New Techniques to Route in Folded-Clos Topology Data Center Networks

Peter Willis, Dr. –Ing. Nirmala Shenoy,
Profs. Yin Pan and Bill Stackpole,
Golisano College of Computing and Information Sciences,

Prof John Hamilton,
School of Mathematics and Statistics

Rochester Institute of Technology, Rochester, New York

Datacenters and Datacenter Networks

- Datacenters comprise of thousands of servers per location - extend to multiple locations
- Data Center Network (DCN) is the backbone that connects servers in a datacenter
- Over the years - Increase in load and performance demands
  - Research on DCN architectures and topologies continue
- Growing energy and carbon footprint concerns,
  - Costs – OPEX and CAPEX from RFC 7938
- High maintenance and troubleshooting efforts / costs
- Several other challenges …
- Research Focus– Can we simplify the protocols and hence router operations in a DCN?
  - And perhaps address several of the challenges

Networks Today

- Follow well-defined architectures and topologies
- Data Center Networks (DCN) –
  - use folded-Clos, VL2, Dcell, Bcube etc.
  - Symmetrical, high redundancy, very structured topologies
- Can we leverage this structure to simplify routing ++

For Our Investigations

- We adopted the popular DCN folded-Clos topology
- And the commonly used protocol (suite) in folded-Clos topology
  - Border Gateway Protocol (BGP), for routing
  - Equal Cost Multipath Routing (ECMP) for load balancing,
  - Bidirectional Forwarding Detection (BFD) to speed up failure detection.

Our Research

- We designed a new protocol to route and forward in a folded–Clos topology DCN
  - a clean slate approach
- Initial study focus – performance of proposed vs current protocol suite to validate the new protocol :-
  - Coded the proposed protocol and compared with BGP from FRRouting (frrouting.org)
  - Deployed the protocols on Fabric test bed (portal.fabric_testbed.net)
  - Assessed multiple performance metrics on an interface failure – multiple points
  - Compared configuration needs, routing tables

Protocols for Datacenter Networks (DCN)

- Several protocol suites have been investigated for folded-Clos topology DCNs.
- A popular protocol suite used in folded-Clos topology DCN
  - Border Gateway Protocol (BGP) – for routing
  - Equal Cost Multipath protocol (ECMP) for load balancing
  - Bidirectional Forwarding Detection (BFD) to speed up failure detection
  - BGP requires Transport Control Protocol (TCP) for its operation and
  - BFD requires User Datagram Protocol (UDP) for its operation
  - We also need Internet Protocol and Address Resolution Protocol!
- Seven different protocols -> Increased number of protocols -> increased operational complexity -> increased configurational, troubleshooting and management needs.
- Increased energy, cooling and equipment cost

Routing in a DCN – A New Approach

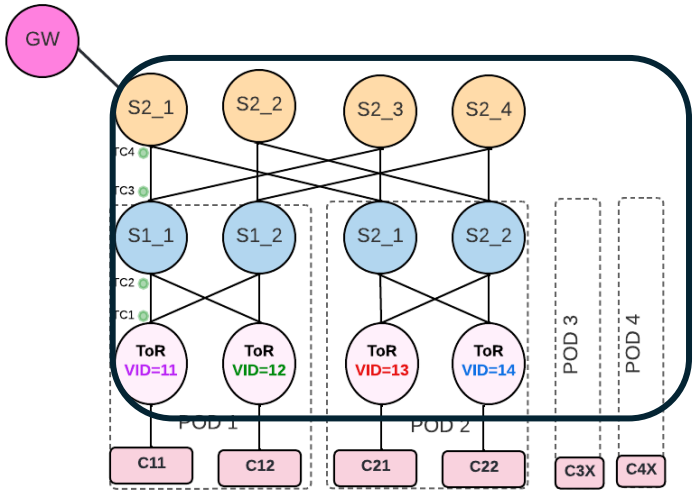
- WHAT IS NEEDED? - To route traffic between servers in a datacenter, the routers require information about the server networks (IP addresses) and how to reach them

Proposed Multi-Root Meshed Tree Protocol (MR-MTP)

- Used the structure in the Folded-Clos topology to simplify route establishment without using routing protocols, IP address, AS number etc.
- Establishes all loop-free paths from ToRs (top of rack) switches to all top tier spines
- Multiple Trees start (rooted) at ToRs - and mesh at the upper tier spines (no loops)

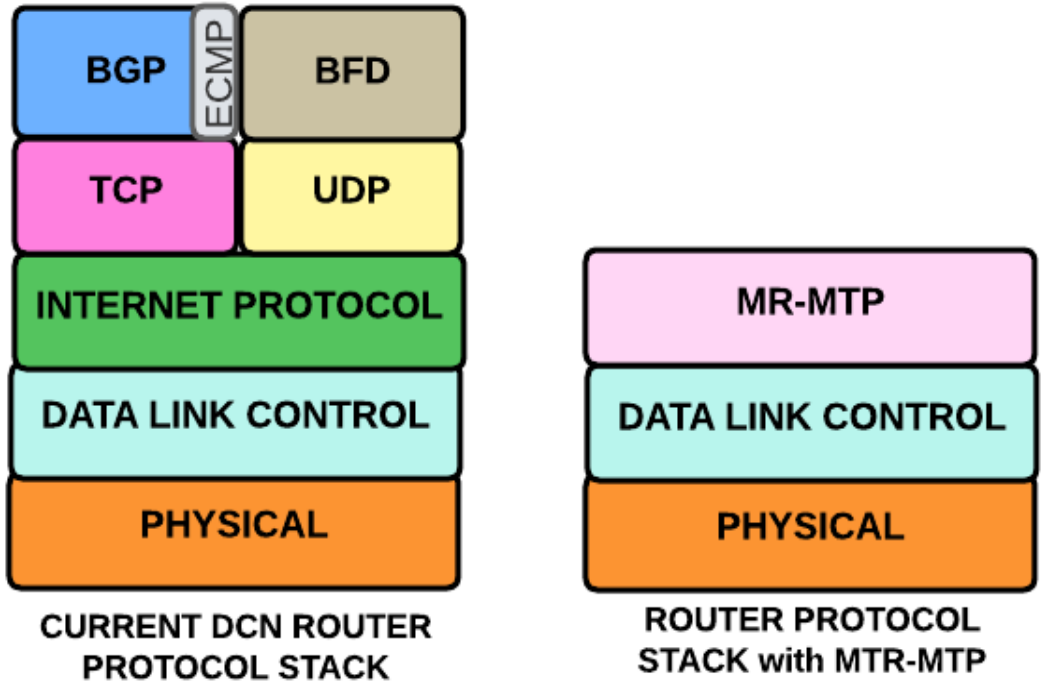
The Multi-Root Meshed Tree Protocol (MR-MTP)

- MR-MTP establishes all routes from ToRs to top tier spines using Virtual IDs (VIDs)
  - VIDs are auto assigned by MR-MTP
- MR-MTP unifies routing, load balancing and fast failure detection in a single protocol
- MR-MTP encapsulates and forwards IP packets between servers.
- MR-MTP is independent of Layer 3 i.e. it is Layer 3 agnostic
- MR-MTP defines its own headers (introduced later)
- MR-MTP is backward compatible to Ethernet (MR-MTP messages are carried in Ethernet frames) and IP (forwards IP packets between servers)
- No changes to the servers
- Communicates with an IP/BGP gateway



MR-MTP Protocol Stack

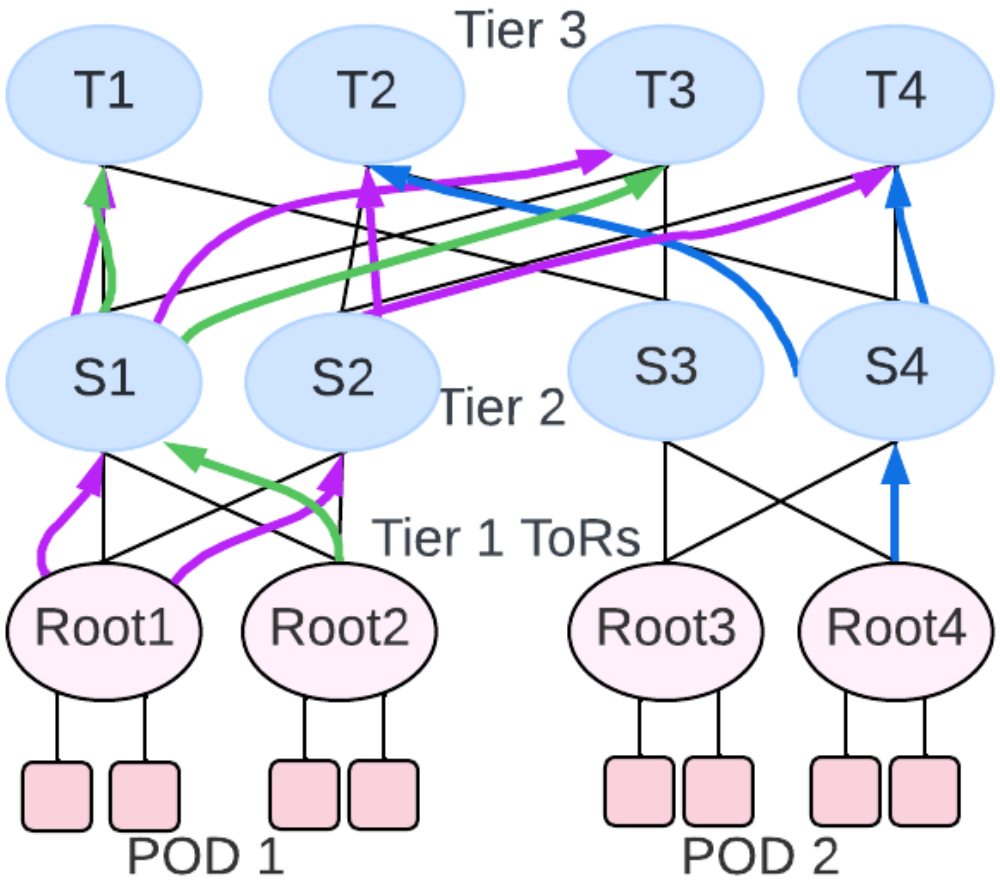
- In its current version MR-MTP replaces BGP, ECMP, BFD and IP. It also avoids the need for TCP (required by BGP) and UDP (required by BFD)
- The protocol stack at the router is cut down significantly
- The benefits ->
  - Reduced operational complexity
  - Reduced configuration needs
  - Reduced troubleshooting
  - Reduced management
  - Reduced energy, cooling and equipment cost



