Packet Spraying in Geneve Overlay Network

draft-xiang-nvo3-geneve-packet-spray-00

Haizhou Xiang , Huawei
Yolanda Yu, Huawei
Paul Congdon , Tallac Networks
Jianglong Wang , China Telecom

IETF 101, March 2018, London

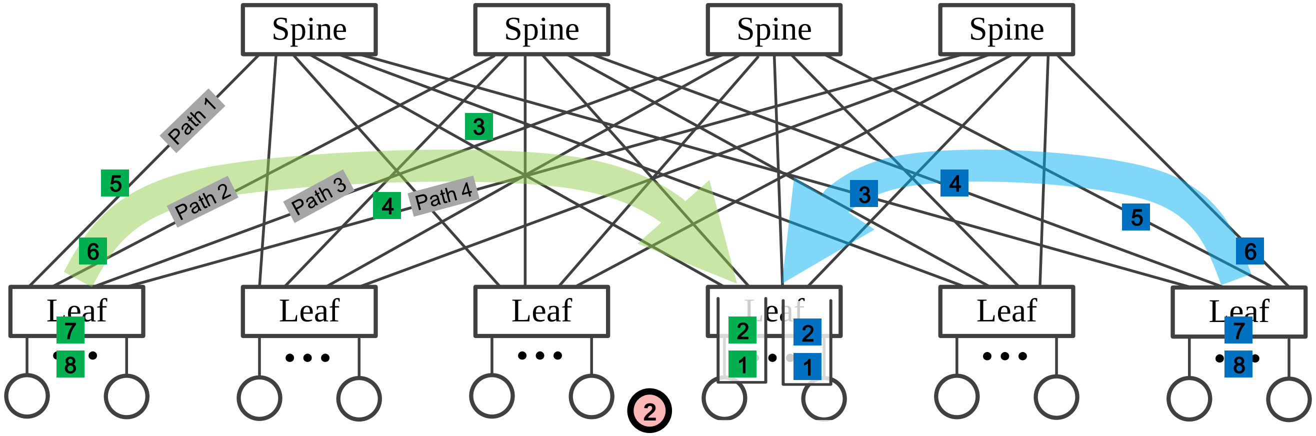
In-network Congestion

- In-network congestion : occurs within the interconnection network channels, due to poor traffic spraying.
- Path selection can be treated as load balancing issue
    - Load balancing technologies are used to solve in-network congestion: such as ECMP, Flowlet, Packet Spraying
    - Packet is both finer granularity and suitable for open system.
    - Packets belong to the same flow may go through different paths, which may lead to packets out of order.

Coping with In-network Congestion

- Packet Spraying (PS) = Packet Spraying + Reordering

Distributed | Finer Granularity | In-Ordering



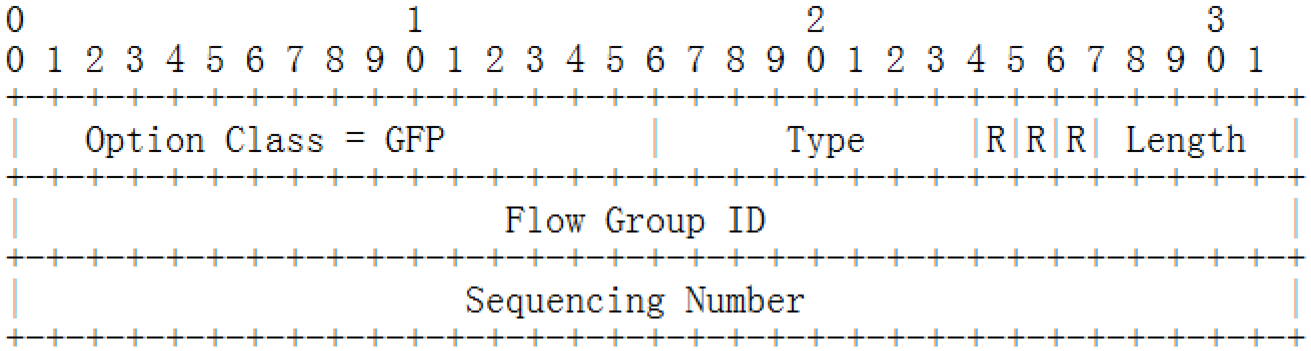
1 Spray packets over paths

2 Reordering @ Dst-end

- ○ Packet spraying at Src-end (Leaf Switch or Server)
  - No need to modify Spine switch
  - Use Geneve to encapsulate the packet Sn

- ○ Packet re-ordering at Dst-end (Leaf Switch or Server)
  - For those (protocol or OS), who can't tolerate packet reordering

