Corigine

WHITE PAPER

Corigine 25GbE SmartNICs with Open vSwitch Hardware Offload Drive Unmatched Cloud and Data Center Infrastructure Performance

NETRONOME AGILIO CX 25GBE SMARTNICS SIGNIFICANTLY OUTPERFORM MELLANOX CONNECTX-5 25GBE NICS UNDER HIGH-STRESS CLOUD WORKLOAD CONDITIONS

CONTENTS

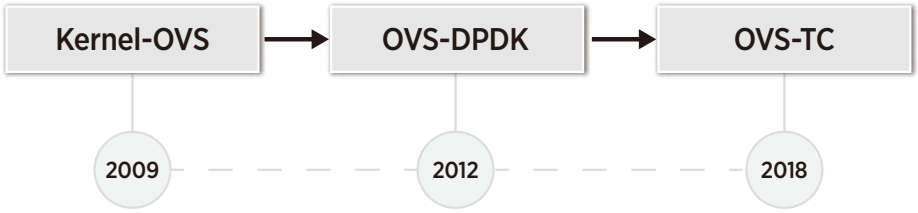
EXECUTIVE SUMMARY

The exponential rise in data volumes and variety of application workloads creates significant performance demands on virtualized cloud and data center infrastructures. In order to support this growth at scale, cloud and data center networks must be able to efficiently handle tens-of-thousands to millions of network flows and rules. These performance, efficiency and scalability requirements have pushed cloud providers to quickly move their Ethernet networking infrastructures from 10Gb/s to 25Gb/s with its better cost-to-bandwidth ratio and simplicity. Server-based networking utilizing 25GbE SmartNICs with full data plane offload capabilities further frees CPU resources to support additional users and process more data.

Open vSwitch (OVS) is the industry's open source standard for server-based networking and available for all major enterprise Linux distributions. OVS uses the Linux traffic control (TC) kernel packet classification engine to transparently offload flows and traffic control actions directly to hardware and SmartNICs, relieving the CPU of the added network burden. Only Corigine Agilio CX 25GbE SmartNICs can provide the high-performance offload required for software defined switching with the complexity of tens-of-thousands to hundreds-of-thousands of network flows and rules. This paper provides detailed benchmarks showcasing how Corigine Agilio SmartNICs with OVS-TC hardware offload outperform the competition.

ABOUT OVS

OVS is a widely deployed example of an SDN-controlled virtual switch for server-based networking. The benefits of OVS for server-based networking deployments have been well established: software-defined flexibility and control of datapath functions, fast feature rollouts, and the benefits of open source ecosystems.



The OVS kernel module (Kernel-OVS) is the most commonly used OVS datapath. Kernel-OVS is implemented as a match/action forwarding engine based on flows that are inserted, modified or removed by user space. In 2012, OVS was further enhanced with another user space datapath based on the data plane development kit (DPDK). The addition of OVS-DPDK improved performance but created some challenges. OVS-DPDK bypasses the Linux kernel networking stack, requires third-party modules and defines its security model for user space access to networking hardware. DPDK applications are more difficult to configure optimally and while OVS-DPDK management solutions do exist, debugging can become a challenge without access to the tools generally available for the Linux kernel networking stack. It has become clear that a better solution is needed.

OVS using traffic control (TC) is the newest kernel-based approach and improves upon Kernel-OVS and OVS-DPDK by providing a standard upstream interface for hardware acceleration. This paper will discuss how an offloaded OVS-TC solution performs against software-based OVS-DPDK.

OVS SOFTWARE-BASED SOLUTIONS: THE CPU BOTTLENECK

Kernel OVS

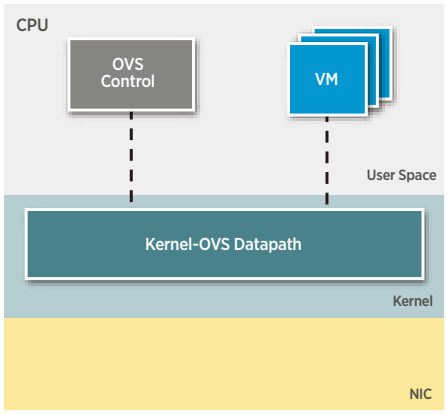


Figure 1. Kernel-OVS datapath

Following its initial release as an open source project in 2009, OVS has become the most ubiquitous virtual switch (vSwitch) in Linux deployments. The standard OVS architecture consists of user space and kernel space components. The switch daemon runs in user space and controls the switch while the kernel module implements the OVS packet datapath.

