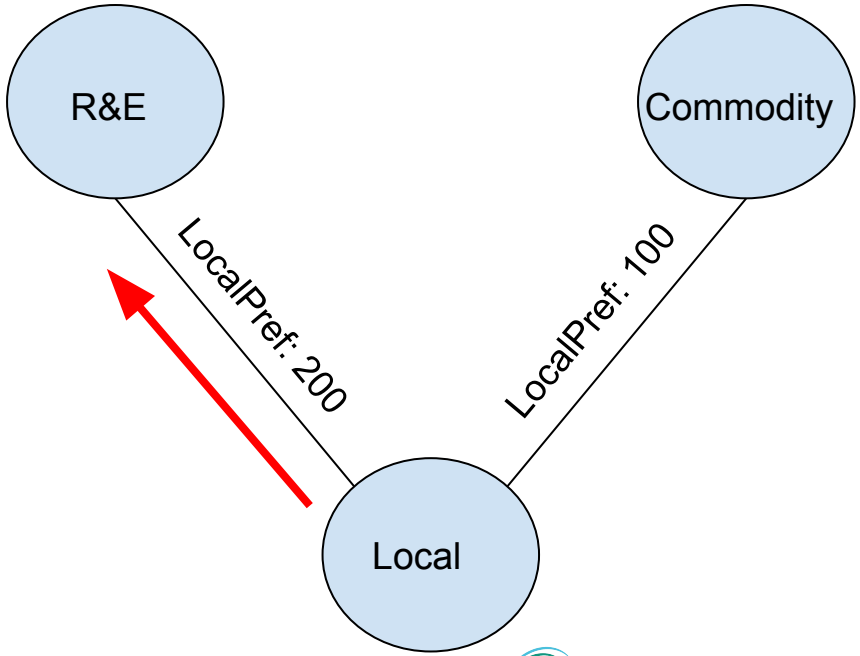
BGP Care and Feeding

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

EPOC
Engagement and Performance Operations Center

LocalPref

- Per prefix
- Modifies path for outbound traffic
- Higher preferred



EPOC
Engagement and Performance Operations Center

BGP Community Strings

- 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

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

AS Path Padding

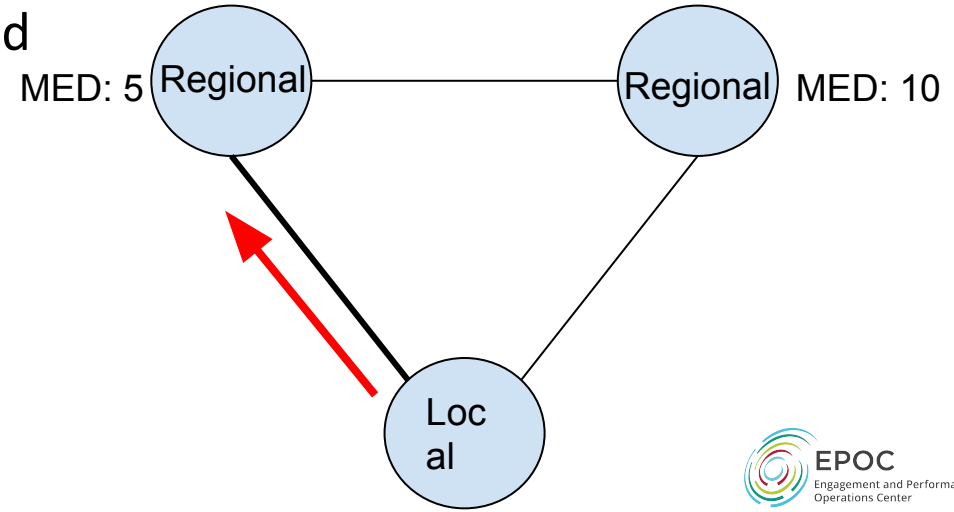
- BGP will choose shortest AS Path
- Add one or more copies of your AS# to prefixes advertised to specific neighbors.

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

EPOC
Engagement and Performance Operations Center

Multi Exit Discriminator (MED)

- Useful when you have N+1 connections to a network
- Indication to external peers of the preferred path into network
- Lowest number preferred



EPOC
Engagement and Performance Operations Center

Questions?

Transfer Performance problems? EPOC is here to help!

- epoc@iu.edu
- https://epoc.global/

NSF Award: 1826994

EPOC
Engagement and Performance Operations Center