Investigating Data Center Network Protocols

Peter Willis, Nirmala Shenoy – ISchool

Yin Pan, Bill Stackpole, Dept. of Cybersecurity

Golisano College of Computing and Information Sciences

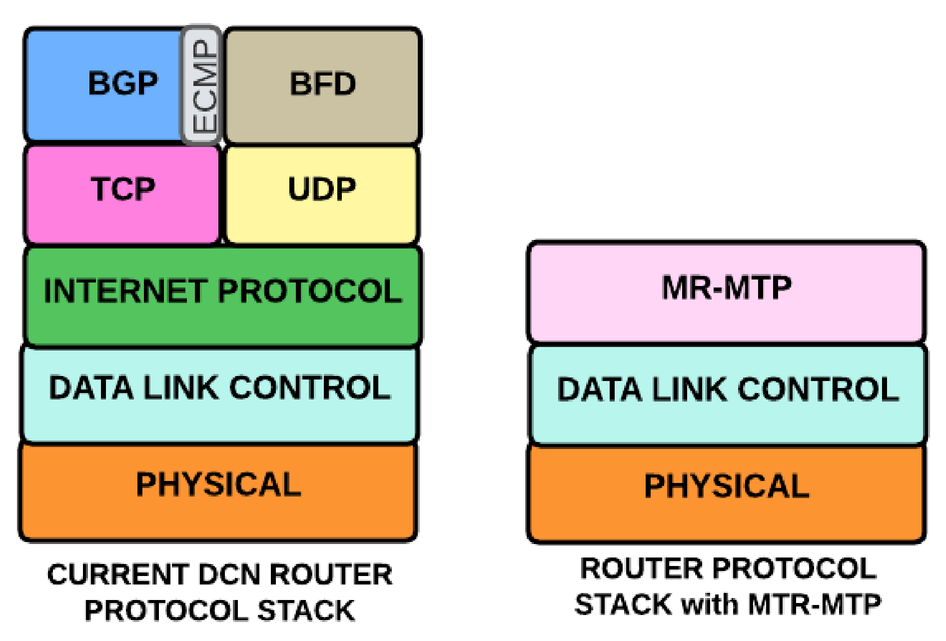
Rochester Institute of Technology, Rochester, New York

Data center networks (DCN)

- **Growing DCN sizes**
- **Increasing operational demands and complexity**
  - ***Multiple protocols, variations***
- **Severe energy and carbon footprint concerns**
- **Security**
- **Configuration**
- **Research - new architectures and topologies**
- **Protocols – variations of current routing protocols**

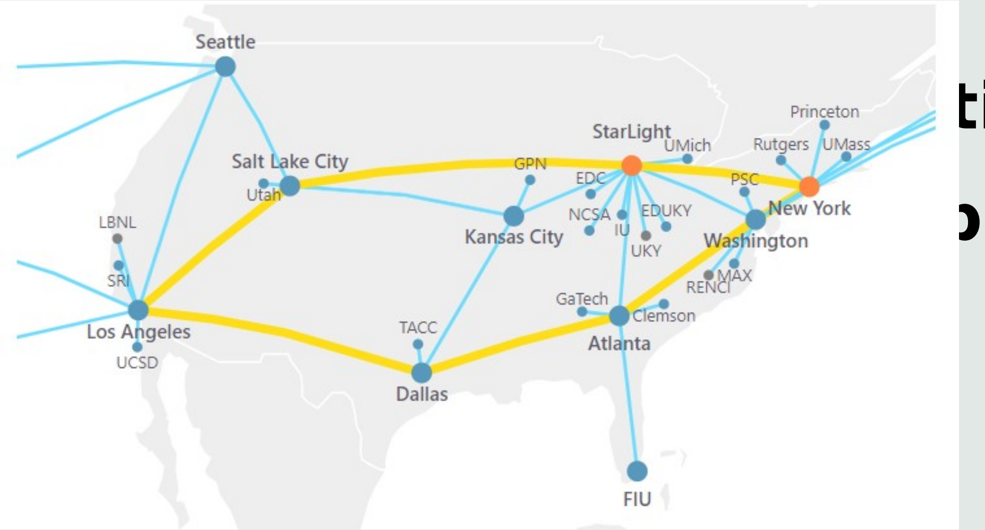
Research Focus

- **Simplify DCN protocols**
- **GOAL: Routers route traffic between servers**
- **PROJECT FOCUS**
- **TOPOLOGY: Folded Clos Topology**
- **PROTCOLS: BGP - routing, ECMP – multipath load balancing, BFD – speed up Failure detection**
- **A SINGLE SIMPLE protocol to route, load balance, speed up failure detection, forward IP Packets**
  - ***Compatible with IPv4, IPv6, Ethernet …..***

