BERKELEY LAB
Bringing Science Solutions to the World

U.S. DEPARTMENT OF ENERGY
Office of Science

Network Tapping for Zeek
A Deep Dive

Michael Smitasin
Cyber Security Engineer
security@lbl.gov

March 17, 2023

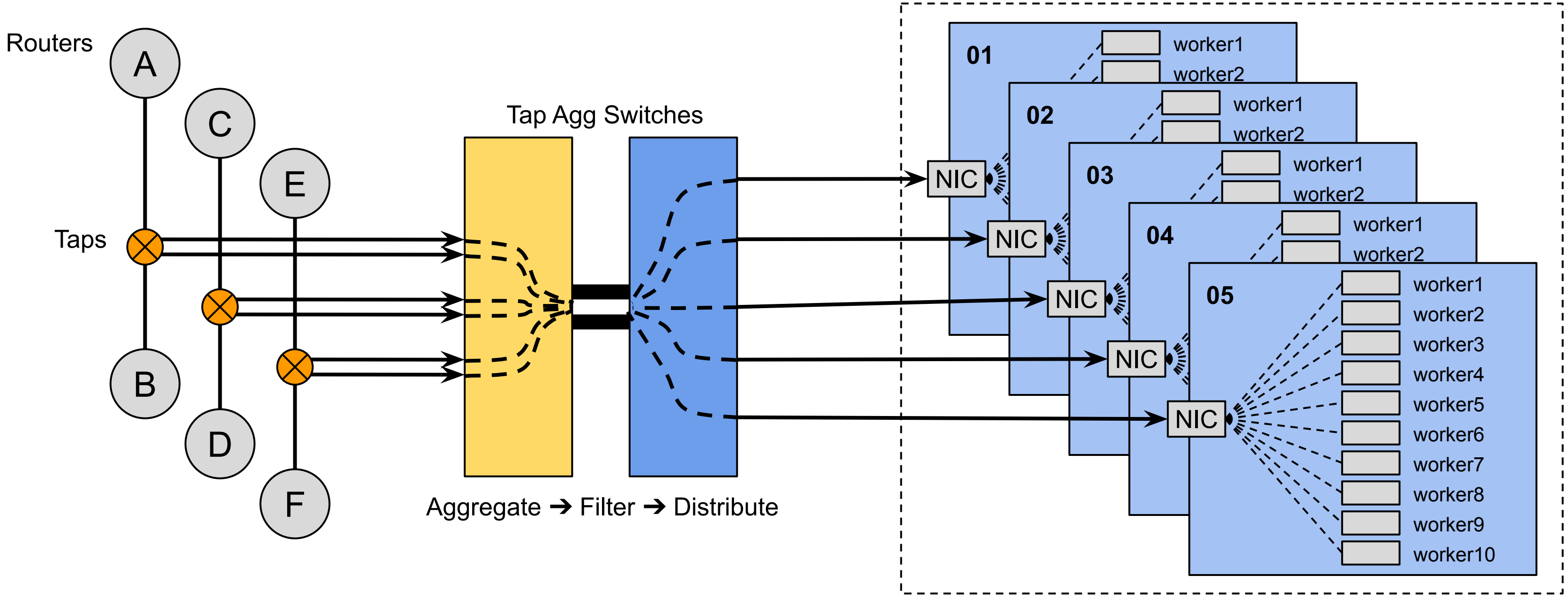
Groundwork Definitions

Taps: make a copy of traffic

Tap Agg: aggregate + manipulate copies

Zeek: distill information, take action

The Formula



Tapped links

Tap Agg

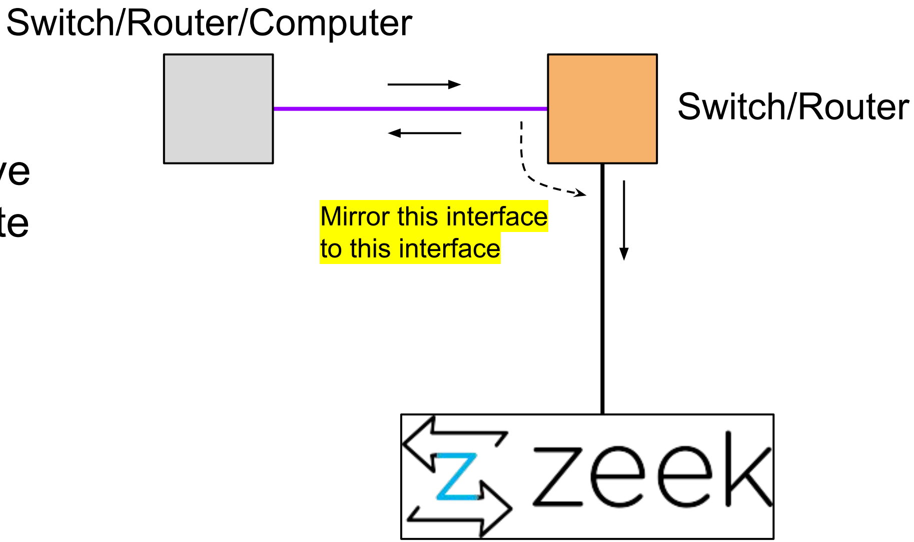
Zeek Cluster

Outline

- Taps
- Tap Aggregation
- LBNL's Environment
- Scaling Up with Load Balancing
- Static Traffic Filter
- Dynamic Filtering (Shunting)
- TCAM Limitations
- Ingress/Egress ACL Workaround
- Identity VLANs
- Tapping Cloud email
- Visibility in the Cloud
- Tapping & Tap Agg @ 400G

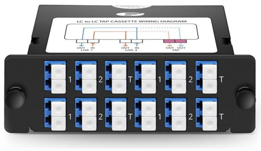
Mirror / Monitor / SPAN* Ports

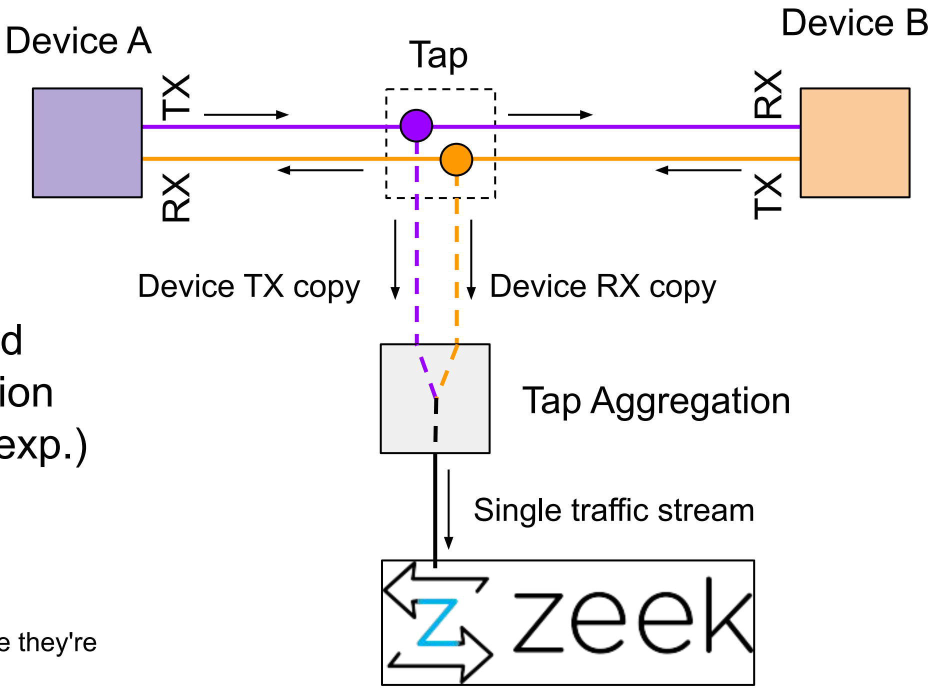
- On-Device Packet Replication
- (+) Free?
- (+) Can filter at source
- (+) Non-disruptive add/change/remove
- (+) RSPAN/Lawful Intercept for remote capture
- (-) In-band / Resource contention?
- (-) Hardware limits
  - Ex: max 2 SPAN ports
- (-) Potential oversubscription
  - (1G TX, 1G RX = 2G tapped)



*Switch Port ANalyzer

Taps



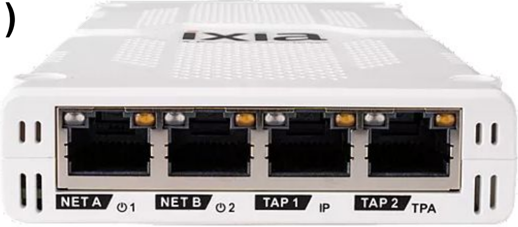


- (+) Out-of-Band
- (+) Fiber taps can be passive/unpowered
- (+) Fiber taps: all light, no oversubscription
- (+) Passive taps: Highly reliable (in our exp.)
- (+)/(-) Price
- (-) Disruptive add/change/remove*

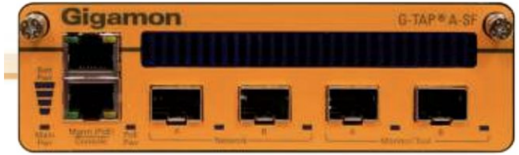
*(there are maybe things like non-disruptive microbend taps, I'm not sure they're commercially available)

Tap Hardware

(1)

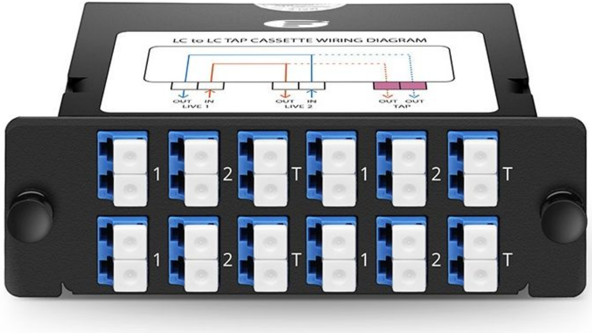


(2)



- Different Flavors of Taps
  1. Copper Taps
  2. Active Optical Taps
  3. Passive Optical Taps
  4. Fiber Patch Tap Cables

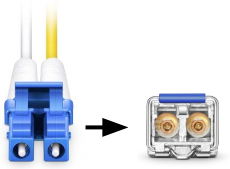
(3)



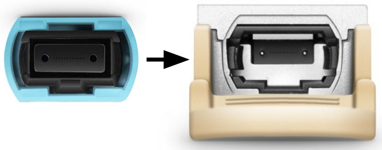
(4)

Fiber

Common Fiber connectors

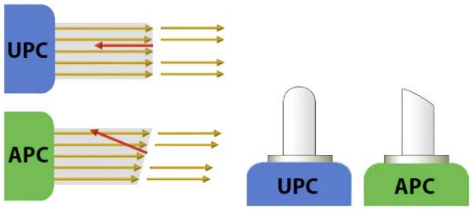


LC



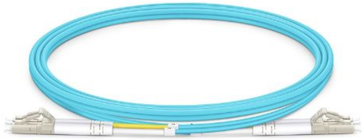
MPO (aka MTP®)

UPC vs APC
Don't mix these!



