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
• PROTOCOLS: 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 details 1. SIGCOMM FIRA 2022, 2023, 2. NANOG 91.
  - *A Simplified Data Center Network Protocol for Folded-Clos Topology* (youtu.be.com)

- **BGP for DCN – from FRRouting**

- FABRIC testbed - portal.fabric-testbed.net
FABRIC testbed

- **Customized scripts** - [https://github.com/pjw7904/FABRIC-Automation](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
BGP configuration at one router

MR-MTP 4-POD configuration file – for the topology
### T-1 Routing Table

<table>
<thead>
<tr>
<th>Network</th>
<th>Device</th>
<th>Protocol</th>
<th>Scope</th>
<th>Source IP</th>
<th>Metric</th>
<th>Next Hop</th>
<th>Device</th>
<th>Metric</th>
</tr>
</thead>
<tbody>
<tr>
<td>10.30.0.0/19</td>
<td>eth0</td>
<td>kernel</td>
<td>link</td>
<td>10.30.8.203</td>
<td>100</td>
<td>10.30.6.11</td>
<td>eth0</td>
<td>100</td>
</tr>
<tr>
<td>169.254.169.254</td>
<td></td>
<td>dhcp</td>
<td></td>
<td>10.30.8.203</td>
<td></td>
<td>10.30.6.11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>172.16.0.0/24</td>
<td>eth4</td>
<td>kernel</td>
<td></td>
<td>172.16.0.1</td>
<td></td>
<td>172.16.0.2</td>
<td>eth3</td>
<td>20</td>
</tr>
<tr>
<td>172.16.8.0/24</td>
<td></td>
<td>kernel</td>
<td></td>
<td>172.16.0.1</td>
<td></td>
<td>172.16.1.2</td>
<td>eth2</td>
<td>20</td>
</tr>
<tr>
<td>172.16.16.0/24</td>
<td></td>
<td>kernel</td>
<td></td>
<td>172.16.0.2</td>
<td></td>
<td>172.16.1.2</td>
<td>eth1</td>
<td>20</td>
</tr>
<tr>
<td>192.168.0.0/24</td>
<td></td>
<td>bgp</td>
<td></td>
<td>172.16.0.1</td>
<td></td>
<td>172.16.0.1</td>
<td>eth3</td>
<td>1</td>
</tr>
<tr>
<td>192.168.1.0/24</td>
<td></td>
<td>bgp</td>
<td></td>
<td>172.16.0.1</td>
<td></td>
<td>172.16.1.2</td>
<td>eth2</td>
<td>1</td>
</tr>
<tr>
<td>192.168.2.0/24</td>
<td></td>
<td>bgp</td>
<td></td>
<td>172.16.0.1</td>
<td></td>
<td>172.16.1.2</td>
<td>eth1</td>
<td>1</td>
</tr>
<tr>
<td>192.168.3.0/24</td>
<td></td>
<td>bgp</td>
<td></td>
<td>172.16.0.1</td>
<td></td>
<td>172.16.1.2</td>
<td>eth3</td>
<td>1</td>
</tr>
<tr>
<td>192.168.4.0/24</td>
<td></td>
<td>bgp</td>
<td></td>
<td>172.16.0.1</td>
<td></td>
<td>172.16.1.2</td>
<td>eth2</td>
<td>1</td>
</tr>
<tr>
<td>192.168.5.0/24</td>
<td></td>
<td>bgp</td>
<td></td>
<td>172.16.0.1</td>
<td></td>
<td>172.16.1.2</td>
<td>eth1</td>
<td>1</td>
</tr>
</tbody>
</table>

### VID Table at T-1

<table>
<thead>
<tr>
<th>Interface</th>
<th>VID 33.1.1</th>
<th>VID 34.1.1</th>
<th>VID 35.1.1</th>
<th>VID 36.1.1</th>
<th>VID 37.1.1</th>
<th>VID 38.1.1</th>
<th>VID 39.1.1</th>
<th>VID 40.1.1</th>
</tr>
</thead>
<tbody>
<tr>
<td>eth1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>eth2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>eth3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>eth4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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

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
Blast Radius – Routers Updating Routing Tables