perfSONAR

EPOC: NARO/SARAO Case

Doug Southworth ▪ Indiana University ▪ dojosout@iu.edu

perfSONAR is developed by a partnership of ESnet GÉANT INDIANA UNIVERSITY INTERNET2 RNP UNIVERSITY OF MICHIGAN

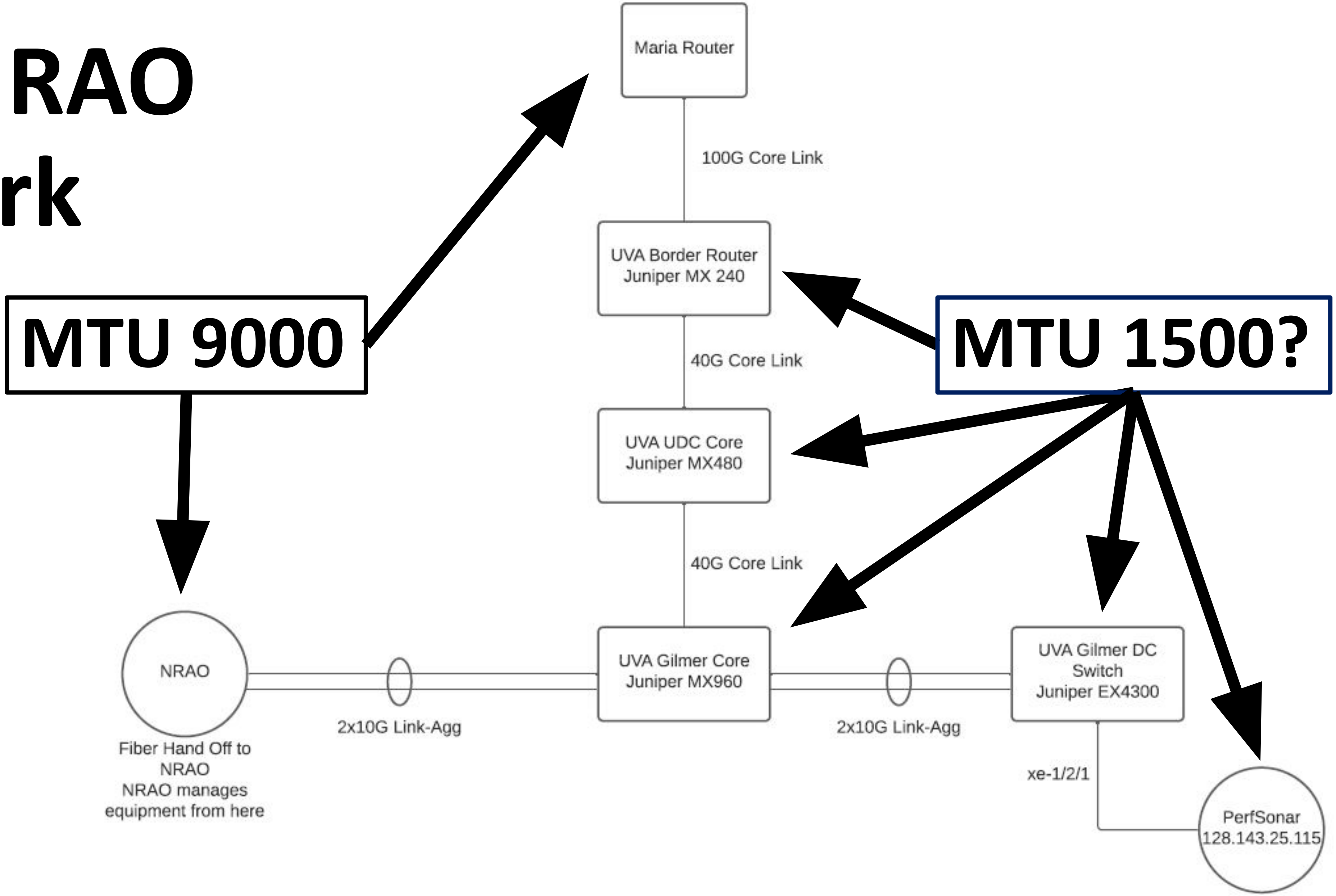
NRAO/UVA <> SARAO Performance Problem

- Data sharing from the National Radio Astronomy Observatory, located on the University of Virginia campus, to the South African Radio Astronomy Observatory
  - Low performance – 4.8Mbps
- Initial testing from the South African side revealed a few potential problems, such as asymmetric routing and paths with unnecessarily circuitous routes.
  - These were identified using normal traceroutes and quickly corrected
  - No appreciable change in performance