Common Fiber cables

OS2 (Optical Singlemode)
Long distance, any speed



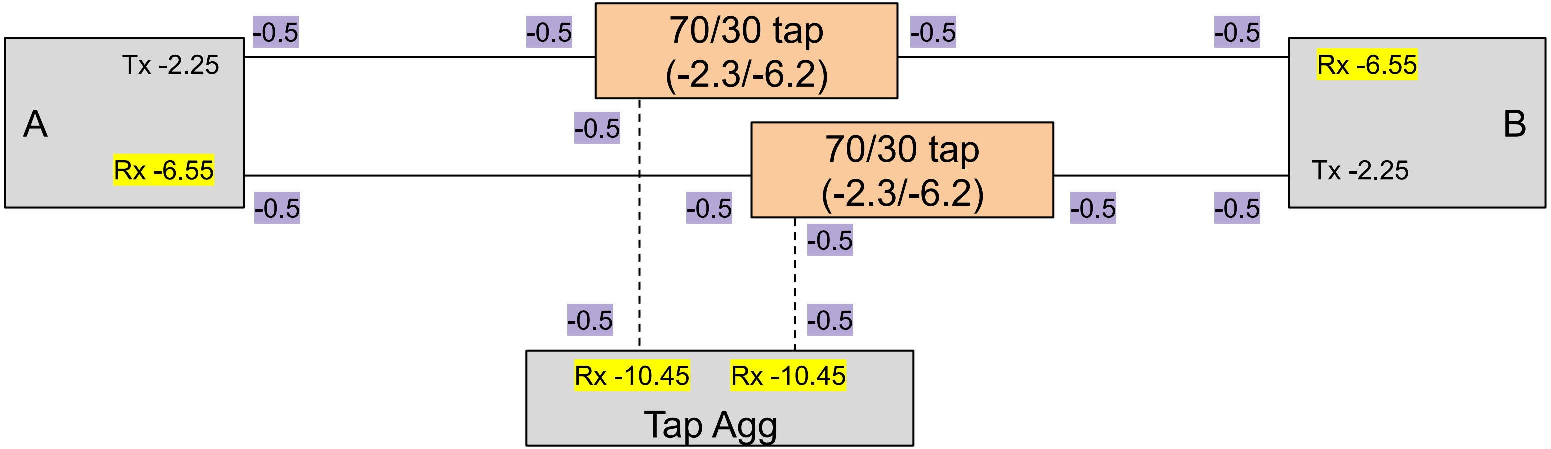
OM4 (Optical Multimode)
LC/LC connectors
Short dist., lower speeds

OM4 (Optical Multimode)
MPO-MPO connectors
(Polarity Type B)
Short dist., higher speeds

Calculating Light Budget

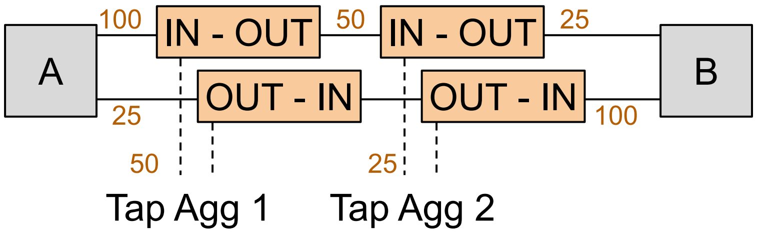
- Light split ratios: 50/50, 70/30, 80/20
  - ○ Do you have enough light budget?

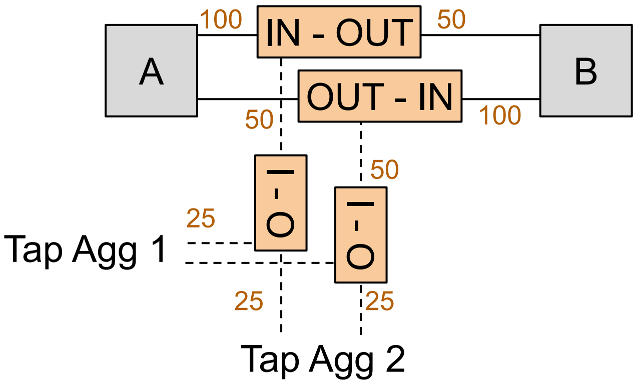
-0.5 = connector loss



Calculating Light Budget

- Multiple taps for multiple locations





Checking Light Levels

- Thresholds: device output below, or check the optical modules specs/data sheet, something like "Receiver Sensitivity" or "Receive Power" max/min.

- Cisco C6800s

```
#show interfaces Te1/1 transceiver detail
```

  "Optical Receive Power (dBm)"

- Arista 7280s

```
#show int et25/1 transceiver detail
```

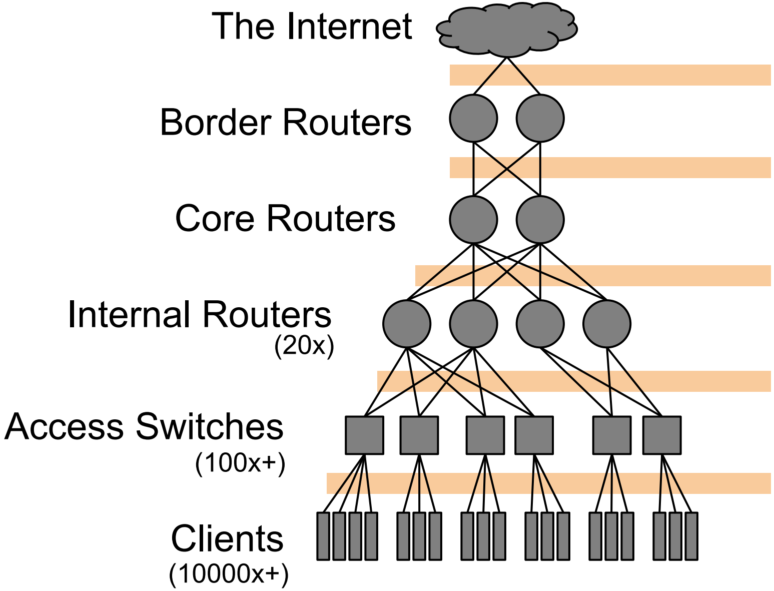
  "Rx Power (dBm)"

- Juniper MX/EX

```
> show interfaces diagnostics optics et-1/0/2
```

  "Laser receiver power"

Hierarchical Network Tapping



External conns, lots of scanning noise, cheap

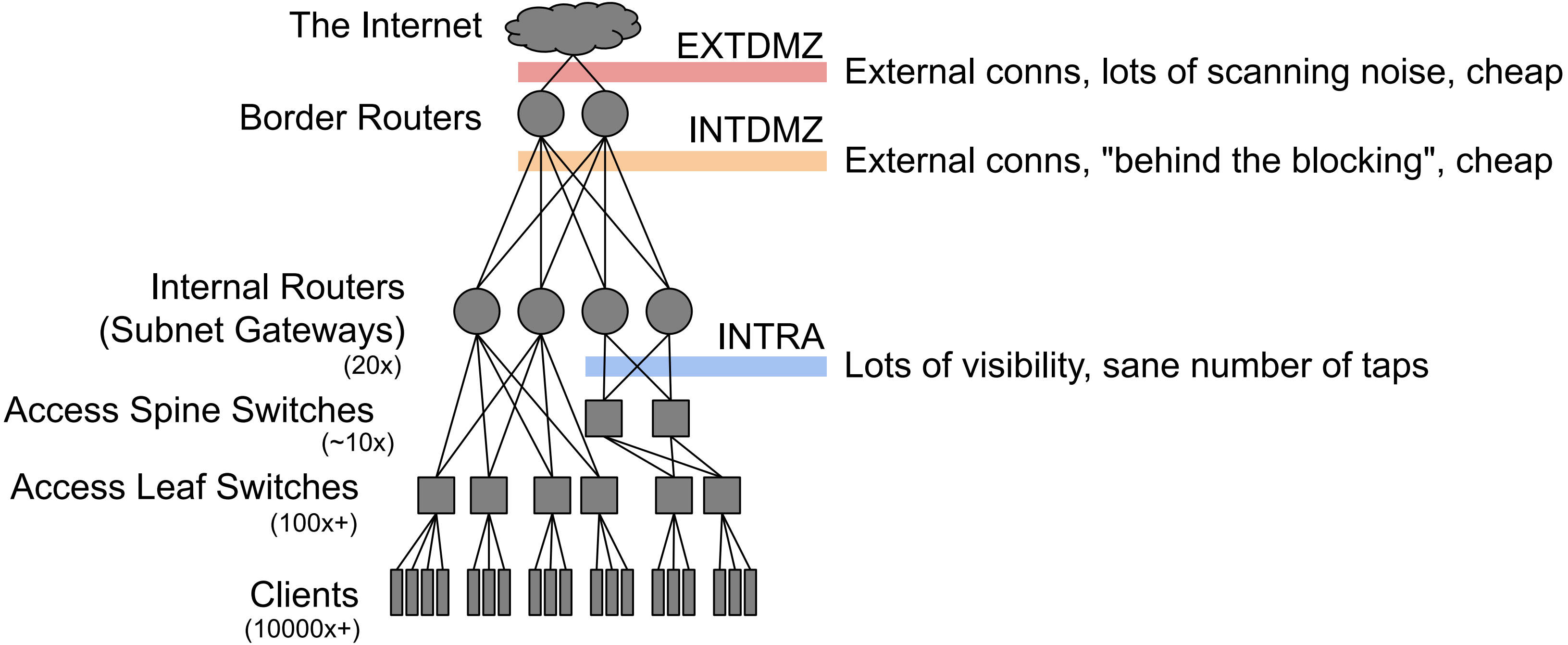
External conns, "behind the blocking", cheap

Some internal visibility, but more expensive

Lots of internal visibility, but cost prohibitive

All the internal visibility, impractical and cost prohibitive

LBNL's Current Tapping



The Internet

Border Routers

Internal Routers
(Subnet Gateways)
(20x)

Access Spine Switches
(~10x)

Access Leaf Switches
(100x+)

Clients
(10000x+)

