APPLICATION NOTE

Intel Corporation

Intel® Ethernet Controller 700 Series - Open vSwitch Hardware Acceleration Application Note

Authors

Mesut Ergin

Harry Van Haaren

Charlie Tai

1 Overview

New generation high speed data plane libraries, such as Data Plane Development Kit (DPDK), and their applications specializing in switching and routing software [for example, Open vSwitch* (OVS*)] have improved efficiency of software-defined networking significantly, yielding better utilized server platforms. Increasing the number of cores available in a typical server CPU, applications requiring higher packet switching performance result in a non-trivial amount of data communication load within the boundaries of the platform. These packets are traversing in and out of network interfaces (such as, North-South traffic), as well as across cores (such as, East-West traffic) at rates of hundreds of millions per second. Decomposition trends in modern software design, like the examples seen in Cloud Native applications, will impose additional challenges for data communication efficiency.

This application note describes using the capabilities of the Intel® Ethernet Controller 700 Series based Network Interface Cards (NICs) to accelerate packet switching performance using the flow hardware offload feature of OVS with DPDK.

Note: In this document, the term acceleration describes a method of processing data on application-specific hardware, instead of on a generic CPU.

The phrase flow hardware offload refers to a certain feature of OVS with DPDK, which uses the `rte_flow` API to process data on the Intel® Ethernet Controller 700 Series based NIC.

Section 2 describes the design of OVS DPDK flow hardware offload, and Section 3 provides details on how to enable OVS DPDK flow hardware offload using Intel® Ethernet Controller 700 Series based NICs.

This document is part of the Network Transformation Experience Kit, which is available at: https://networkbuilders.intel.com/network-technologies/network-transformation-exp-kits

Table of Contents

Figures

Tables

1.1 Terminology

Table 1. Terminology

ABBREVIATION	DESCRIPTION
API	Application Programming Interface
dpcls	Datapath Classifier
DPDK	Data Plane Development Kit
EMC	Exact Match Cache
FD	Flow Director
I/O	Input/Output
NIC	Network Interface Card
OVS	Open vSwitch
OVSDB	OVS Rule Database
PMD	Poll-Mode Drivers
RSS	Receive Side Scaling
SMC	Signature Match Cache
VF	Virtual Function

1.2 Reference Documents

Table 2. Reference Documents

REFERENCE	SOURCE
Flow Hardware Offload in the Using Open vSwitch with DPDK Documentation	http://docs.openvswitch.org/en/latest/howto/dpdk/]
Intel® Ethernet Controller XL710 datasheet	https://www.intel.com/content/dam/www/public/us/en/documents/datasheets/xl710-10-40-controller-datasheet.pdf
Intel® Ethernet Controller 700 series firmware version 6.01 or newer	https://downloadcenter.intel.com/product/75021/Intel-Ethernet-Controller-XL710-Series
Intel® Network Adapter Driver for PCIe 40 Gigabit Ethernet Network Connections Under Linux	https://downloadcenter.intel.com/download/24411/Intel-Network-Adapter-Driver-for-PCIe-40-Gigabit-Ethernet-Network-Connections-Under-Linux-

2 OVS* with DPDK: Flow Hardware Offload

2.1 OVS Cache Overview

OVS uses DPDK to implement fast packet Input/Output (I/O) through Poll-Mode Drivers (PMD) in the userspace and pushes these received packets through a multi-tier optimized lookup and action processing pipeline in the software. Switching rules defined by OVS users or network management software are installed into OVS rule database (OVSDB) in the form of Open Flow rules. Received packets are checked against these installed rules for a potential match. If the first packet of a given flow matches a rule, OVS furnishes multiple levels of lookup-cache with pertinent information, such that subsequent packets of the flow can be matched much quicker than the first one.

Currently, there are two levels of caching implemented in OVS lookup pipeline. Exact Match Cache (EMC) and Signature Match Cache (SMC) constitute the first level of caches. The first-level cache provides non-wildcard type matches, and offers fast turnaround if total number of flows present on the network does not exceed the cache capacity. Datapath Classifier (dpcls) is the second level caching, and implements tuple-space search with arbitrary bitwise matching on packet header fields. This classifier is also referred to as OVS mega-flow cache. Packets that miss both levels of the cache are subject to full Open Flow rule processing (such as, ofproto classification) for looking up a potential match.

