Multipath Traffic Engineering
TEAS WG, IETF 112, Bangkok

Kireeti Kompella
Juniper Networks

JUNIPER NETWORKS | Driven by Experience™

New(ish) Concepts

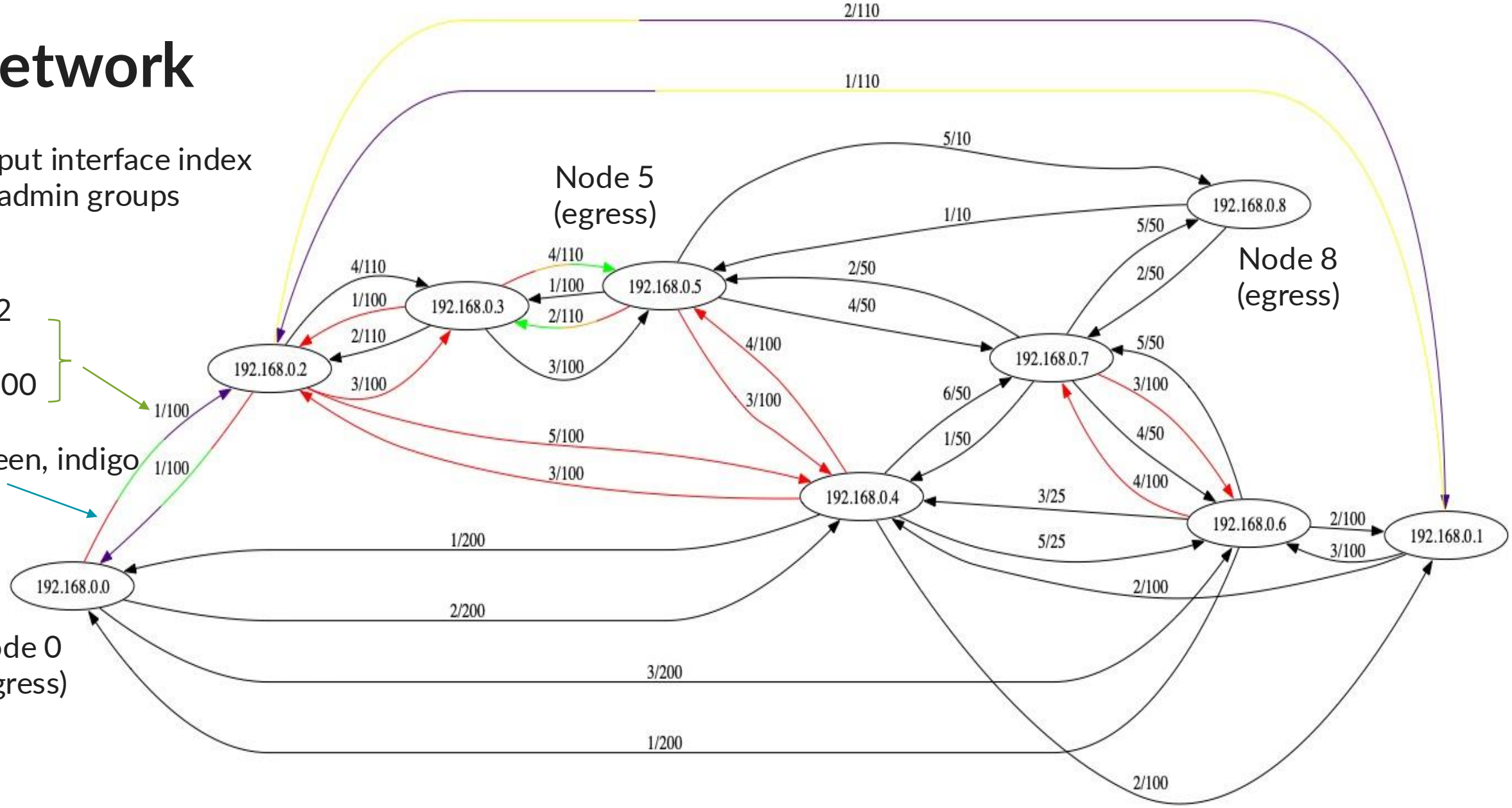
- TE with ECMP (multipath TE or MPTE) (TE++ on steroids)
- nECMP “with slack x”: non-ECMP, within x of shortest path
- Multi-egress (think multipath iBGP)
- Multi-ingress (needs more thinking)
- Unequally weighted load balancing