EXTDMZ

External conns, lots of scanning noise, cheap

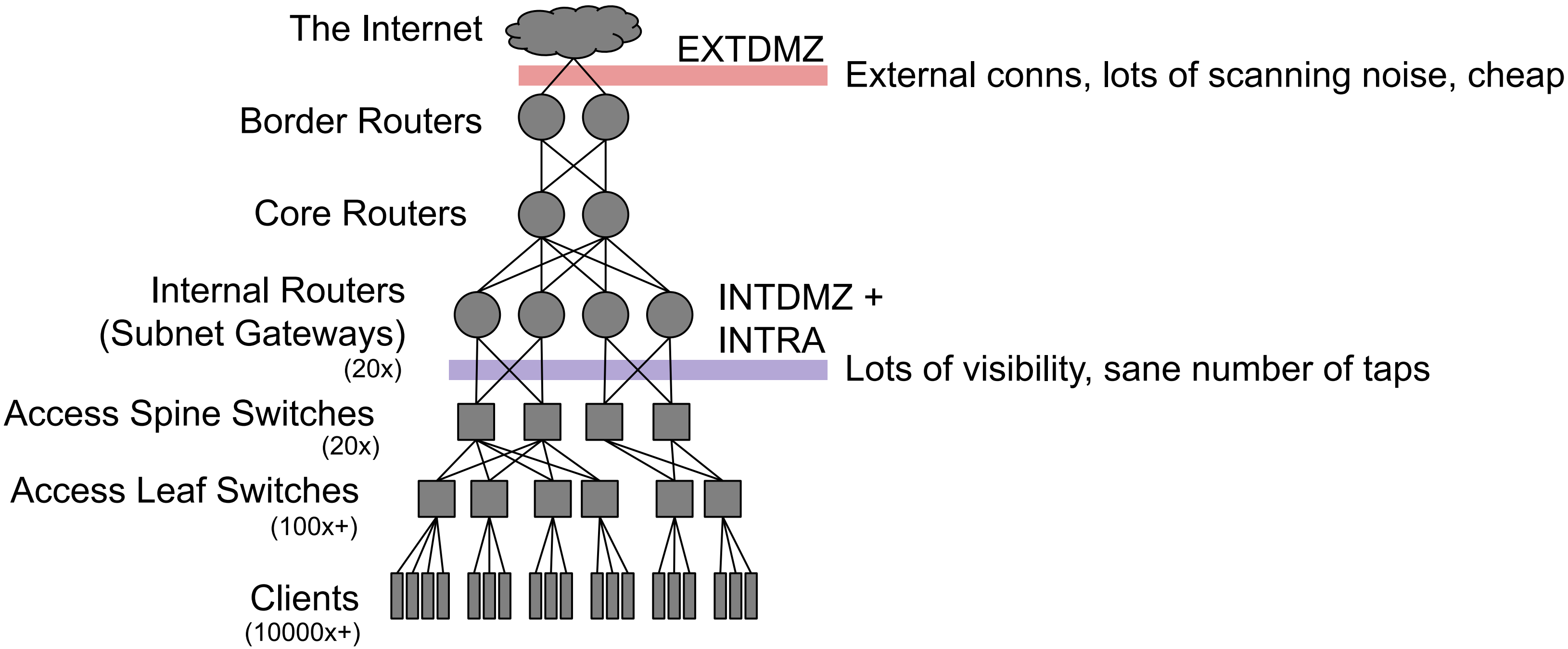
INTDMZ

External conns, "behind the blocking", cheap

INTRA

Lots of visibility, sane number of taps

LBNL's Future Tapping



The Internet

EXTDMZ

External conns, lots of scanning noise, cheap

Border Routers

Core Routers

Internal Routers (Subnet Gateways) (20x)

INTDMZ + INTRA

Lots of visibility, sane number of taps

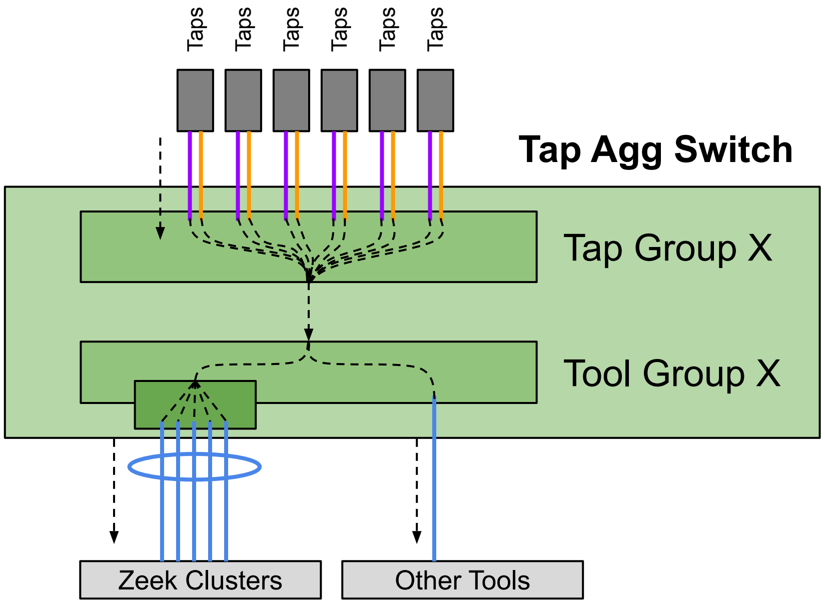
Access Spine Switches (20x)

Access Leaf Switches (100x+)

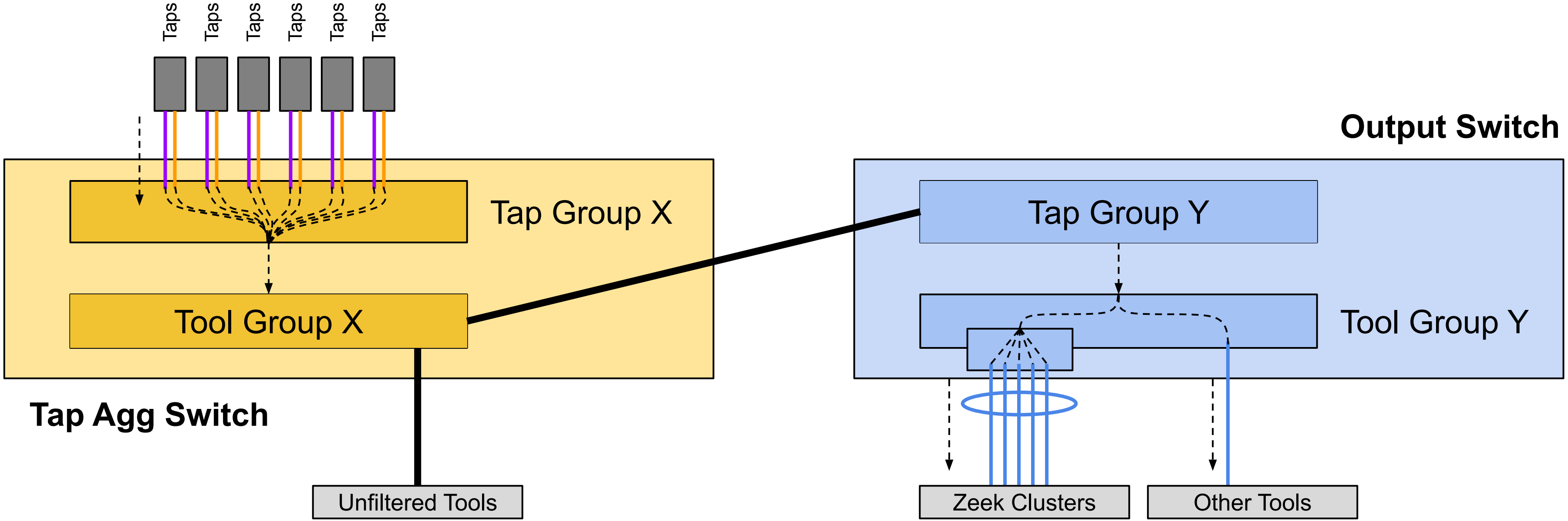
Clients (10000x+)

Tap Agg Concepts (1)

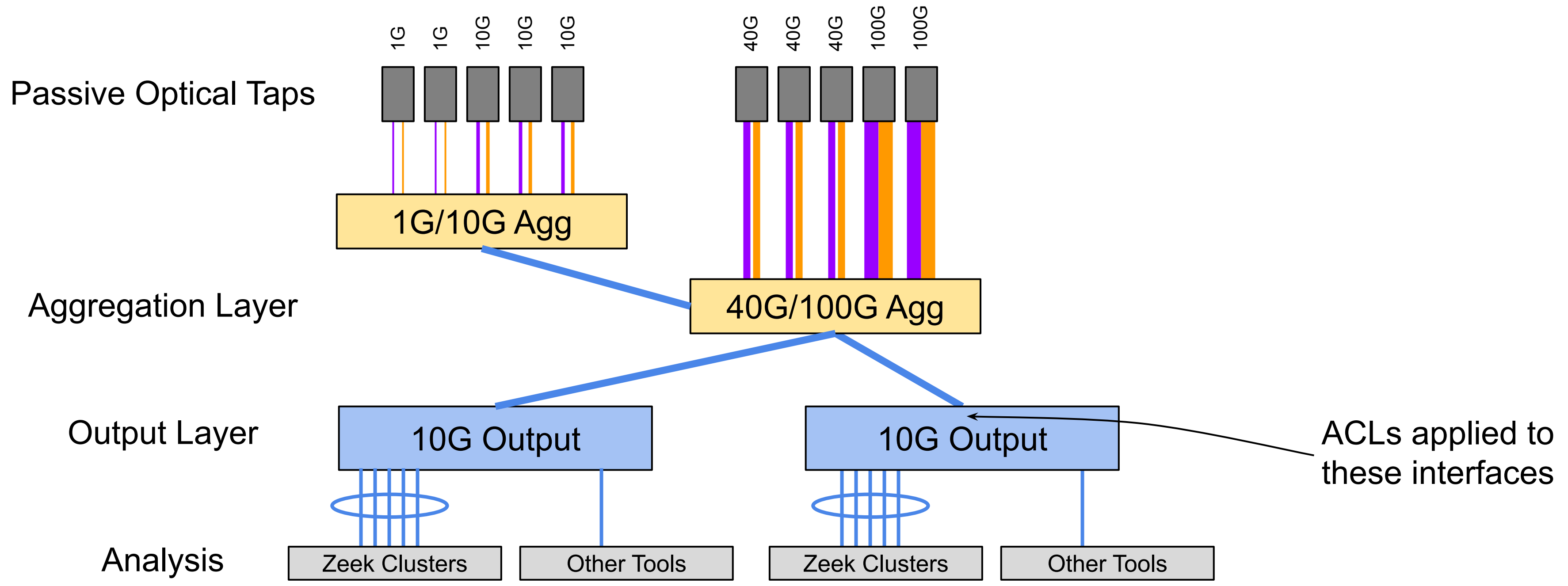
- "Tap Agg Switch"
  AKA "Network Packet Broker"
- **Aggregate** taps to traffic streams
- **Filter** out traffic you don't want
- **Replicate** copies to different tools
- **Distribute** across cluster nodes



