# IETF Hackathon

## -MSR6 TE

IETF 116

March 25-26, 2023

# Hackathon Plan

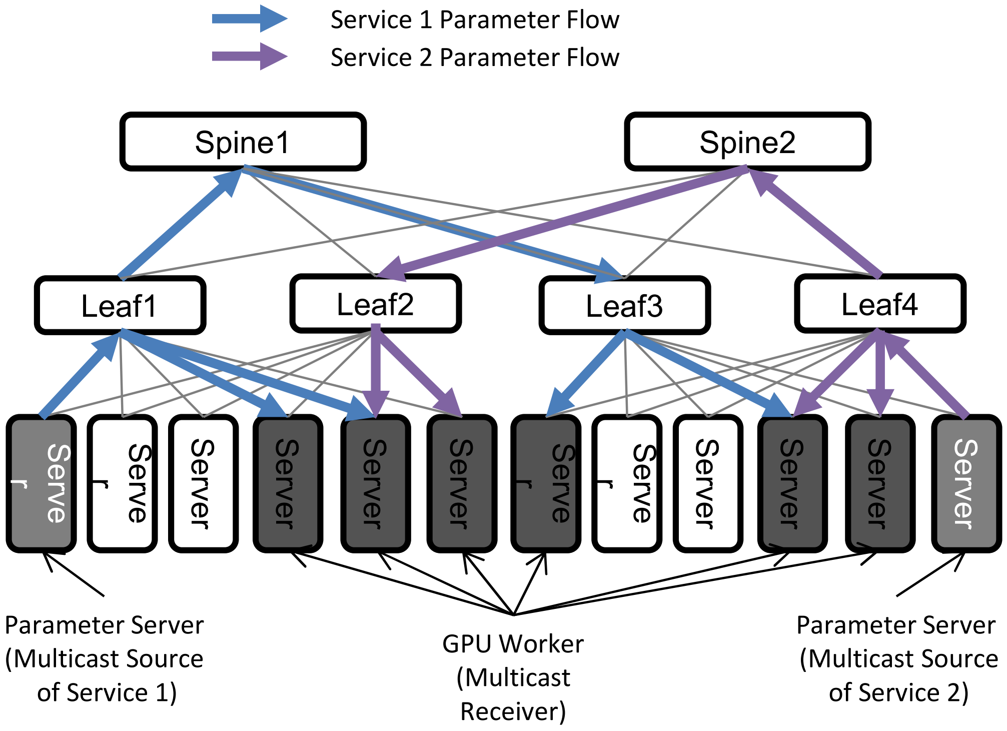
- Implemented a demo for MSR6 TE (Multicast Source Routing Traffic Engineering) and RLB (Replication through Local Bitstring) based on *P4*

- Conducted some simulations of these demos based on *real P4 switches*

- Documents
  - https://datatracker.ietf.org/doc/draft-eckert-msr6-problem-statement
  - https://datatracker.ietf.org/doc/draft-cheng-msr6-design-consideration/
  - https://datatracker.ietf.org/doc/draft-geng-msr6-traffic-engineering/02
  - https://datatracker.ietf.org/doc/draft-geng-msr6-rlb-segment/01

# New requirements: Multicast in large-scale DC Network

- Large-Scale DC Network could have numerous multicast services running, such as AI training, HPC (High Performance Computing) and SAN (Storage Area Network) scenarios in DCN
- Following is a use case for AI training



– In AI training scenarios, parameter server will push the parameters to all GPU workers after gradient descent.

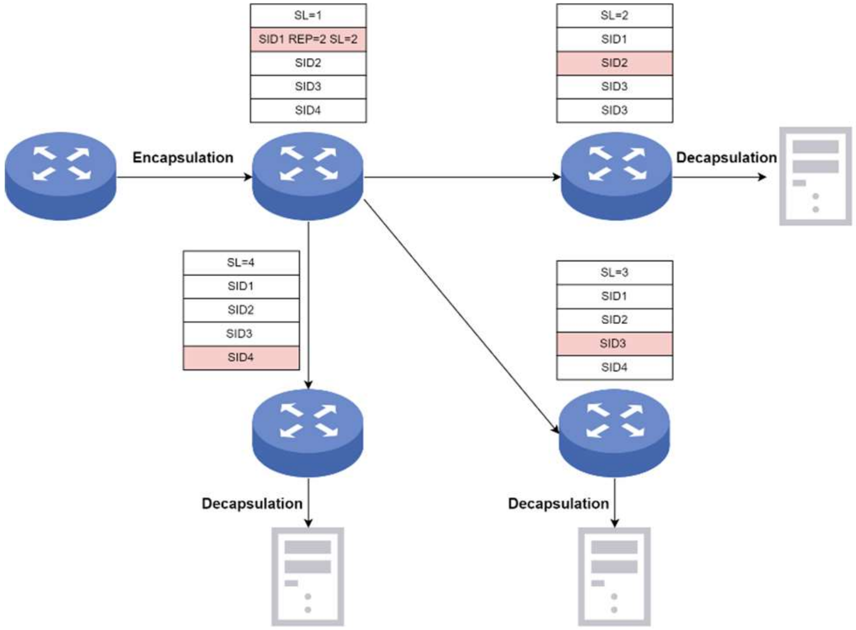
– Multicast can improve bandwidth usage and training speed in AI training scenarios:

  - **Network scale: Switches ~ 3k, links ~ 60k;**
  - **Large number of potential leaves (10k GPUs);**
  - **Sparse/Dense trees** (Number of GPUs in one service depends on specific trainning requirements**);**
  - **Large number of services** (Cloud AI).

# Potential Solution: Muticast Source Routiing over IPv6

- MSR6 leverages the benefits of source routing over IPv6 data plane to provide simplified multicast
- MSR6 TE is a TE solution for high quality traffic such as AI training, SAN in data center network

- Without unnecessary multicast tree status and complex control plane protocols.
- Provide traffic engineering capability.



- MSR6 TE has two implementations called TE and RLB. Different from TE, RLB uses a bitstring to indicate the forward path.

# Demo overview:MSR6 TE based on P4 Swiches

- We've implemented the demo based on P4, and conducted some experiments based on Tofino switches.

- Functions in Demo
  1. **Encapsulation:** The encapsulation of MRH(Multicast Routing Header) for specified packets
  2. **Replication:** Transit Nodes read MRH, clone packets and forward packets respectively
  3. **Decapsulation:** Leaf Nodes receive MSR6 packets and decapsulate them to original format
  4. **TE:** Path switching by modifying MRH encapsulated at Ingress Node.
  5. **Video Experiment:** Video stream replication and transportation.
  6. **Infiniband Experiment:** InfiniBand packets replication and transportation.

# Outcomes

- Simulation Videos:

We’ve uploaded the simulation videos of MSR6 TE to Youtube, you can get them through the following link and QR code.

MSR6 TE : https://youtu.be/m2L9BEwFKCA

QR code:

# Future Plan

- Form a testbed with RDMA devices.

Simulations:

1. Test the compatibility of MSR6 TE with real RDMA devices
2. Test the performance of MSR6 TE on P4 devices

# Wrap Up

Team members:

- Weihong Wu (wuweihong@bupt.edu.cn)
- Jiang Liu (liujiang@bupt.edu.cn)
- Jing Jia (jiajing@bupt.edu.cn)
- Yunyi Tang (tangyunyi0708@bupt.edu.cn)
- Sijia Li (lisijia@bupt.edu.cn)
- Yuxin Jiang (jyxin@bupt.edu.cn)
- Weiqiang Cheng（chengweiqiang@chinamobile.com）