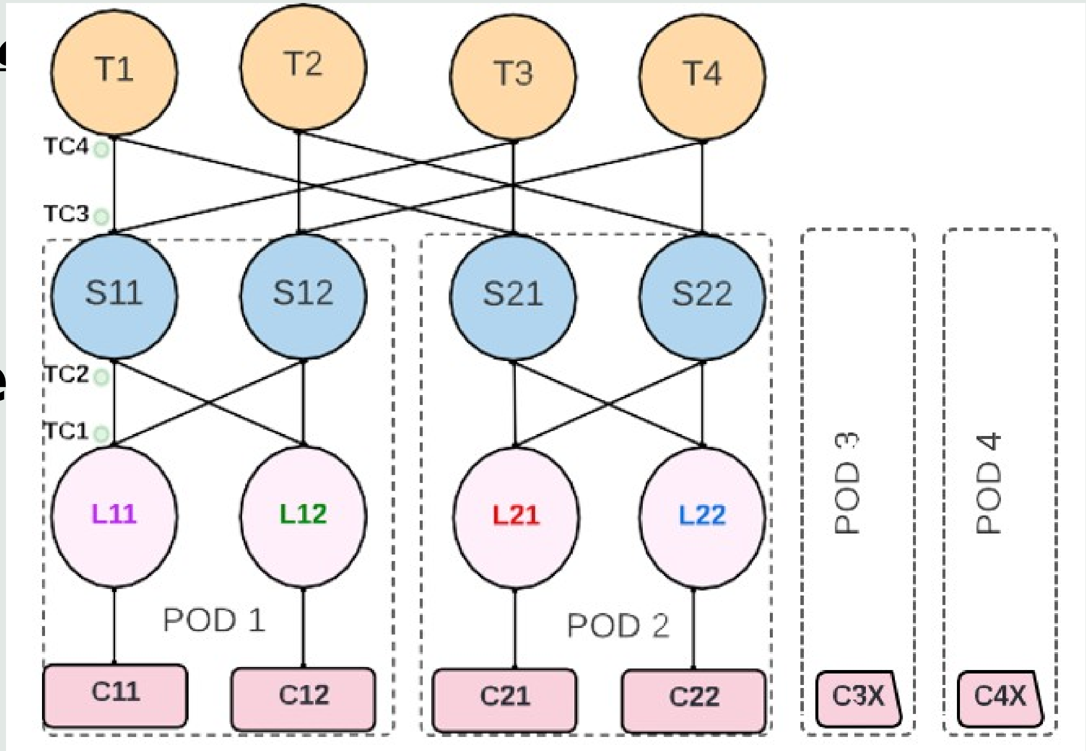


Testing

- **Proposed protocol – Multi Root Meshed Tree Protocol (MR-MTP) – C coded**
  - ***Available: https://github.com/pjw7904/CMTP***
  - ***Published and more détails 1. SIGCOMM FIRA 2022, 2023, 2. NANOG 91.***
  - ***A Simplified Data Center Network Protoco*** ***be.com)***

