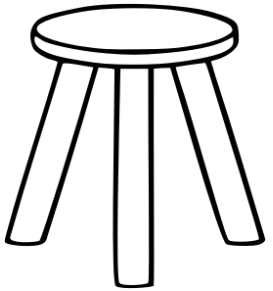
Layer-transcending traceroute - LIME implications?

draft-nordmark-nvo3-transcending-traceroute-01

Erik Nordmark
nordmark@arista.com

Three Main Points



1. Existing traceroute shows the path even though the user can’t control the devices nor the path
2. Overlay networks might hide underlay path
   - We should separate policy and mechanism
3. *Not here - details specified for VXLAN in draft*

Revealing the path

```
bash$ traceroute www.dn.se
traceroute to a1910.g1.akamai.net (63.150.12.17), 64 hops max, 52 byte packets
 1  cs2-wifi-epool-vl1070.aristanetworks.com (172.22.227.3)  1.717 ms  6.007 ms  2.755 ms
 2  us-ca-sc1-paf001-vl3316.aristanetworks.com (172.22.199.45)  2.118 ms  2.269 ms  2.339 ms
 3  us-ca01-01-sw7124-01-vl550.aristanetworks.com (162.210.130.1)  2.260 ms  3.126 ms  2.697
ms
 4  10ge8-4.core3.fmt2.he.net (216.218.196.189)  4.334 ms  5.611 ms  3.930 ms
 5  10ge10-1.core1.sjc2.he.net (184.105.222.14)  16.673 ms  5.709 ms  12.821 ms
 6  sjo-b21-link.telia.net (213.248.67.105)  4.875 ms  7.110 ms  5.097 ms
 7  qwest-ic-300327-sjo-b21.c.telia.net (62.115.12.94)  4.848 ms  6.488 ms  8.788 ms
 8  * * *
 9  * * *
```

- Useful information for a trouble ticket
- Policy? Could filter or not send ICMPs

Overlay providing L2 service

- Overlay traceroute and ping show nothing
- Overlay ARP may or may not time out
- Need access to ingress NVE to
  - Inspect tables - Mac address to NVE address? Port, vlan to vni id mapping?
  - Run underlay ping/traceroute to destination NVE
- Without NVE access trouble ticket is empty
  - Difficult to troubleshoot temporary conditions

Overlay tunnel model

- IETF has developed a pipe and a uniform tunnel model (for diffserv and ttl)
- Pipe tunnel model is commonly used
  - Ingress NVE uses a fixed outer ttl
  - Egress NVE doesn’t look at outer ttl
- Uniform tunnel model counts underlay hops
  - Ingress NVE sets outer ttl to (inner ttl - 1)
  - Egress NVE sets inner ttl to (outer ttl - 1)

ICMP error handling

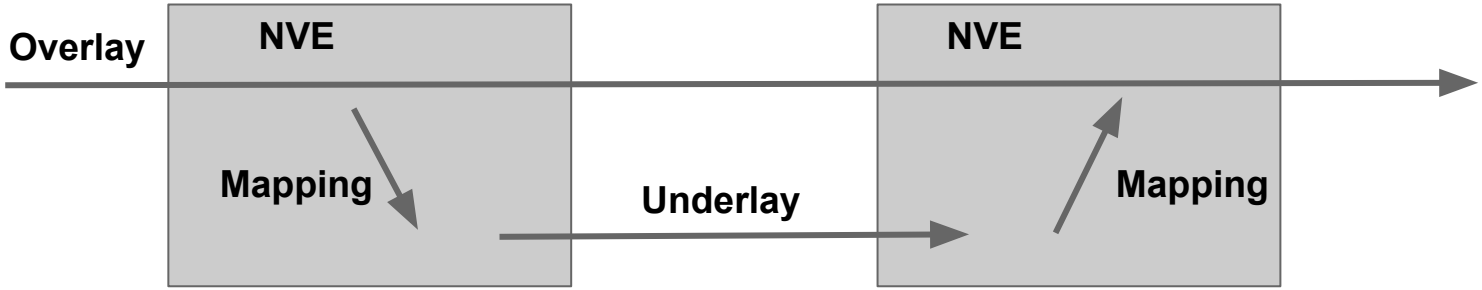
- Based on idea going back to RFC1933
- Underlay routers will send ICMP error back to outer IP source
  - Standard IP behavior - RFC1812
  - ICMP error gets delivered to ingress NVE
- Added behavior at NVE
  - Use such ICMP errors to form ICMP errors for original source

Potential Surprises

- The underlay IP addresses are unrelated to overlay
  - Different IP address realm
  - Could be IPv6 and IPv4 combinations
- Example for IPv6 over IPv4:

```
traceroute to 2000:0:0:40::2, 30 hops max, 80 byte packets
 1  ::2.0.1.1 (::2.0.1.1)  1.231 ms  1.004 ms  1.126 ms
 2  ::2.0.1.2 (::2.0.1.2)  1.994 ms  2.301 ms  2.016 ms
 3  ::2.0.2.1 (::2.0.2.1)  18.846 ms  30.582 ms  19.776 ms
 4  2000:0:0:40::2 (2000:0:0:40::2)  48.964 ms  60.131 ms  53.895 ms
```

Upleveling to LIME



- Does the LIME model prevent layer transcendance as described here?
- Should LIME be open to layer transcendance, subject to policy?
- Failures in mappings?