Lisbon, OpenWrt Summit 2018

State of fast path networking in Linux

Damir Samardžić

sartura

October 29, 2018

Agenda

- Linux kernel networking
- Linux kernel bypass
  - DPDK, netmap, Snabb, PF_RING
  - VPP (FD.io), Open vSwitch, OpenDataPlane, OpenFastPath
- Linux kernel fast path
  - eBPF, XDP

Linux kernel networking

- Linux kernel networking stack is complex
  - Big SKBs, interrupts, context switching – a lot of CPU cycles
- 10GbE, 25GbE, 40GbE and counting
- Offload techniques and optimization are not enough
- Filter traffic as early as possible or bypass the kernel?

Linux kernel bypass

- Idea – improve networking performance by going around the Linux networking stack
- Requires:
  - Modified device drivers and/or additional kernel modules
  - Handling upper protocol layers in your app

DPDK

- Data Plane Development Kit
- Multi-vendor, supports x86, ARM and PowerPC
- Runs in user space
- Memory huge pages, kernel UIO/VFIO module and poll mode drivers (PMDs)
- Open vSwitch, VPP (FD.io), OpenFastPath

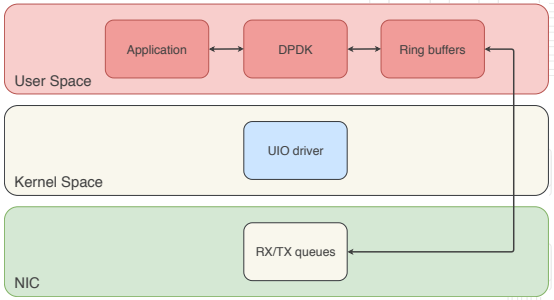

Image source: Understanding DPDK (Denys Haryachyy, February 2015)

VPP (FD.io)

- FD.io – Open source version of Cisco's Vector Packet Processing (VPP)
- Stack for commodity hardware (supports x64, ARM support WIP)
- Runs in user space, modular ***packet processing graph*** approach
- KVM and ESXi support, Vhost-user, netmap, virtio paravirtualized NICs, tun/tap drivers, DPDK PMDs

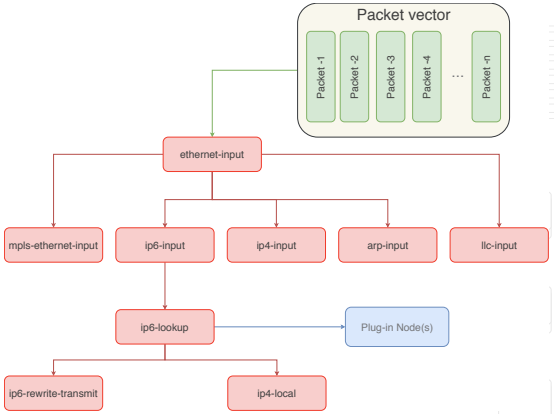

Image source: wiki.fd.io

Open vSwitch

- Open vSwitch
- Multi-layer virtual switch
- Supports DPDK and Linux devices
- Transparent distribution across multiple physical servers by creating cross-server switches
- Used in virtualization platforms

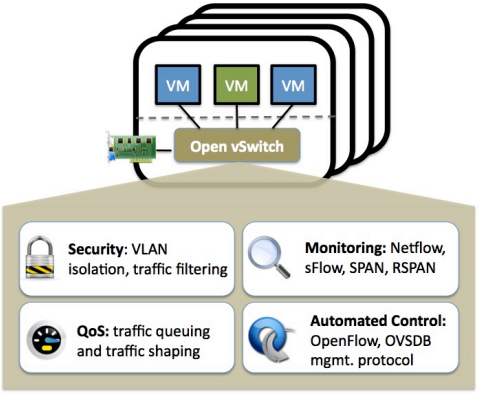

Image source: openvswitch.org

OpenDataPlane

- OpenDataPlane
- Open Source, cross-platform set of APIs for networking data plane
- Supports standard Linux API, DPDK, vendor-specific implementation
- ARMv7, ARMv8, MIPS64, PowerPC, x86