MR-MTP Features

- MR-MTP routers require tier information be configured
  - ToRs at tier 1, spines at tier 2, 3 etc
  - ToRs need a Virtual ID – currently auto derived from server subnet
- MR-MTP (C code) executable code size currently is 40 Kbytes.
  - https://github.com/pjw7904/CMTP
  - FABRIC testbed scripts - https://github.com/pjw7904/FABRIC-Automation
- MR-MTP can be turned off to fall back to current protocols
  - This will help in incremental deployment
- MR-MTP in one location can communicate with BGP in another location

Meshed Trees in a Folded-Clos topology – The Concept



Picture shows meshed trees constructed by protocol

- ToRs are roots of the meshed trees
- Note the purple tree from Root1
- A partial green tree from Root2

  --- so on
- A partial blue tree from Root4
- All trees mesh at all upper tier spines.
- Each tree from a ToR spans to all top tier spines.
- The meshed trees cover all loop-free paths from each ToR to every top tier spine.
- How to implement this?

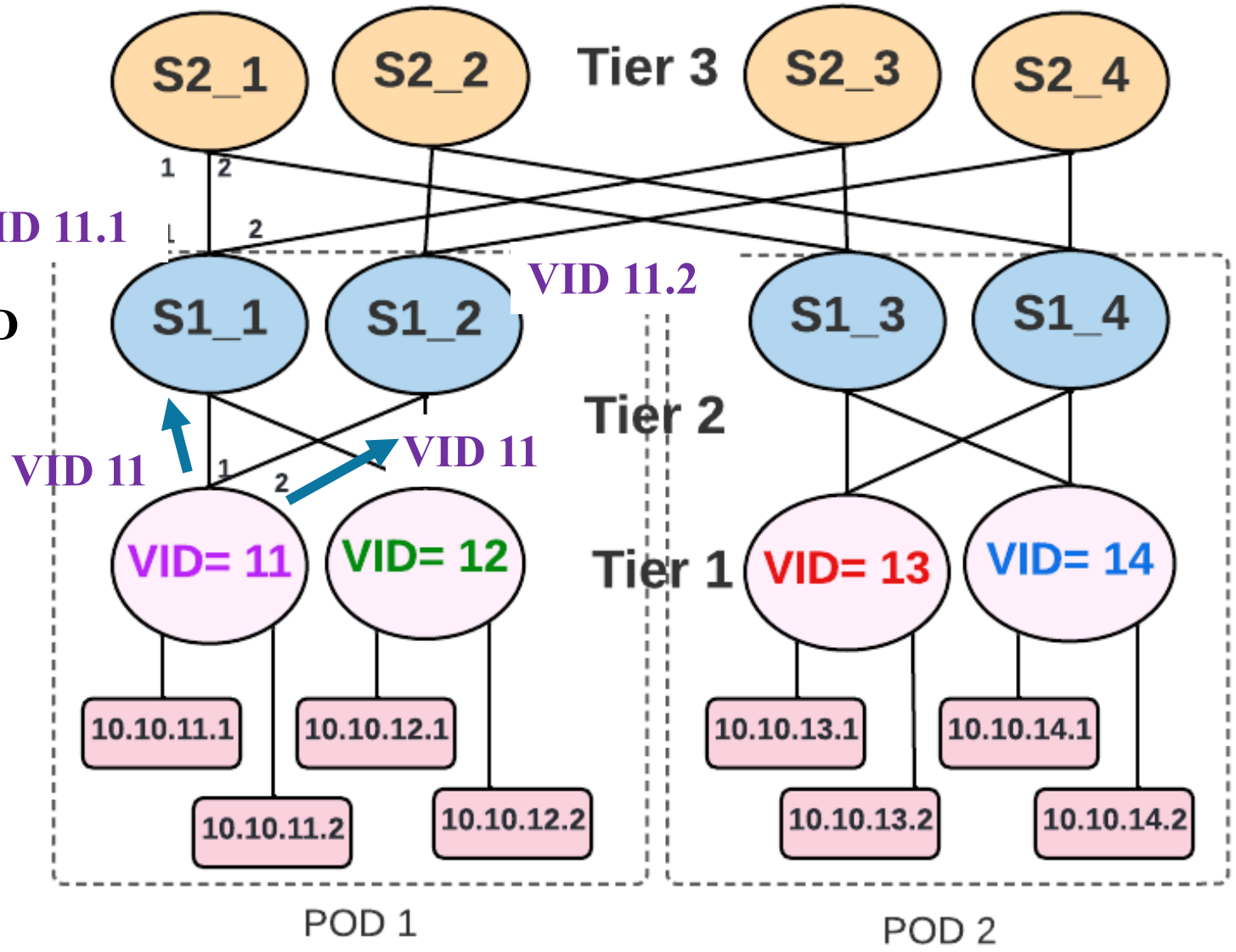
Establishing Meshed Trees with Virtual IDs – MR-MTP Operation

And then

Spines S1_1, S1_2 send in a request. ToRs assign VIDs 11.1 and 11.2 by appending the port number to their VID