In a software-based solution, the kernel is not ideal for virtual switching because it grants short time quanta to the processes and treats them like any other process in the system. As a result, it imposes excessive contention, of which the latency is greater than the actual runtime of a service. Additionally, virtual switching requires frequent, per-packet system calls that cause the vSwitch to yield the CPU to the operating system (OS) so that the OS can perform the necessary I/O operations. This leads to degraded network performance.

OVS-DPDK

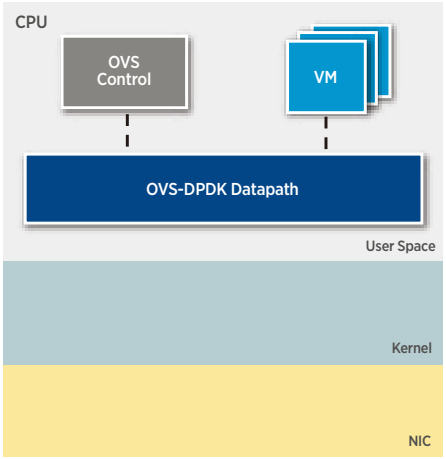


Figure 2. OVS-DPDK datapath

OVS-DPDK is a noteworthy attempt to address the fundamental limitations of Kernel-OVS. By implementing the OVS datapath in user space, the DPDK poll mode driver delivers packets directly into the dedicated user space application, bypassing the Linux kernel stack altogether. This eliminates unnecessary overhead in the stack and can enable additional optimizations for the vSwitch, such as loading packets directly into caches and batch processing. The DPDK community regularly provides optimizations and tuning for OVS with upstreamed DPDK setup for network interface cards (NICs).

OVS-DPDK enables software acceleration, and in a few cases, the user can tune it to be an adequate "workaround" for the major Kernel-OVS performance bottlenecks. DPDK generally uses dedicated logical cores to gain sufficient networking performance, which places a limit on the scalability of DPDK as a software-accelerated virtual switching solution. Moreover, OVS-DPDK is not part of the Linux kernel, and this solution is rendered cumbersome with higher operational overhead.

It is important to note that for OVS-DPDK to run closer to line-rate performance, it has to consume more CPU cores. This is the true hidden cost of deploying OVS-DPDK. As shown in the benchmarks, even when sacrificing precious CPU cores, this solution is not capable of performing in a scaled-up data center environment.

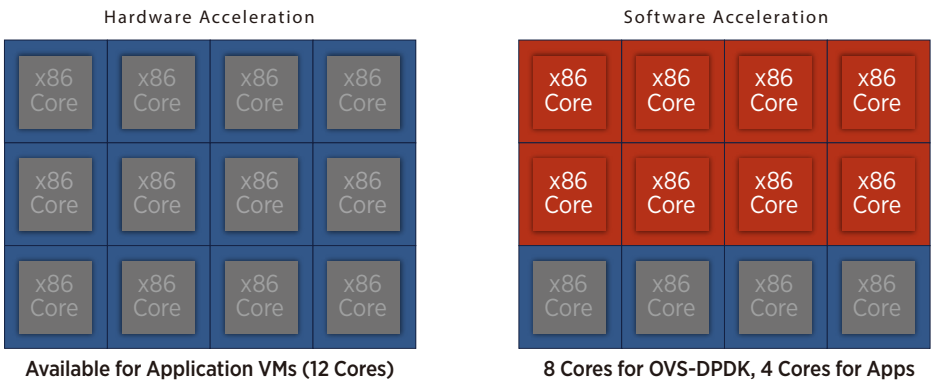


Figure 3. True performance cost of software acceleration on a single CPU socket

OVS-TC

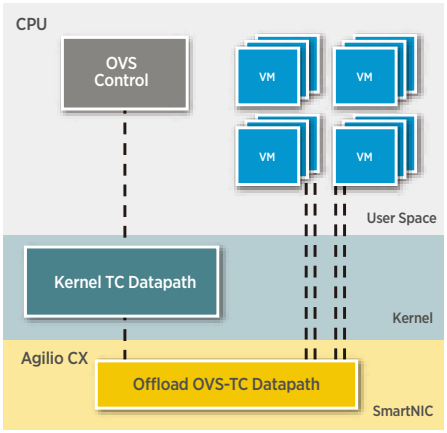


Figure 4. OVS-TC datapath

Linux TC is the kernel packet classification engine and TC Flower is the extension of TC that enables it to offload flows, and TC actions directly to hardware. OVS-TC allows matching on a variety of predefined flow keys. The user can match IP addresses, UDP/TCP ports, meta-data and more. Similar to OVS, OVS-TC includes an action side which allows packets to be modified, forwarded or dropped. The user can influence what is offloaded or not down to a per flow basis.