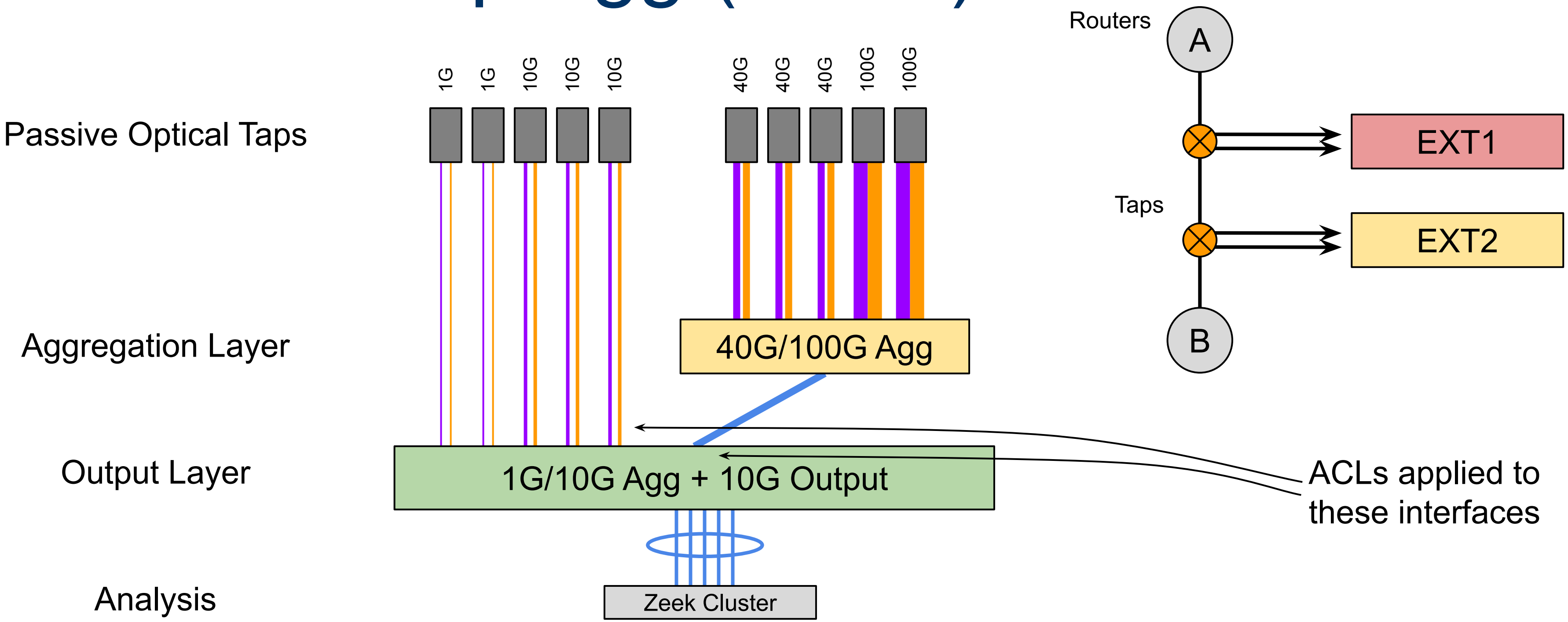
Tap Agg Concepts (2)



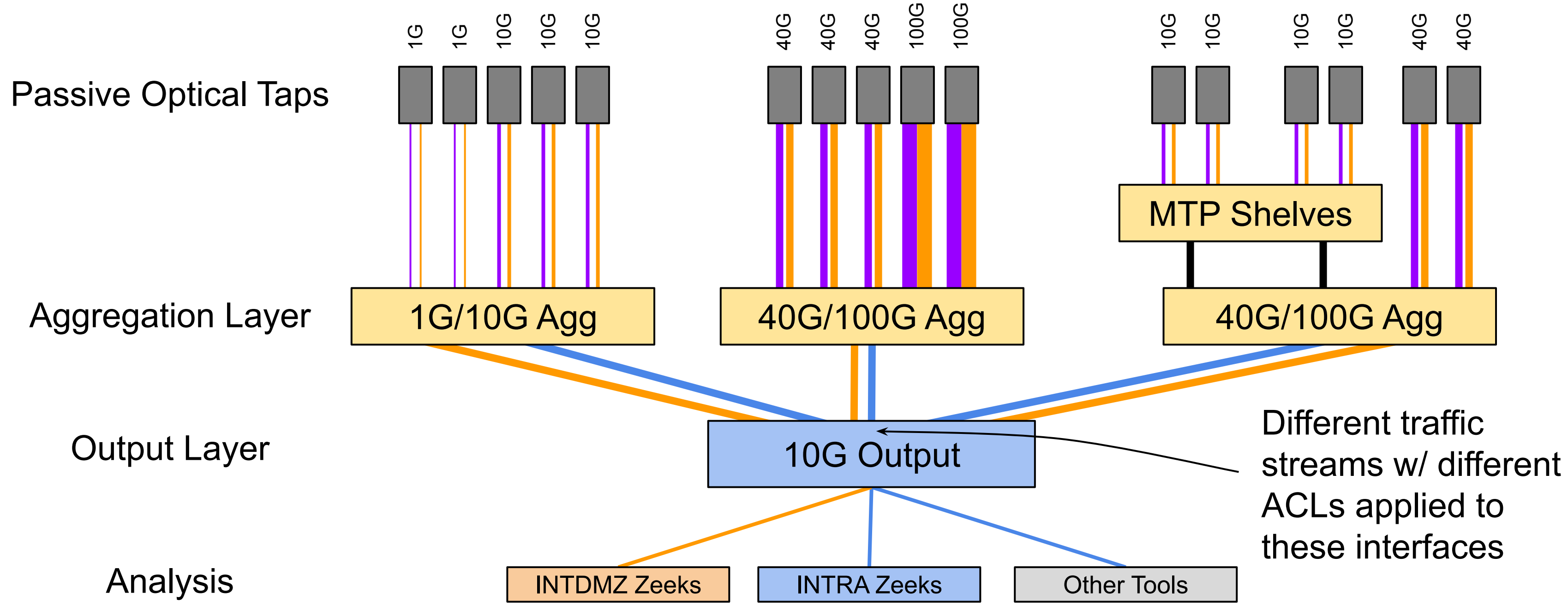
LBNL's Tap Agg (EXT1)



LBNL's Tap Agg (EXT2)



LBNL's Tap Agg (INTDMZ + INTRA)



Different traffic streams w/ different ACLs applied to these interfaces

TLP:WHITE

UNCLASSIFIED

Tap Agg Hardware

- Tap Agg Switches

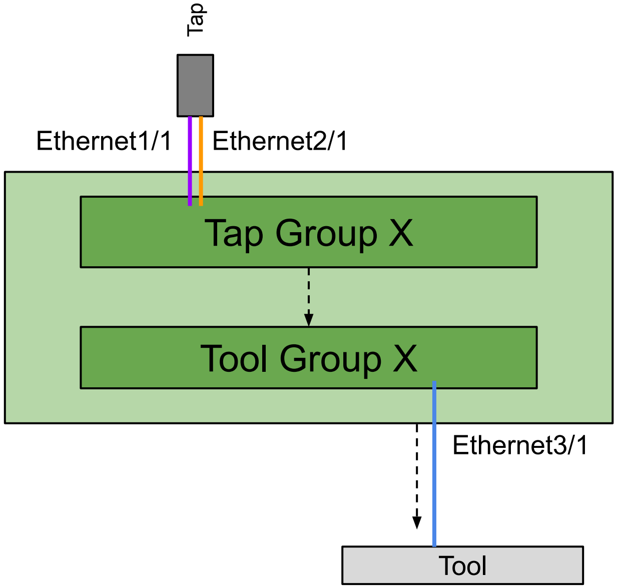
Arista 7280SR(A)-48C6	48x1G/10G + 6x40G/100G
Arista 7280SR3-48YC8	48x1G/10G/25G + 8x40G/100G
Arista 7280QR-C36	24x40G + 12x40G/100G
Arista 7280CR3-32D4	32x100G + 4x400G

  - Need "Tap Agg Mode" Licenses

- Zeek Node NICs
  - Myricom 2x10G SFP+ w/ Sniff License (10G-PCIE2-8C2-2S+SNF3)
  - Intel X710 2x10G SFP+ w/ AF_Packet

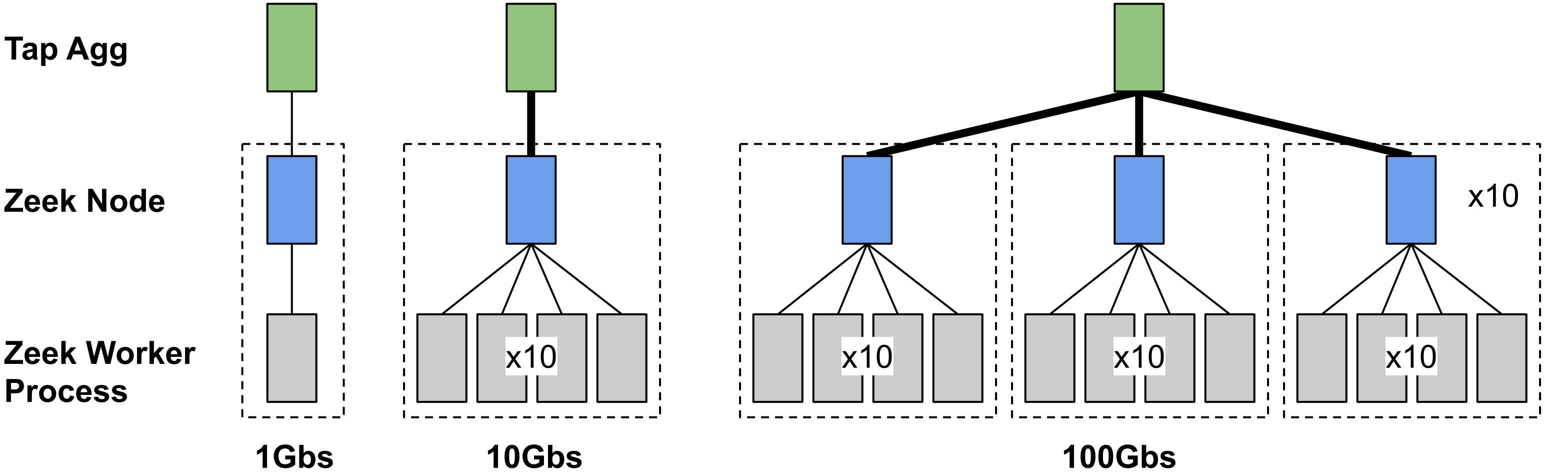
Minimum Arista Tap Agg Config

(No ACLs, no port channels to clusters)

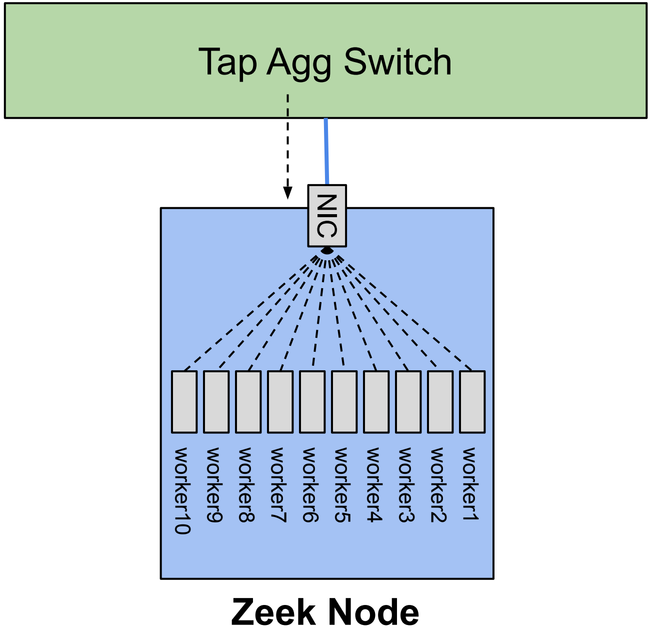


```
tap aggregation
   mode exclusive

interface Ethernet1/1
   description "TX Tap Input"
   switchport mode tap
   switchport tap default group X

interface Ethernet2/1
   description "RX Tap Input"
   switchport mode tap
   switchport tap default group X

interface Ethernet3/1
   description "Output to Tool"
   switchport mode tool
   switchport tool group set X
```