ToRs advertise their VIDs

Assume ToRs have assigned VIDs – such as 11, 12, 13, 14

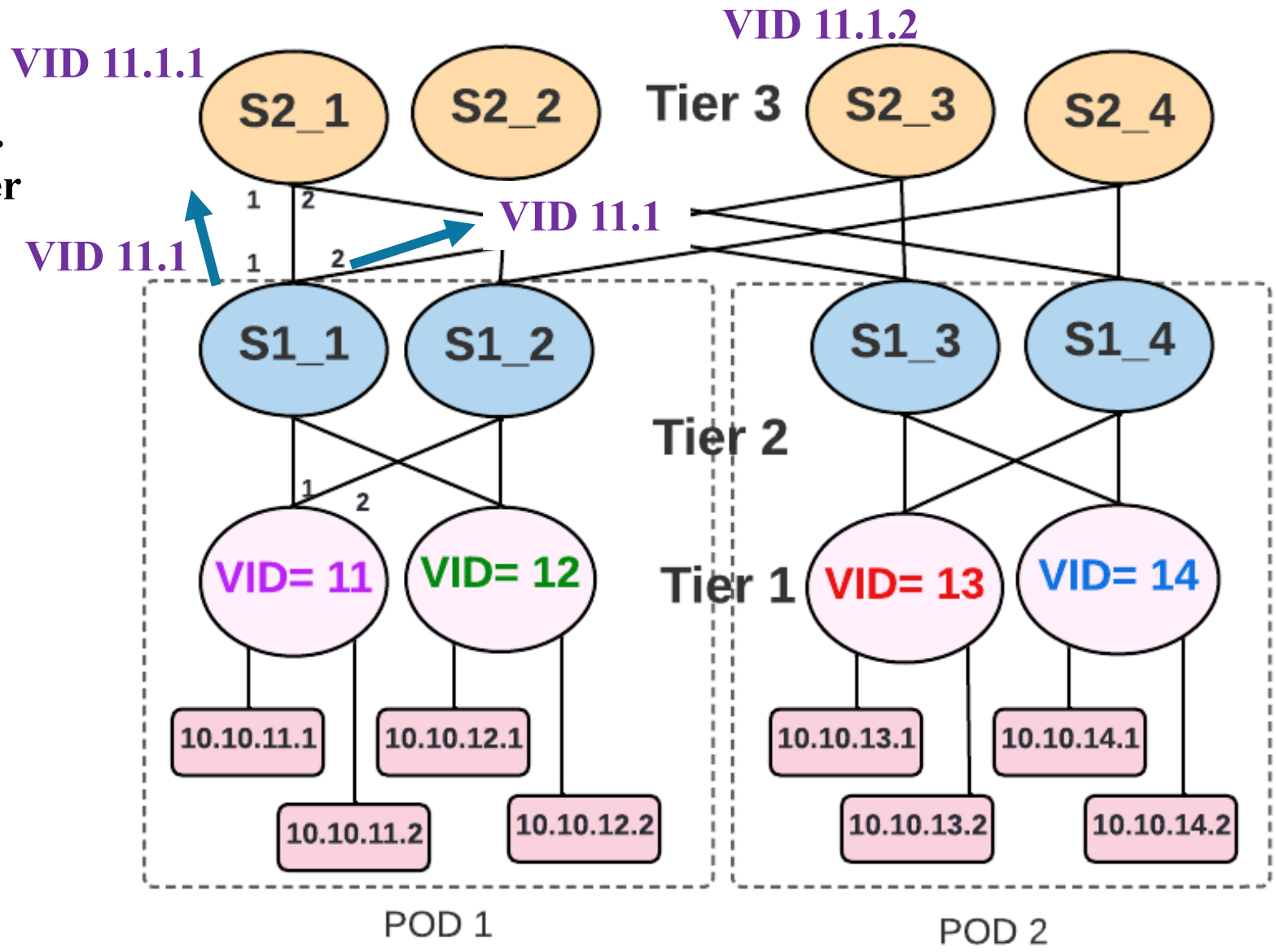


Establishing Meshed Trees with Virtual IDs

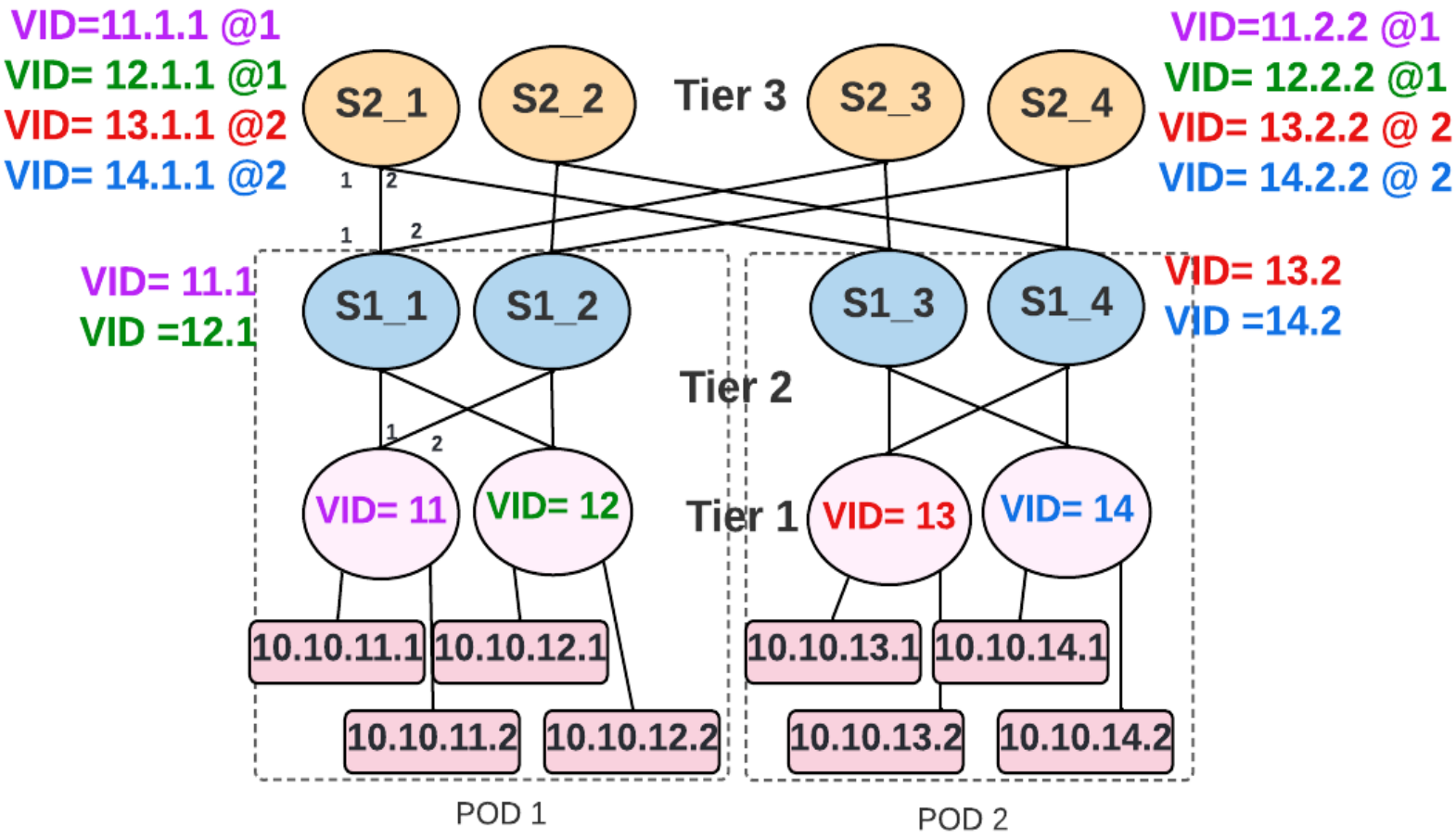
Spines S2_1, S2_3 send in a request. S1_1 assigns VIDs 11.1.1, 11.1.2 after appending the port number (on which the request arrived), to their VID

Spines S1_1advertise its VIDs

ToRs derive a unique VID from the subnet IP address (other secure algorithms to auto derive ToR VIDs possible



Virtual IDs Maintain Routing Paths



Spines store acquired VIDs @ ports of acquisition

All route paths are established using simple VIDs. Trace the color.

No routing protocols,

No route or network reachability dissemination

No IP addresses to networks, devices.

THE ONLY COFIGURATION REQUIRED IS TIERS OF THE DEVICES

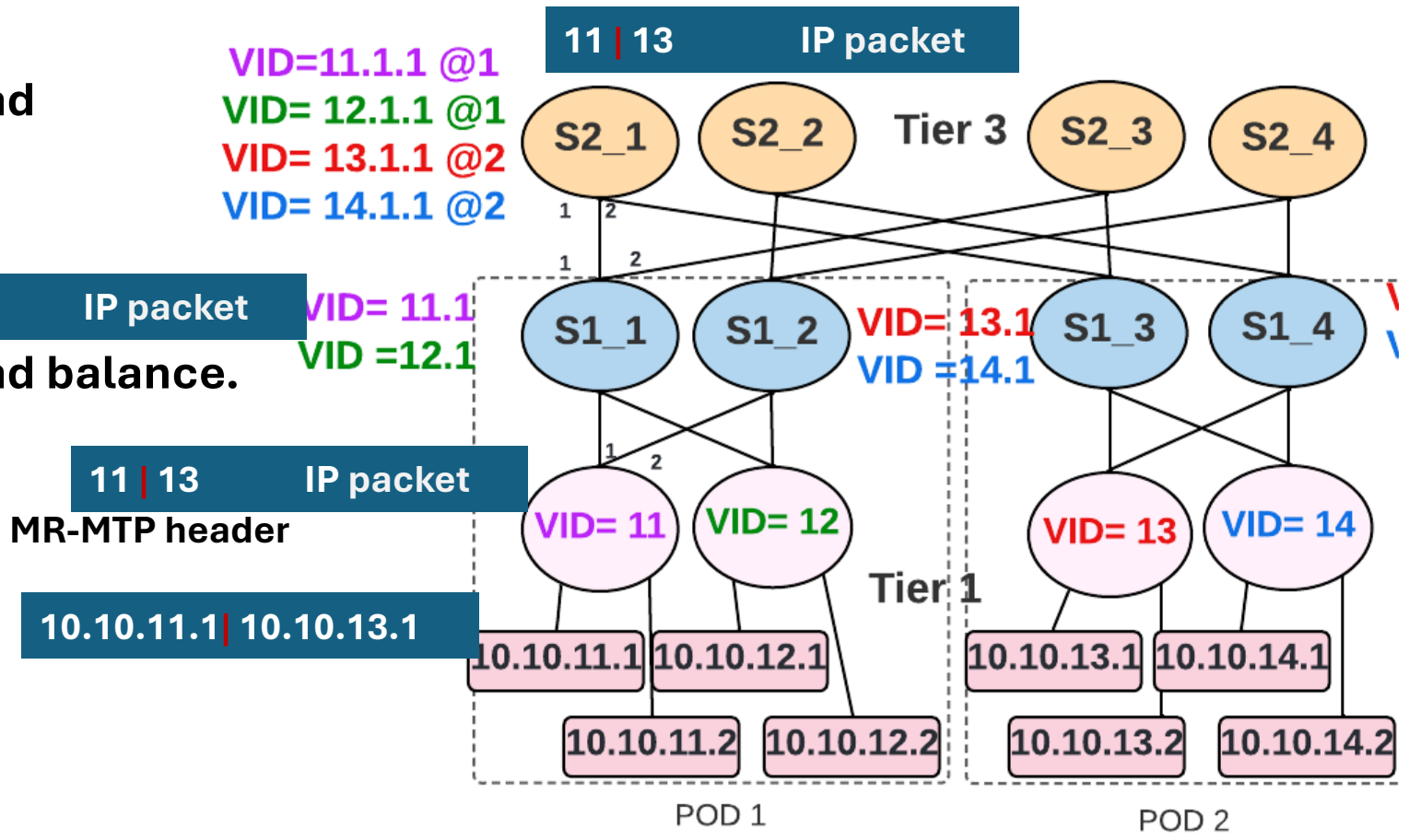
IP Packet Forwarding Between Servers

Spine S1_1 checks its VID table.
No entry for dst VID 13.
Default - send to upper tier after load balance. Send to S2_1

ToR 11 Checks VID table.
No entry for dst VID 13.
Default - send to upper tier after load balance.
Send to S1_1