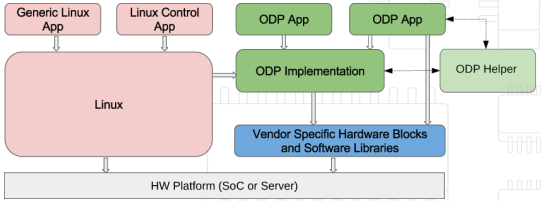

Image source: opendataplane.org

OpenFastPath

- OpenFastPath
- Open Source implementation of a high-performance TCP/IP stack
- Library to fast path applications that use OpenDataPath and DPDK
- Linux integration via TAP (slow), routes and MAC in sync with netlink
- x86, ARM, PowerPC, MIPS

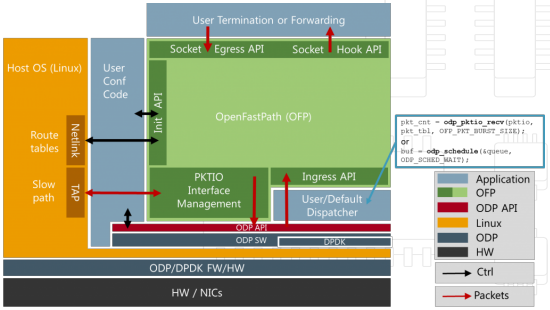

Image source: openfastpath.org

netmap

- netmap
- In kernel mode – comes as several kernel modules
- Pre-allocated fixed size buffers
- Uses memory region shared by user processes
- Zero-copy between interfaces

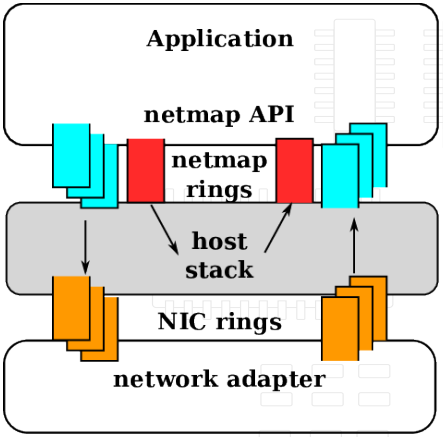

Image source: info.iet.unipi.it

Snabb

- Snabb
- Toolkit for developing network functions in user space
- Linux x86/64 supported
- User space drivers (apps) for supported NICs
- A program (LUA) formed as graph of applications

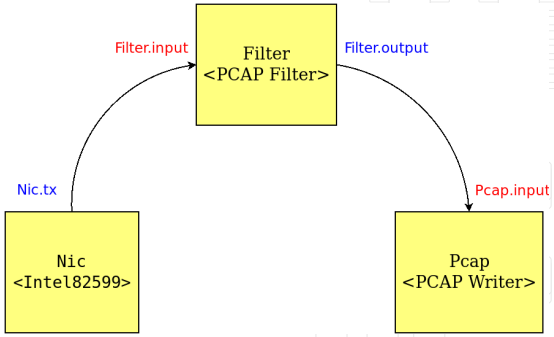

Image source: Snabb explained in less than 10 minutes

PF_RING

- PF_RING - ntop
- Separate Linux kernel module
- Set of drivers for several NICs
- Packet capturing, active traffic analysis and manipulation
- Zero-copy possible (Intel), requires huge pages support
- Alternative BPF - nBPF

