EPOC
Engagement and Performance Operations Center

Routing, Buffers, and Network Performance, Oh my.

Brenna Meade

Hans Addleman

ESnet
ENERGY SCIENCES NETWORK

INDIANA UNIVERSITY

Agenda

- EPOC Overview
- R&E Networks or Commodity and why does it matter?
  - Example routing asymmetry commodity
- How do you identify network performance issues?
- What can affect network performance
- BGP / Routing Steering mechanisms
- Routing Working Group and getting involved

Engagement and Performance Operations Center

- Joint project between Indiana University and ESnet
  - co-PI Zurawski (ESnet) and Jent (IU GlobalNOC)
- Part of CC* program for domestic science support
  - Program Officer: Kevin Thompson
  - Award #1826994, $3.5M over 3 years
- Partnerships with regional, infrastructure, and science communities that span the NSF and DOE continuum of funding

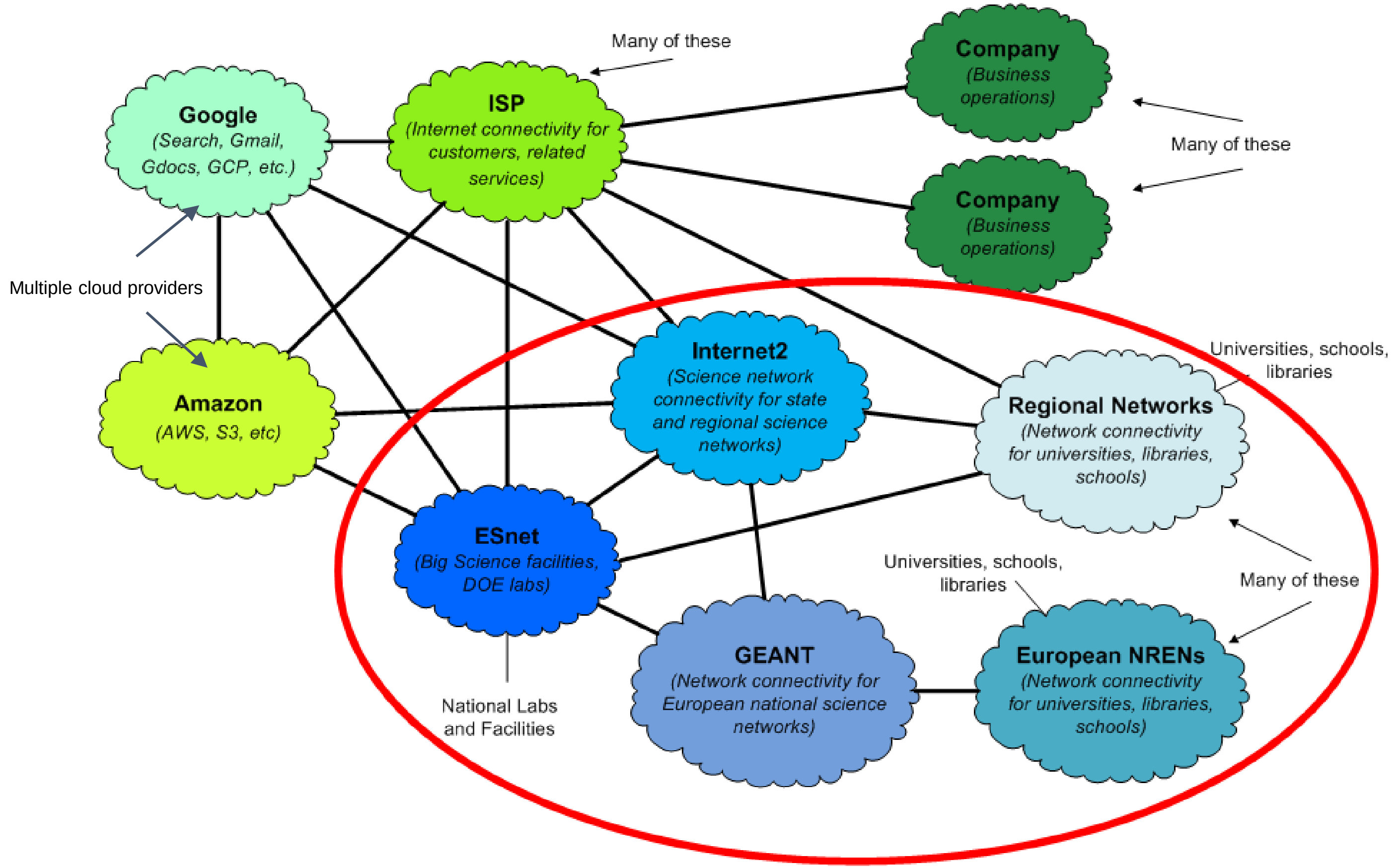
EPOC
Engagement and Performance Operations Center

Why an Engagement Operations Center?

- Today's science is collaborative science
- Collaborative science
  - Multiple partners
  - Multiple data sets
  - Many points of connection
  - Cross agency cooperation
- With better access to data we ask harder questions
- Interactive data sources change the science we do

EPOC
Engagement and Performance Operations Center

R&E vs. Commodity: What is the difference?



EPOC
Engagement and Performance Operations Center

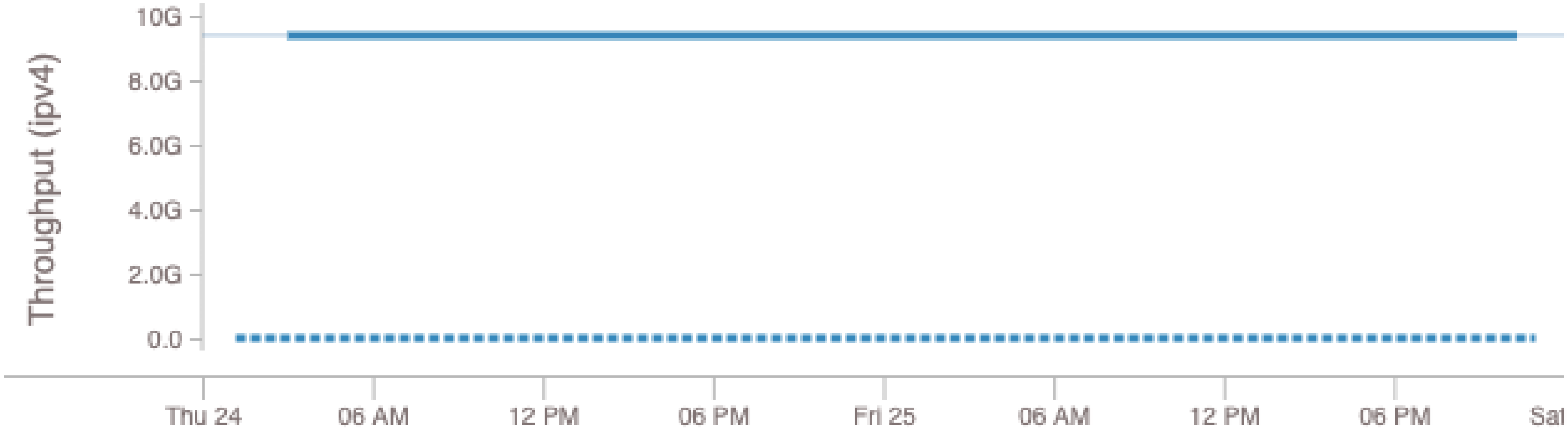
R&E Routing Architecture Vs. Commodity.