Scaling Up with Load Balancing

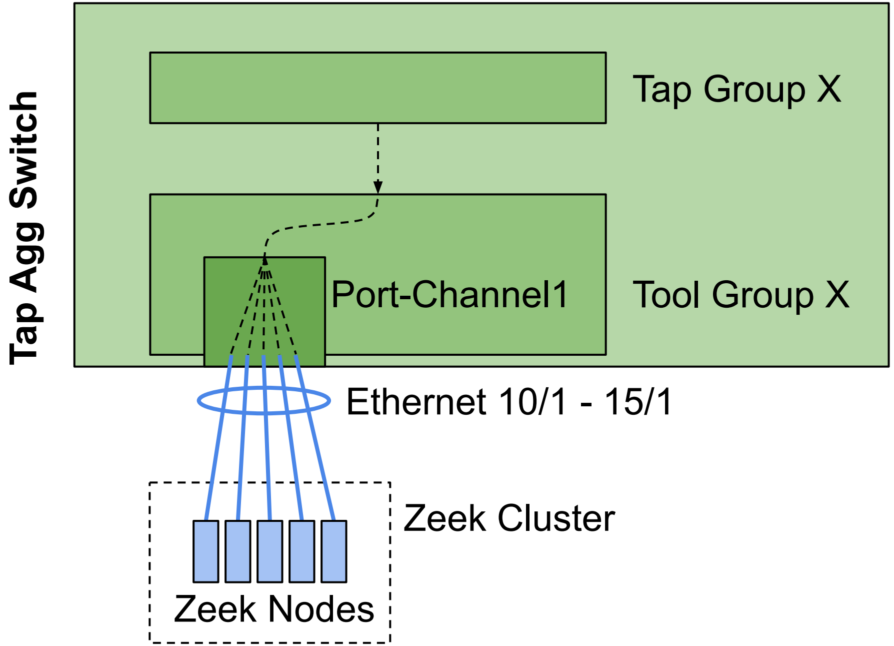


Distribute to Zeek Workers



```
# (zeekpath)/host/etc/node.cfg

# Myricom Sniffer Driver
lb_method=myricom
lb_procs=10
pin_cpus=3,5,7,9,11,13,15,17,19,21
env_vars=LD_LIBRARY_PATH=/usr/local/opt/snf/lib:/usr/local/
lib:$PATH,  SNF_DATARING_SIZE=0x80000000,
SNF_NUM_RINGS=10, SNF_FLAGS=0x1, SNF_APP_ID=1


# AF_Packet
lb_method=custom
lb_procs=10
pin_cpus=2,4,6,8,10,12,14,16,18,20
af_packet_fanout_id=23
af_packet_fanout_mode=AF_Packet::FANOUT_HASH
af_packet_buffer_size=128*1024*1024
```

Distribute to a Zeek Cluster



```
load-balance policies
  load-balance sand profile symmetric
    no fields mac
    fields ipv4 symmetric-ip
    fields ipv6 symmetric-ip
    fields l4 symmetric-ports
    no fields mpls
    fields symmetric-hash
    port-channel ip ip-tcp-udp-header

port-channel load-balance sand profile symmetric

interface Port-Channel1
  switchport mode tool
  switchport tool group set X

interface Ethernet 10/1 - 15/1
  channel-group 1 mode on
```

Static Traffic Filtering

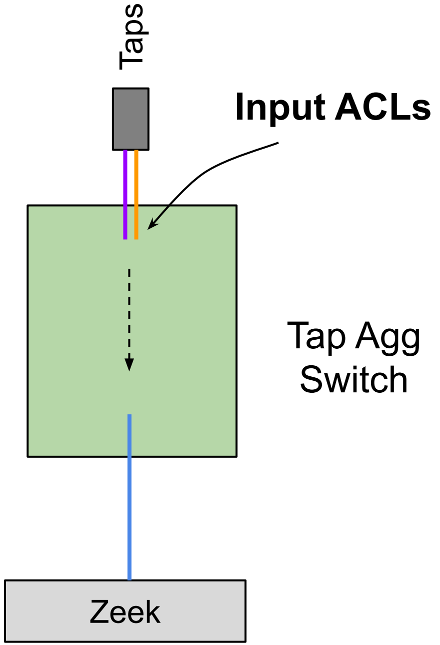
- **Why: filter out specific traffic from being analyzed**
  - Protect low capacity tools
  - Drop uninteresting traffic

- IP addresses, subnets/prefixes

- Ports/protocols

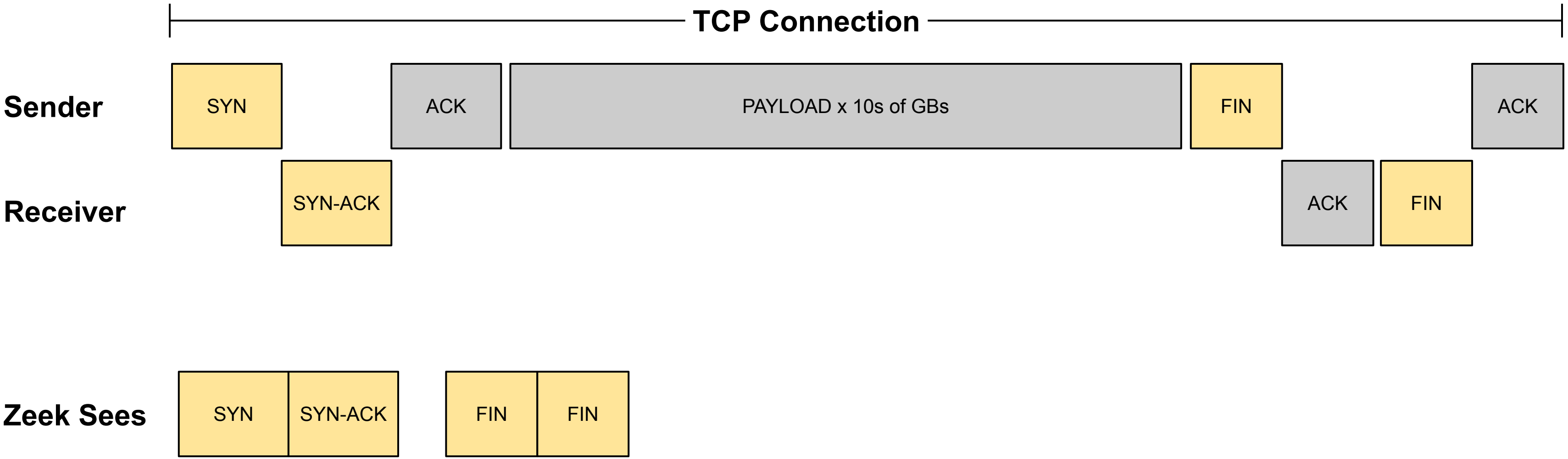
- Only "Control Packets"