Proposed Packet Spraying Format over Geneve



```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|   Option Class = GFP          |      Type     |R|R|R| Length  |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                        Flow Group ID                          |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                      Sequencing Number                        |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Packet Spraying Format over Geneve

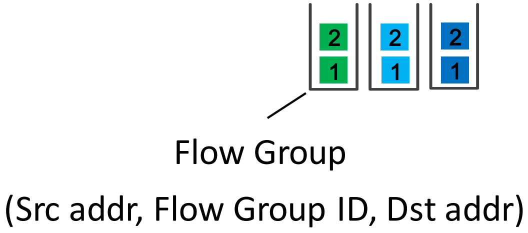
- Option Class = Geneve Forwarding Policy(suggested), to be assigned by IANA (TBA).
- Type = TBA.
- Length = 2 (8 byte)
- Flow Group ID: identifies a group of flows within the same reorder sequence space between a Src/Dst pair. A Flow Group is uniquely identified by the 3 tuple that includes Src address, Dst address and Flow Group ID.
- Sequence Number: value ranges from 0 to (2**32)-1

Packet Spraying function @ Src

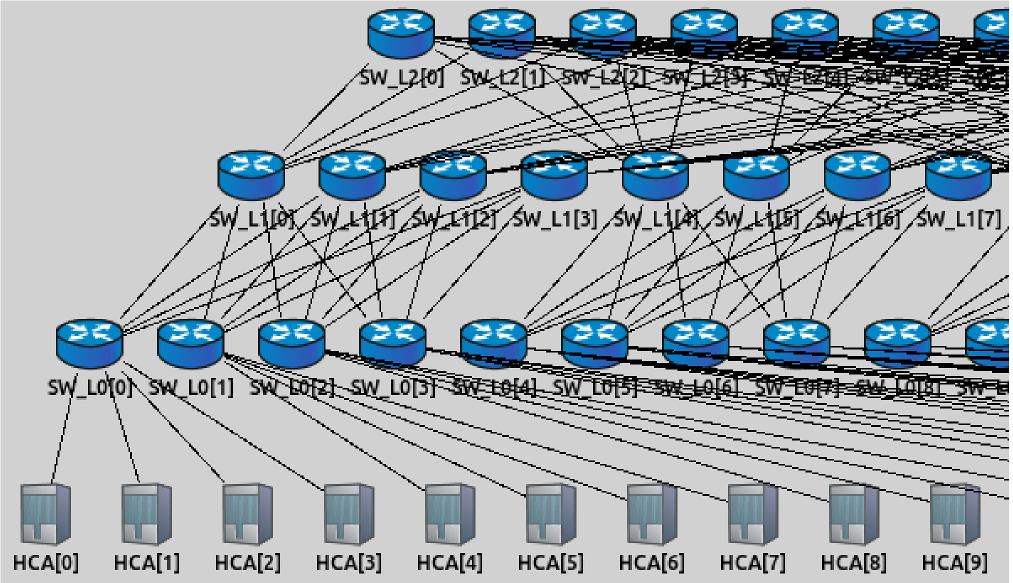
- The Flow Group ID may correspond to an individual flow, some subset of flows, or even all flows between the Src/Dst pair.
- How the flow corresponds to the Flow Group ID is not defined by this draft.
- The source node allocates the sequence number according to the order packets are sent for flows of the same Flow Group.

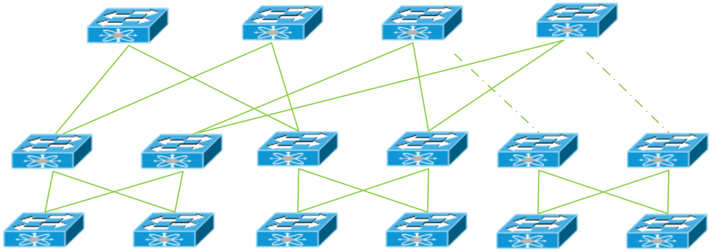
Reordering function @ Dst

- The destination perform reordering to the packet with same 3 tuple(Src addr, Dst addr, Flow Group ID) by sequence number.
- The destination needs to notify the capability (reorder queues assigned to the peer) to the source.