- 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

Commodity vs R&E Example: OSC to ESnet

- New perfSONAR node installed at OSC and was getting terrible performance to an ESnet pS node in one direction



EPOC
Engagement and Performance Operations Center

Commodity vs R&E OSC Troubleshooting 2

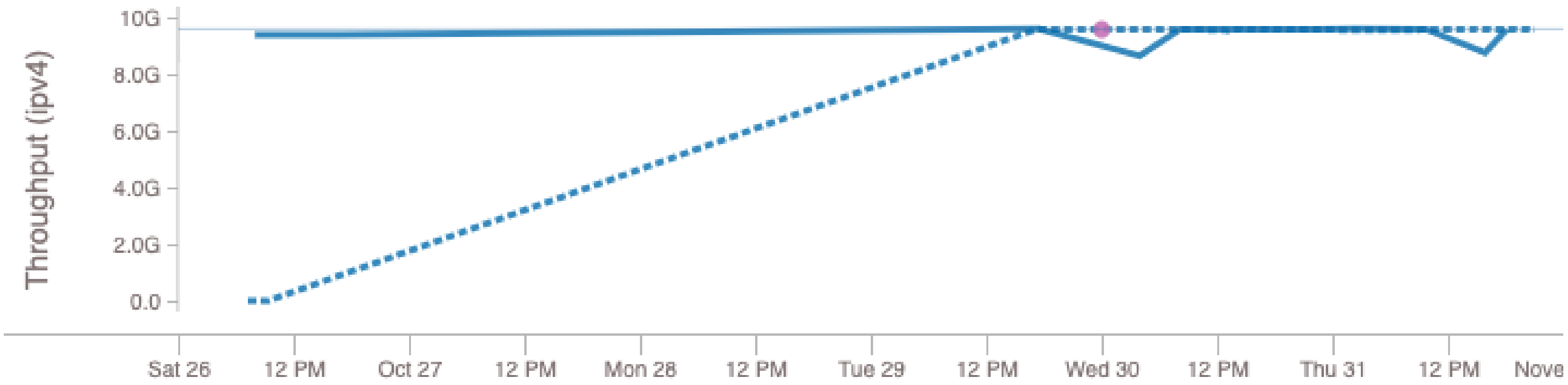
- OSC Engineer found a memory allocation issue on border router causing the routing table to not fully populate.
    - This kept the best path to ESnet out of the table
- ESnet engineer found an out of date routing configuration as well
- These fixes allowed for a R&E symmetric path for the transfer

```
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
```

EPOC
Engagement and Performance Operations Center

Commodity vs R&E: OSC Results

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



EPOC
Engagement and Performance Operations Center

Identifying Network Performance issues: Hard vs. Soft Failures

- Hard failures are the kind of problems every organization understands
  - Fiber cut
  - Power failure takes down routers
  - Hardware ceases to function
- Classic monitoring systems are good at alerting hard failures
  - i.e., NOC sees something turn red on their screen
  - Engineers paged by monitoring systems
- Soft failures are different and often go undetected
  - Basic connectivity (ping, traceroute, web pages, email) works
  - Performance is just poor

EPOC
Engagement and Performance Operations Center

Network Performance: Soft Network Failures

- Soft failures are where basic connectivity functions, but high performance is not possible.
- TCP was intentionally designed to hide all transmission errors from the user:
  - "As long as the TCPs continue to function properly and the internet system does not become completely partitioned, no transmission errors will affect the users." (From IEN 129, RFC 716)
- Some soft failures only affect high bandwidth long RTT flows.
- Hard failures are easy to detect & fix
  - soft failures can lie hidden for years!
- One network problem can often mask others

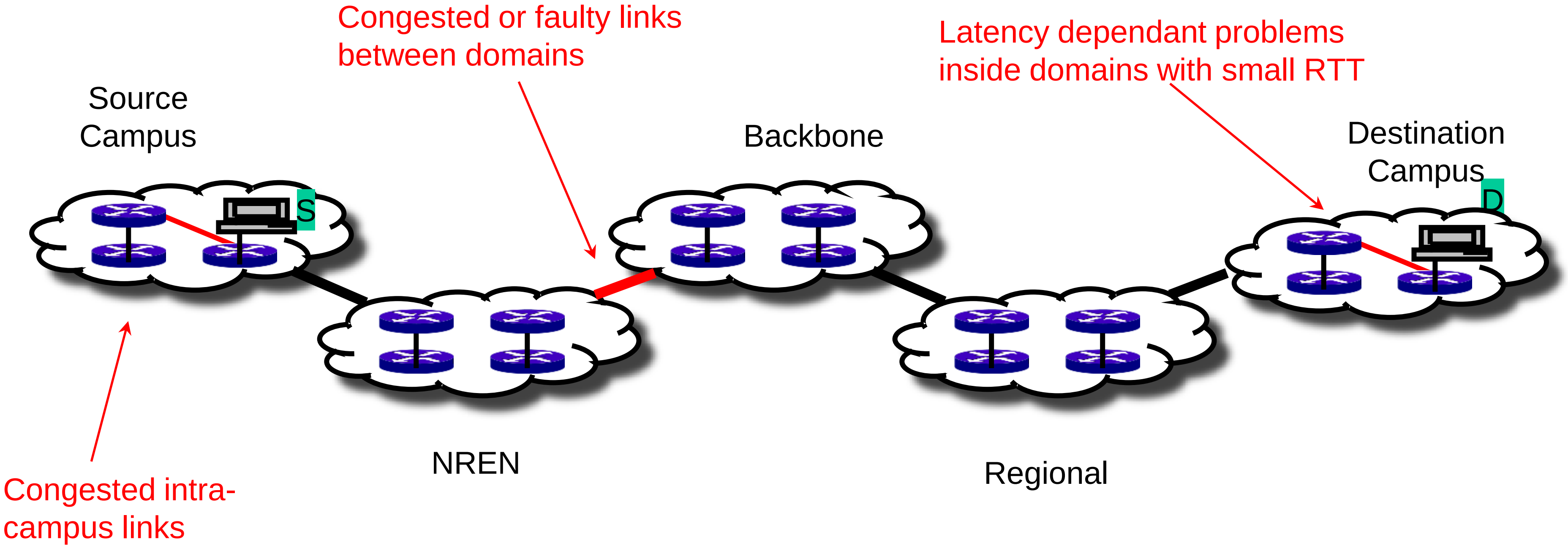
EPOC
Engagement and Performance Operations Center