- Packet Truncation



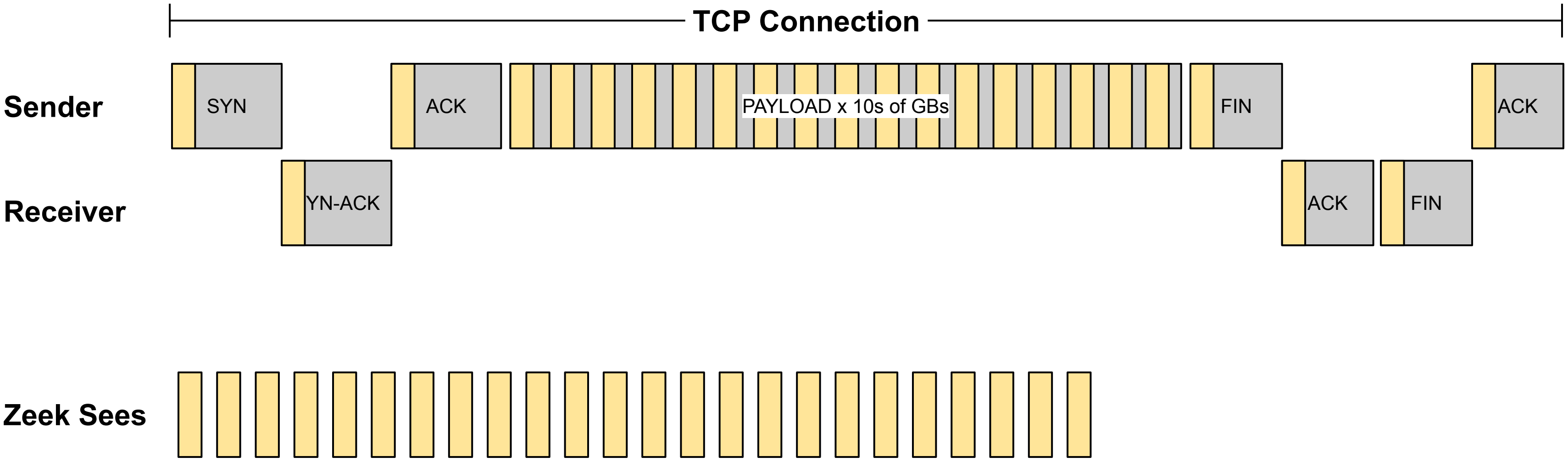
Static: Control Packets

- Permit TCP SYN/FIN/RST/FRAG, UDP + GRE + ICMP



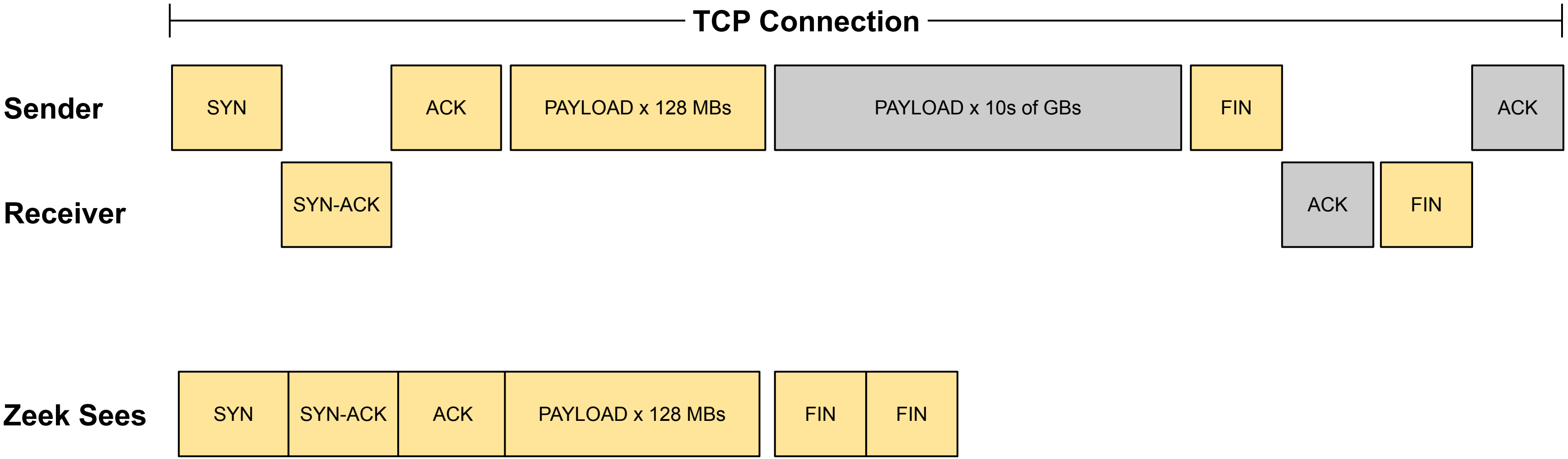
Static: Packet Truncation

- Each packet is truncated to X bytes



Dynamic Traffic Filtering

- Accept Control Packets + up to 128MBs total connection size



TCP Connection

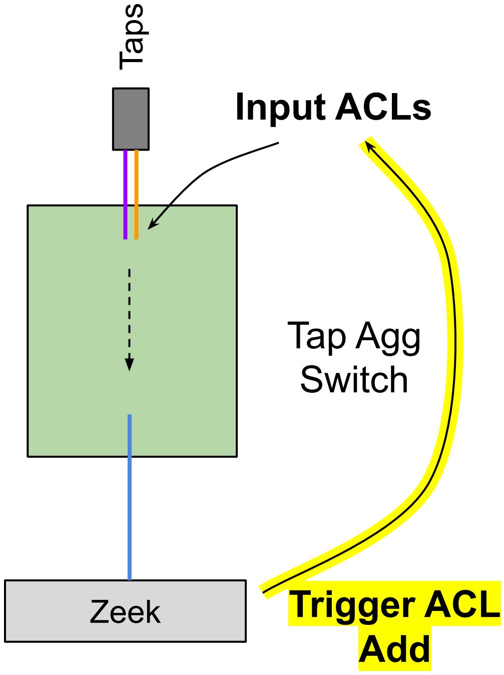
Sender: SYN | ACK | PAYLOAD x 128 MBs | PAYLOAD x 10s of GBs | FIN | ACK

Receiver: SYN-ACK | ACK | FIN

Zeek Sees: SYN | SYN-ACK | ACK | PAYLOAD x 128 MBs | FIN | FIN

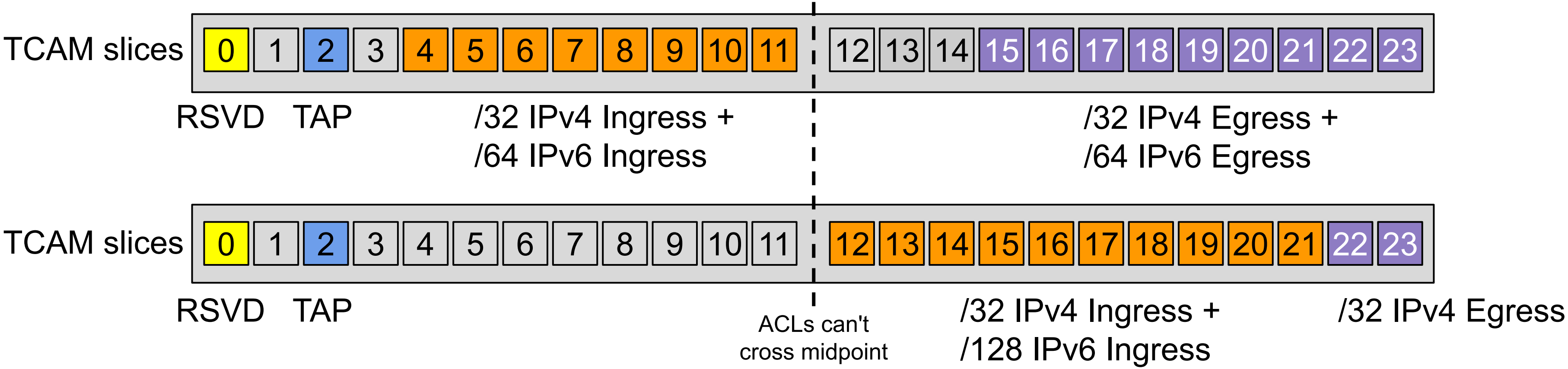
Dynamic ACLing

- Dynamically "shunt" big (elephant) flows' payloads
- Use lots of match criteria
  - Connection size (1 direction, both directions)
  - Packets (1 direction, both directions)
  - Protocol, Port Numbers
  - Country code?
- On match, trigger adding 5-tuple ACL
- conn-bulk.zeek -> dumbno.py -> API -> tap agg switch



TCAM Limitations

- Memory that (some) network devices like tap agg use
- May limit what kinds of ACLs you can do
- We ran into this on Arista 7150: `show platform fm6000 tcam usage`
  - Could do /32 Ingress + /128 Ingress + /32 Egress, but NOT /128 Egress
- Arista 7280s use Virtual Output Queues instead, haven't run into this there (but other limits)



TCAM slices 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23

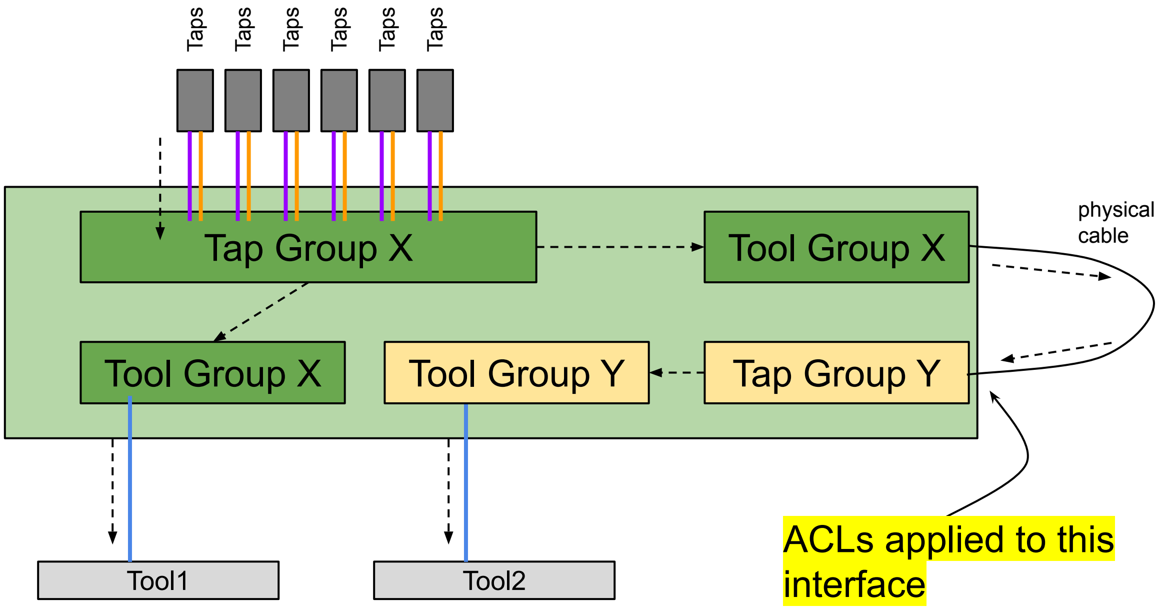
RSVD TAP /32 IPv4 Ingress + /64 IPv6 Ingress /32 IPv4 Egress + /64 IPv6 Egress

TCAM slices 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23

RSVD TAP ACLs can't cross midpoint /32 IPv4 Ingress + /128 IPv6 Ingress /32 IPv4 Egress

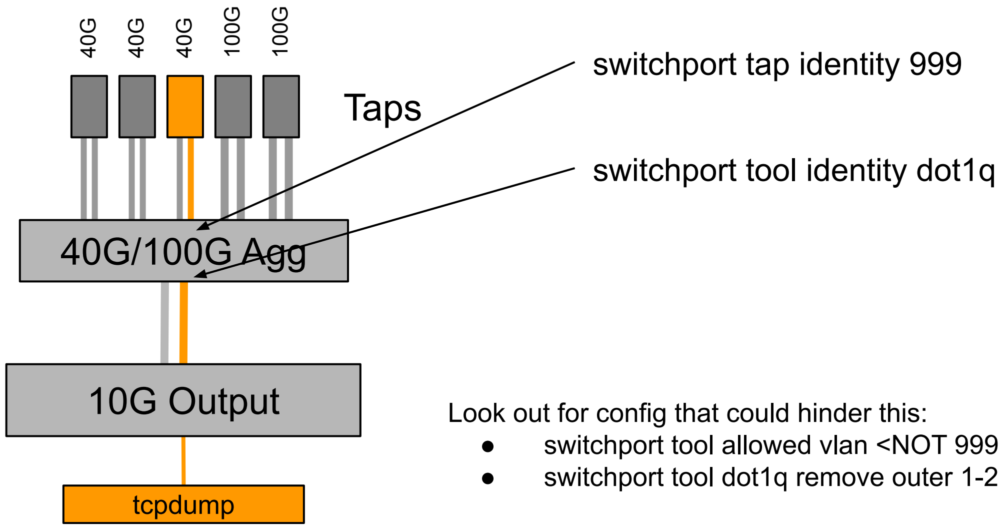
Ingress/Egress ACL Workaround

- Need to send **different** traffic to **different** tools
- Used to be able to do Egress ACLs, now spotty support
- Loop a cable and apply it as an Ingress ACL
- It burns 2x more ports
- It's hacky but works



Identity VLANs

- You need to be able to separate a specific link's traffic from everything else
- VLAN tag it at tap agg ingress
- Basically Q-in-Q (802.1ad)
- $ tcpdump -e -i en0 'vlan 999'



switchport tap identity 999

switchport tool identity dot1q

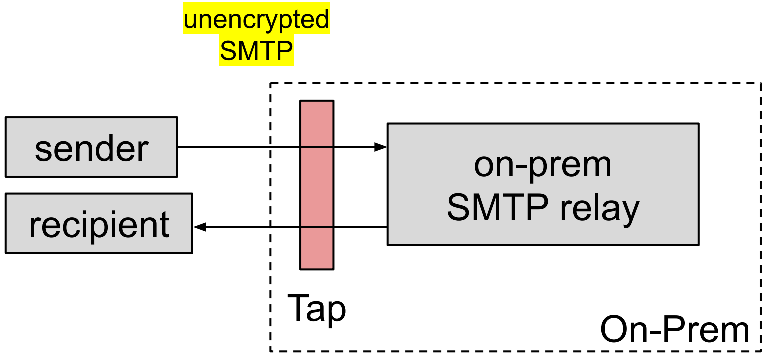
Look out for config that could hinder this:

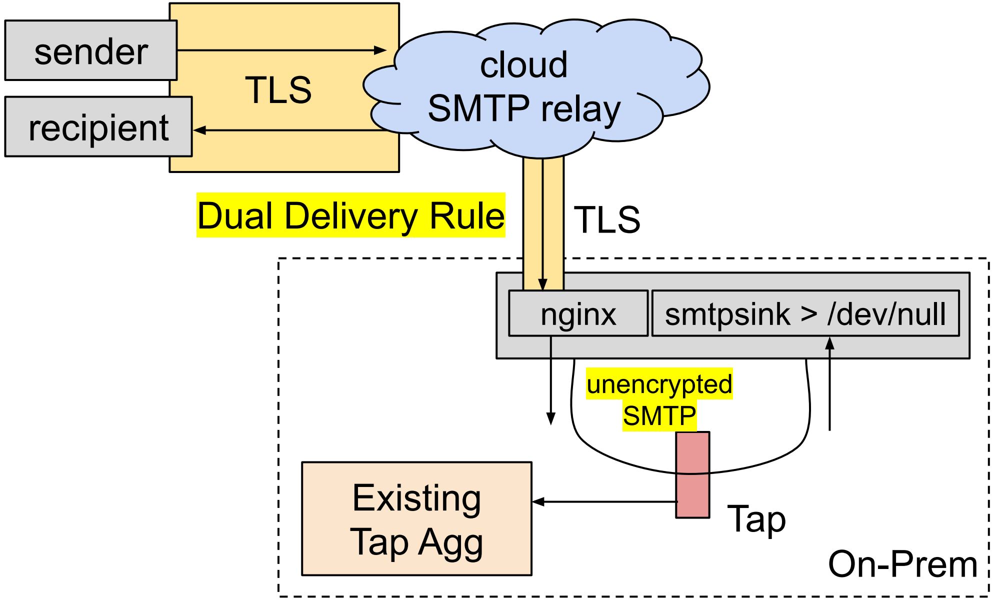
- switchport tool allowed vlan <NOT 999>
- switchport tool dot1q remove outer 1-2

Example packet capture:

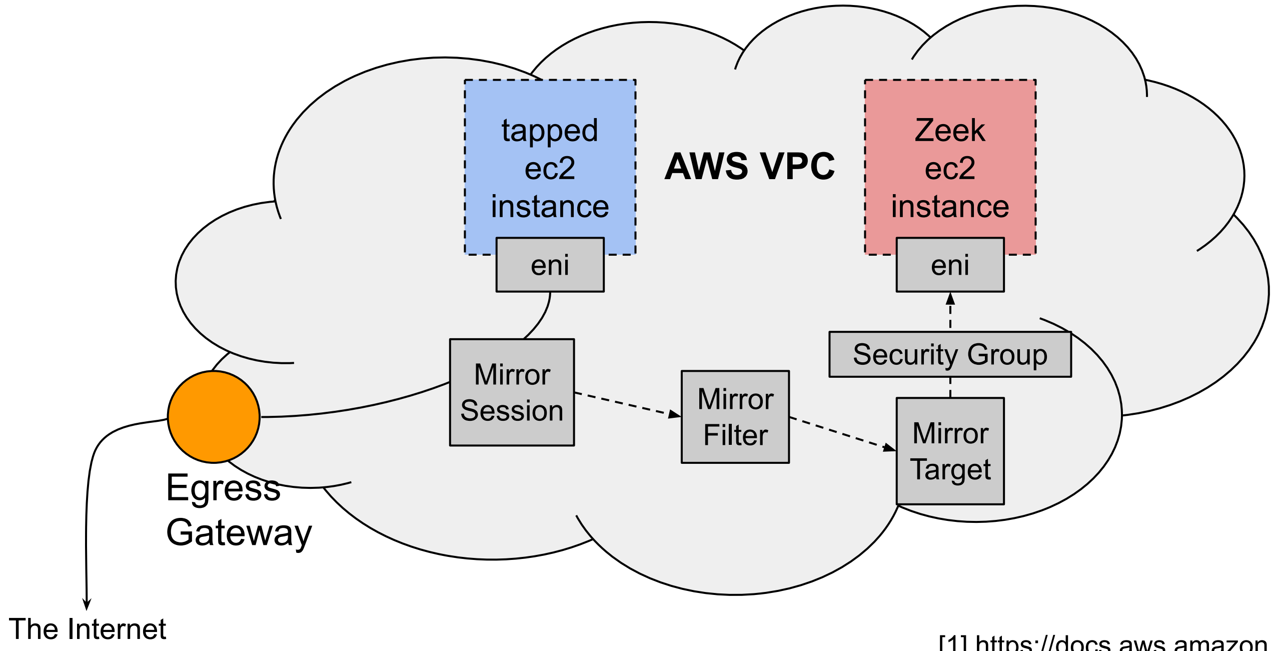
```
16:10:57.778824 00:53:00:e4:3d:3a > 00:53:ff:00:00:05, ethertype 802.1Q (0x8100), length 102: vlan 999, p 0,
ethertype 802.1Q, vlan 53, p 6, ethertype IPv4, vlan53.ir998.lbl.gov > ospf-all.mcast.net: OSPFv2, Hello, length 60
```

Tapping Email: Cloud+STARTTLS





Visibility in the Cloud



- A filter is required, even if it just allows-all
- You may need to create a Security Group to allow VXLAN
- You may want to disable checksum offloading
- There are limitations[1]

[1] https://docs.aws.amazon.com/vpc/latest/mirroring/traffic-mirroring-limits.html

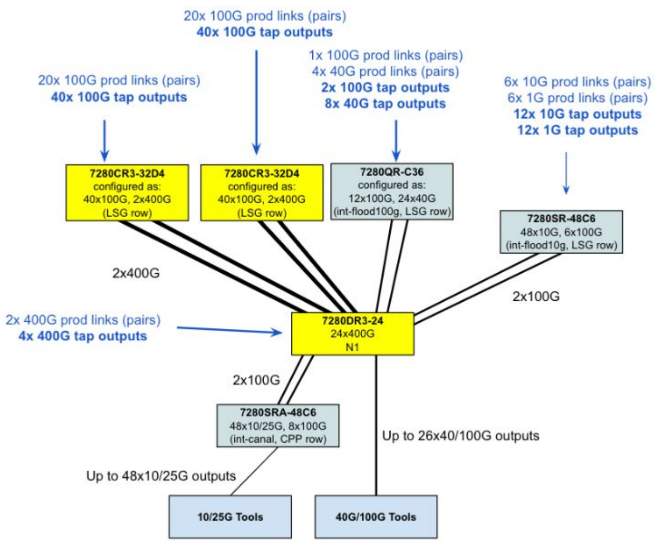
Tapping 400G Ethernet

- We don't have this online yet, but it's coming
- Arista 7280R3 line has 400G ports + LANZ
    - 7280CR3-32D4 = 32x100G + 4x400G
    - 7280DR3-24 = 24x 400G
- Unsure on feature parity… 100G took a while to catch-up to 10G
- Possible gotcha[1]:
    - QSFP-DD 400G-LR4 may work with 100G-rated taps but @ 400G
    - But some tap manufacturers may say to use QSFP-DD 400G-**P**LR4 and break it out to 4x100G taps

[1] Courtesy of Ryan Walker @ University of Illinois

Our 400G Tap Agg Plans

100G/400G INTDMZ/INTRA Tap Agg



20x 100G prod links (pairs)
40x 100G tap outputs

20x 100G prod links (pairs)
40x 100G tap outputs

1x 100G prod links (pairs)
4x 40G prod links (pairs)
2x 100G tap outputs
8x 40G tap outputs

6x 10G prod links (pairs)
6x 1G prod links (pairs)
12x 10G tap outputs
12x 1G tap outputs

7280CR3-32D4
configured as:
40x100G, 2x400G
(LSG row)

7280CR3-32D4
configured as:
40x100G, 2x400G
(LSG row)

7280QR-C36
configured as:
12x100G, 24x40G
(int-flood100g, LSG row)

7280SR-48C6
48x10G
(int-flood10g, LSG row)

2x400G

2x100G

2x 400G prod links (pairs)
4x 400G tap outputs

7280DR3-24
24x400G
N1

2x100G

7280SRA-48C6
48x10/25G, 8x100G
(int-canal, CPP row)

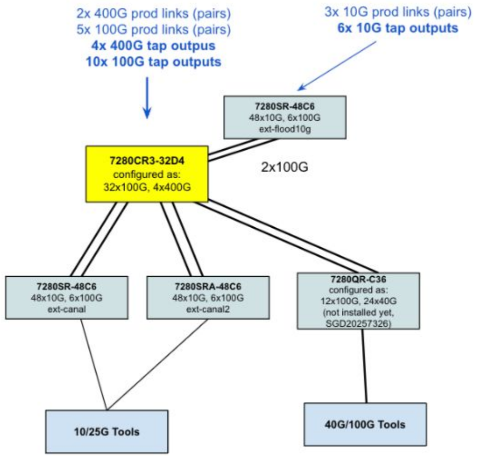
Up to 26x40/100G outputs

Up to 48x10/25G outputs

10/25G Tools

40G/100G Tools

100G/400G Node 1 EXTDMZ Tap Agg



2x 400G prod links (pairs)
5x 100G prod links (pairs)
4x 400G tap outpus
10x 100G tap outputs

3x 10G prod links (pairs)
6x 10G tap outputs

7280SR-48C6
48x10G, 6x100G
ext-flood10g

7280CR3-32D4
configured as:
32x100G, 4x400G

2x100G

7280SR-48C6
48x10G, 6x100G
ext-canal

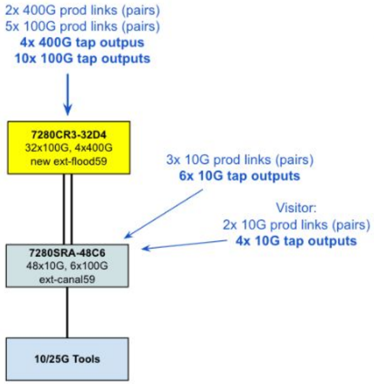
7280SRA-48C6
48x10G, 6x100G
ext-canal2

7280QR-C36
configured as:
12x100G, 24x40G
(not installed yet,
SGD20257326)

10/25G Tools

40G/100G Tools

100G/400G Bldg 59 EXTDMZ Tap Agg



2x 400G prod links (pairs)
5x 100G prod links (pairs)
4x 400G tap outpus
10x 100G tap outputs

7280CR3-32D4
32x100G, 4x400G
new ext-flood59

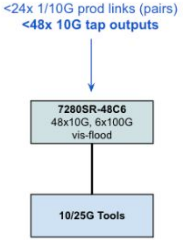
3x 10G prod links (pairs)
6x 10G tap outputs

Visitor:
2x 10G prod links (pairs)
4x 10G tap outputs

7280SRA-48C6
48x10G, 6x100G
ext-canal59

10/25G Tools

Node 1 Visitor Tap Agg



<24x 1/10G prod links (pairs)
<48x 10G tap outputs

7280SR-48C6
48x10G, 6x100G
vis-flood

10/25G Tools

400G-capable switches

See Appendix for more examples

Questions? Suggestions?