- MPLS *and* IP tunnels

Network

oif = output interface index
colors = admin groups

Link 0→2
oif: 1
metric: 100
colors:
red, green, indigo

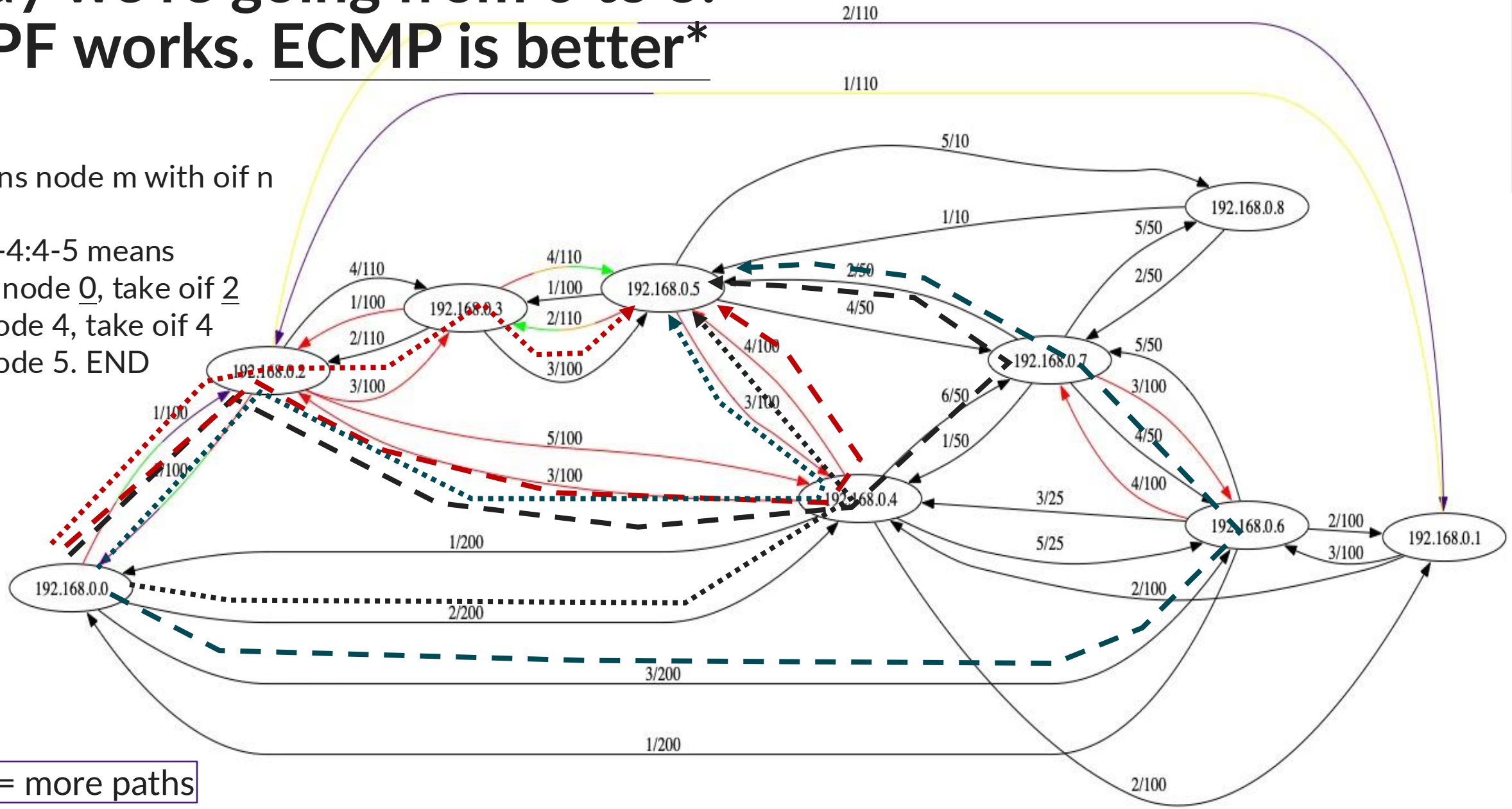


Say we're going from 0 to 5. SPF works. ECMP is better*

6 paths to 5 (dist 300-300)
0:2-4:4-5 – 300

m:n means node m with oif n

path 0:2-4:4-5 means
start at node 0, take oif 2
go to node 4, take oif 4
go to node 5. END



*better = more paths

What does link metric measure? What does path length measure?