Active vs. Passive Monitoring

- Passive Monitoring
  - SNMP polling
  - Netflow/sflow
  - Logs
- Active Monitoring
  - perfSONAR

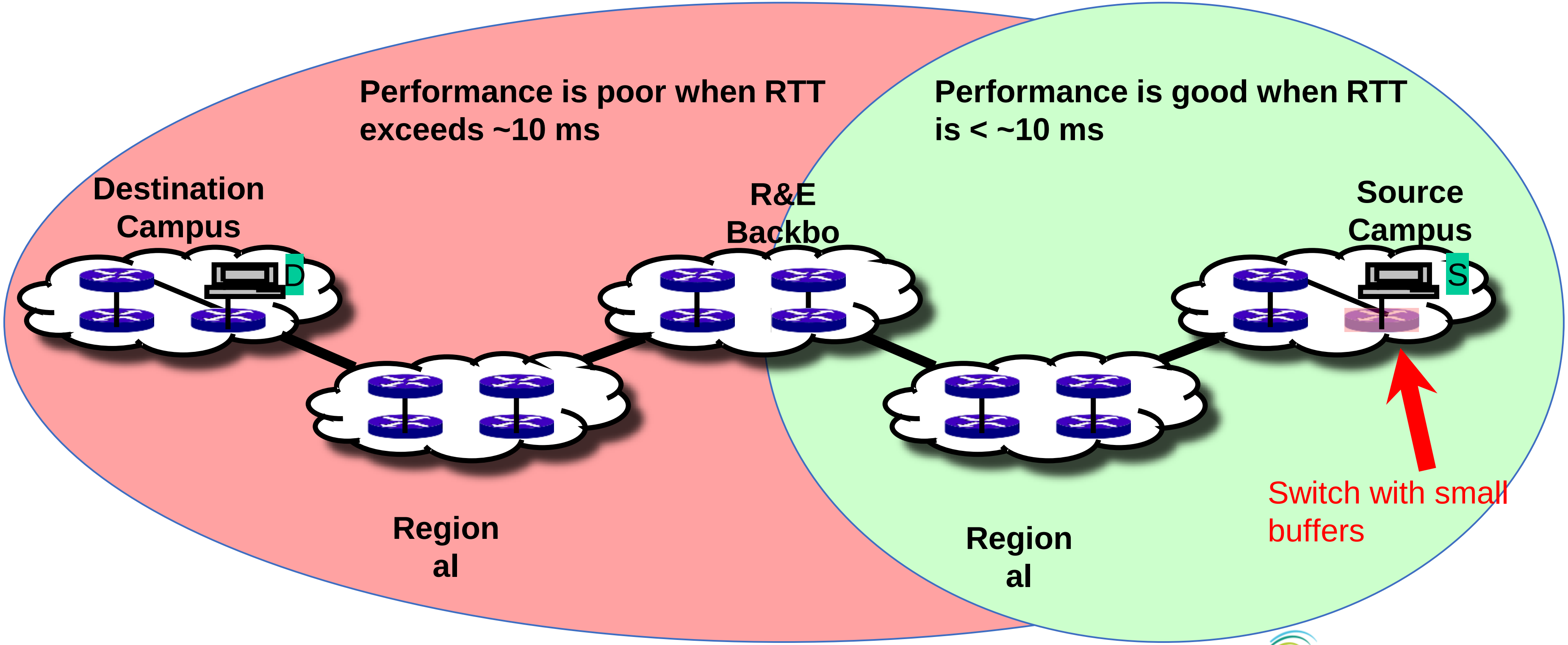
EPOC
Engagement and Performance Operations Center

Active Monitoring - Why?



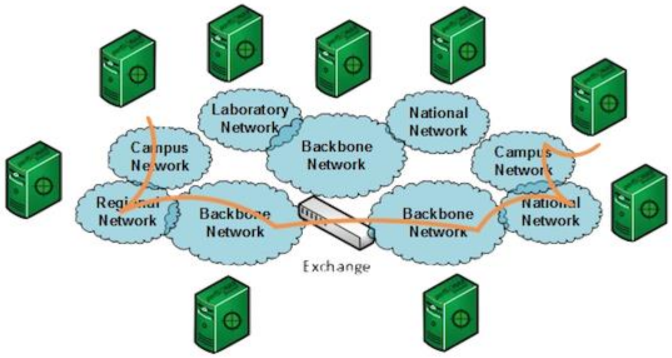
EPOC
Engagement and Performance Operations Center

Active Monitoring - Why - 2?



Active Monitoring - perfSONAR

- Consistent behavior requires clean path
- A clean path requires the ability to find and fix problems
- ***You can't fix what you can't find***
- ***You can't find what you can't see***
  - ***perfSONAR lets you see***