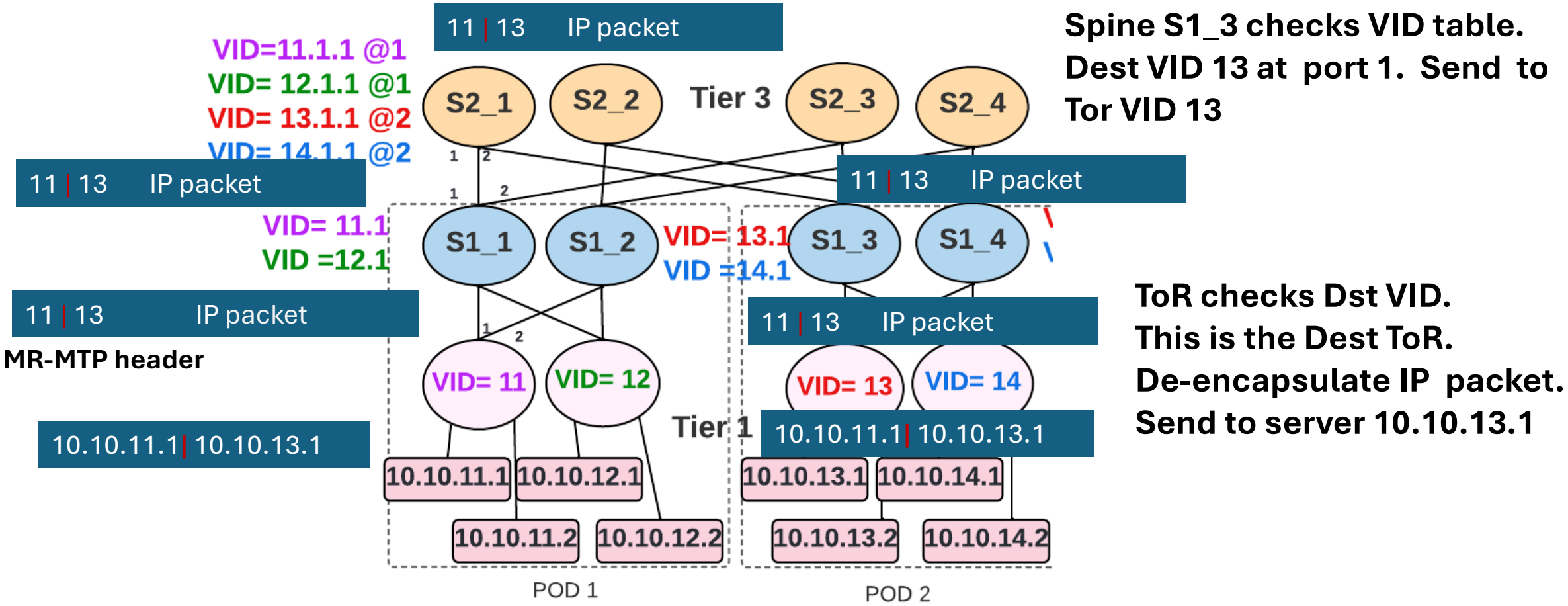
ToR encapsulates with headers.
Src, dst VID derived from subnet address

IP packet arrives at ToR 11.
Src=10.10.11.1, Dst = 10.10.13.1



IP Packet Forwarding Between Servers

Spine S2_1 checks its VID table. Dest VID 13 at port 2. Send to S1_3



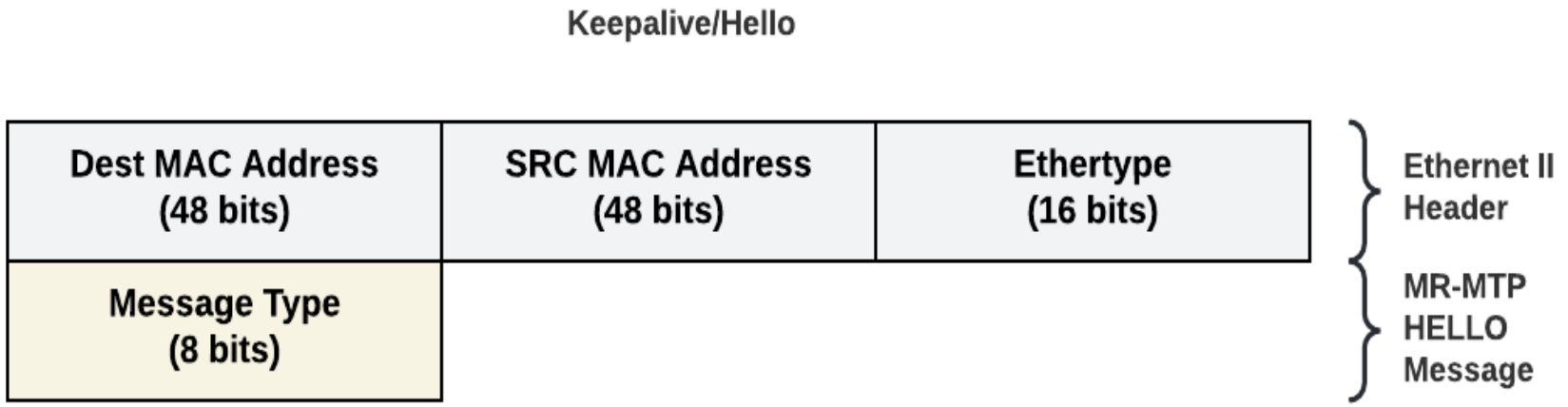
Spine S1_3 checks VID table. Dest VID 13 at port 1. Send to Tor VID 13

ToR checks Dst VID. This is the Dest ToR. De-encapsulate IP packet. Send to server 10.10.13.1

No IP addresses to route. No routing protocol.
The whole IP packet (with IP addresses) can be encrypted.

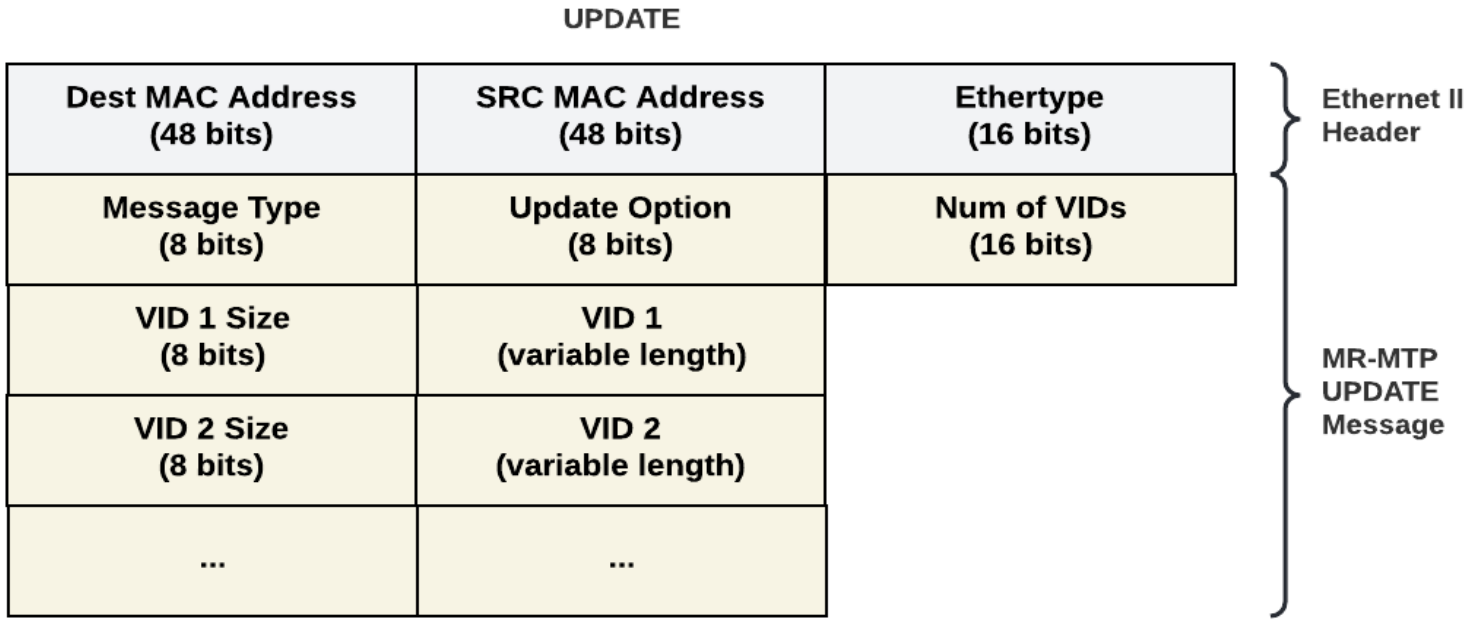
Improved Network Availability With MR-MTP

- QUICK TO DETECT- SLOW TO ACCEPT
- All MR-MTP messages are keep-alive – they have an MR-MTP header
  - They are all handled by MR-MTP
- If there are NO MR-MTP messages to send for the duration of ‘hello timer’ send a 1-byte hello message
- Missing messages for 1.5 * hello timer – assume neighbor down.
- Speeds up failure detection – QUICK TO DETECT



Improved Network Availability With MR-MTP

- To handle route /interface flapping and dampening – SLOW TO ACCEPT
  - After receiving three consecutive messages – assume neighbor up
  - Benefits - failure detection is 3 times faster.
- Update dissemination message carries only lost/added VID
- On receiving an update, a node notes
  - A Dest VID is inaccessible on the port on which the failure message was received.
  - No routing table to update
  - Low control overhead