- “fiber miles” – end-to-end distance
- “minimum delay” – end-to-end delay

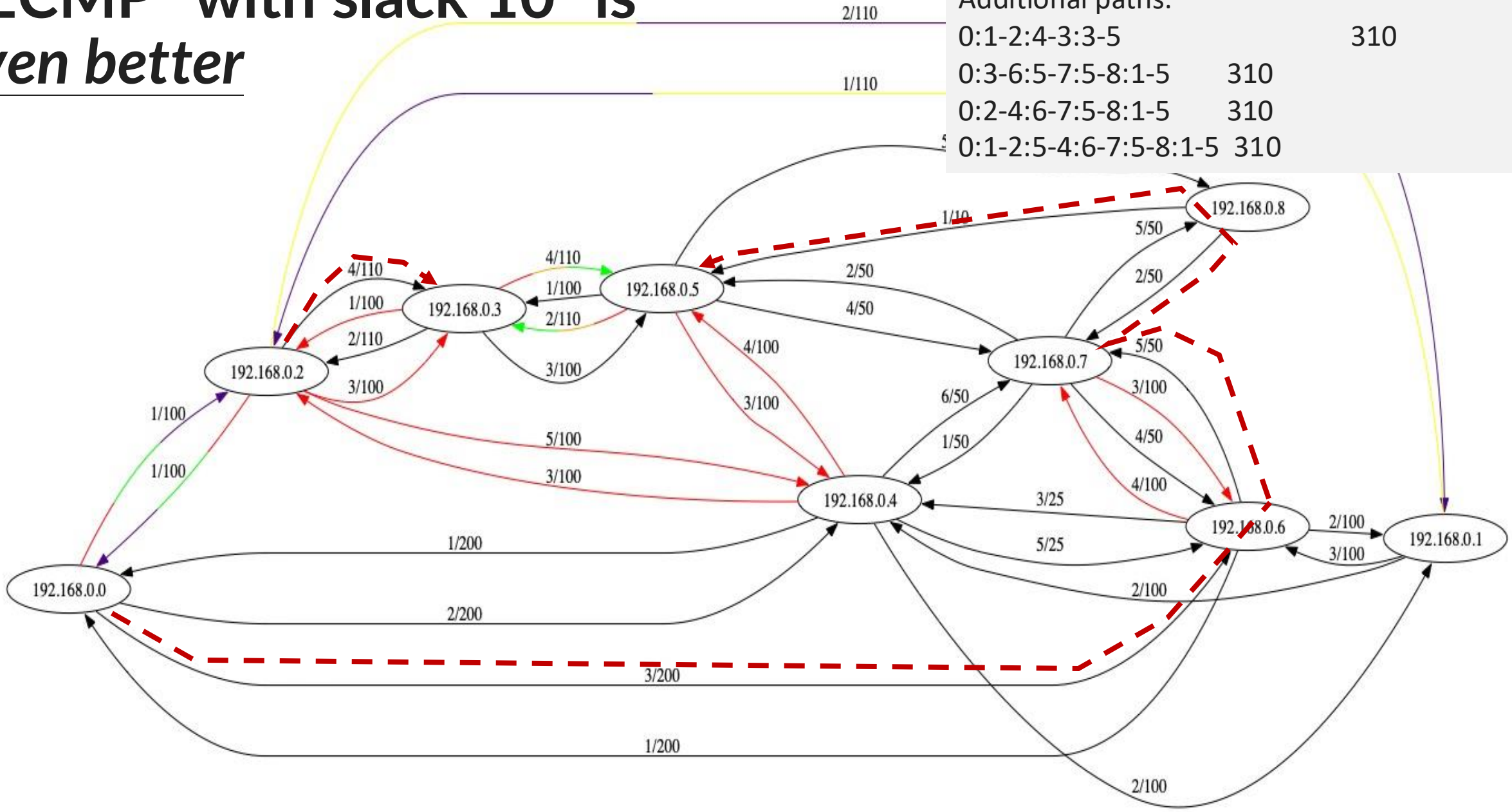
- In either case, does it matter if a path length is 300 or 310? Should paths that are a *bit* longer count as “almost” ECMP?