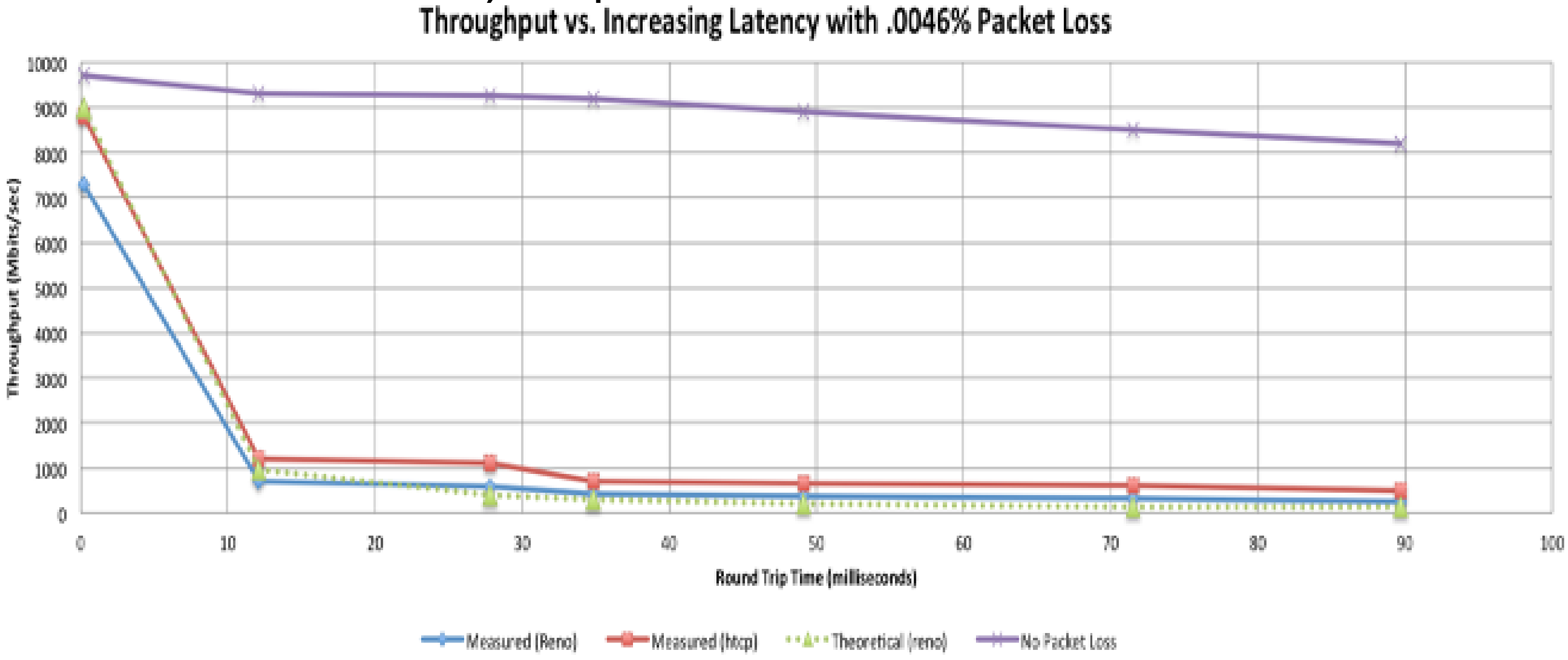


- Especially important when deploying high performance services
  - If there is a problem with the infrastructure, need to fix it
  - If the problem is not with your stuff, need to prove it
- Many players in an end to end path
- Ability to show previous patterns aids in problem localization
- Adhoc testing along trouble path available.

EPOC
Engagement and Performance Operations Center

What affects network performance: Packet Loss

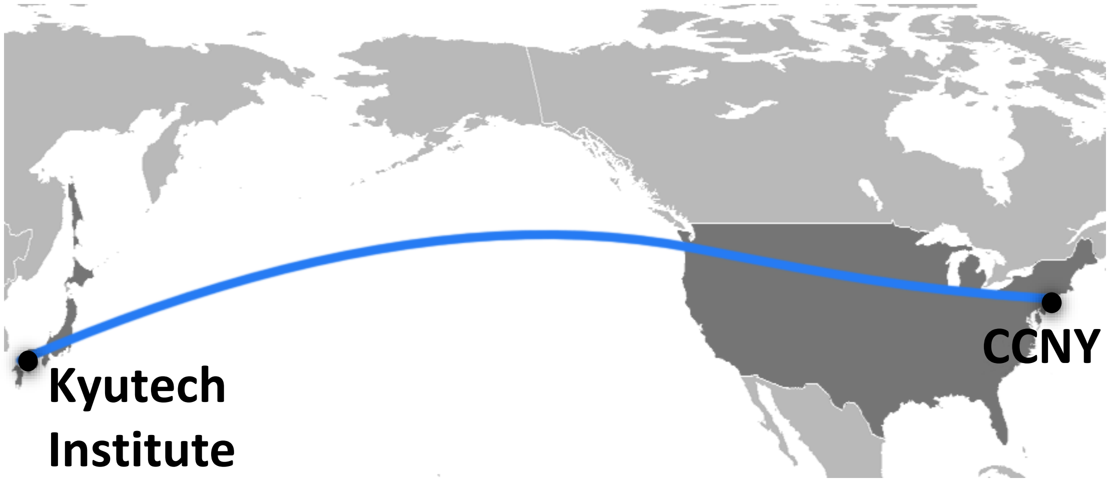
- .0046% = 1 out of 22,000 packets



Packet Loss Example - CCNY to Kyutech

Reported asymmetric, poor performance across GRE tunnel

- JGN to CCNY (TCP)
  - No packet loss
  - 79Mbps throughput
- CCNY to JGN (TCP)
  - 0.082% packet loss
  - 8Mbps throughput

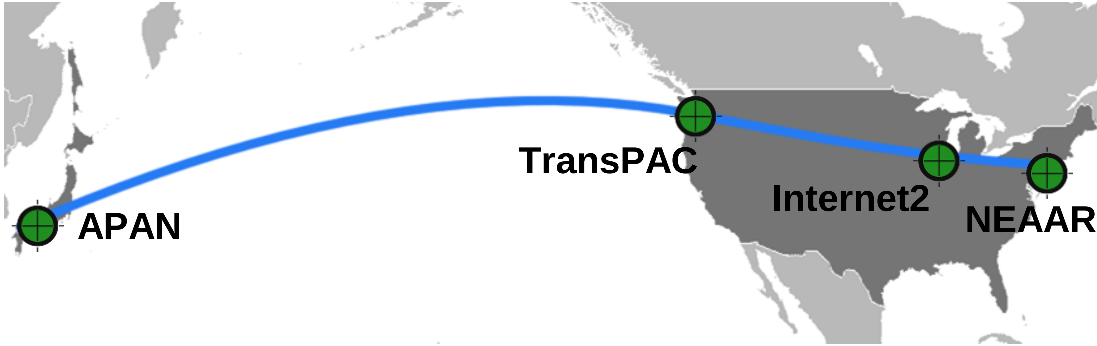


Tested UDP performance, however, was symmetric at 90Mbps either direction

Packet Loss Example - CCNY to Kyutech Troubleshooting

Used perfSONAR nodes along the path to test to closest open node available at MAN LAN

- 3rd Party ad hoc pS testing crucial



Nodes located at

- APAN/Tokyo
- TransPAC/Seattle
- Internet2/Chicago
- NEAAR/ManLan

Testing to NYC showed good performance and no packet loss- indicating problem was likely within CCNY