Initial problem isolation

- Tests from various domestic and international perfSONAR nodes to UVAs campus were telling:
  - CHPC South Africa -> Internet2 Washington - 6.67 Gbps
  - Internet2 Atlanta -> Internet2 Albany (last hop) - 9.893 Gbps
  - Internet2 Washington -> NRAO - 3.31 Gbps (lots of retries)
  - Internet2 Washington -> HPC University Virginia - 2.21 Gbps (lots of retries)
  - NRAO -> HPC University Virginia - 6.64 Gbps (lots of retries)

UVA/NRAO Network



MTU 9000

MTU 1500?

Maria Router

100G Core Link

UVA Border Router
Juniper MX 240

40G Core Link

UVA UDC Core
Juniper MX480

40G Core Link

UVA Gilmer Core
Juniper MX960

NRAO

2x10G Link-Agg

Fiber Hand Off to
NRAO
NRAO manages
equipment from here

2x10G Link-Agg

UVA Gilmer DC
Switch
Juniper EX4300

xe-1/2/1

PerfSonar
128.143.25.115

Path MTU Discovery (PMTUD)

- Is a layer 3 construct
- Requires UDP and ICMP to function
  - UDP packets larger than the MTU setting of the receiving router interface will trigger an ICMP “unreachable” message back to the sending router, which in turn causes a renegotiation to a lower MTU
- All is not lost if PMTUD doesn’t work
  - Smart transfer tools can figure out a common MTU, at the cost of time
  - Packets sent at 9K can be fragmented to adhere to a smaller MTU, at the cost of performance…unless the no-fragment flag is set
  - Neither of these scenarios is good for high performance. PMTUD should be made to work and common MTUs enforced wherever possible

Further isolation

Working inward from a known good ESnet perfSONAR node to UVA:

```
Interval      Throughput    Retransmits    Current Window
0.0 - 1.0     9.13 Gbps   22           33.17 MBytes
1.0 - 2.0     9.35 Gbps   0            33.58 MBytes
2.0 - 3.0     9.38 Gbps   0            33.58 MBytes
3.0 - 4.0     9.38 Gbps   0            33.58 MBytes
4.0 - 5.0     9.38 Gbps   0            33.58 MBytes
5.0 - 6.0     9.35 Gbps   0            33.58 MBytes
6.0 - 7.0     9.36 Gbps   0            33.58 MBytes
7.0 - 8.0     9.38 Gbps   0            33.58 MBytes
8.0 - 9.0     9.37 Gbps   0            33.58 MBytes
9.0 - 10.0    9.37 Gbps   0            33.58 MBytes

Summary
Interval      Throughput    Retransmits    Receiver Throughput
0.0 - 10.0     9.35 Gbps         22               9.25 Gbps
```

This test looks good, because the hosts successfully negotiate 1500 MTU

Negotiations break down

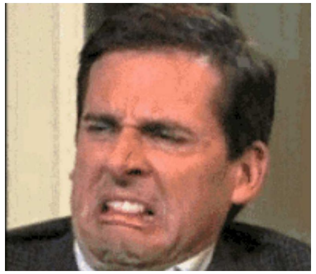
Working inward from a known good ESnet perfSONAR node to NRAO:
(Keep in mind, we know MTU 9000 on both ends, but with a step down to 1500 in the middle of the UVA campus)

```
Interval     Throughput    Retransmits   Current Window
0.0 - 1.0    2.54 Mbps   2            8.95 KBytes
1.0 - 2.0    0.00bps     1            8.95 KBytes
2.0 - 3.0    0.00bps     0            8.95 KBytes
3.0 - 4.0    0.00bps     31           3.07 KBytes
4.0 - 5.0    0.00bps     67           5.12 KBytes
5.0 - 6.0    4.45 Mbps   2            17.41 KBytes
6.0 - 7.0    8.26 Mbps   0            33.79 KBytes
7.0 - 8.0    23.28 Mbps 0             94.21 KBytes
8.0 - 9.0    51.75 Mbps 0             218.11 KBytes
9.0 - 10.0   83.88 Mbps 0             392.19 KBytes

Summary
Interval     Throughput    Retransmits   Receiver Throughput
0.0 - 10.0    17.42 Mbps    103               10.29 Mbps
```