mnsmitasin@lbl.gov

security@lbl.gov

Appendix

Tap Install Checklist

❑ Check with policy / legal counsel

❑ Identify which specific link(s) you want to tap

❑ Note the link type: copper/fiber, Singlemode/Multimode, connector type, speed (1G/10G/40G/100G)

❑ Fiber: Check light levels, select appropriate ratio (80/20, 70/30, 50/50)

❑ Plan what will plug-in where

❑ Schedule a maintenance window (the link will go down)

❑ Disconnect, clean connectors, add new cable, add tap

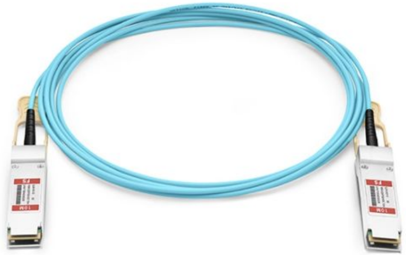
❑ Confirm link comes up, check light levels after

❑ Plumb the output to your tap agg or Zeek

Appendix

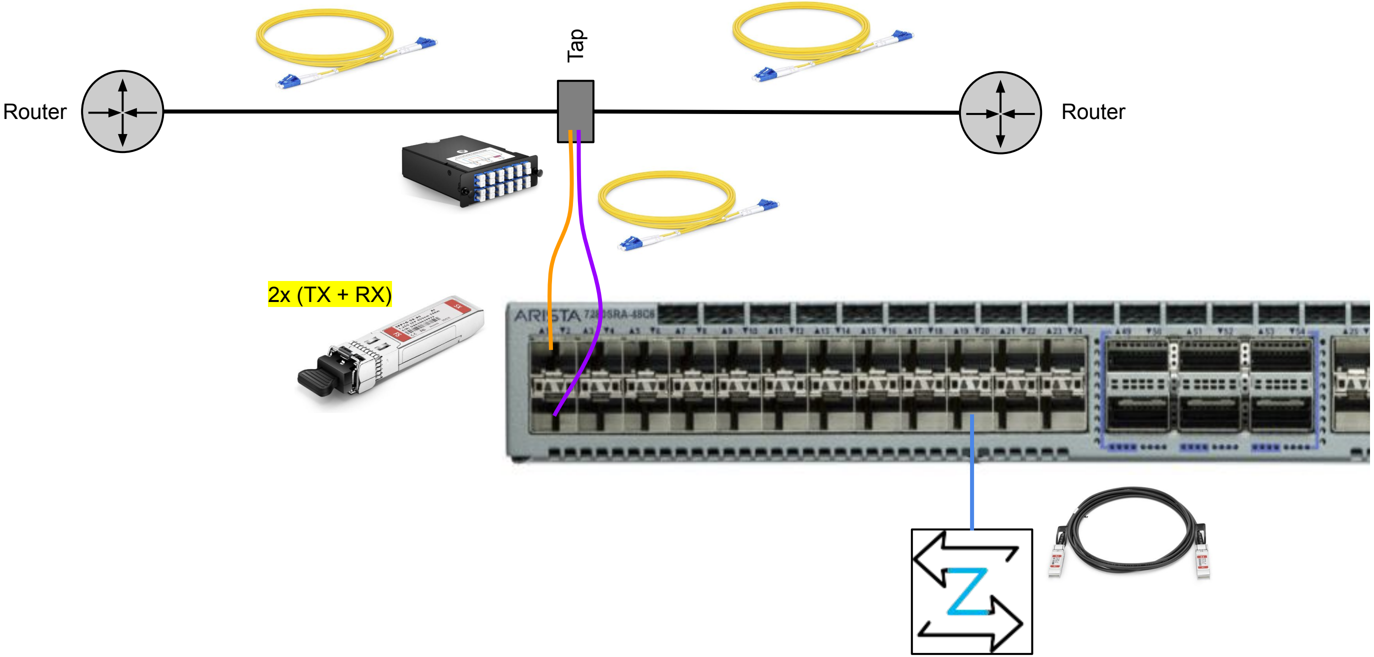
Transceivers

- Optical Modules
  - SFP = 1Gbps (most commonly, technically other speeds)
  - SFP+ = 10Gbps
  - SFP28 = 25Gbps
  - QSFP+ = 40Gbps
  - QSFP28 = 100Gbps
  - QSFP-DD = 400Gbps
- Cables
  - DAC = Direct Attached Copper
  - AOC = Active Optical Cable



Appendix

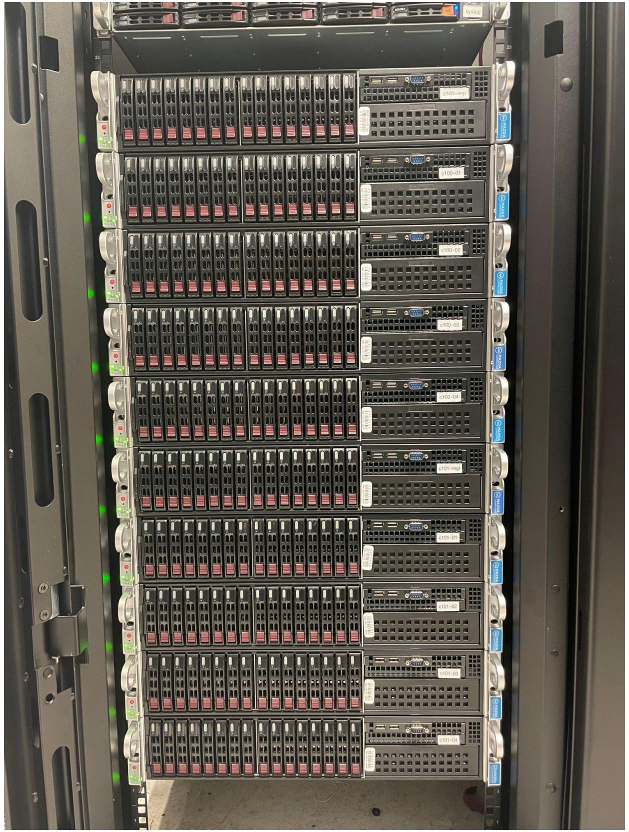
Hardware Example Install



Appendix

Zeek Cluster Hardware

- Zeek Cluster Nodes
  - (1x) Manager
    - Supermicro 2216RSJ2L-2T chassis
    - 2x Intel Xeon 6230, 20x cores @ 2.10GHz
    - 512GB (16x32GB) DDR4 RAM
    - 2x1TB NVMe (OS - Intel P4510), 4x3.8TB SSD (Data - Intel D3-S4610)
    - 1x Intel X710-DA2 10G NIC
  - (4-5x) Worker Nodes
    - Supermicro 2216RSJ2L-2T chassis
    - 2x Intel Xeon 6230, 20x cores @ 2.10GHz
    - 256GB (8x32GB) DDR4 RAM
    - 2x1TB NVMe (OS - Intel P4510)
    - 1x Intel X710-DA2 10G NIC



(this is 2 clusters)

Static ACLing

TCP control packets + similar

```
ip access-list <ACLNAME>
   counters per-entry
   10 permit tcp any any syn
   20 permit tcp any any fin
   30 permit tcp any any rst
   40 permit tcp any any fragments
   50 permit udp any any
   60 permit gre any any
   70 permit icmp any any
   [...]
   100 deny ip host <perfsonar> any
   110 deny ip any host <perfsonar>
   [...]
```

PerfSonar Nodes

xRootd data transfer nodes

```
200 deny tcp any 131.243.135.0/26 range 1090 1100
210 deny tcp 131.243.135.0/26 range 1090 1100 any
220 deny tcp any range 1090 1100 131.243.135.0/26
230 deny tcp 131.243.135.0/26 any range 1090 1100
240 deny tcp any 131.243.135.0/26 range 10900 10910
250 deny tcp 131.243.135.0/26 range 10900 10910 any
260 deny tcp any range 10900 10910 131.243.135.0/26
270 deny tcp 131.243.135.0/26 any range 10900 10910
[...]
1000 deny tcp any host <SMTPSINK> eq smtp
1010 deny tcp host <SMTPSINK> eq smtp any
[...]
500001 permit ip any any
```

Encrypted SMTP

Appendix

Dynamic ACLing :: ACL example

```
ip access-list bulk_1
   counters per-entry
   10 permit tcp any any fin
   20 permit tcp any any syn
   30 permit tcp any any rst
   40 permit tcp any any fragments
   50 permit udp any any
   60 permit gre any any
   70 permit icmp any any
   80 deny pim any any
```

Accept TCP control packets + similar

```
[...]
   36075 deny tcp host 192.0.2.32 eq ssh host 203.0.113.5 eq 44144
   44051 deny tcp host 203.0.113.150 eq 62218 host 192.0.2.15 eq 50935
   44053 deny tcp host 203.0.113.150 eq 62220 host 192.0.2.15 eq 50935
   44057 deny tcp host 203.0.113.150 eq 62221 host 192.0.2.15 eq 50935
   44059 deny tcp host 203.0.113.150 eq 62222 host 192.0.2.15 eq 50114
   44623 deny tcp host 192.0.2.32 eq 53526 host 203.0.113.104 eq https
   45255 deny tcp host 192.0.2.116 eq 53042 host 203.0.113.188 eq https
```

Big Shunted Payloads

```
[...]
   500001 permit ip any any
```

Appendix

Dynamic ACLing :: conn-bulk.zeek

```
export {
    const size_threshold = 134217728 &redef; #128 megabytes


if ((( c$orig$size > size_threshold || c$resp$size > size_threshold ) && c$orig$num_pkts > 100 && c$resp$num_pkts > 100))
        event Bulk::connection_detected(c);
        return -1sec;
 }
```

Connection cut-off in bytes

You could use other criteria here too:

- orig_pkts
- resp_pkts
- IPs
- ports/protocols
- country code
- If it's a Zeek field, you can probably use it

Dynamic ACLing :: dumbno.cfg

```
[switch]
ip = <Tap Agg mgmt IP>
user = <APIUSER>
password = <APIPASSWORD>

[ports]
Ethernet1 = <Dynamic ACL name applied to ingress Tap ports>

[egress_ports]
Ethernet2 = tool1
```

Input port(s) from taps

Output port(s) that goes to Zeek

Appendix

Dynamic ACLing :: T-Shooting logs

- Zeek :: conn_bulk.log

```
1663570498.392966     Coqv5l3qjHNZjqN1ag     192.0.2.70   44470   203.0.113.63  443     tcp     ssl
1.688105       625     445138831      SF      F       T       0       ShADdFafRR 14      1197    8       2687    -
worker-2-1      LK      US
```

- /var/log/dumbno/

```
@400000006323aaee121ea624 INFO:dumbno:op=ADD seq=32905 rule='tcp host 192.0.2.70 eq 44470
host 203.0.113.63 eq 443'
@400000006323aaf4267ad28c INFO:dumbno:op=REMOVE acl=bulk_1 family=ip seq=32905 rule='tcp
host 192.0.2.70 eq 44470 host 203.0.113.63 eq 443'
```

- /var/log/dumbno-stats/

```
@40000000632738e330d4639c INFO:dumbno_stats:mbps: in=3633 out=1852 filtered=1780
```