EPOC
Engagement and Performance Operations Center

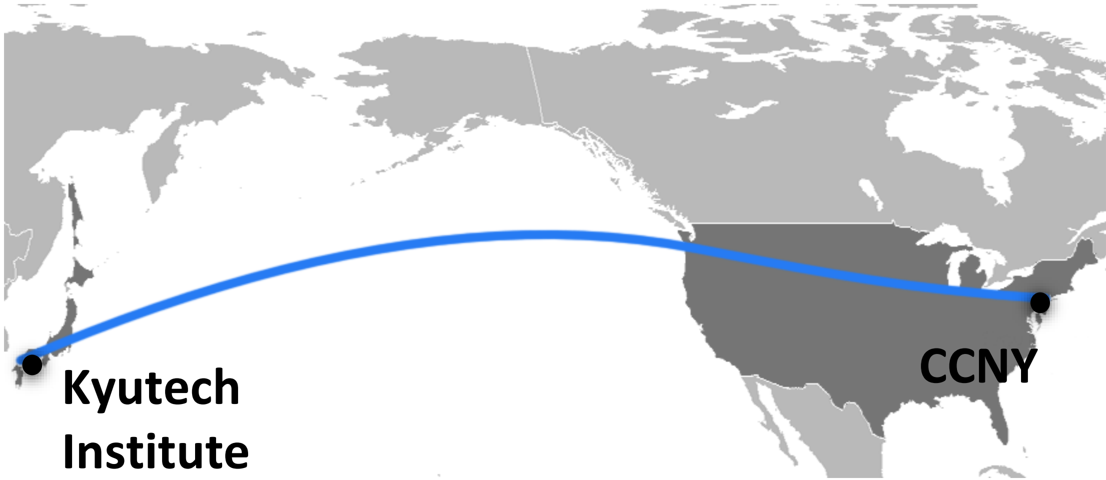
Packet Loss Example - CCNY to Kyutech Troubleshooting

- NYSERNet
  - Regional network for NY
  - Provides R&E connectivity for CCNY
  - Engineers installed a new CCNY pS node at campus edge
- Testing at regional edge to lab
  - Packet fragmentation and MTU issues on the ingress path to CCNY
  - Packet loss isolated to specific segment of the CCNY campus network

EPOC Engagement and Performance Operations Center

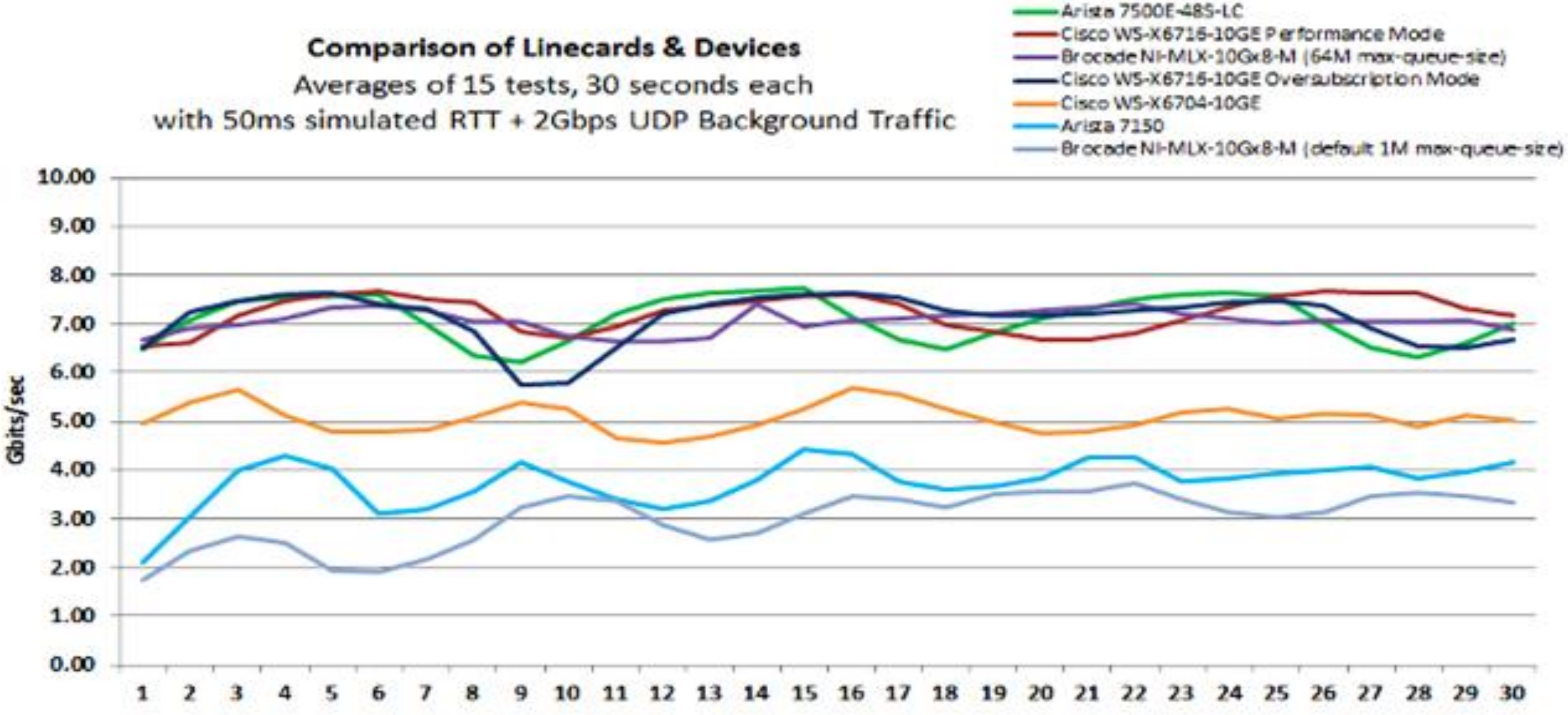
Packet Loss Example: CCNY to Kyutech Final Results

- CCNY replaced an old security appliance.
- CCNY/JGN GRE tunnel shows consistent, symmetric performance
- JGN -> CCNY (TCP)
  - No packet loss
  - 80Mbps throughput
- CCNY -> JGN (TCP)
  - **No packet loss**
  - 85Mbps throughput
  - **10-fold improvement**