The TC command line provides a common set of tools for configuring queuing disciplines, classifiers and actions. The TC Flower classifier, combined with actions, may be used to provide match/action behavior similar to Kernel-OVS and OVS-DPDK. OVS leverages the TC Flower datapath to gain hardware acceleration.

Service providers need a scalable vSwitch, and now there is an open source, upstreamed and kernel-compliant solution with OVS-TC which maintains all the benefits of Kernel-OVS and OVS-DPDK. In addition, hardware-accelerated OVS-TC provides better CPU efficiency, lower complexity, enhanced scalability and increased network performance.

	KERNEL-OVS	OVS-DPDK	OVS-TC OFFLOAD
CPU Efficiency	✗	✗	✔
Low Optimization Complexity	✔	✗	✔
Scalability	✗	✔	✔
High Performance	✗	✗	✔

UPSTREAMED OVS-TC OFFLOAD: THE SMARTNIC REVOLUTION

OPEN SOURCE AND UPSTREAMED SOLUTIONS LIKE OVS-TC REPRESENT THE FUTURE OF OUR INDUSTRY BECAUSE THEY ARE EASIER TO IMPLEMENT AND NON-PROPRIETARY.

Corigine Agilio CX SmartNICs enable transparent offload of the TC datapath. While OVS software still runs on the server, the OVS-TC datapath match/action modules are synchronized down to the Agilio SmartNIC via hooks provided in the Linux kernel. The 60 cores (480 threads) on the Agilio CX SmartNIC provide industry-leading hardware acceleration, consume less than 25W and deliver groundbreaking ROI.

OVS-TC hardware acceleration on the Agilio SmartNIC is supported with a wide range of features. As the OVS community adds more features to OVS-TC, Corigine enables acceleration of those features with firmware upgrades. In the appendix of this whitepaper, there is a comprehensive list of the match/action-supported features for OVS-TC.

Open source and upstreamed solutions like OVS-TC represent the future of our industry because they are easier to implement and non-proprietary. The traffic classification subsystem contained on TC makes it possible to use other types of classifiers to implement the matching of packets. An example of this would be BPF, which uses a bpfilter to match packets.

AGILIO TRANSPARENT OVS OFFLOAD ARCHITECTURE

By running the OVS-TC data functions on the Agilio SmartNIC, the TC datapath is dramatically accelerated while leaving higher-level functionality and features under software control. Hardware acceleration achieves significant performance improvements while retaining the leverage and benefits derived from a sizable open source development community.

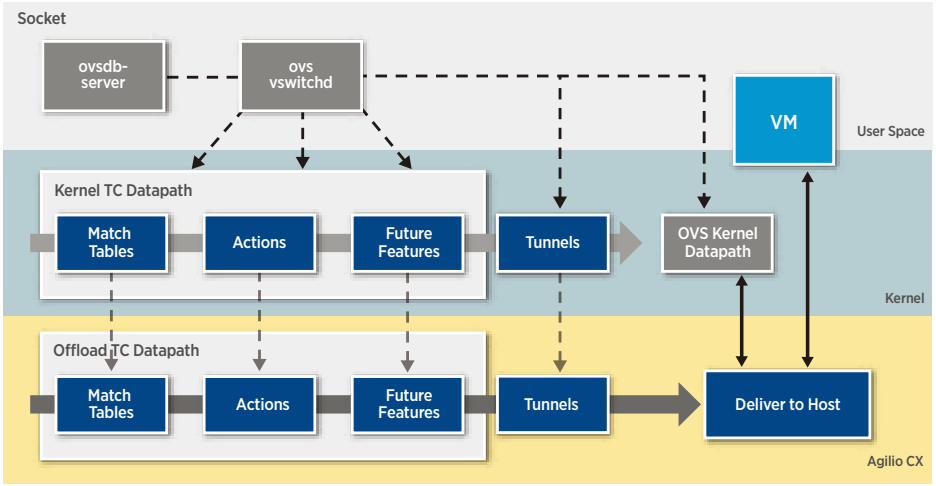


Figure 5. The Agilio transparent OVS-TC offload architecture

When using OVS-TC, there are three datapaths present:

1. SmartNIC datapath
2. TC kernel datapath
3. OVS kernel datapath

OVS contains a user space-based agent and a kernel-based datapath. The user space agent is responsible for switch configuration and flow table population. As observed in Figure 5, it is broken up into two fundamental components: ovs-vswitchd and ovsdb-server. The user space agent can accept configuration via a local command line (e.g. ovs-vsctl) and from a remote controller. When using a remote controller, it is common to use ovsdb to interact with ovsdb-server for vSwitch configuration. The kernel TC datapath is where the Agilio offload hooks are inserted. With this solution, the OVS software still runs on the server, but the OVS-TC datapath match/action modules are synchronized down to the Agilio SmartNIC via hooks provided in the Linux kernel.

BENCHMARK METHODOLOGY AND TEST SETUP

(PHY-VM-PHY)

Corigine has put together a set of tests that provide performance data comparing Netronome Agilio CX 2x25GbE SmartNICs, and Mellanox ConnectX-5 2x25GbE NICs offloading OVS solutions. A PHY-VM-PHY with and without VXLAN topology was tested on both NICs. The fundamental goal of this benchmarking was to quantify performance capabilities of the offload datapath on real-world scenarios with many flows and OVS rules.

The following hardware and software were used for the test:

Server: Dell PowerEdge R730

CPU: 2X Intel® Xeon® CPU E5-2630 v4 @ 2.20GHz | 10 Cores per CPU

Software:
OS Version: Ubuntu 18.04.1, RHEL 7.5
Kernel Version: v4.15 (Ubuntu), v3.10 (RHEL)
Open vSwitch: v2.9.2 (https://github.com/openvswitch/ovs.git)
DPDK Release: 17.11.2 (git://dpdk.org/dpdk-stable)

NICs:
a) Corigine Agilio CX 2x25GbE SmartNIC
b) Mellanox ConnectX-5 EN 2x25GbE NIC

Test Generators:
TRex Realistic Traffic Generator (DPDK-based, open source)
PROX (Packet pROcessing eXecution Engine) (DPDK-based, open source)
2X TX/RX (phy,vhost)

The benchmarking setups used for accelerated OVS-TC are shown in Figure 6:

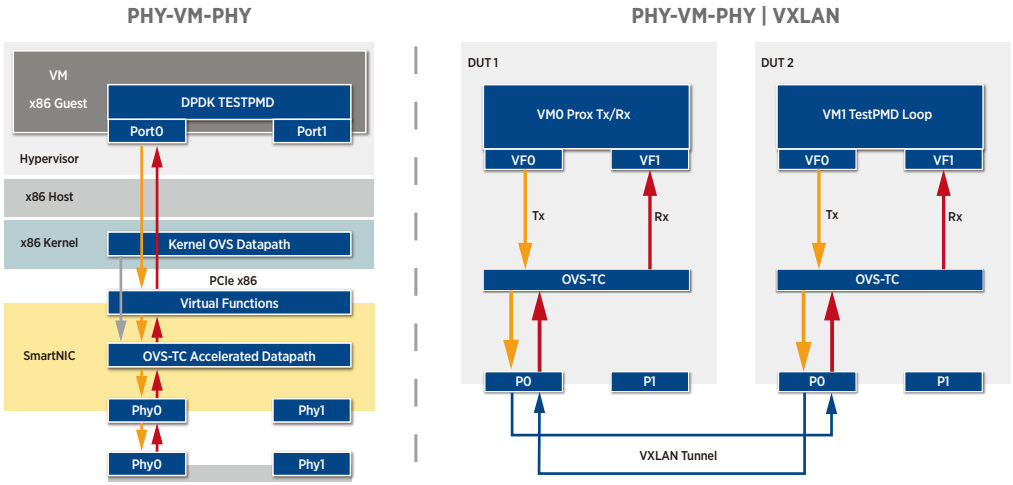


Figure 6. OVS-TC benchmark setups connected to DPDK-based traffic generator

TRex and Prox were used as external traffic generators. TRex is open source and provides a fully scalable software implementation that addresses real-traffic challenges and the enormous traffic flows that modern data centers face today. Prox is an open source, highly configurable DPDK-based application which allows creating software architectures for granular testing.