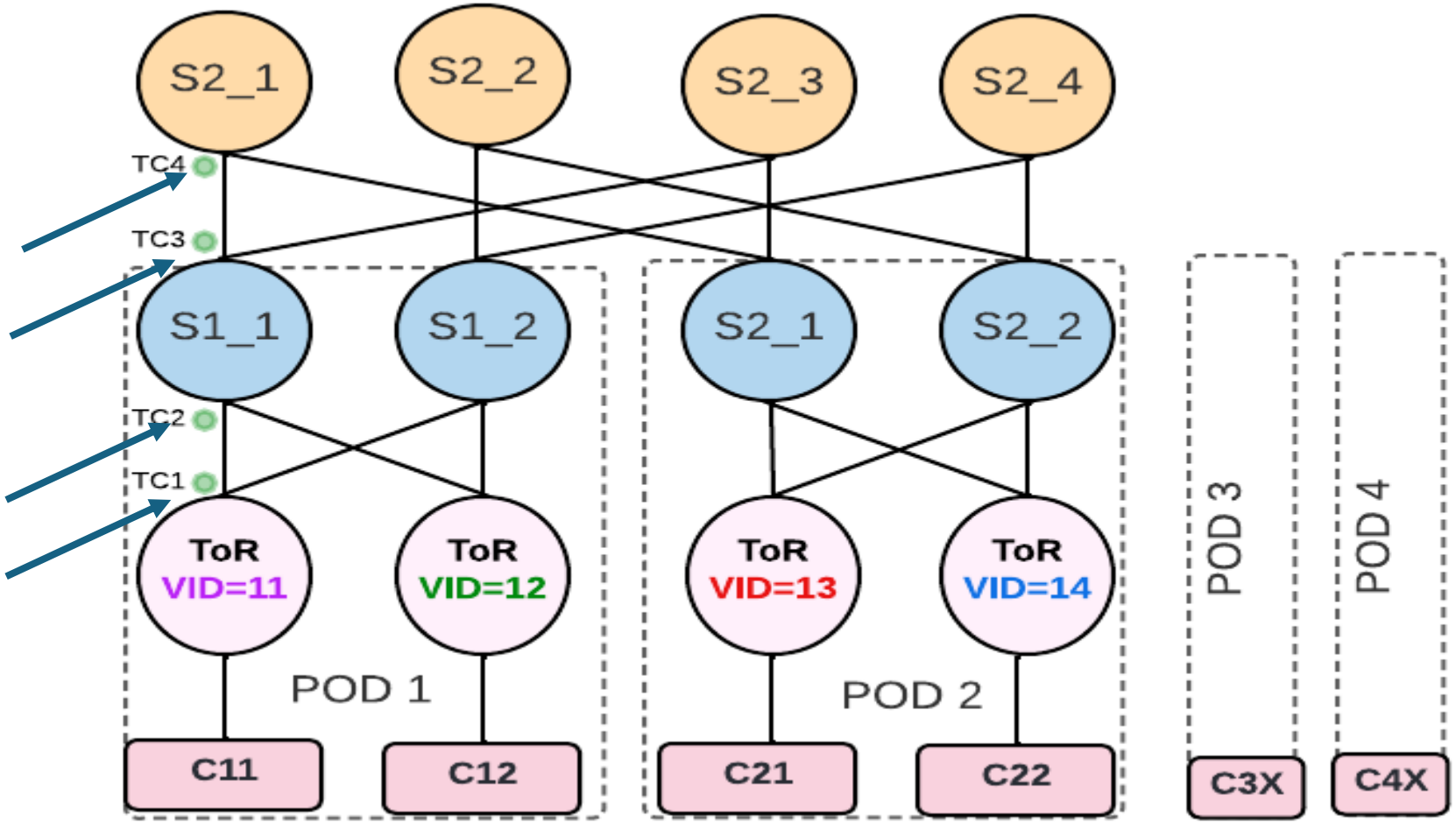
MR-MTP Operation Summary

- MR-MTP operates over Ethernet (Layer 2) – it is a Layer 3 protocol
  - Establishes routes
  - Forwards IP packets between servers
- Agnostic to layer 3 protocols.
- Replaces BGP, ECMP, BFD, TCP, UDP, IP
  - Heavy reduction in operational complexity and memory needs
- Backward compatible with IP (v4, v6) and Ethernet
  - Communicate with another DCN running IP, Ethernet etc.
- MR-MTP can be turned on/ off

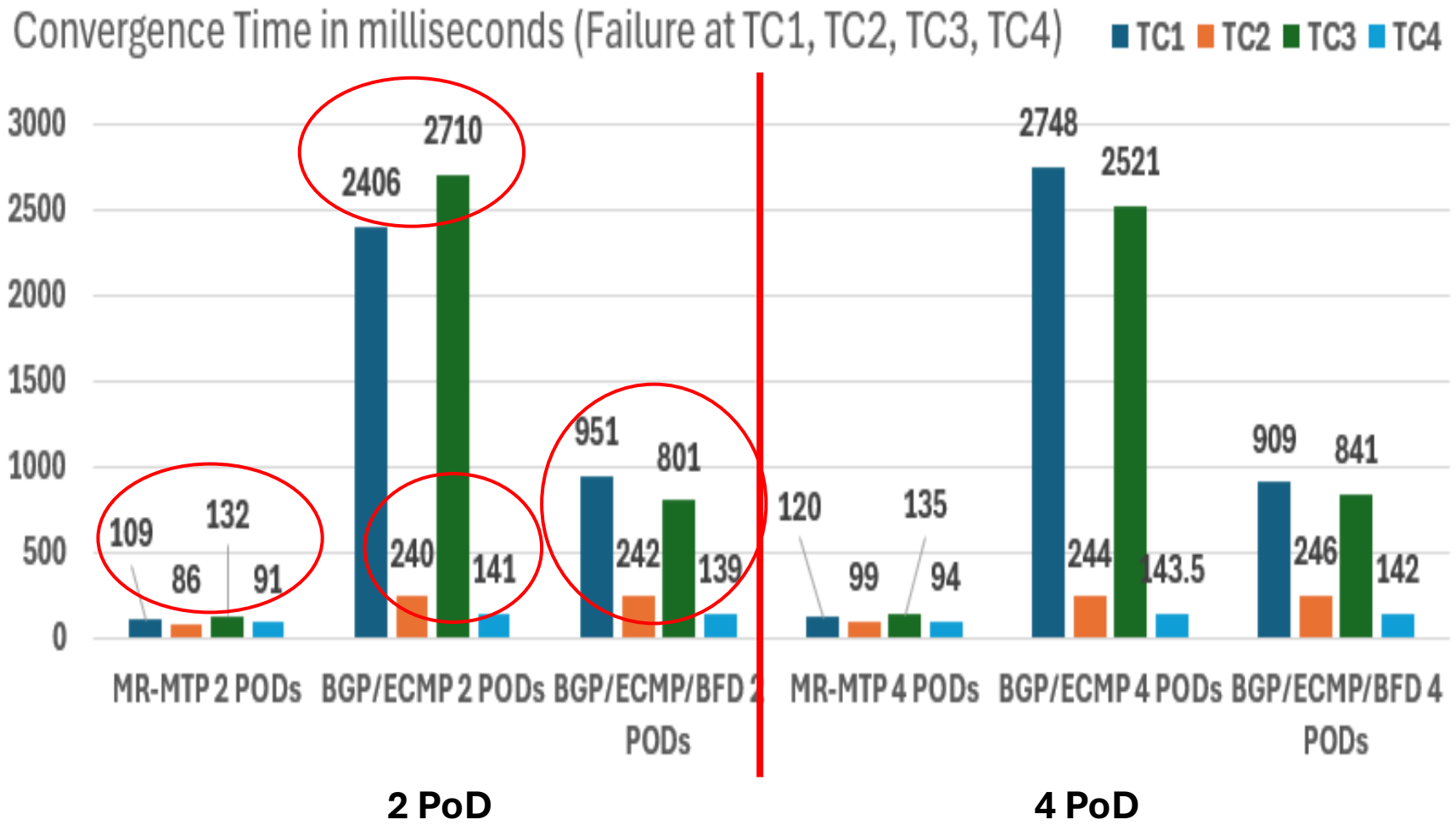
Performance Comparison

- Test Topology - 2 PoD and 4 PoD folded-Clos topology
- MR-MTP as defined
- BGP/ECMP/BFD protocol suite used for comparison studies
  - BGP modified to work on folded-Clos topology, adjusted AS numbers – requires TCP
  - Used Bidirectional Forwarding Detection to speed up failure detection – requires UDP
  - Equal Cost Multipath Protocol – used with BGP for load balancing.
  - IP for Packet Forwarding
- Presented work – BGP for DCN as per RFC7938
- Performance metrics – convergence time, control overhead, churn and packet loss on single interface failure (multiple points)
- Router configuration requirements, routing tables

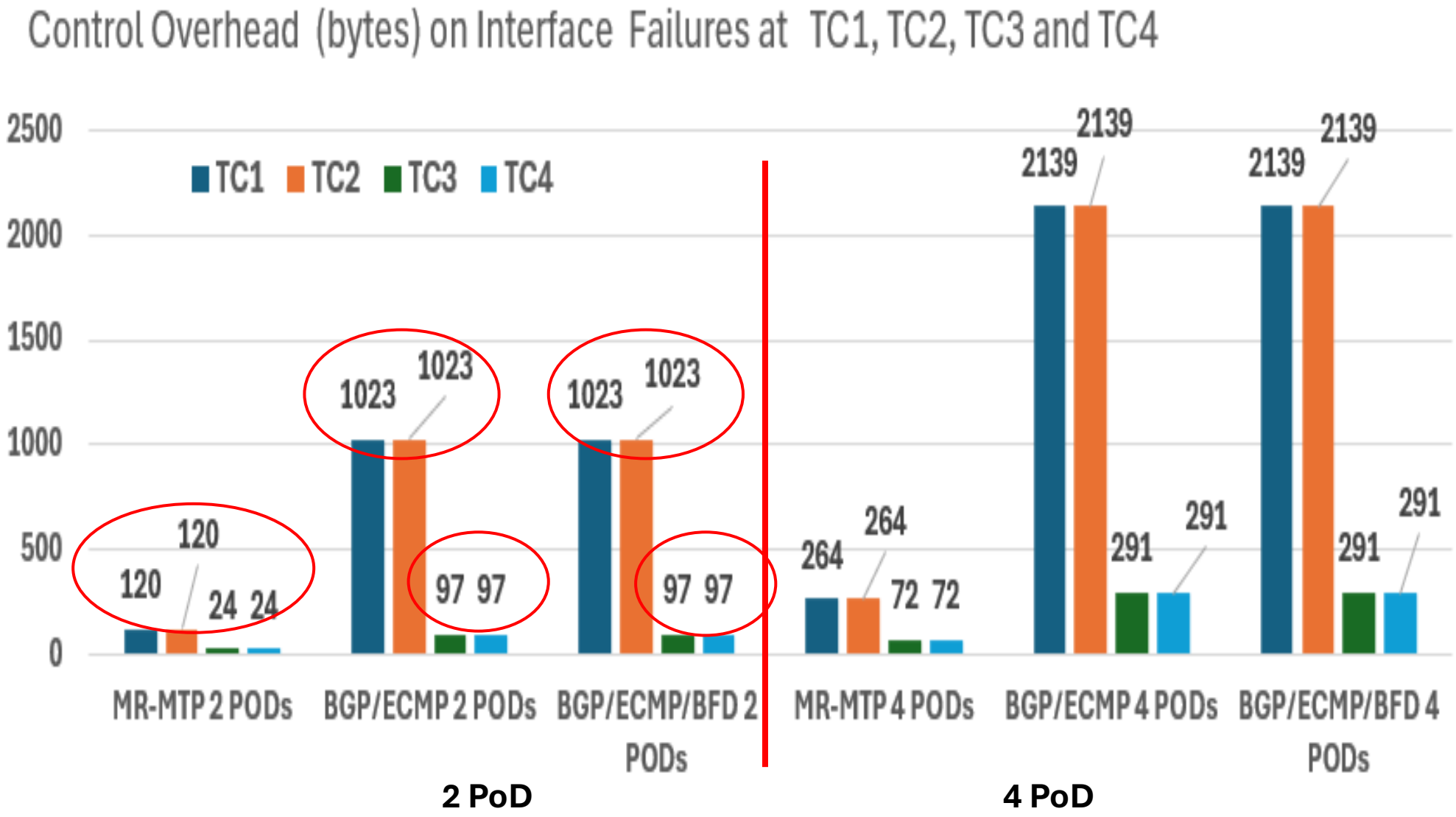
Performance – Test Topology and Test cases