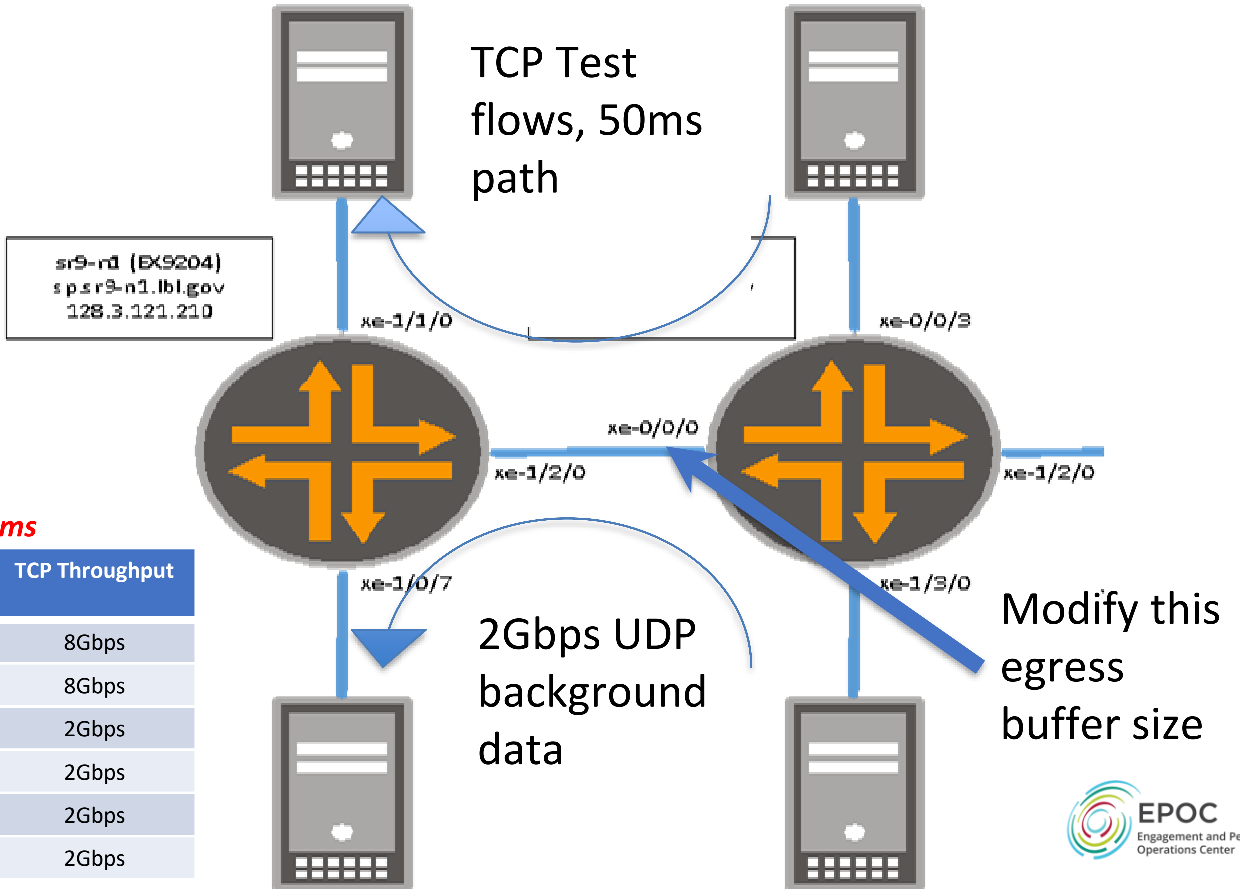


EPOC
Engagement and Performance Operations Center

Network Performance: Switch/Router Buffers



Network Performance: Switch/Router Buffers 2



30 Second test, 2 TCP streams

Buffer Size	Packets Dropped	TCP Throughput
120 MB	0	8Gbps
60 MB	0	8Gbps
36 MB	200	2Gbps
24 MB	205	2Gbps
12 MB	204	2Gbps
6 MB	207	2Gbps

EPOC Engagement and Performance Operations Center

Network Performance: BDP and the Host

The Bandwidth Delay Product

- The amount of “in flight” data for a TCP connection (BDP = bandwidth * round trip time)
- Example: 10Gb/s cross country, ~100ms
  - 10,000,000,000 b/s * .1 s = 1,000,000,000 bits
  - 1,000,000,000 / 8 = 125,000,000 bytes
  - 125,000,000 bytes / (1024*1024) ~ ***125MB***
- As the speed increases, there are more packets.
- If there is not memory, we drop them, and that makes TCP sad.

Network Performance: MTU

- Transfer performance can be impacted by MTU
  - MTU: Maximum Transmission Unit
  - MTU mismatches between networks AND internal to networks
  - Non standard MTU changes made or required by commercial DDOS scrubbing services
  - Path MTU Discovery blocked by security appliances and ACL’s

- EPOC wrote a quick guide to explain and help fix: https://epoc.global/wp-content/uploads/About-MTUs.pdf

MTU Example: Traceroute: ESnet to NRAO

```
traceroute to perfsonar-10.cv.nrao.edu (198.51.208.55), 30 hops max, 60 byte packets
 1  esneteastrt1-eastdcpt1.es.net (198.124.238.37)  0.549 ms  0.544 ms  0.547 ms
 2  newycr5-ip-a-esneteastrt1.es.net (198.124.218.17)  1.969 ms  1.963 ms  1.953 ms
 3  aofacr5-ip-a-newycr5.es.net (134.55.37.77)  2.330 ms 2.304 ms  2.313 ms
 4  et-2-1-5.197.rtsw.newy32aoa.net.internet2.edu (64.57.28.14)  2.323 ms  2.324 ms  2.327 ms
 5  ae-3.4079.rtsw.wash.net.internet2.edu (162.252.70.138)  7.571 ms  7.672 ms  7.528 ms
 6  ae-0.4079.rtsw2.ashb.net.internet2.edu (162.252.70.137)  8.095 ms  8.077 ms  8.061 ms
 7  ae-2.4079.rtsw.ashb.net.internet2.edu (162.252.70.74)  28.089 ms  18.414 ms  18.454 ms
 8  192.122.175.14 (192.122.175.14)  8.221 ms  8.179 ms  8.205 ms
 9  br01-udc-et-1-0-0-20.net.virginia.edu (192.35.48.33)  10.310 ms  10.310 ms  10.383 ms
10  cr01-udc-et-4-2-0.net.virginia.edu (128.143.236.6)  12.609 ms  12.603 ms  12.638 ms
11  cr01-gil-et-7-0-0.net.virginia.edu (128.143.236.89)  12.407 ms  12.403 ms  12.393 ms
12  perfsonar-10.cv.nrao.edu (198.51.208.55)  10.058 ms  10.032 ms  10.022 ms
```