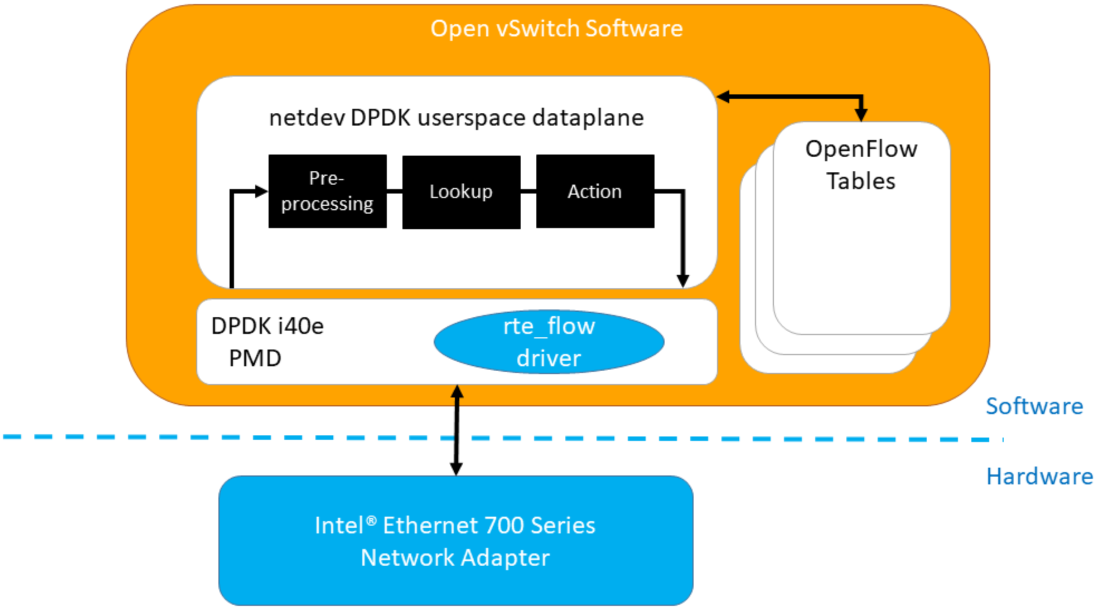

Figure 1. OVS Packet Processing Overview

2.2 OVS Miniflow Extraction

Before any of the caches are looked up, all received packets are subject to OVS's pre-processing called miniflow extraction. This process parses packet header fields and extracts them to store into a miniflow data structure, which is a compressed representation to reduce the memory footprint and optimize CPU cache access. Once miniflow representation of a packet is created, lookups can proceed, starting with the first level cache (such as EMC/SMC).

2.3 DPDK Generic Flow API: rte_flow

DPDK provides a generic way to configure supported hardware to match specific ingress or egress traffic, alter its fate and query related counters according to any number of user-defined rules. The API `rte_flow` allows users to request matches on packet data (protocol headers, payload) and properties (for example, associated physical port, virtual device function ID). Once matched, possible operations include dropping traffic, diverting it to specific queues to virtual/physical device functions or ports, performing tunnel offloads, adding marks, etc.

2.4 rte_flow Support for Intel® Ethernet Controller 700 Series

As part of the DPDK userspace device driver for the Intel® Ethernet Controller 700 Series, the `rte_flow` API is supported and exposes various match and actions capabilities in the NIC.

For the most typical Ethernet/IP/UDP/TCP based flow matching, the i40e `rte_flow` driver relies on Intel® Ethernet Flow Director (FD) capabilities of the Intel® Ethernet Controller 700 Series. Intel® Ethernet FD supports advanced filters that can match N-tuple flows, and carry out various actions.

2.5 OVS DPDK and rte_flow API

OVS DPDK has recently added support for flow offloads to hardware by making use of the `rte_flow` API. The feature is experimental and disabled by default as of this writing.

For details, refer to "Flow Hardware Offload" in the Using Open vSwitch with DPDK documentation (refer to Table 2).

