Contribution ID: **156** Type: **not specified**

Bringing the Power of eBPF to Open vSwitch

Wednesday, 14 November 2018 11:35 (45 minutes)

Among the various ways of using eBPF, OVS has been exploring the power of eBPF in three: (1) attaching eBPF to TC, (2) offloading a subset of processing to XDP, and (3) by-passing the kernel using AF_XDP. Unfortunately, as of today, none of the three approaches satisfies the requirements of OVS. In this presentation, we'd like to share the challenges we faced, experience learned, and seek for feedbacks from the community for future direction.

Attaching eBPF to TC started first with the most aggressive goal: we planned to re-implement the entire features of OVS kernel datapath under net/openvswitch/* into eBPF code. We worked around a couple of limitations, for example, the lack of TLV support led us to redefine a binary kernel-user API using a fixed-length array; and without a dedicated way to execute a packet, we created a dedicated device for user to kernel packet transmission, with a different BPF program attached to handle packet execute logic. Currently, we are working on connection tracking. Although a simple eBPF map can achieve basic operations of conntrack table lookup and commit, how to handle NAT, (de)fragmentation, and ALG are still under discussion.

Moving one layer below TC is called XDP (eXpress Data Path), a much faster layer for packet processing, but with almost no extra packet metadata and limited BPF helpers support. Depending on the complexity of flows, OVS can offload a subset of its flow processing to XDP when feasible. However, the fact that XDP has fewer helper function support implies that either 1) only very limited number of flows are eligible for offload, or 2) more flow processing logic needed to be done in native eBPF.

AF_XDP represents another form of XDP, with a socket interface for control plane and a shared memory API for accessing packets from userspace applications. OVS today has another full-fledged datapath implementation in userspace, called dpif-netdev, used by DPDK community. By treating the AF_XDP as a fast packet-I/O channel, the OVS dpif-netdev can satisfy almost all existing features. We are working on building the prototype and evaluating its performance.

RFC patch:
OVS eBPF datapath.
https://www.mail-archive.com/iovisor-dev@lists.iovisor.org/msg01105.html

I agree to abide by the anti-harassment policy

Presenters: TU, William (VMware); STRINGER, Joe (Isovalent); WEI, Yi-Hung (VMware); SUN, Yifeng (VMware)

Session Classification: Networking Track