Convergence time in ms
Routing Table Stabilization time

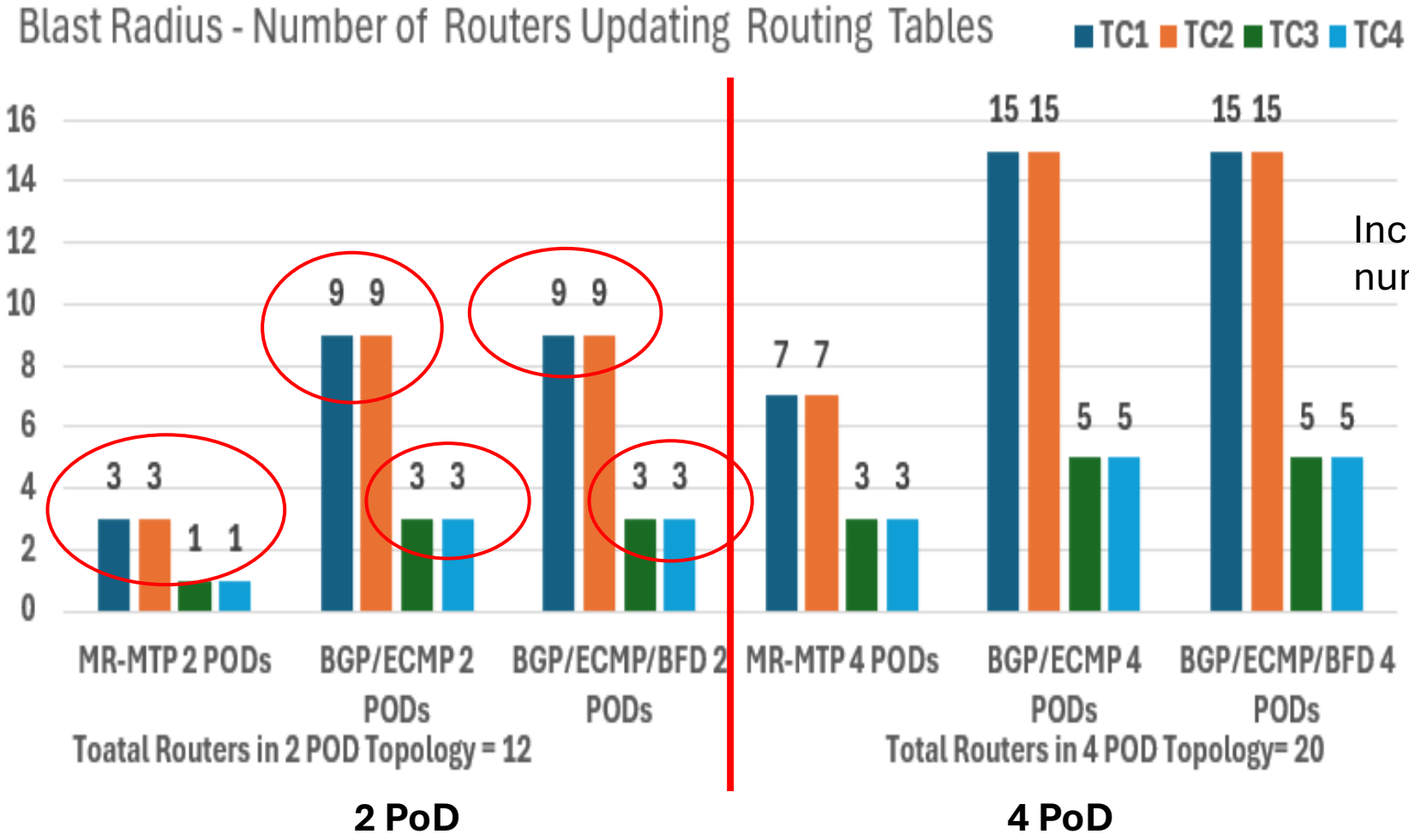


Control Overhead



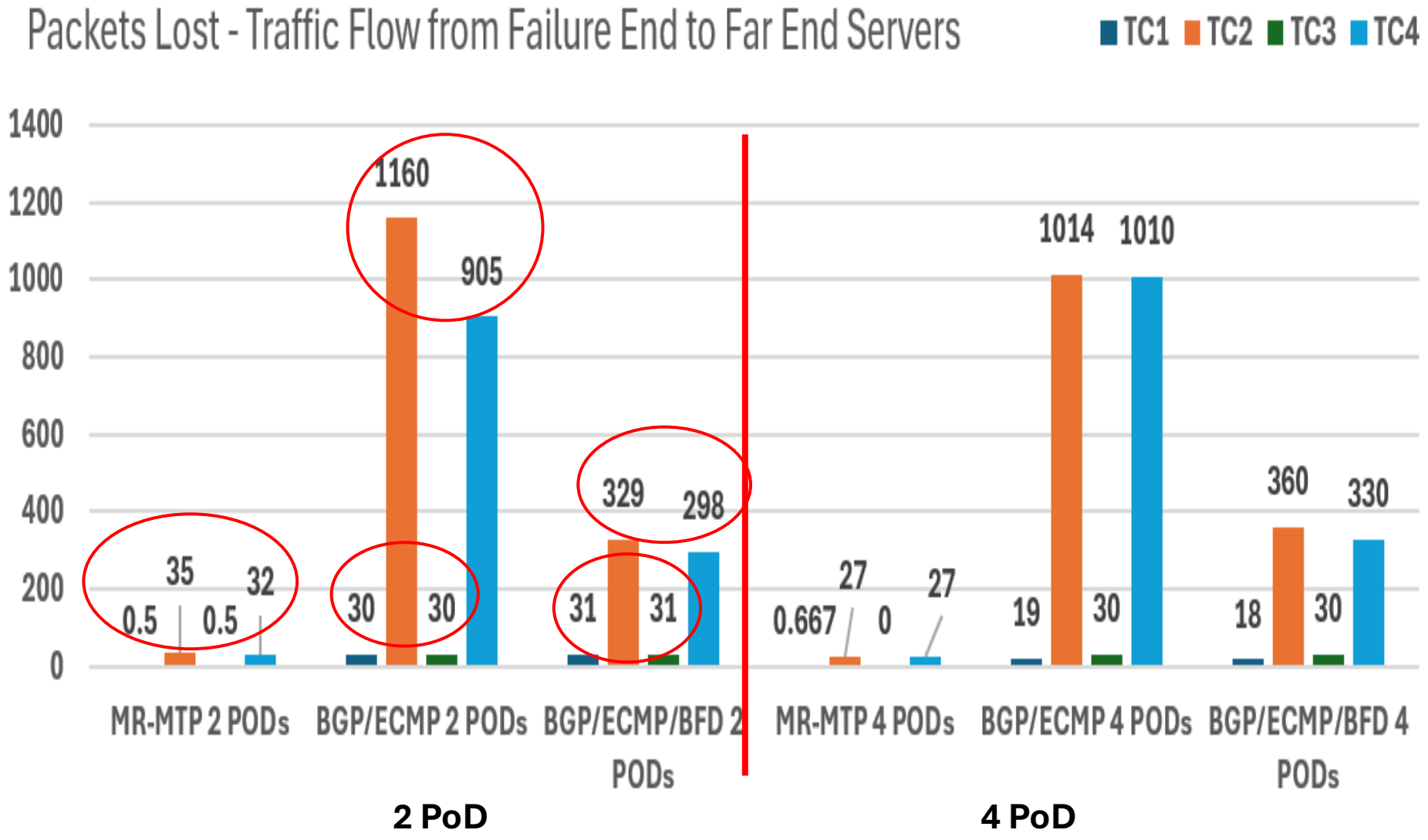
BFD overhead not accounted for

Blast Radius – Routers Updating Routing Tables

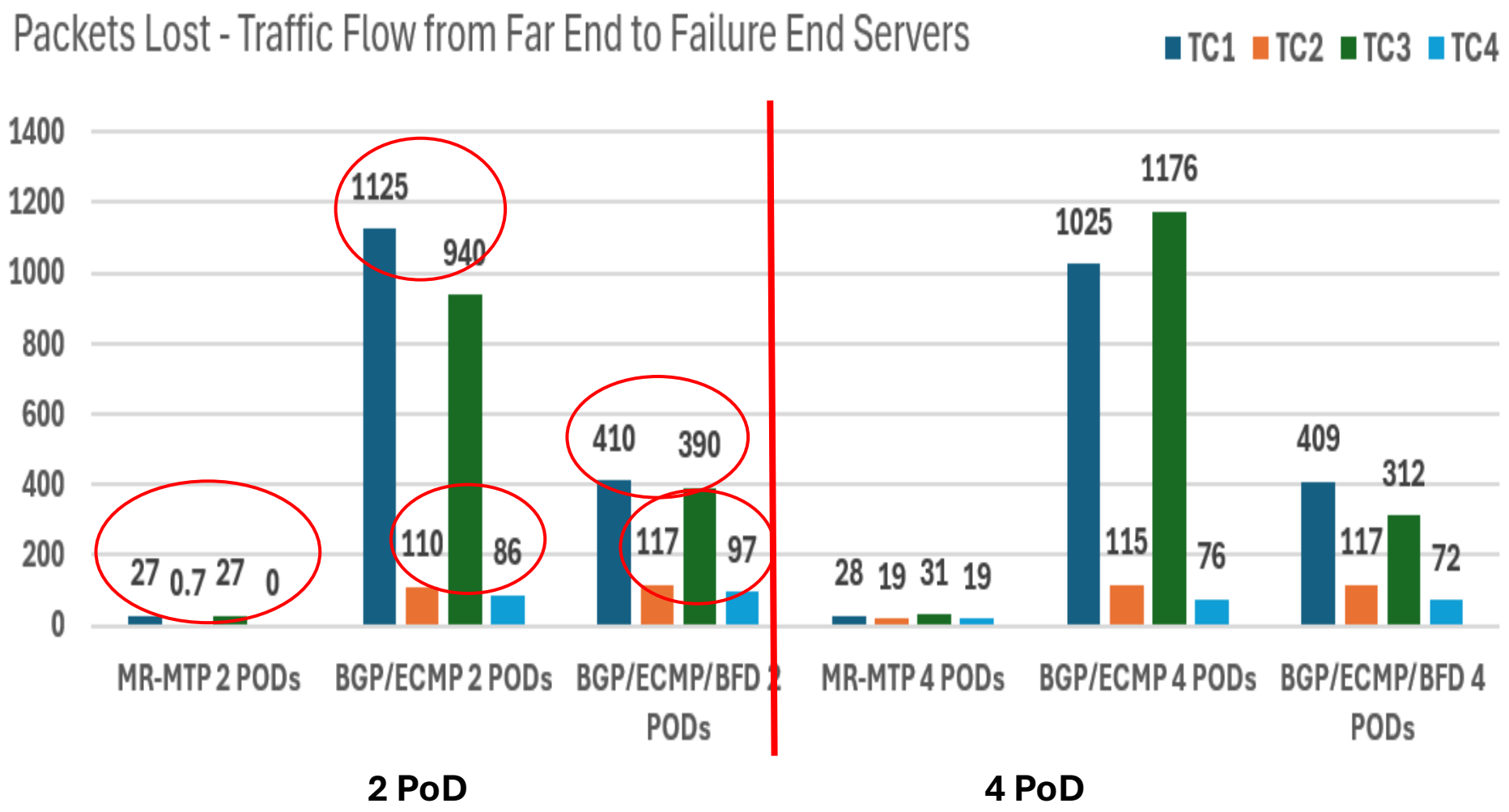