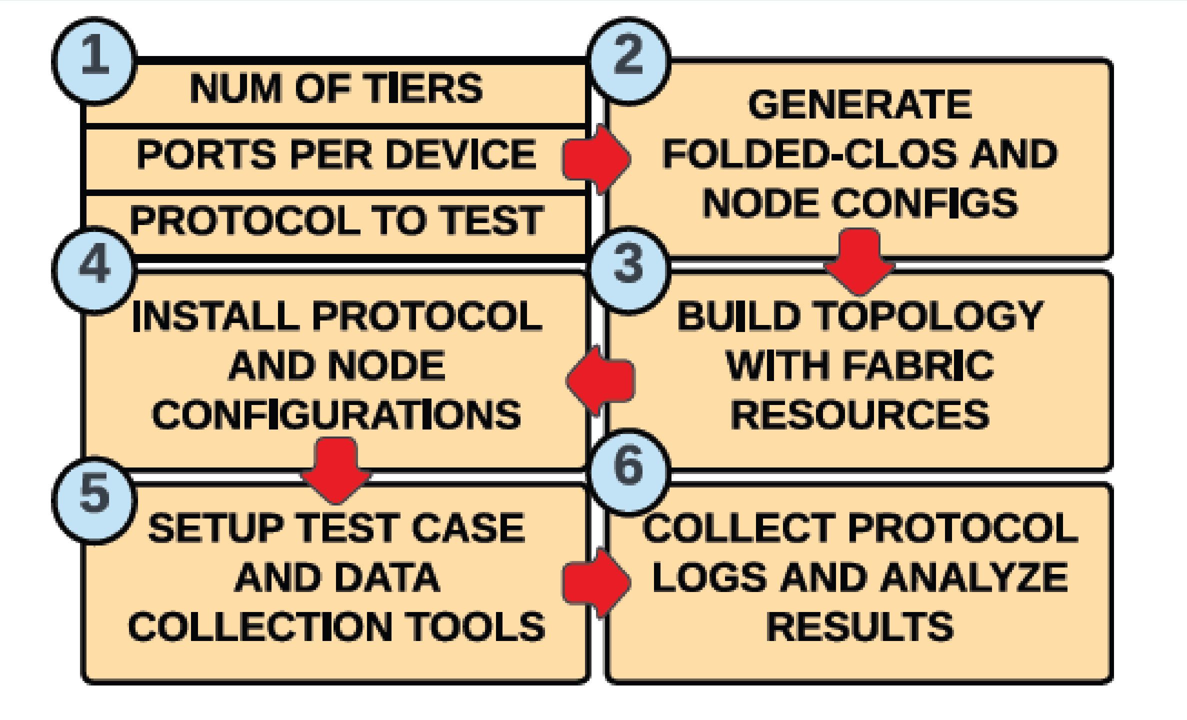


- **ting**
- **bric-te**

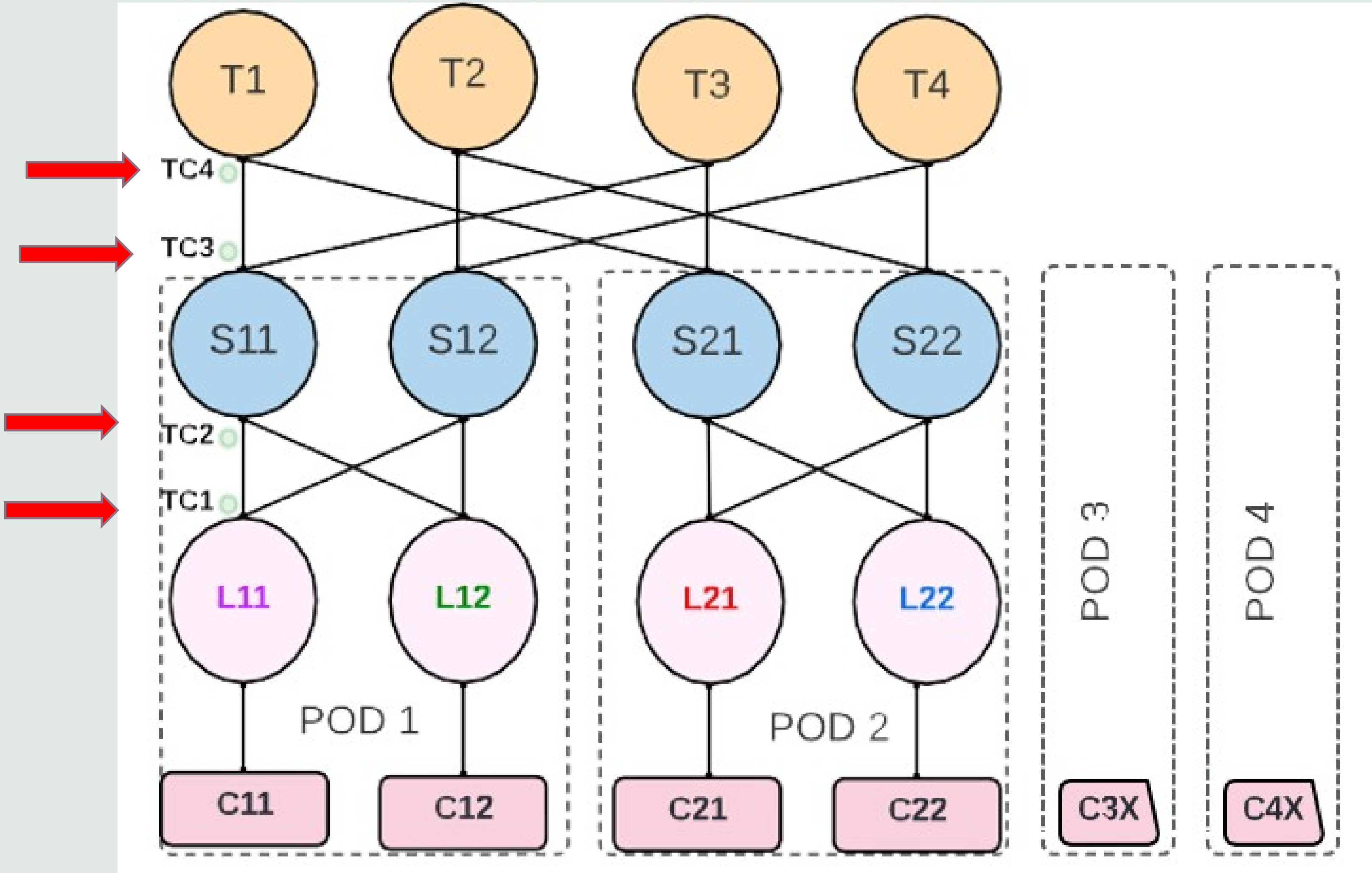


FABRIC testbed

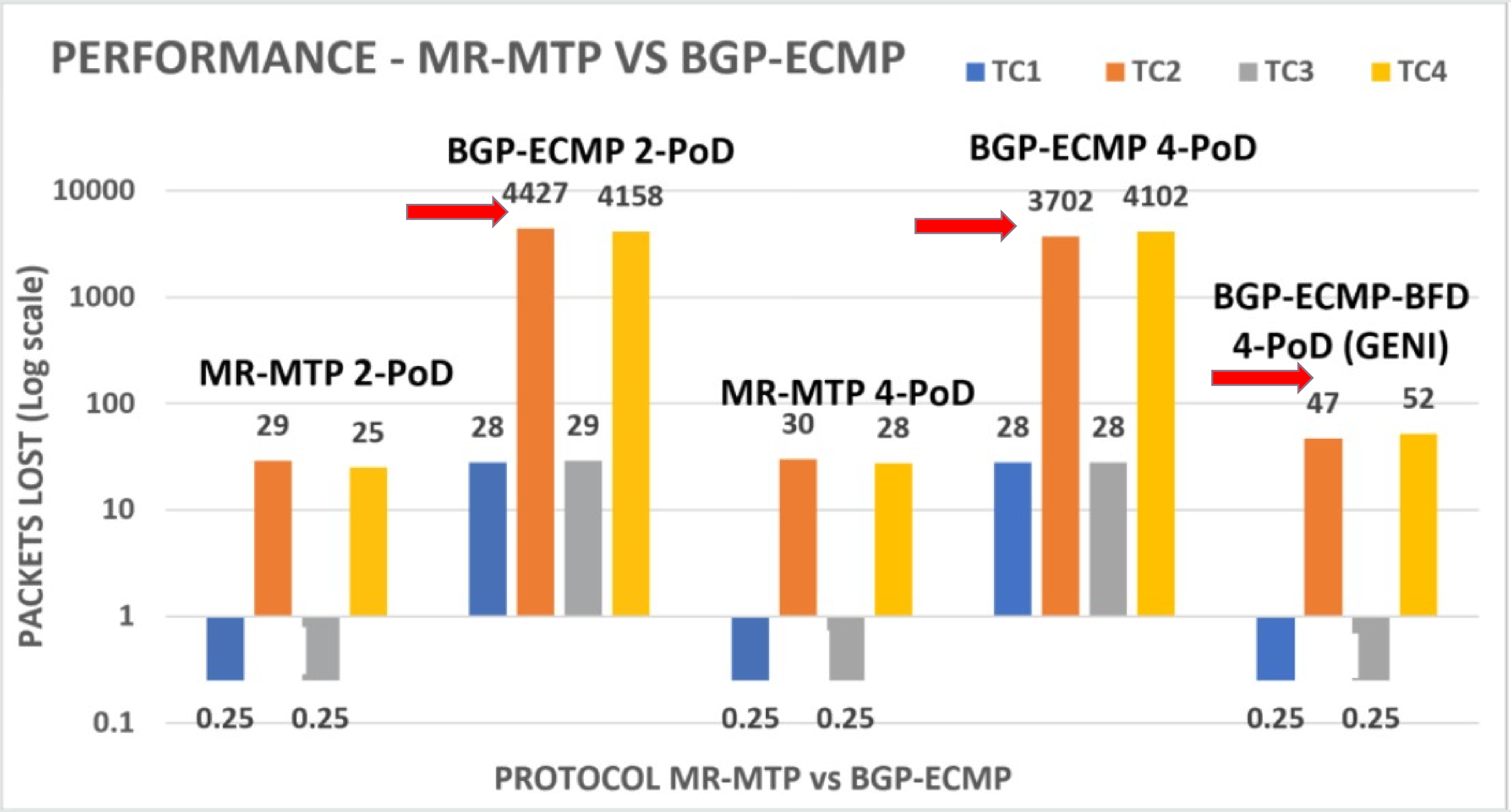
- **Customized scripts -** https://github.com/pjw7904/FABRIC-Automation
- Modular test suite
- Set up any number of tiers
- Set up the clos topology
- Identify the protocol to test
- Setup test cases – run tests
- Collect performance metrics
  - *Convergence time*
  - *Control overhead*
  - *Blast radius*
  - *Packet loss – custom traffic generator*