Specifically, when OVS matches a packet with one of the Open Flow rules for the first time, the flow that is recorded into the megaflow cache is noted as a potential candidate for offloading to the hardware device whose port was used to receive this initial packet. To amortize the cost of configuring I/O devices, OVS will try to service the accumulated offload candidates in a timely manner and utilize `rte_flow` to program them to hardware.

Before programming the flow to the hardware, the OVS creates a 32 bit identifier (such as Mark ID) that can later be referenced as a shortcut for a completed lookup process. The megaflow cache entry that was just inserted can be translated to a `rte_flow` pattern, and the mark ID is passed along as part of a MARK `rte_flow` action object. Once the installation of this `rte_flow` rule to the hardware is successful, the OVS may start finding a non-zero mark ID value in the received packet's metadata structure. This is because the hardware has detected a match to a previously programmed flow. This mark ID can be used to bypass the multi-level software lookup process, and OVS directly moves to the action processing part of the pipeline, since the mark ID was generated as a result of a previous lookup. Because the flow action for packets of the offloaded flow is still carried out by the OVS software pipeline, this acceleration style is also referred to as partial offload. The modified flow processing of OVS DPDK with flow hardware offload is depicted in Figure 2.

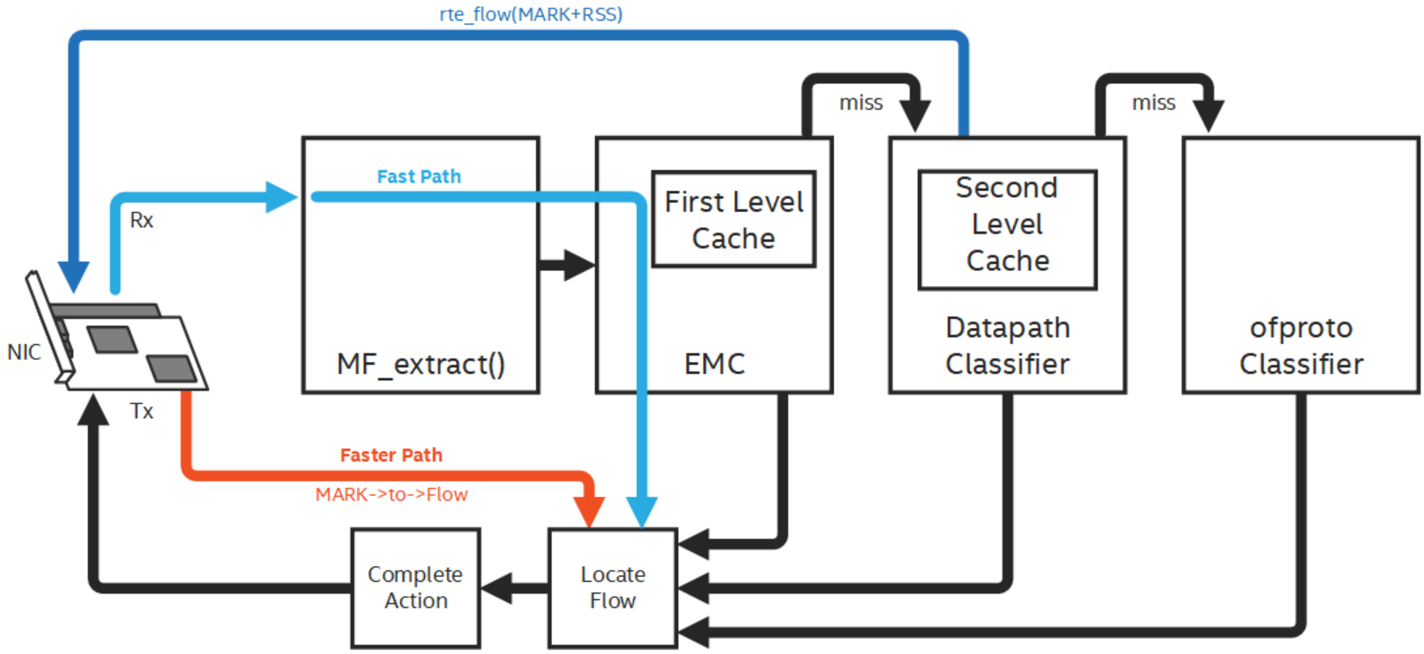

Figure 2. OVS DPDK Flow Offload Provides a Faster Way to Lookup

Due to hardware limitations and capacity, not all flow candidates for offloading may be successfully programmed into the hardware. As of this writing DPDK `rte_flow` doesn't support enumeration of hardware capabilities, therefore, OVS opportunistically tries offloading all flows, once hardware offloading feature is enabled. The flows programmed into hardware successfully are subject to OVS housekeeping. For example, as flows are modified, deleted, or aged in the megaflow cache, OVS must maintain the hardware state accordingly, to mirror the needed changes for correct operation.

3 Usage Details

As of DPDK v19.08 release, the Intel® Ethernet Controller 700 Series device driver i40e has added support for `rte_flow` actions including MARK and RSS (Receive Side Scaling), even at the same time. Therefore, OVS releases after v2.10 can be used along with DPDK v19.08 to benefit from flow hardware offloading with the Intel® Ethernet Controller 700 Series devices. For ease of compiling OVS with the most recent DPDK versions, OVS project maintains a non-release branch named dpdk-latest at: https://github.com/openvswitch/ovs

OVS releases strive to use DPDK LTS versions, and the first LTS release that will support OVS DPDK flow hardware offload will be DPDK 19.11, which will be consumed by OVS 2.13.

In the DPDK v19.08 release, `rte_flow` MARK support is only available for the non-vector receive path. Therefore, before compiling DPDK for OVS purposes, users need to opt-out of vector receive path in i40e PMD, by editing DPDK config/common_base file as follows:

```
CONFIG_RTE_LIBRTE_I40E_INC_VECTOR=n
```

Figure 3. Disabling Vector RX Path in DPDK

Once DPDK is compiled and included in the OVS compilation steps, vswitch configuration can be made to indicate flow hardware offloading is desired, by invoking ovs-vsctl as follows:

```
$OVS_VSCTL --no-wait set Open_vSwitch . other_config:hw-offload=true
```

Figure 4. Enable OVS Hardware Flow Offload

Once an Intel® Ethernet Controller 700 Series device is added as a port to a switch instance, flows will be attempted for programming to hardware, as packets are received and matched to the Open Flow rules programmed in the switch. Further debugging output for flows as they are processed for programming to the hardware can be obtained by enabling the following syslog debug options through `ovs-appctl` as follows:

```
$OVS_APPCTL vlog/set dpdk:syslog:dbg
$OVS_APPCTL vlog/set netdev_dpdk:syslog:dbg
```

Figure 5. Enable Syslog Debugging Using ovs-appctl

4 Summary

As network software requires higher packet switching performance, high performance data plane processing elements, such as those provided by DPDK, offer various acceleration capabilities.

In this application note, the Intel® Ethernet Controller 700 Series based NICs are shown to accelerate the OVS with DPDK using its flow hardware offload feature.

By utilizing hardware capabilities in Intel 700 Series Ethernet devices, we are accelerating classification and lookup process for network packets, as it is at the very heart of software packet switching.

You may not use or facilitate the use of this document in connection with any infringement or other legal analysis concerning Intel products described herein. You agree to grant Intel a non-exclusive, royalty-free license to any patent claim thereafter drafted which includes subject matter disclosed herein.

No license (express or implied, by estoppel or otherwise) to any intellectual property rights is granted by this document.

Intel disclaims all express and implied warranties, including without limitation, the implied warranties of merchantability, fitness for a particular purpose, and non-infringement, as well as any warranty arising from course of performance, course of dealing, or usage in trade.

This document contains information on products, services and/or processes in development. All information provided here is subject to change without notice. Contact your Intel representative to obtain the latest forecast, schedule, specifications and roadmaps.

The products and services described may contain defects or errors known as errata which may cause deviations from published specifications. Current characterized errata are available on request. No product or component can be absolutely secure.

Copies of documents which have an order number and are referenced in this document may be obtained by calling 1-800-548-4725 or by visiting www.intel.com/design/literature.htm

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Learn more at http://www.intel.com/ or from the OEM or retailer.