- nECMP “with slack n” → non-ECMP paths within n of shortest path
- Yields more usable paths to play with

nECMP “with slack 10” is *even better*

10 paths to 5 (dist 300-310)
Additional paths:

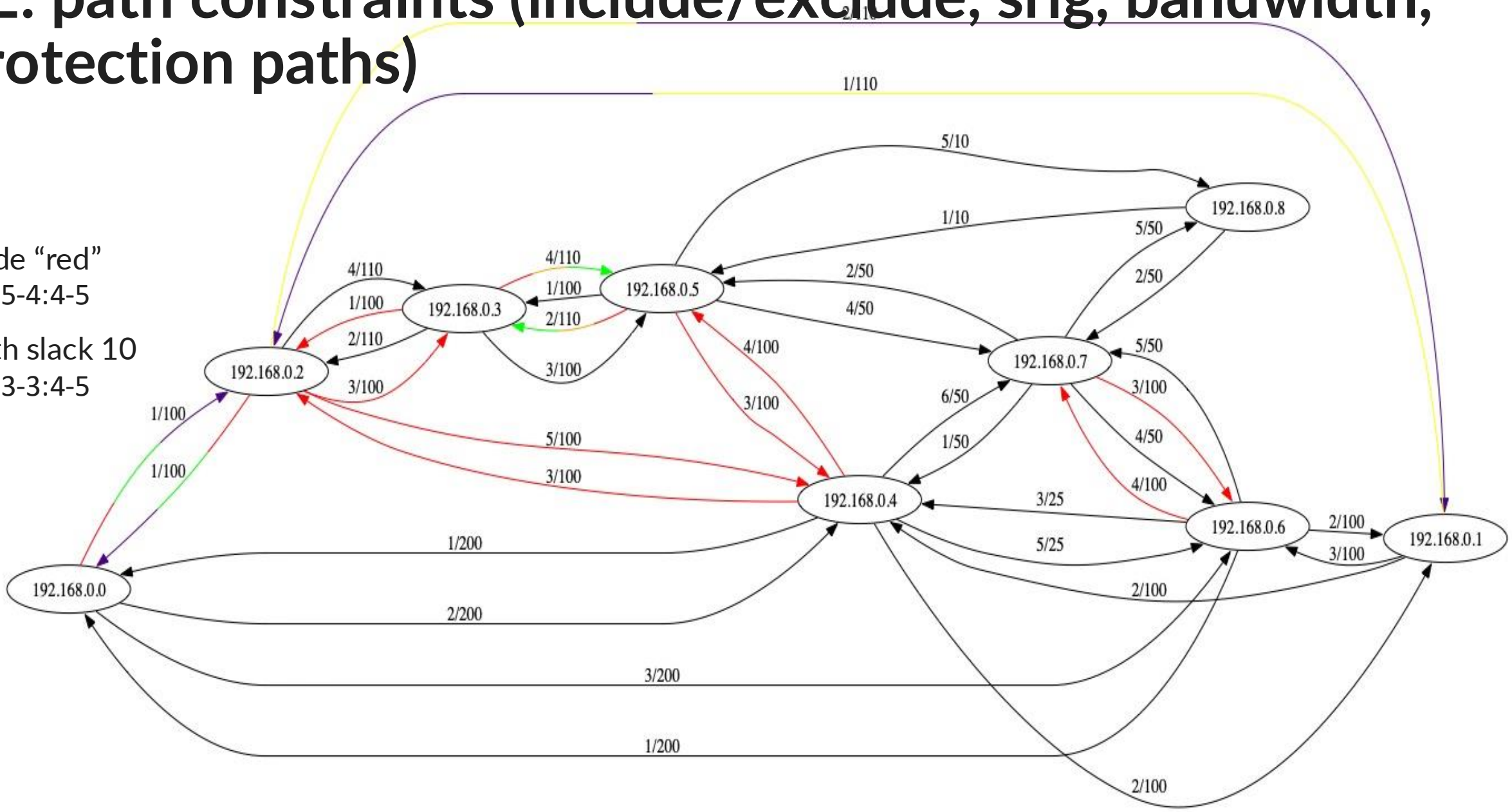
0:1-2:4-3:3-5	310
0:3-6:5-7:5-8:1-5	310
0:2-4:6-7:5-8:1-5	310
0:1-2:5-4:6-7:5-8:1-5	310