Traffic Profiles

The following traffic profiles were used:

- Frame sizes (bytes): 64, 128, 256, 512, 768, 1024, 1280, 1518
- 1-Port FWD, ethipv4_1p-fwd | PHY-VM-PHY | VXLAN
- Flows: 8,000, 16,000, 32,000 64,000, 128,000
- Rules: 8,000, 16,000, 32,000 64,000, 128,000

Performance Tuning

BIOS tuning
Host boot kernel tuning
CPU isolation
VM/workload CPU scheduling

BENCHMARK RESULTS

NICs deployed in today's data centers manage very complex network traffic with tens-of-thousands of flows. To support the complexity and performance requirements of modern applications, a NIC must handle tens or hundreds-of-thousands of flows and provide advanced programming capabilities. For each of the test cases, the applied load is 25Gb/s for packet sizes ranging from 64Byte to 1518Byte with 8,000 up to 128,000 flows/rules. Packets are injected from the traffic generator at the network interface and into the datapath for all cases. The following graphs display the datapath performance for hardware accelerated OVS-TC in a scale-up test using a single port.

Some NICs claim high performance, but this is only for minimal flows and not indicative of a

real-world use case. Only Corigine Agilio SmartNICs perform in true cloud and data center networks with tens-of-thousands to hundreds-of-thousands of network flows and rules

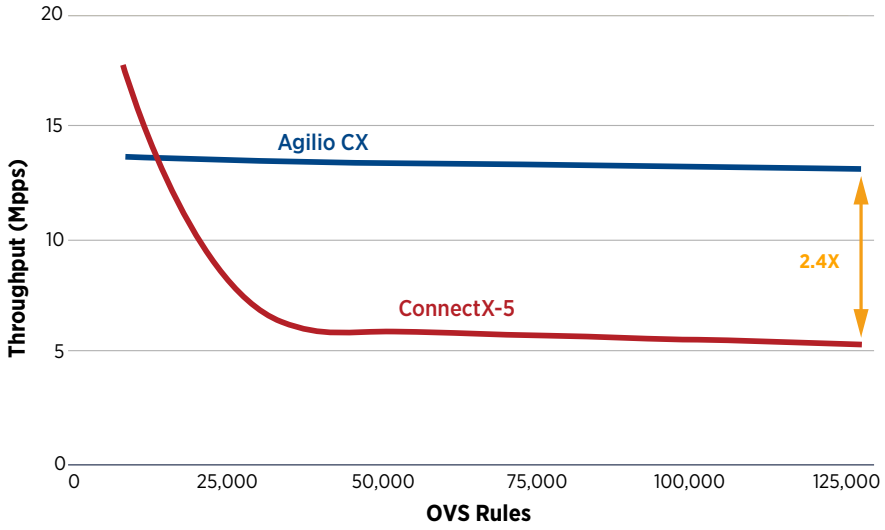


Figure 7. Red Hat PVP – Rule scalability at 64B packet size test results

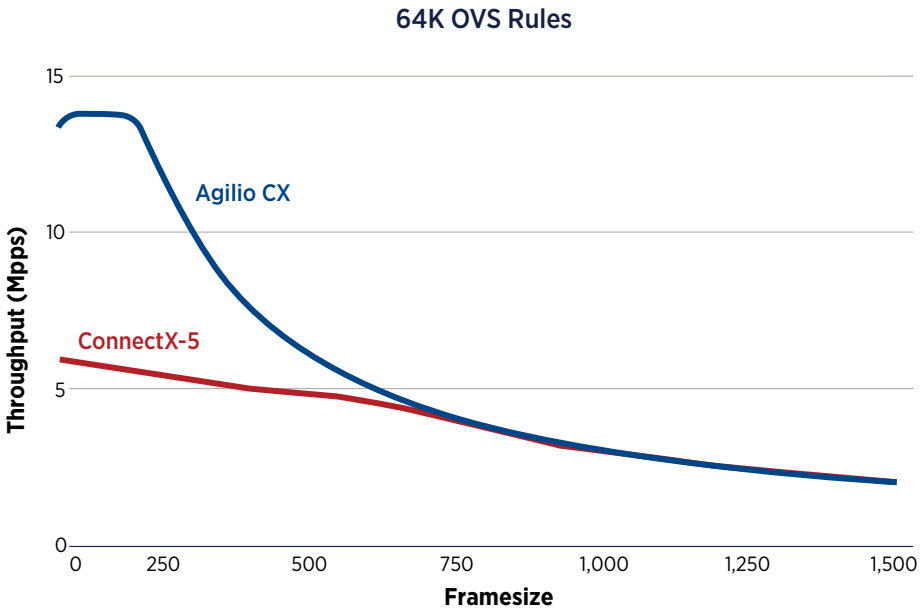


Figure 8. Mellanox ConnectX-5 performance drops as we scale to higher flows/rules count