9000B packets failing

1500B packets after re-negotiation



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

Tracepath: ESnet to NRAO

```
1?: [LOCALHOST]                                          pmtu 9000
 1:  esneteastrt1-eastdcpt1.es.net                       0.788ms
 1:  bnlmr2-bnlpt1.es.net                                0.728ms
 2:  no reply
 3:  aofacr5-ip-b-newycr5.es.net                         2.411ms asymm  2
 4:  et-2-1-5.197.rtsw.newy32aoa.net.internet2.edu  2.468ms asymm  3
 5:  ae-3.4079.rtsw.wash.net.internet2.edu              8.176ms asymm  4
 6:  ae-0.4079.rtsw2.ashb.net.internet2.edu                  8.889ms asymm  5
 7:  ae-2.4079.rtsw.ashb.net.internet2.edu              8.242ms asymm  6
 8:  192.122.175.14                                      8.522ms asymm  7
 9:  no reply
10:  no reply
11:  no reply
12:  no reply
```

Traceroute works, but tracepath doesn't??

Different Tools, Different Packets

- Traceroute uses small 60B UDP packets
- Tracepath uses larger 64KB UDP packets

So, somewhere we have a roadblock. Small packets can make it through, but larger ones are dropped (not fragmented).

How do we figure out the max size? Trial and error. Start at 9K and cut the size in half until you get a response, then sneak back up until the packets disappear again.

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
9:  no reply
```

← **MARIA**

← **UVA**

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
```

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

Ping verification

```
ping -s 8972 -M do -c 4 perfsonar-10.cv.nrao.edu (don’t fragment)

PING perfsonar-10.cv.nrao.edu (198.51.208.55) 8972(9000) bytes of data.
From cr01-gil-et-7-0-0.net.virginia.edu (128.143.236.89) icmp_seq=1 Frag needed and DF set (mtu = 1500)
ping: local error: Message too long, mtu=1500
ping: local error: Message too long, mtu=1500
ping: local error: Message too long, mtu=1500
```

```
ping -s 8972 -M dont -c 4 perfsonar-10.cv.nrao.edu (do fragment)

PING perfsonar-10.cv.nrao.edu (198.51.208.55) 8972(9000) bytes of data.
8980 bytes from perfsonar-10.cv.nrao.edu (198.51.208.55): icmp_seq=1 ttl=55 time=10.3 ms
8980 bytes from perfsonar-10.cv.nrao.edu (198.51.208.55): icmp_seq=2 ttl=55 time=10.2 ms
8980 bytes from perfsonar-10.cv.nrao.edu (198.51.208.55): icmp_seq=3 ttl=55 time=10.2 ms
8980 bytes from perfsonar-10.cv.nrao.edu (198.51.208.55): icmp_seq=4 ttl=55 time=10.2 ms
```

Yeah, yeah, but what about performance??

Before:

```
pscheduler task throughput --source cpt-chpc-10g.perfsonar.ac.za --dest
perfsonar-10.cv.nrao.edu
Summary
Interval      Throughput     Retransmits    Receiver Throughput
0.0 - 10.0     380.37 Kbps      58        108.18 Kbps
```

After:

```
pscheduler task throughput -t 30 --source cpt-chpc-10g.perfsonar.ac.za --dest
perfsonar-10.cv.nrao.edu
Summary
Interval      Throughput     Retransmits    Receiver Throughput
0.0 - 30.0     2.67 Gbps        0           2.62 Gbps
```

perfSONAR

Thanks!

For more information,
please visit our web site:
https://www.perfsonar.net

Thanks icon by priyanka from The Noun Project

perfSONAR is developed by a partnership of ESnet GÉANT INDIANA UNIVERSITY INTERNET2 RNP UNIVERSITY OF MICHIGAN

EPOC: NARO/SARAO Case

Doug Southworth ▪ Indiana University ▪ dojosout@iu.edu

perfSONAR is developed by a partnership of