A Hierarchical SW Load Balancing Solution for Cloud Deployment

Hongjun Ni
Intel
FD.io Sweetcomb PTL
FD.io NSH_SFC PTL

Acknowledgement:
Ray Kinsella, Steve Liang @Intel
Pierre Pfister, Jerome Tollet @Cisco
Ping Yu, Jeff Shaw @Intel
Agenda

• Why Choosing FD.io

• Option 1: Three-Level Load Balancing Solution
  • Router, Load Balancer and Service Proxy Implementation

• Option 2: Two-Level Load Balancing Solution

• Key Takeaway
Option 1: Three-Level Load Balancing Solution

- Router, Load Balancer and Service Proxy could be implemented based on VPP-DPDK.
- Supports IPv4 and IPv6.
- They run as typical K8s networking.
Option 1: Router on VPP-DPDK - Injecting Mode

- Creates a tap for each interface.
- Injects locally destined, broadcast and multicast traffic to host stack.
- BIRD receives and processes traffic using sockets on the taps.
- Uses Netlink to interact with Linux kernel stack.
- Rtinject plugin listens for Netlink address, link, neigh, and route messages.
- Mirror configuration onto VPP’s FIB.
- Supports IPv4, IPv6 and MPLS.
Option 1: Load Balancer on VPP-DPDK

- Distributes traffic among K8s nodes
- Consistent Hashing ensures resilience to K8s node changes.
- Connection Tracking supports connection persistence.
- Supports L3 and L4 load balancing.
- Supports two encapsulation types
  - GRE tunnel
  - IPIP tunnel
- Integrated with OpenStack as LBaaS and deployed at Yahoo Japan.
Option 1: Service Proxy on VPP-DPDK

- Distributes traffic among Pods
- Supports two interface types
  - vhost and virtio-user
  - memif
- Supports three service types
  - ClusterIP
  - NodePort
  - External LoadBalancer
- DNAT translates Service IP to Pod IP. SNAT does the opposite.

K8s-node

Backend Pod 1
Pod IP: x.x.x.x
- virtio-user
- vhost
- memif

Backend Pod 2
Pod IP: y.y.y.y
- virtio-user
- memif
- vhost
- memif

VPP-DPDK
- DNAT
- Connection Tracking
- SNAT
- Load balancing
- eth
Option 1: Data Flow

- VPP Router enables internal load balancing feature.
- VPP Load Balancer distributes traffic and encapsulates packets via GRE tunnels. A specific flow will be sent to the same K8s node.
- On K8s node, it removes GRE tunnel and goes through Service Proxy and performs DNAT, and then distributes traffic to selected pods.
- Return traffic will also pass through Service Proxy performing SNAT.
- Pod IPs are not visible to VPP LB.
Option 2: Two-Level Load Balancing Solution

- Each LB and K8s node runs iBGP
- LBs and K8s nodes do full mesh.
- Pod IPs are visible to VPP LB and K8s nodes.
Option 2: Data Flow

- VPP Router enables internal load balancing feature.

- VPP Load Balancer selects pod as per Service IP and encapsulates packets through IPIP tunnels.

- On K8s node, vRouter just forwards IPIP traffic to selected Pod.

- On each Pod, it removes IPIP tunnel, manipulates packets, swaps source IP and destination IP, and then sends packets directly to clients (Direct Server Return).
Key Takeaway

- A User Space solution to enable high performance Cloud Networking.

- Provides Three-Level load balancing solution for typical K8s usage.

- VPP-DPDK can implement Router, Load Balancer and Service Proxy.

- Provides Two-Level load balancing solution for Direct Server Return.
Thank you!

Q & A

Email: hongjun.ni@intel.com