Repeat – see test cases



Performance – Scale?



Takeaway

- Do we need routing protocols?
- Simple automated techniques can establish paths.
- Benefits of auto-configuration and auto address assignment
- Non-IP based solutions can be very efficient and be backward compatible with IP and Ethernet.
  - *Communicate with IPv4, IPv6, limited domains, special addresses*
- Better ways to cut down on costs – energy, equipment and maintenance
- No BGP, TCP, IP –> improves security
- ------

Thank You

Questions?

Extended results

From Fabric testbed – Configuration

```
BGP: SHOW RUNNIGN CONF
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
```

BGP configuration at one router

```
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 ]
            ]
    }
}
```

MR-MTP 4-POD configuration file – for the topology

From FABRIC Testbed – Routing Tables

T-1 Routing table

10.30.0.0/19 dev eth0 proto kernel scope link src 10.30.8.203 metric 100
169.254.169.254 via 10.30.6.11 dev eth0 proto dhcp src 10.30.8.203 metric 100
172.16.0.0/24 dev eth4 proto kernel scope link src 172.16.0.1
172.16.1.0/24 dev eth2 proto kernel scope link src 172.16.1.1
172.16.2.0/24 dev eth3 proto kernel scope link src 172.16.2.1
172.16.3.0/24 dev eth1 proto kernel scope link src 172.16.3.1
192.168.0.0/24 via 172.16.0.2 dev eth4 proto bgp metric 20
192.168.1.0/24 via 172.16.0.2 dev eth4 proto bgp metric 20
192.168.2.0/24 via 172.16.1.2 dev eth2 proto bgp metric 20
192.168.3.0/24 via 172.16.1.2 dev eth2 proto bgp metric 20
192.168.4.0/24 via 172.16.2.2 dev eth3 proto bgp metric 20
192.168.5.0/24 via 172.16.2.2 dev eth3 proto bgp metric 20
192.168.6.0/24 via 172.16.3.2 dev eth1 proto bgp metric 20
192.168.7.0/24 via 172.16.3.2 dev eth1 proto bgp metric 20