Well, that looks good. Let's try tracepath and see where the MTU changes

EPOC
Engagement and Performance Operations Center

MTU Example: Tracepath: ESnet to NRAO, 1509 bytes

1: esneteastrt1-eastdcpt1.es.net 0.340ms
2: no reply
3: aofacr5-ip-a-newycr5.es.net 2.279ms asymm 2
4: et-2-1-5.197.rtsw.newy32aoa.net.internet2.edu 2.310ms asymm 3
5: ae-3.4079.rtsw.wash.net.internet2.edu 7.574ms asymm 4
6: ae-0.4079.rtsw2.ashb.net.internet2.edu 9.422ms asymm 5
7: ae-2.4079.rtsw.ashb.net.internet2.edu 7.986ms asymm 6
8: 192.122.175.14 8.123ms asymm 7 ← **MARIA**
9: no reply ← **UVA**

EPOC Engagement and Performance Operations Center

Tracepath: ESnet to NRAO, 1508 bytes

1: bnlmr2-bnlpt1.es.net 0.327ms
2: no reply
3: aofacr5-ip-b-newycr5.es.net 2.332ms asymm 2
4: et-2-1-5.197.rtsw.newy32aoa.net.internet2.edu 2.338ms asymm 3
5: ae-3.4079.rtsw.wash.net.internet2.edu 7.668ms asymm 4
6: ae-0.4079.rtsw2.ashb.net.internet2.edu 9.833ms asymm 5
7: ae-2.4079.rtsw.ashb.net.internet2.edu 7.872ms asymm 6
8: 192.122.175.14 8.166ms asymm 7 ← **MARIA**
9: br01-udc-et-1-0-0-20.net.virginia.edu 9.998ms asymm 7 ← **UVA**
9?: br01-udc-et-1-0-0-20.net.virginia.edu asymm 7
10: cr01-udc-et-4-2-0.net.virginia.edu 10.470ms asymm 8
11: cr01-gil-et-7-0-0.net.virginia.edu 10.208ms asymm 9
12: cr01-gil-et-7-0-0.net.virginia.edu 10.253ms pmtu 1500
12: perfsonar-10.cv.nrao.edu 10.154ms !H
Resume: pmtu 1500

MTU Example: Problem located

- The issue was between the MARIA router and the UVA router
  - The MARIA interface was configured for MTU 9192
  - The UVA interface was configured for MTU 1518
- With PMTUD broken there was no hope for external MTU 9000 equipment to negotiate an appropriate MTU with the NRAO node
- UVA changed the MTU on their router interface to match that of MARIA, while keeping their downstream equipment at their campus standard MTU 1500

EPOC
Engagement and Performance Operations Center

Network Performance: Asymmetric Routing

- Transfer performance can be impacted by asymmetric routing
  - Can reduce flow throughput
  - Large latency differences between routes
  - Round Trip Time (RTT) impacts performance

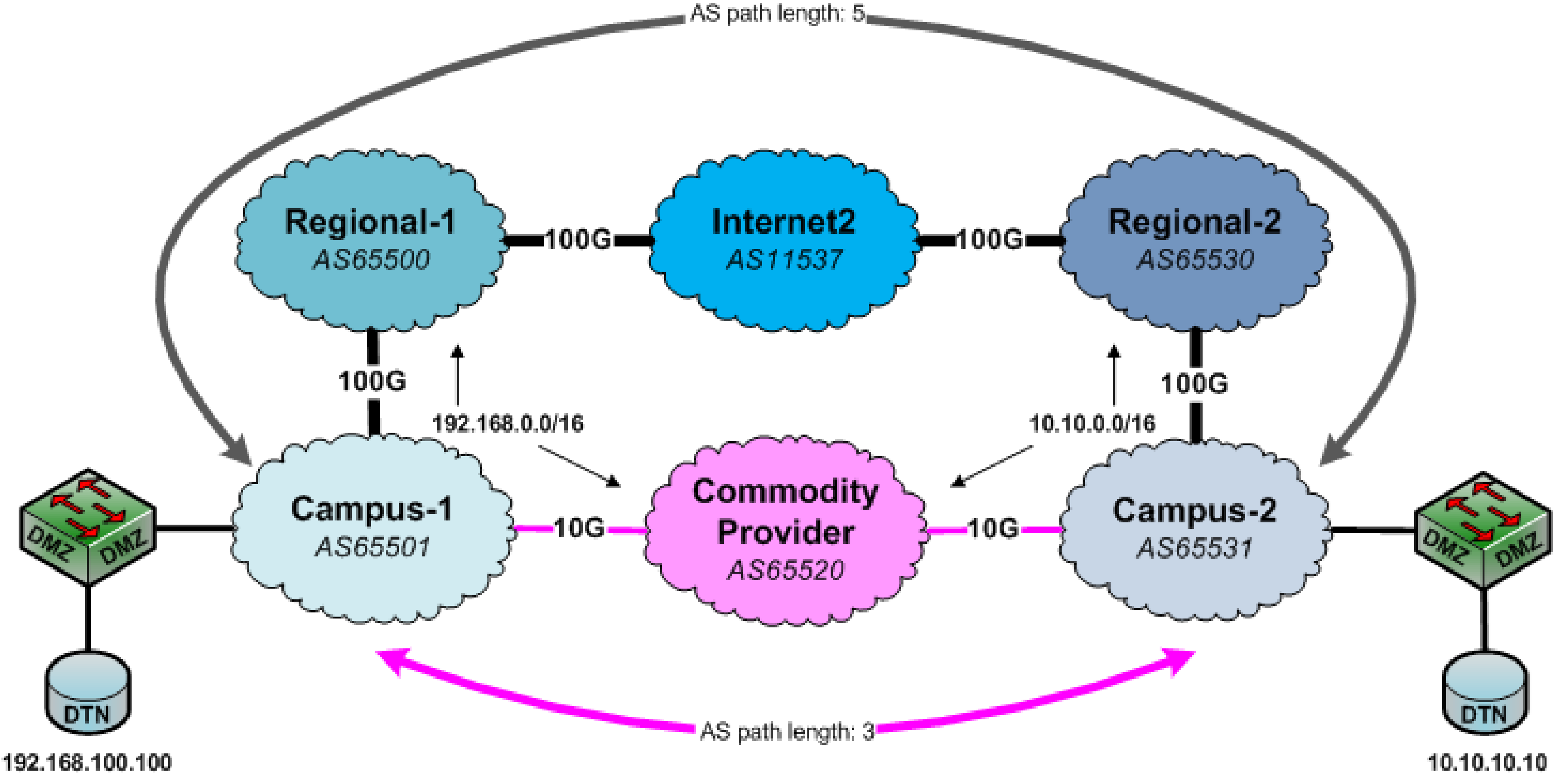
Asymmetric Routing - Example 2

Internal to Asia traffic traversing the US

- Singapore to Taiwan
  - SINGAREN (Singapore) - APAN (Asia Pacific Advanced Network) - ASGCNET Academia Sinica Grid Computing Center (Taiwan)
    - perfSONAR PS test result: 2.06 Gbps
- Taiwan to Singapore
  - ASGCNET Academia Sinica Grid Computing Center (TAIWAN) - INTERNET2-RESEARCH-EDU, (US, CHICAGO) - INTERNET2-RESEARCH-EDU(US, LA) - SINGAREN
    - perfSONAR PS test result: 815.53 Mbps
- Result of fixed asymmetrical routing
  - round trip time dropped from 290 ms to ~49ms
  - Consistent performance between 1.5 gbps and 2 gbps each direction

BGP AS Path Length Illustrated

- Hop count is a legacy metric!