TE: path constraints (include/exclude, srlg, bandwidth, protection paths)

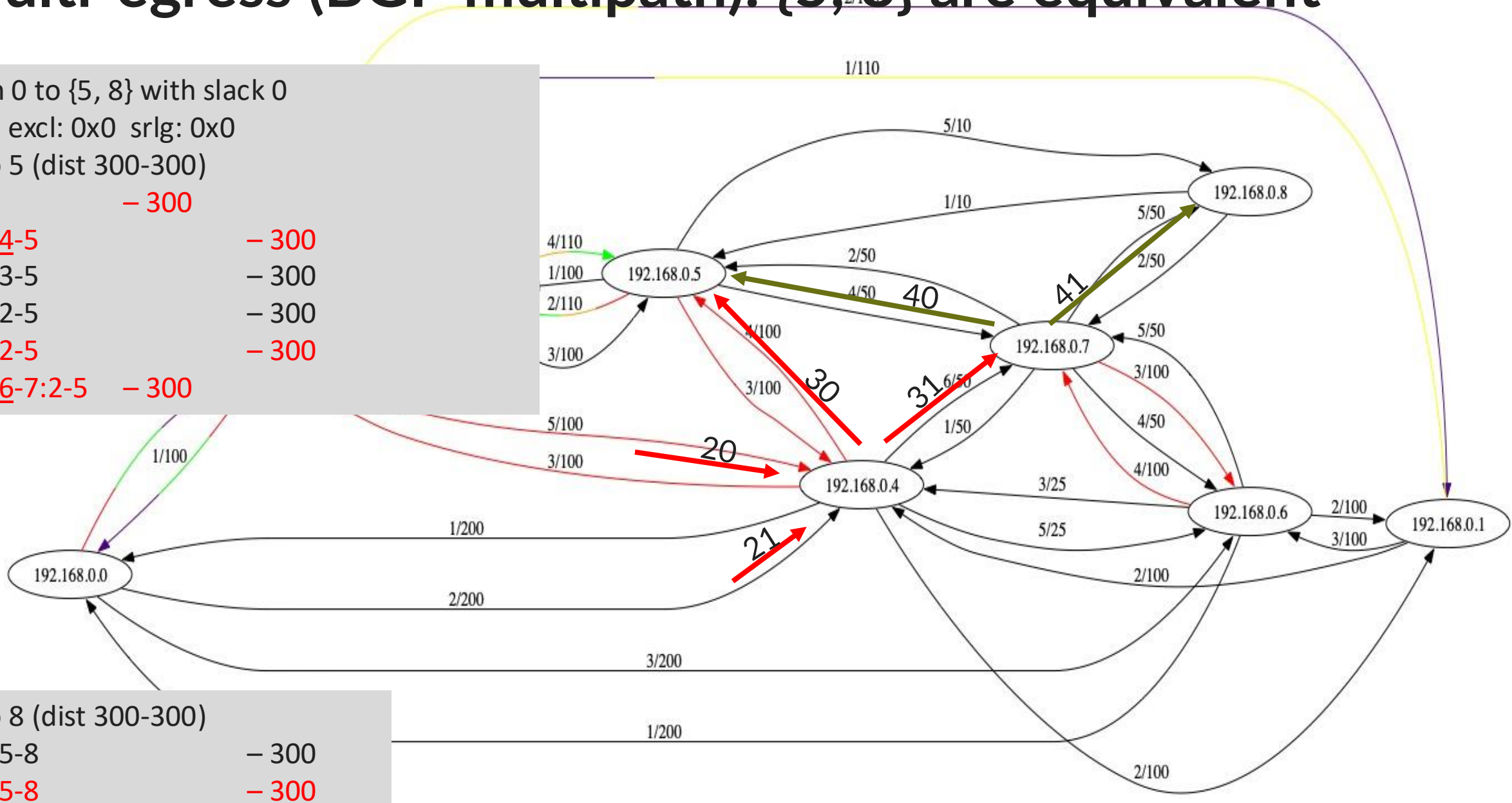
Include “red”
0:1-2:5-4:4-5

… with slack 10
0:1-2:3-3:4-5



Multi-egress (BGP multipath): {5, 8} are equivalent

CSPF from 0 to {5, 8} with slack 0
incl: 0x0 excl: 0x0 srlg: 0x0
6 paths to 5 (dist 300-300)
0:2-4:4-5 – 300
0:1-2:5-4:4-5 – 300
0:1-2:3-3:3-5 – 300
0:3-6:5-7:2-5 – 300
0:2-4:6-7:2-5 – 300
0:1-2:5-4:6-7:2-5 – 300



3 paths to 8 (dist 300-300)
0:3-6:5-7:5-8 – 300
0:2-4:6-7:5-8 – 300
0:1-2:5-4:6-7:5-8 – 300

Key Notion: MPTED

- Multipath TE DAG
  - Result of a CSPF with nECMP from ingress to 1 or more egresses

- An MPTED is broken up into junctions