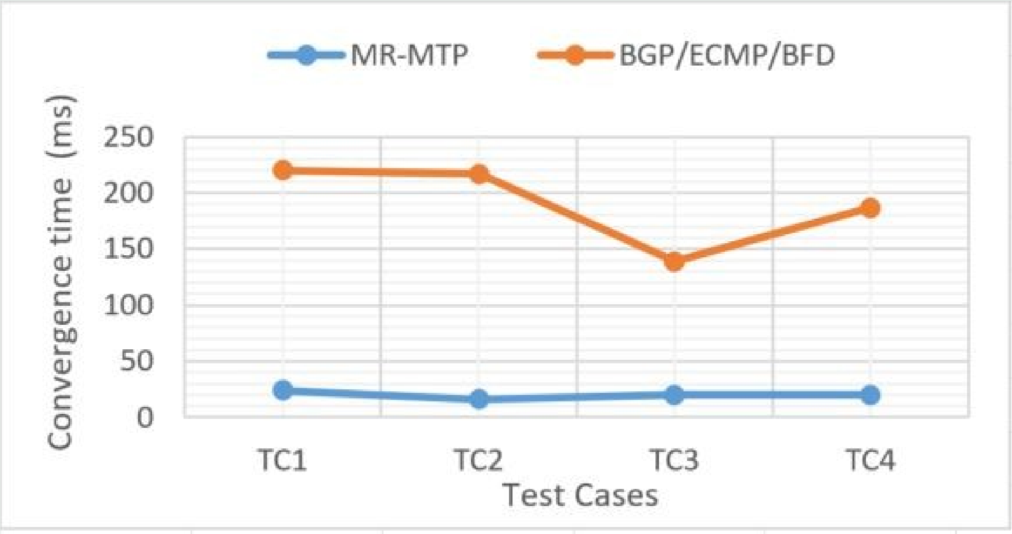
VID table at T-1

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

S-1-1 Routing Table

10.30.0.0/19 dev eth0 proto kernel scope link src 10.30.6.239 metric 100
169.254.169.254 via 10.30.6.11 dev eth0 proto dhcp src 10.30.6.239 metric 100
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 src 172.16.16.1
172.16.17.0/24 dev eth1 proto kernel scope link src 172.16.17.1
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

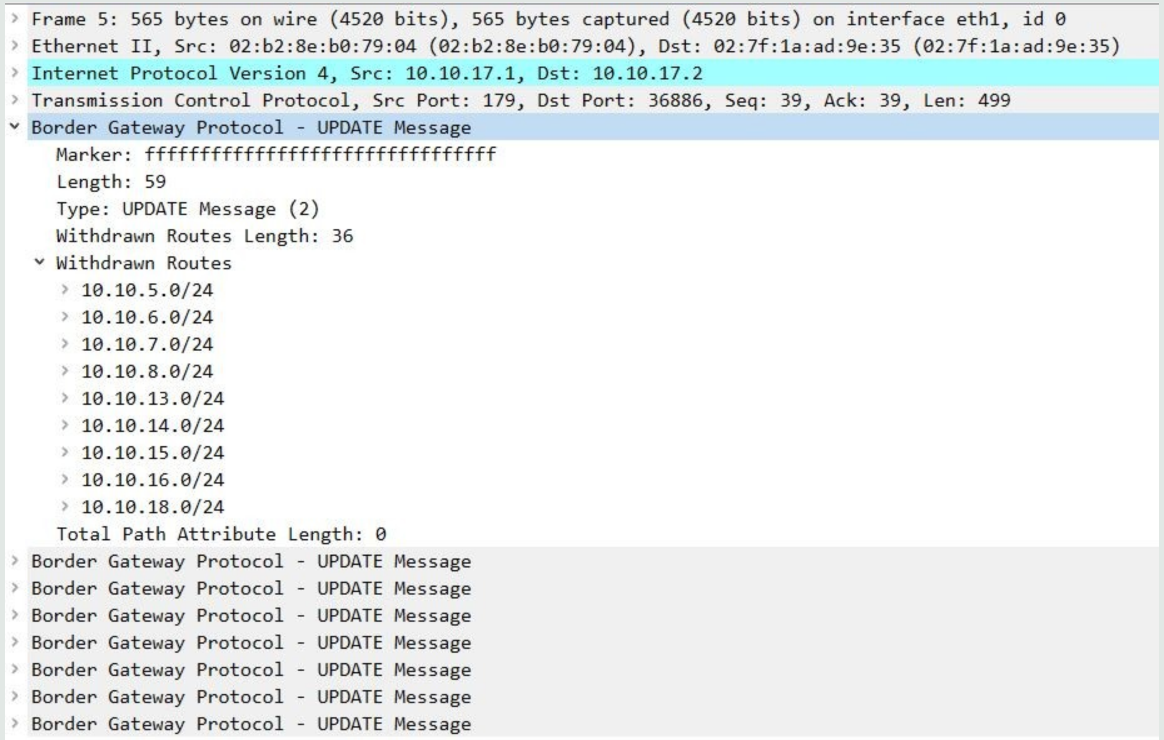
Convergence in milliseconds – Routing Table Stabilization time



BGP/ECMP/BFD convergence time (**140 to 220 ms**)
MR-MTP – convergence time (**around 25 ms**)