Compared to Mellanox OVS offload, Corigine OVS-TC hardware acceleration has very different performance behaviors with increased flows/rules. The tests highlight that the increased number of flows and rules has a negative impact on the Mellanox throughput. As the flows and rules increase, the frame rate drops. At 64B packet size, Mellanox ConnectX-5 delivers 11.8Mpps on 16K flows/rules, but when scaled to 64K flows/rules the performance drops by 1.9X to approximately 5.7Mpps (Figure 7). As the amount of flows/rules increases, the Mellanox Accelerated Switching and Packet Processing (ASAP^2) performance degrades dramatically. At a 64B packet size, Corigine Agilio CX with OVS-TC delivers 13Mpps for

64K flows/rules. OVS-TC running on the Agilio CX SmartNIC performs 2.4X better than the Mellanox ASAP^2 on ConnectX-5.

The VXLAN-based, PHY-VM-PHY test with ipv4tcp at 78B packet size, scaled from 8,000 up to 64,000 flows/rules. A 1:2 mapping (flows:OVS-rules) is used, with two OVS-rules for every flow generated - one flow going in and the other going out. There is a 1:1 mapping for OVS-rules to VNIs in use. Media aggregated throughput performance was measured for both Mellanox ConnectX-5 25GbE NIC and Corigine Agilio CX 25GbE SmartNIC. The most important observation from these tests is that even with a relatively small number of flows and rules (16K), Corigine outperforms Mellanox by more than 50%. As the number of flows and rules increases, the performance of ConnectX-5 degrades. At 64,000 flows and rules with VXLAN (VNI), Agilio CX delivers 2X better performance than ConnectX-5.

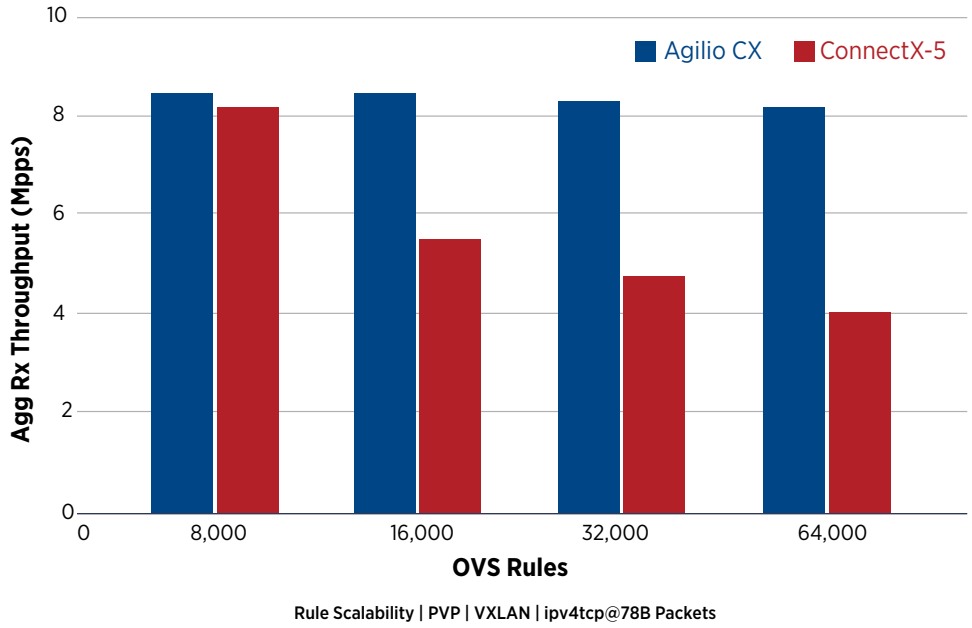


Figure 9. VXLAN Performance – Mellanox ConnectX-5 vs. Corigine Agilio CX

CONCLUSION

Corigine Agilio CX 25GbE SmartNICs with OVS-TC significantly outperform Mellanox ConnectX-5 25GbE NICs with ASAP^2. The packet-per-second throughput of the Agilio CX hardware-accelerated OVS-TC is 2.4X higher than the ConnectX-5 NIC at scale. The Corigine Agilio transparent offload architecture with its programmable datapath for virtual switching acceleration delivers improved server efficiency and is the best SmartNIC available for today's complex networks that must support tens-of-thousands to hundreds-of thousands of concurrent flows.