Increasing with more number of routers.

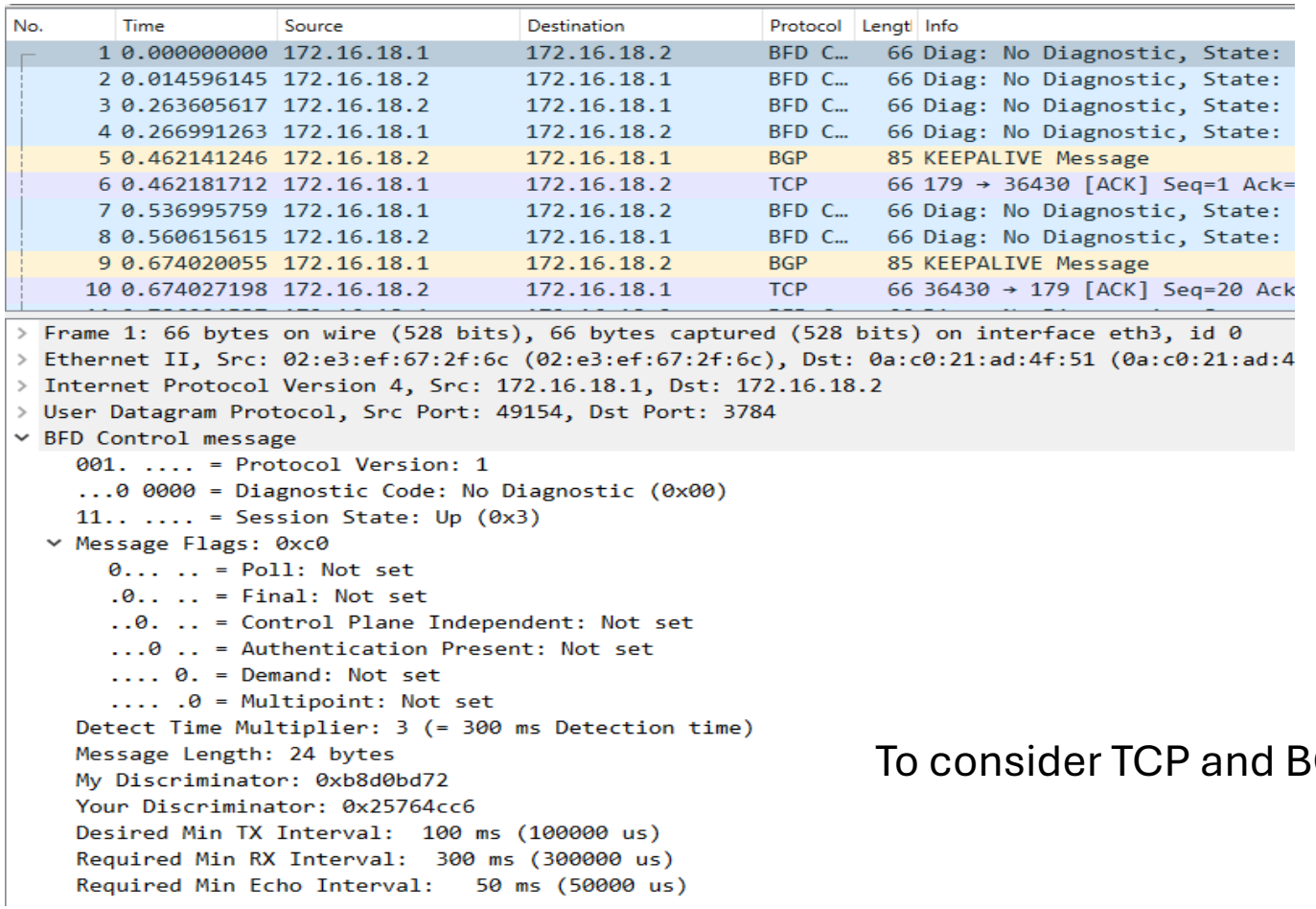
Packet Loss – Traffic from Failure End to Far End



Packet Loss – Traffic from Far End to Failure End

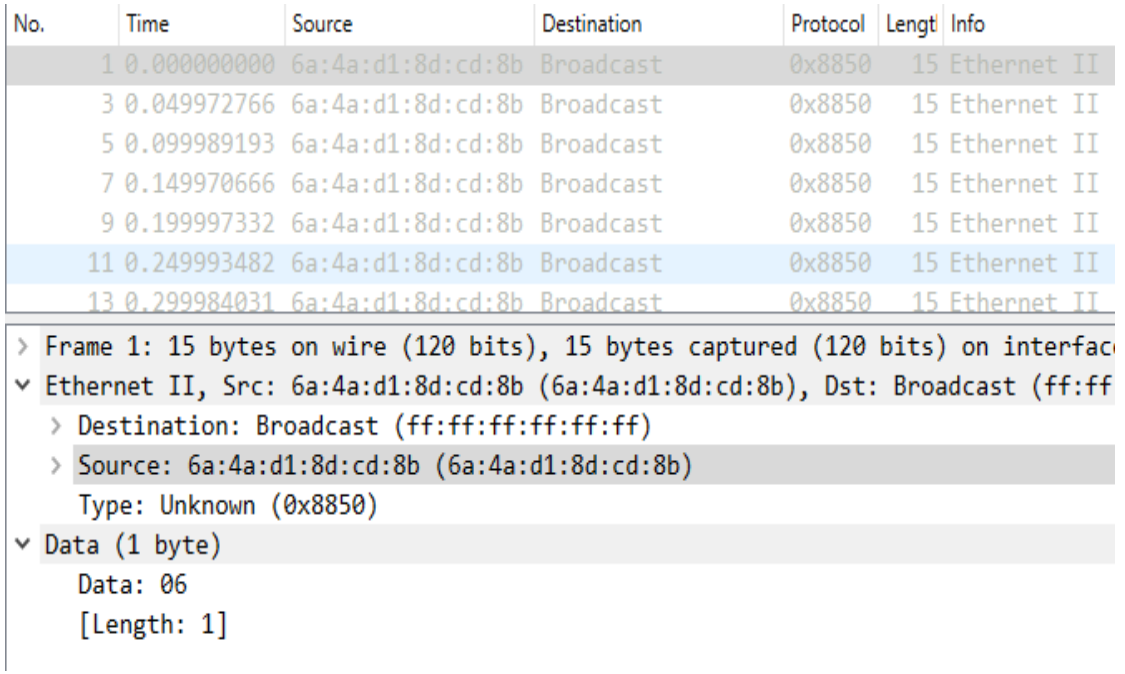


Keep – Alive Overhead (wireshark captures)