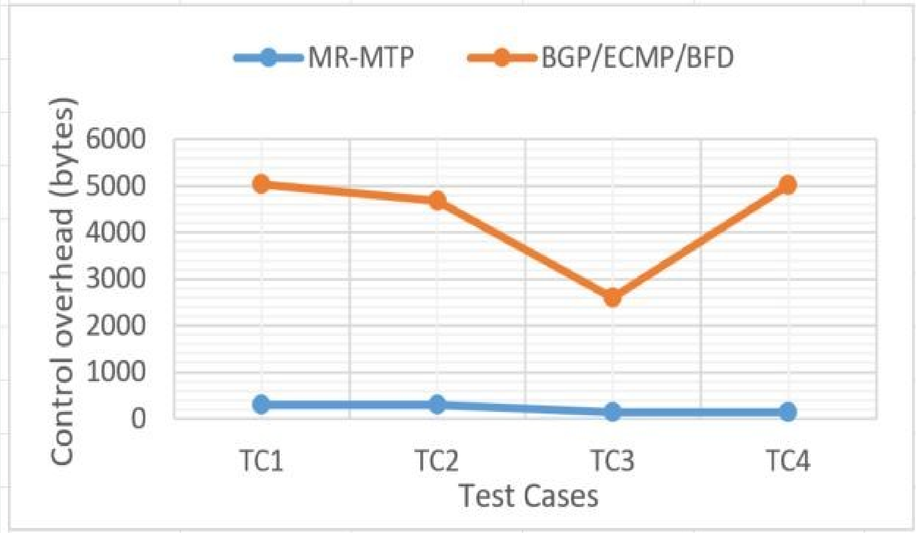
VM limitations and false failures

Control Overhead



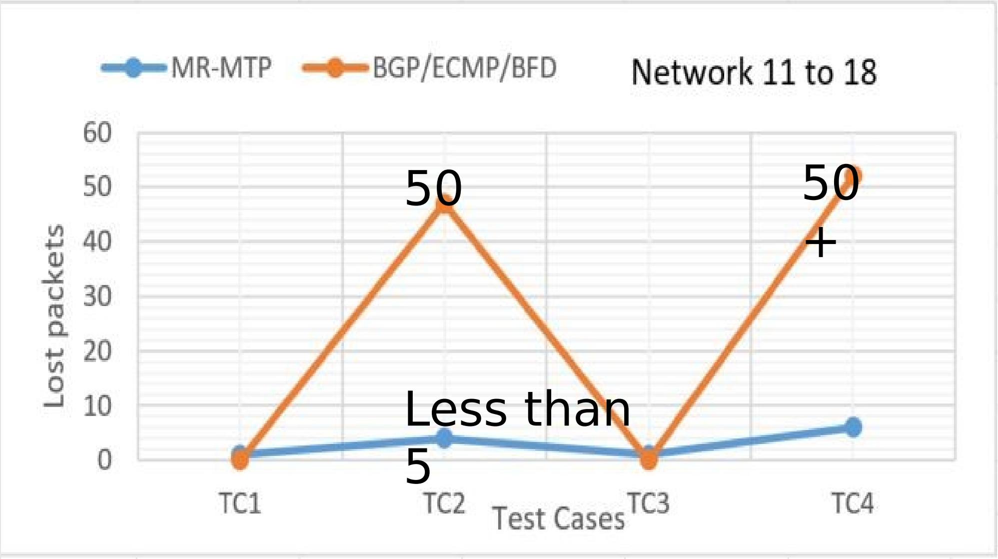
```
> Frame 5: 565 bytes on wire (4520 bits), 565 bytes captured (4520 bits) on interface eth1, id 0
> Ethernet II, Src: 02:b2:8e:b0:79:04 (02:b2:8e:b0:79:04), Dst: 02:7f:1a:ad:9e:35 (02:7f:1a:ad:9e:35)
> Internet Protocol Version 4, Src: 10.10.17.1, Dst: 10.10.17.2
> Transmission Control Protocol, Src Port: 179, Dst Port: 36886, Seq: 39, Ack: 39, Len: 499
v Border Gateway Protocol - UPDATE Message
    Marker: ffffffffffffffffffffffffffffffff
    Length: 59
    Type: UPDATE Message (2)
    Withdrawn Routes Length: 36
  v Withdrawn Routes
    > 10.10.5.0/24
    > 10.10.6.0/24
    > 10.10.7.0/24
    > 10.10.8.0/24
    > 10.10.13.0/24
    > 10.10.14.0/24
    > 10.10.15.0/24
    > 10.10.16.0/24
    > 10.10.18.0/24
    Total Path Attribute Length: 0
> Border Gateway Protocol - UPDATE Message
> Border Gateway Protocol - UPDATE Message
> Border Gateway Protocol - UPDATE Message
> Border Gateway Protocol - UPDATE Message
> Border Gateway Protocol - UPDATE Message
> Border Gateway Protocol - UPDATE Message
> Border Gateway Protocol - UPDATE Message
```