APPENDIX A

OVS-TC Supported Hardware-Accelerated Match/Action Features on Agilio CX SmartNICs

KERNEL DATAPATH MATCH: **ACCELERATED MATCH	OVS-DPDK	OVS-TC
OVS_KEY_ATTR_PRIORITY	✗	✗
OVS_KEY_ATTR_IN_PORT	✔	✔
OVS_KEY_ATTR_ETHERNET	✔	✔
OVS_KEY_ATTR_VLAN	✔	✔
OVS_KEY_ATTR_ETHERTYPE	✔	✔
OVS_KEY_ATTR_IPV4	✔	✔
OVS_KEY_ATTR_IPV6	✔	✔
OVS_KEY_ATTR_TCP	✔	✔
OVS_KEY_ATTR_UDP	✔	✔
OVS_KEY_ATTR_ICMP	✔	✔
OVS_KEY_ATTR_ICMPV6	✔	✔
OVS_KEY_ATTR_ARP	✗	✗
OVS_KEY_ATTR_ND	✗	✗
OVS_KEY_ATTR_SKB_MARK	✗	✗
OVS_KEY_ATTR_TUNNEL	✔	✔
OVS_KEY_ATTR_SCTP	✔	✔
OVS_KEY_ATTR_TCP_FLAGS	✗	✗
OVS_KEY_ATTR_DP_HASH	✗	✗
OVS_KEY_ATTR_RECIRC_ID	✗	✗
OVS_KEY_ATTR_MPLS	✔	✔
OVS_KEY_ATTR_CT_STATE	✔	✗
OVS_KEY_ATTR_CT_ZONE	✔	✗
OVS_KEY_ATTR_CT_MARK	✔	✗
OVS_KEY_ATTR_CT_LABELS	✔	✗
ADDITIONAL ACTIONS (OUTSIDE VANILLA OVS)		
OVS_KEY_ATTR_NSH	✗	✗

KERNEL DATAPATH ACTIONS: **ACCELERATED ACTIONS	OVS-DPDK	OVS-TC
OVS_ACTION_ATTR_OUTPUT	✔	✔
OVS_ACTION_ATTR_SET_PRIORITY	✔	✗
OVS_ACTION_ATTR_SET_SKB_MARK	✗	✗
OVS_ACTION_ATTR_SET_TUNNEL_ INFO	✔	✔
OVS_ACTION_ATTR_SET_ETHERNET	✔	✔
OVS_ACTION_ATTR_SET_IPV4	✔	✔
OVS_ACTION_ATTR_SET_IPV6	✔	✔
OVS_ACTION_ATTR_SET_TCP	✔	✔
OVS_ACTION_ATTR_SET_UDP	✔	✔
OVS_ACTION_ATTR_SET_MPLS	✔	✗
OVS_ACTION_ATTR_SET_CT_STATE	✔	✗
OVS_ACTION_ATTR_SET_CT_ZONE	✔	✗
OVS_ACTION_ATTR_SET_CT_MARK	✔	✗
OVS_ACTION_ATTR_SET_CT_LABELS	✔	✗
OVS_ACTION_ATTR_PUSH_VLAN	✔	✔
OVS_ACTION_ATTR_POP_VLAN	✔	✔
OVS_ACTION_ATTR_RECIRC	✗	✗
OVS_ACTION_ATTR_HASH	✗	✗
OVS_ACTION_ATTR_PUSH_MPLS	✔	✗
OVS_ACTION_ATTR_POP_MPLS	✔	✗
OVS_ACTION_ATTR_SET_MASKED	✔	✔
OVS_ACTION_ATTR_SAMPLE	✗	✗
OVS_ACTION_ATTR_CT	✔	✗
OVS_ACTION_ATTR_DROP	✔	✔

FEATURES	OVS-DPDK	OVS-TC
Breakout cable support	✔	✔
Balance SLB	✔	✔
VXLAN	✔	✔
VLAN	✔	✔
VXLAN+VLAN+LAG	✔	✗

Corigine

Email: sales@corigine.com
www.corigine.com.cn

TB-XVIO-1/17