No.	Time	Source	Destination	Protocol	Lengt	Info
1	0.000000000	172.16.18.1	172.16.18.2	BFD C…	66	Diag: No Diagnostic, State:
2	0.014596145	172.16.18.2	172.16.18.1	BFD C…	66	Diag: No Diagnostic, State:
3	0.263605617	172.16.18.2	172.16.18.1	BFD C…	66	Diag: No Diagnostic, State:
4	0.266991263	172.16.18.1	172.16.18.2	BFD C…	66	Diag: No Diagnostic, State:
5	0.462141246	172.16.18.2	172.16.18.1	BGP	85	KEEPALIVE Message
6	0.462181712	172.16.18.1	172.16.18.2	TCP	66	179 → 36430 [ACK] Seq=1 Ack=
7	0.536995759	172.16.18.1	172.16.18.2	BFD C…	66	Diag: No Diagnostic, State:
8	0.560615615	172.16.18.2	172.16.18.1	BFD C…	66	Diag: No Diagnostic, State:
9	0.674020055	172.16.18.1	172.16.18.2	BGP	85	KEEPALIVE Message
10	0.674027198	172.16.18.2	172.16.18.1	TCP	66	36430 → 179 [ACK] Seq=20 Ack

```
> Frame 1: 66 bytes on wire (528 bits), 66 bytes captured (528 bits) on interface eth3, id 0
> Ethernet II, Src: 02:e3:ef:67:2f:6c (02:e3:ef:67:2f:6c), Dst: 0a:c0:21:ad:4f:51 (0a:c0:21:ad:4
> Internet Protocol Version 4, Src: 172.16.18.1, Dst: 172.16.18.2
> User Datagram Protocol, Src Port: 49154, Dst Port: 3784
v BFD Control message
    001. .... = Protocol Version: 1
    ...0 0000 = Diagnostic Code: No Diagnostic (0x00)
    11.. .... = Session State: Up (0x3)
  v Message Flags: 0xc0
      0... .. = Poll: Not set
      .0.. .. = Final: Not set
      ..0. .. = Control Plane Independent: Not set
      ...0 .. = Authentication Present: Not set
      .... 0. = Demand: Not set
      .... .0 = Multipoint: Not set
    Detect Time Multiplier: 3 (= 300 ms Detection time)
    Message Length: 24 bytes
    My Discriminator: 0xb8d0bd72
    Your Discriminator: 0x25764cc6
    Desired Min TX Interval:  100 ms (100000 us)
    Required Min RX Interval:  300 ms (300000 us)
    Required Min Echo Interval:   50 ms (50000 us)
```



No.	Time	Source	Destination	Protocol	Lengt	Info
1	0.000000000	6a:4a:d1:8d:cd:8b	Broadcast	0x8850	15	Ethernet II
3	0.049972766	6a:4a:d1:8d:cd:8b	Broadcast	0x8850	15	Ethernet II
5	0.099989193	6a:4a:d1:8d:cd:8b	Broadcast	0x8850	15	Ethernet II
7	0.149970666	6a:4a:d1:8d:cd:8b	Broadcast	0x8850	15	Ethernet II
9	0.199997332	6a:4a:d1:8d:cd:8b	Broadcast	0x8850	15	Ethernet II
11	0.249993482	6a:4a:d1:8d:cd:8b	Broadcast	0x8850	15	Ethernet II
13	0.299984031	6a:4a:d1:8d:cd:8b	Broadcast	0x8850	15	Ethernet II

```
> Frame 1: 15 bytes on wire (120 bits), 15 bytes captured (120 bits) on interfac
v Ethernet II, Src: 6a:4a:d1:8d:cd:8b (6a:4a:d1:8d:cd:8b), Dst: Broadcast (ff:ff
  > Destination: Broadcast (ff:ff:ff:ff:ff:ff)
  > Source: 6a:4a:d1:8d:cd:8b (6a:4a:d1:8d:cd:8b)
    Type: Unknown (0x8850)
v Data (1 byte)
    Data: 06
    [Length: 1]
```

To consider TCP and BGP

Router configuration

```
frr version 10.0
frr defaults datacenter hostname T-1
log file /var/log/frr/bgpd.log
log timestamp precision 3
no ipv6 forwarding
debug bgp updates in  debug bgp updates out
debug bgp updates detail
router bgp 64512
 timers bgp 1 3
 neighbor 172.16.0.2 remote-as 64513
 neighbor 172.16.0.2 bfd
 neighbor 172.16.1.2 remote-as 64514
 neighbor 172.16.1.2 bfd
 neighbor 172.16.2.2 remote-as 64515
 neighbor 172.16.2.2 bfd
 neighbor 172.16.3.2 remote-as 64516
 neighbor 172.16.3.2 bfd
bfd
 profile lowerIntervals transmit-interval
  100
 peer 172.16.0.2
  profile lowerIntervals
 peer 172.16.1.2
  profile lowerIntervals
 peer 172.16.2.2
  profile lowerIntervals
 peer 172.16.3.2
  profile lowerIntervals
```

LISTING 1: BGP Configuration at Router T-1

```
topology:
  leaves: [L-1-1,L-1-2,L-2-1,L-2-2,L-3-1,L-3-2,L-4-1,L-4-2],
 leavesNetworkPortDict:
  L-1-1 : eth3,
  L-1-2 : eth3,
  L-2-1 : eth3,
  L-2-2 : eth3,
  L-3-1 : eth1,
  L-3-2 : eth3,
  L-4-1 :      eth3,
  L–4-2 :      eth2
   topSpines : [ T-1 , T-2 , T-3 , T-4 ],
    pods : [
          topSpines : [ S-1-1 , S-1-2 ]
          topSpines : [ S-2-1 , S-2-2 ]
          topSpines : [ S-3-1 , S-3-2 ]
          topSpines : [ S-4-1 , S-4-2 ]
          ]
```

LISTING 2: MR-MTP 4-PoD json file to configure all Routers

Routing Table Size

```
172.16.0.0/24 dev eth3 proto kernel scope link src 172.16.0.2
172.16.8.0/24 dev eth4 proto kernel scope link src 172.16.8.2
172.16.16.0/24 dev eth2 proto kernel scope link src172.16.16.1
172.16.17.0/24 dev eth1 proto kernel scope link src172.16.17.1
192.168.0.0/24 via 172.16.16.2 dev eth2 proto bgp metric 20
192.168.1.0/24 via 172.16.17.2 dev eth1 proto bgp metric 20
192.168.2.0/24 proto bgp metric 20
  nexthop via 172.16.0.1 dev eth3 weight 1
  nexthop via 172.16.8.1 dev eth4 weight 1
192.168.3.0/24 proto bgp metric 20
  nexthop via 172.16.0.1 dev eth3 weight 1
  nexthop via 172.16.8.1 dev eth4 weight 1
192.168.4.0/24 proto bgp metric 20
  nexthop via 172.16.0.1 dev eth3 weight 1
   nexthop via 172.16.8.1 dev eth4 weight 1
192.168.5.0/24 proto bgp metric 20
  nexthop via 172.16.0.1 dev eth3 weight 1
  nexthop via 172.16.8.1 dev eth4 weight 1
192.168.6.0/24 proto bgp metric 20
  nexthop via 172.16.0.1 dev eth3 weight 1
   nexthop via 172.16.8.1 dev eth4 weight 1
192.168.7.0/24 proto bgp metric 20
  nexthop via 172.16.0.1 dev eth3 weight 1
  nexthop via 172.16.8.1 dev eth4 weight 1
```

LISTING 3: Tier 2 Spine BGP Routing Table

```
eth1  33.1.1,  34.1.1
eth2  35.1.1,  36.1.1
eth3  37.1.1,  38.1.1
eth4  39.1.1,  40.1.1
```

LISTING 4: MR-MTP VID table at Router T-1

Benefits

- The cost of the equipment will reduce significantly as the hardware and software requirements to implement an MR-MTP router will reduce.
- The energy consumption per router and by the DCN will reduce significantly.
- Autoconfiguration and auto addressing will reduce the configuration steps and this will reduce human errors and misconfigurations.
- The above benefits will increase multiplicatively as the DCN size increases.
- The performance benefits will be more significant as the size of the DCN increases.
- The MR-MTP DCN routers are not running BGP, TCP and IP – which will reduce the possibility of security attacks on the DCN. Simple rules to allow only IP traffic at interfaces connecting to compute nodes and gateways can protect the DCN

Future work

- Scale the folded clos topology to multiple spine tiers - Mininet
- Tuning of timers
- Extended failure test cases
- More server traffic in the network
- Use an algorithm that can be seeded to generate ToR VIDs – secure
- Encrypt IP packets originating from servers.
- Security testing – no BGP, TCP, IP
- Impact on energy consumption and carbon footprint, CPU and memory utilization
- Impact on cost and investment
- Future tests will also include overhead calculations of using the MR-MTP header for every IP packet and overhead calculations due to all protocols such as BGP, TCP, BFD and UDP which will be considered for comparison.
- Every MR_MTP message is a keep alive, cut down on the keep alive overhead incurred in current protocols
- Interested in Collaborations - security, sustainability, hw & sw, security, economics benefits

Takeaway

- Do we need routing protocols?
- Given structured networks - simple techniques can automatically establish paths.
- Comes with benefits of auto-configuration and auto address assignment
- Non-IP based solutions can be very efficient and be backward compatible with IP and Ethernet.
- Significant reduction in costs – energy, equipment and maintenance
- Highly secure DCN No BGP, TCP, IP

Thank you

Contact: Nirmala Shenoy nxsvks@rit.edu