  - Ingress: 0 phops, 1 or more nhops
  - Transit: 1 or more phops, 1 or more nhops
  - Egress: 1 or more phops, 0 nhops
- phop: list of <pnode, oif> [oif as assigned by prev node]
- nhop: list of <oif, nnode> [oif as assigned by this node]

Key Notion: Junction

- A node [*in this case, node 4*] with
  - phops (0:2, 2:5) (m:n means node m, oif n)
  - nhops (oif4→5;r1, oif6→5;r2) (n→p means oif n going to node p)

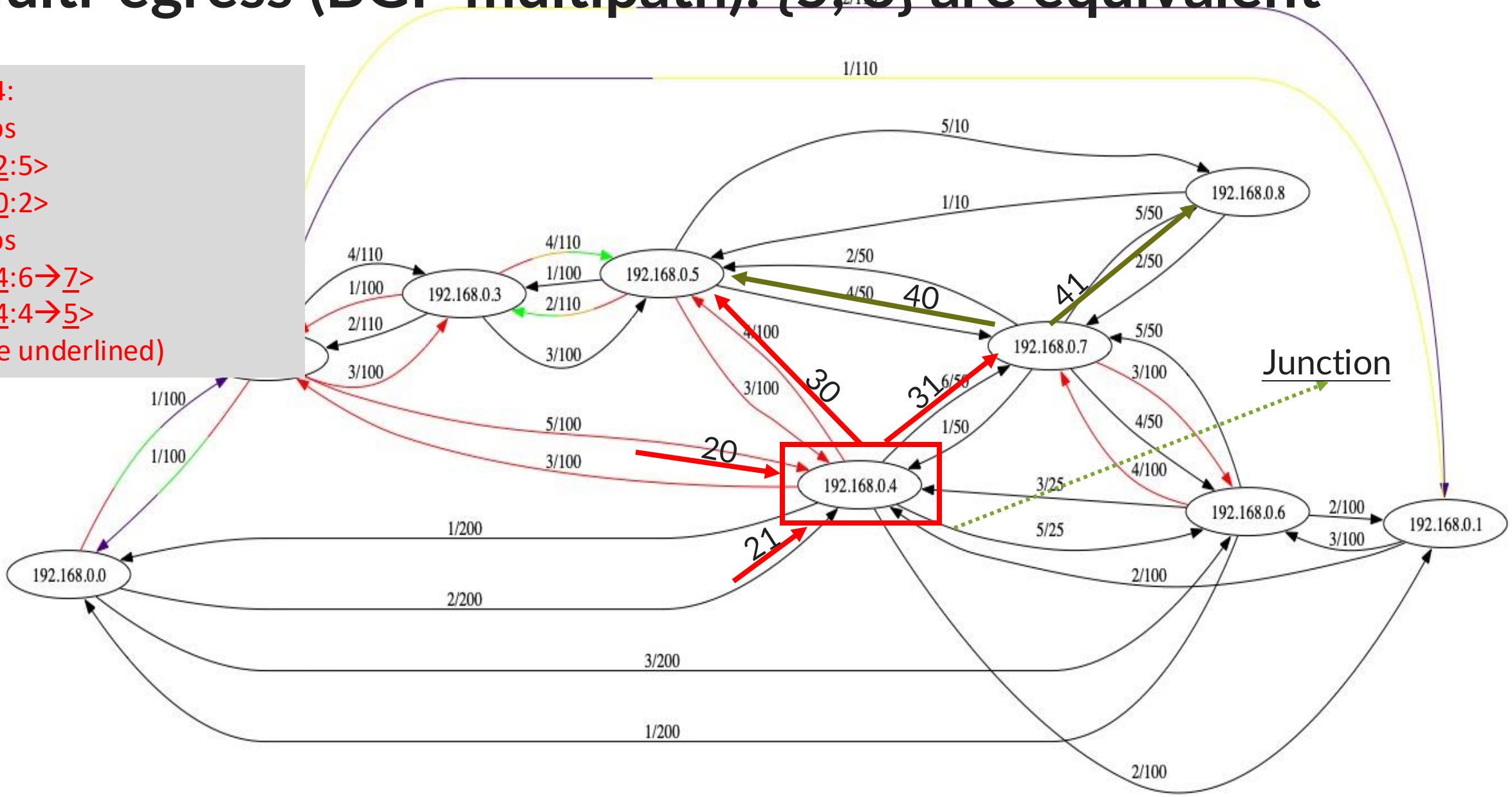
  - For nhops, r1, r2, … are ratios in which the traffic is to be divided
    - oif4 gets r1/(r1+r2+…) of the traffic
    - oif6 gets r2/(r1+r2+…) of the traffic
    - (potentially unequal load balancing)