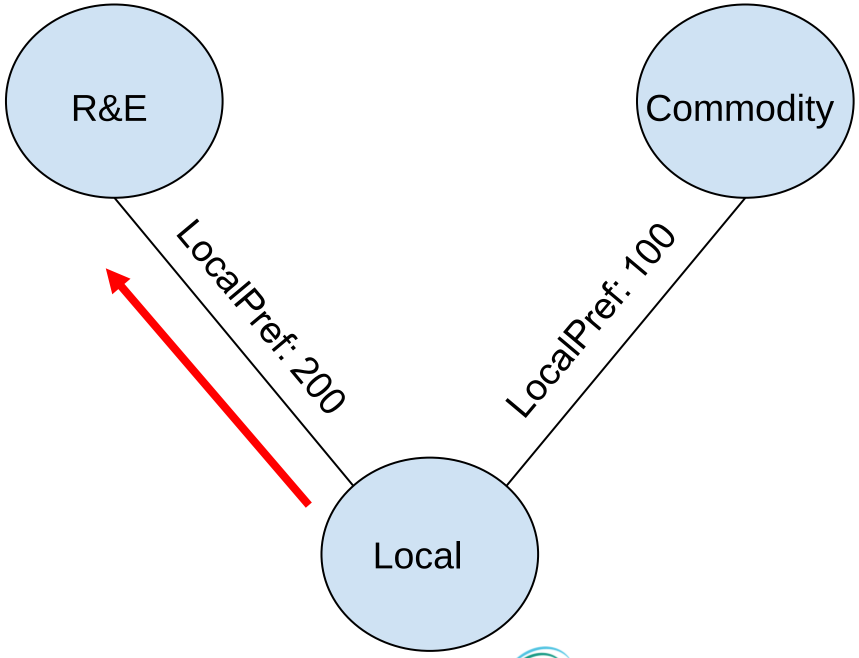
BGP - Care and feeding

- 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

LocalPref

- Per prefix
- Modifies path for outbound traffic
- Higher preferred
- Good tool for keeping R&E traffic on R&E networks



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

EPOC
Engagement and Performance Operations Center

BGP Community Strings offered by Internet2

- 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

https://noc.net.internet2.edu/i2network/maps-documentation/documentation/bgp-communities.html

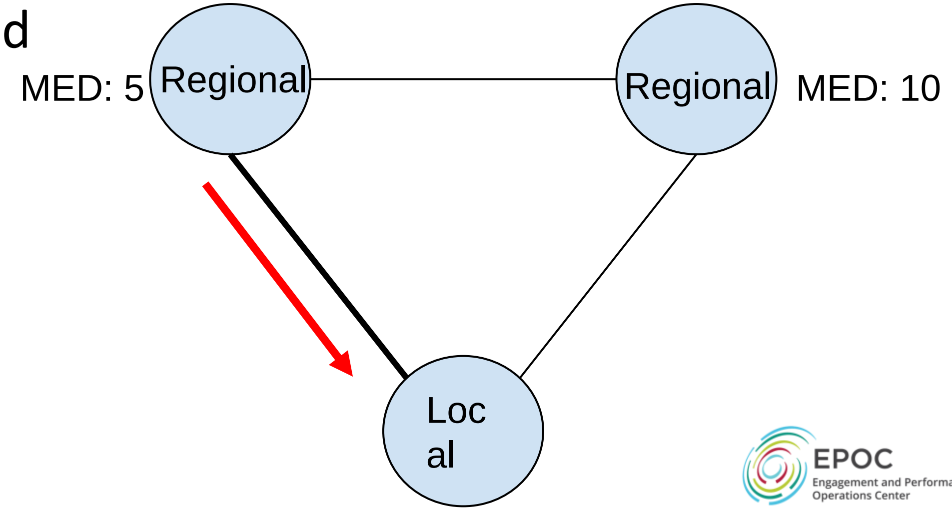
AS Path Padding

- 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
```

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



Takeaways!

Routing will not take care of itself

- Old routes may not work well with new networks
- New routes may not work as planned

How do we address routing anomalies as a community?

The Routing Working Group!

EPOC
Engagement and Performance Operations Center

Routing Working Group - What are the goals?

- **Engineering focus**
  - Document possible erroneous routes
  - Identify teams to address them
  - Check in together as we work through them
- **Policy Focus**
  - Detail routing policies for paths
    - Including preferred backup paths!
  - Verify if policy is being followed

EPOC
Engagement and Performance Operations Center

Routing Working Group

- Asymmetrical routing - meaning a source to a destination takes one path and takes a different path when it returns to the source
- R&E data takes a less efficient route around the world - affecting performance
  - Europe to Asia routes traversing the US
  - Africa to Europe routes traversing the US
- R&E data takes a commodity route when an R&E path is available
- New R&E links are removed or added but routing does not adjust appropriately
- Leaking of Private ASN's into the global routing table by R&E networks
- IP blocks advertised with a Bogon Origin ASN's within R&E routing table

EPOC
Engagement and Performance Operations Center

Submit your cases!

Email the Chairs!

meadeb@iu.edu

addlema@iu.edu

warrick.mitchell@aarnet.edu.au

Join the routing working group!

Mailing list routing-wg@gna-g.net

- Contact Brenna to be added meadeb@iu.edu

Slack

- APAN Slack Instance, Channel: Routing

Web

- https://www.gna-g.net/join-working-group/gna-g-routing-wg/

Contact any of the co-chairs for more information!

More Information

- Single point of contact to help with end-to-end performance issues: epoc@iu.edu
- More about EPOC:
  - http://epoc.global
  - Deep Dive reports: https://epoc.global/materials
- Jennifer Schopf, jmschopf@iu.edu
- Jason Zurawski, zurawski@es.net

EPOC
Engagement and Performance Operations Center