- The source needs to tune the allocation mechanism of Flow Group ID according to the capability of destination
- When the number of Flow Group IDs of received packets exceed the local capability:
  - Discard the Geneve packet for the Flow Group ID that exceeds the local capability
  - Remove the Geneve encapsulation, without performing reordering and pass the packet to higher layer protocol.



Simulation Set-up



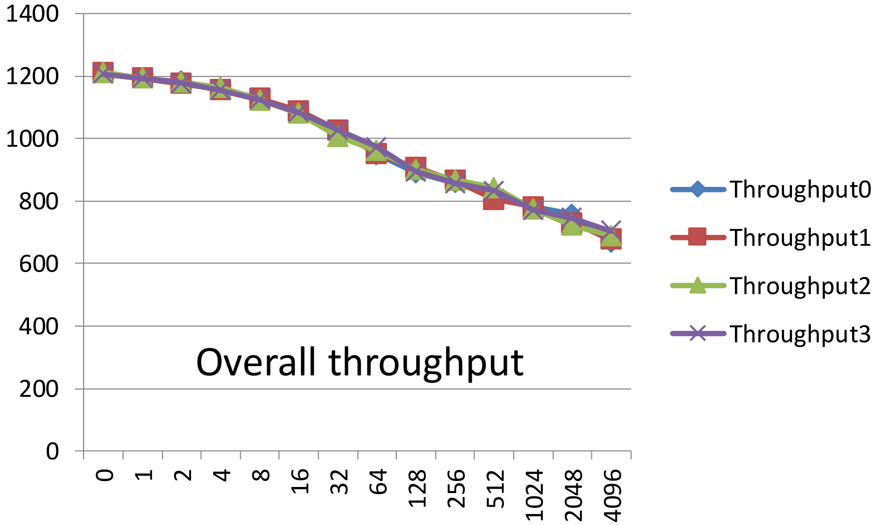


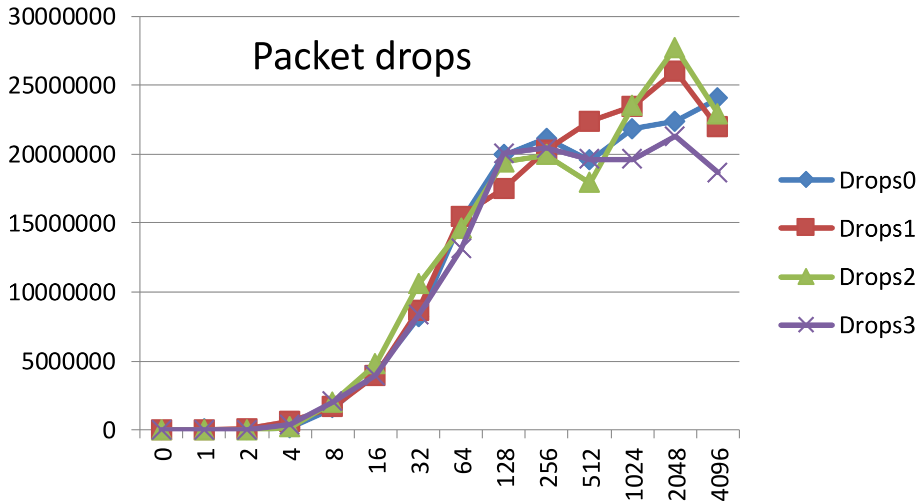
- **Platform:** OMNET++
- **3 Tier CLOS:** 10G interface, 16 Core SW, 32 Edge SW, 32 Leaf SW, 128 Server
- **Traffic Pattern:** UDP, Uniform random destination

Performance Comparison

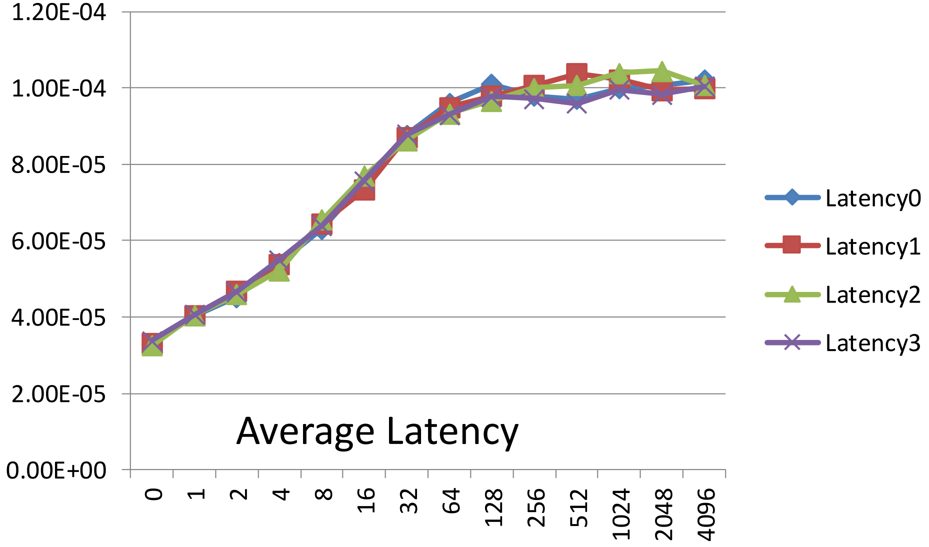
- Load balancing granularity
  - Packet Spray
    - Random select next hop for every packet
  - Sub-flow
    - Random select next hop for every 2^n packets
    - n = (0 ~ 12)
    - When n = 0, equal to packet spray. When n=12, close to ECMP.
  - ECMP
    - Select next hop by 5-tuple hash
- Performance factor
  - Overall throughput
  - Overall drops
  - Average latency

Performance comparison





- 4 rounds with different random seed
- Packet spray achieve best performance
- Sub-flow Random select next hop for every 2^n packets, with n increasing, close to ECMP
- In general, ECMP achieve worst performance, its overall throughput is the lowest.



Next Step

- Seek comments and more collaboration
- Continue the simulation on the packet reordering
- Validate the overall performance under a real test bed