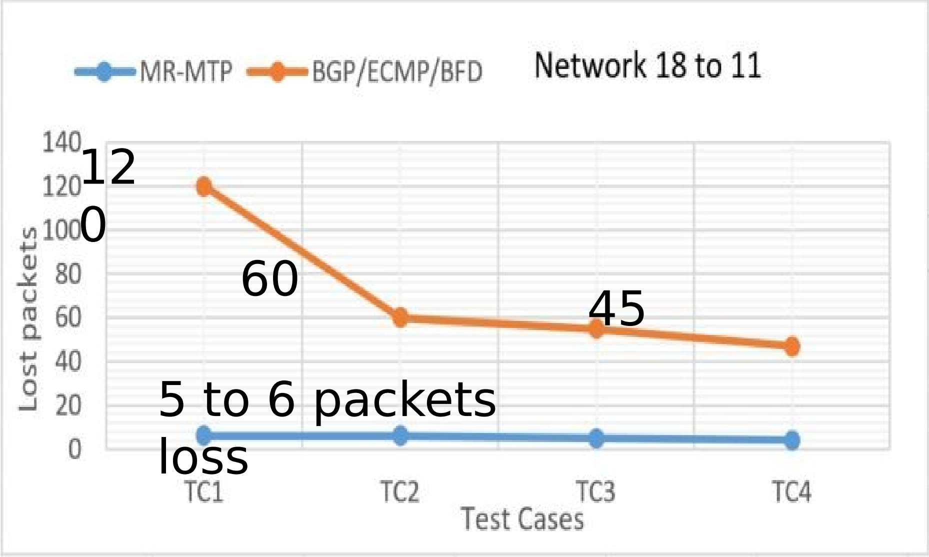
MR-MTP updates – add remove a port against a VID



BGP/ECMP/BFD control overhead (**upto 5000 bytes**)
MR-MTP – control overhead (**below 300 bytes)**
MR-MTP is more stable

Packet Loss – Network 11-18, 18 - 11



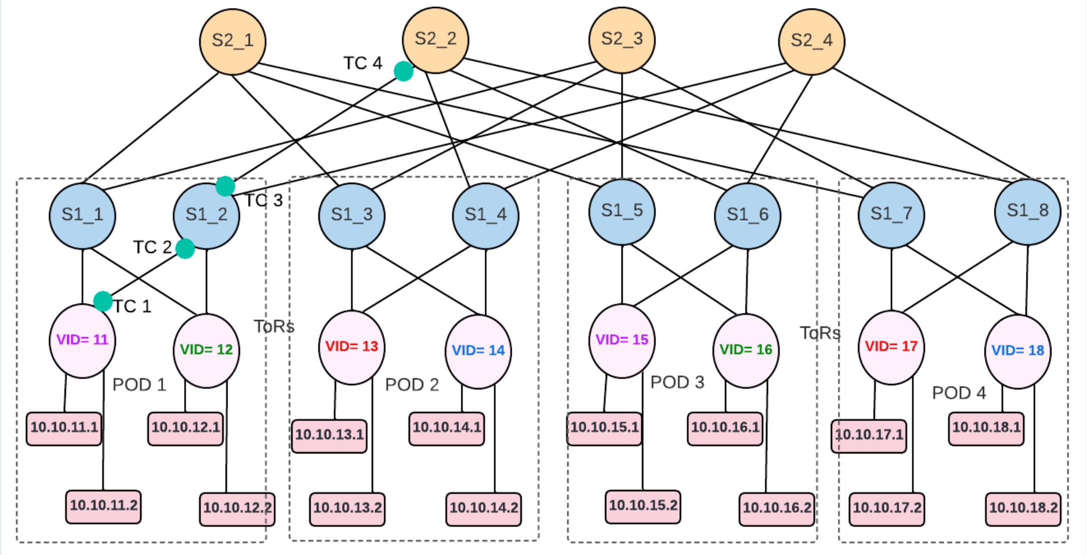


On failure at TC1, TC3, BGP router flips to other interface immediately.

MR-MTP – code in user space (no link layer failure detection)

BGP/ECMP/BFD – kernel space

IMPACT WHEN YOU SCALE



Test Topology and Test Cases (Green Dots)

Blast Radius – Routers Updating Routing Tables