Note that the overall tunnel bandwidth may or may not be specified

Multi-egress (BGP multipath): {5, 8} are equivalent

Junction 4:
- 2 phops
  - <2:5>
  - <0:2>
- 2 nhops
  - <4:6→7>
  - <4:4→5>

(nodes are underlined)



Signaling (1): “BGP-like” i.e., over TCP; a handful of messages*

- OPEN
- HELLO (keepalive)
- JUNCTION
- LABEL
- NOTIFICATION (errors)

- Full mesh of TCP connections among all nodes in the network
  - Notion of “MPTEP reflector” to help scale this
  - Alternatively, PCE connects to all nodes; nodes connect to IGP nbrs

* Many thanks to Haas and Scudder for gyaan on BGP

Signaling (2)

JUNCTION message:

- MPTED ID with version
- ingress; egresses
- tunnel type
  - IP-in-IP; GRE w/o key; GRE w/ key; GRE6; GRE-in-UDP; MPLS-in-UDP; …
  - signaled MPLS; static MPLS (both potentially with entropy label)
  - For IP tunnels, contains source/dest; for static MPLS, contains label
- #phops
  - List of phops
- #nhops
  - List of nhops
- flags

underlined IP tunnels support entropy

Signaling (3)

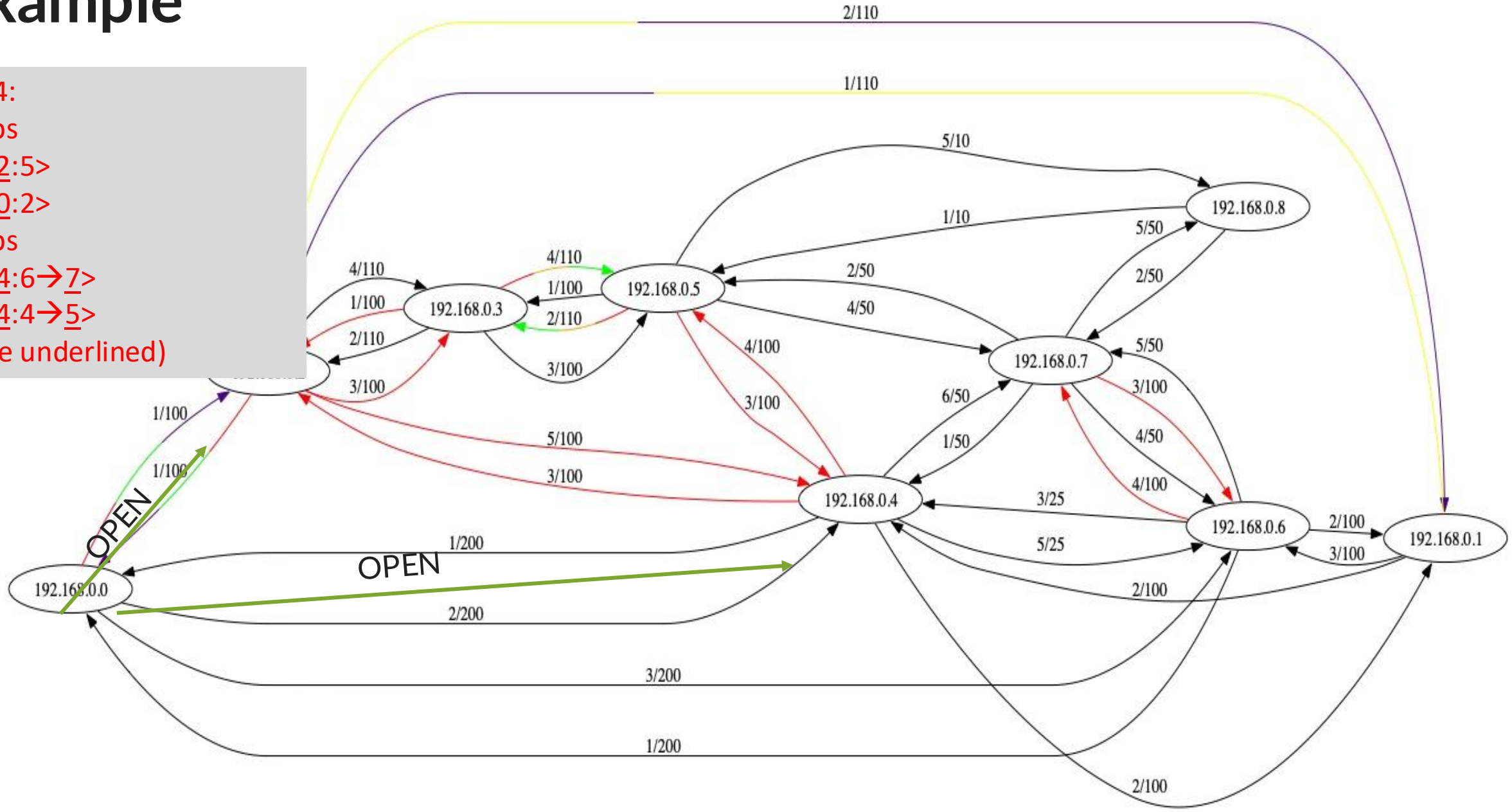
- Ingress sends JUNCTION message directly to all junction nodes
  - No hop-by-hop signaling
- For most tunnel types, junction node can immediately install FIB state; tunnel is “ready” immediately (modulo FIB install time)
- For “signaled MPLS”, LABEL message goes hop-by-hop from egress(es) to ingress [ordered control/RSVP-TE RESV]; tunnel is “ready” when ingress receives LABEL (modulo FIB install time)
  - Only message that goes hop-by-hop

Example

Junction 4:

- 2 phops
  - <2:5>
  - <0:2>
- 2 nhops
  - <4:6→7>
  - <4:4→5>

(nodes are underlined)



Questions

- Is multi-egress (e.g., BGP load balancing) useful?
  - If so, can egress1 be used as a transit to egress2?
- Would a new protocol be acceptable?
  - Alternate: new AFI/SAFI in BGP – “head-of-line” blocking
- Thoughts about “slack”? detours?

Next Steps

- More details on Graceful Restart, FIB updates, IP tunnels
- Yang model for MPTE
- More noodling on signaling protocol
- Explore PCEP-based configuration of MPTE