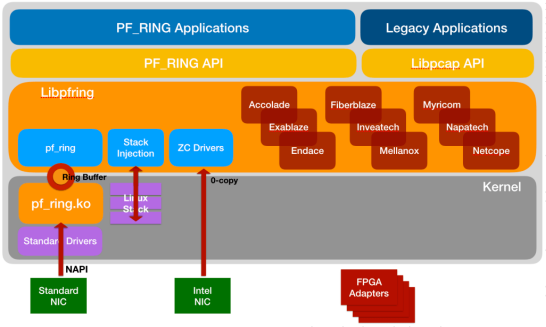

Image source: ntop.org

Linux kernel fast path

- Idea – process as much data early on the data path
  - In Linux driver code, before SKB allocation
  - On NIC itself
- Used for packet inspection and filtering, DoS protection, ACLs
- Linux kernel build-in features: XDP and eBPF

eBPF

- Linux Socket Filtering aka Berkeley Packet Filter (BPF)
- Flexible, efficient, VM-like construct in Linux kernel
- Allows safe bytecode execution at various hook points

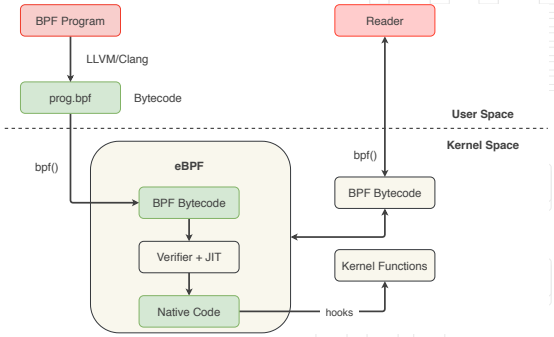

Image source: The BSD Packet Filter (Suchakra Sharma, June 2017)

- Used also for debugging and performance analysis (kprobe, uprobe, perf events)
- eBPF programs can attach to:
  - traffic control (tc) subsystem
  - tunnels
  - earliest networking driver stage via fast data path subsystem called eXpress Data Path (XDP)

XDP

- eXpress Data Path
- Processing RX packet-pages directly out of driver
- Before SKB allocation
- Native (driver supports XDP), offloaded (BPF into NIC), generic
- Uses eBPF programs, in-kernel security model

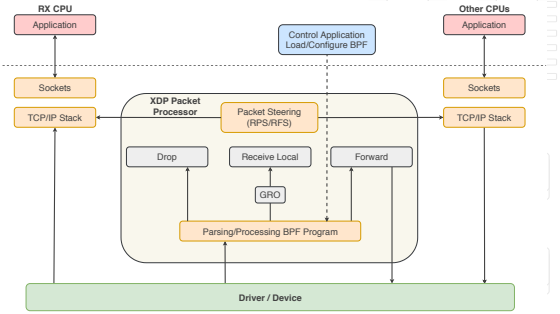

Image source: iovisor.org

- Actions:
  - XDP_PASS, XDP_DROP, XDP_TX, XDP_ABORTED, XDP_REDIRECT
- Use cases:
  - DDoS protection, load balancing, traffic sampling and monitoring
- AF_XDP – new socket for getting packets to user space (via XDP)
  - New allocator UMEM (used in driver) allows zero-copy
  - DPDK PMD for AF_XDP

OpenWrt

- Kernel bypass solution
  - DPDK, VPP and other technologies not packaged for OpenWrt
  - netmap provides Makefile in official source repository
- XDP and eBPF in mainline kernel, but driver support is lacking
  - Most drivers on Linux 4.14 support XDP_PASS, XDP_DROP and XDP_TX
  - Only Intel ixgbe supports XDP_REDIRECT
  - As of Linux 4.19 XDP_REDIRECT supported in Intel i40e, Intel ixgbe and Mellanox mlx5 drivers

Sartura & fast data path

- Measurements – from 1.5x to 5.5x improvement
- eBPF application development
- XDP driver enablement
  - Upcoming mvneta driver XDP support
- Training materials
- ISP collaboration

State of fast path networking in Linux

damir.samardzic@sartura.hr · info@sartura.hr · www.sartura.hr