redhat.

Red Hat's perspective on OVS HW Offload Status

Current state and what is WIP

Rashid Khan
Senior Manager, Networking Services
Nov. 17, 2017

Acknowledgements and Disclaimers

I am presenting the work of many many people… Special thanks to:
Andrew T, Franck B, Eelco C, Marcelo L, Paolo A, Flavio L, Kevin T

Performance numbers shown in this presentation are based on test results from running a specific series of tests in our labs.

Test results vary from one setup to another and based on different use cases.
Any test results mentioned are for example-only scenarios and are not conclusive nor a recommendation of one vendor's solution over another.

AGENDA

Why offload ?

Does it look promising?

What is left to do?

Backup / more info

Please view Franck's presentation from Thursday 11:30am:

OVS-DPDK for NFV: go live feedback!

Please view Aaron Conole's presentation from Thursday 3:30pm:

Conntrack + OvS

Why not just SW?

Simply way too many cores needed

- 4 Mpps/core with expert level tuning
  - Yes even with DPDK !

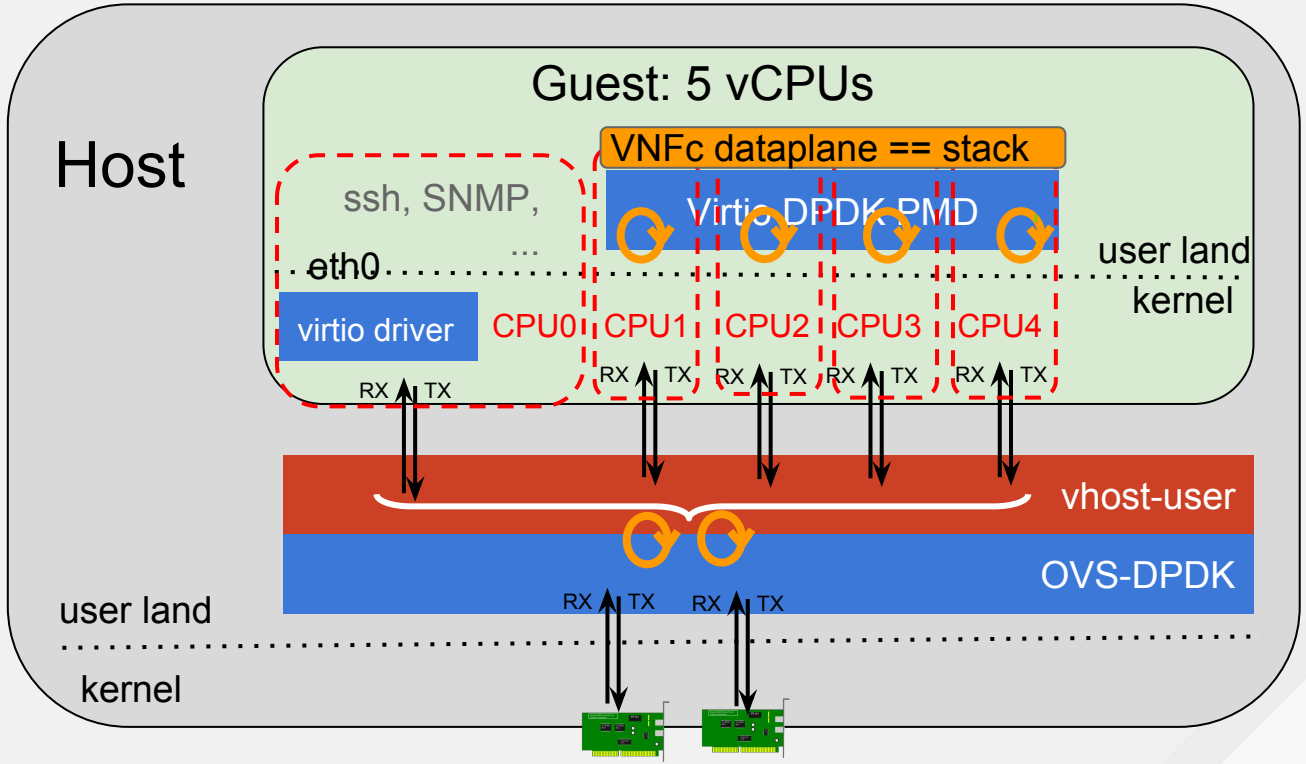
- Does not scale to 25G, 40G, 100G ....

OVS-DPDK: virtio, vhost-user, virtio PMD

DPDK
DATA PLANE DEVELOPMENT KIT

ACTIVE LOOP

```
while (1) {
        RX-packet()
    forward-packet()
}
```

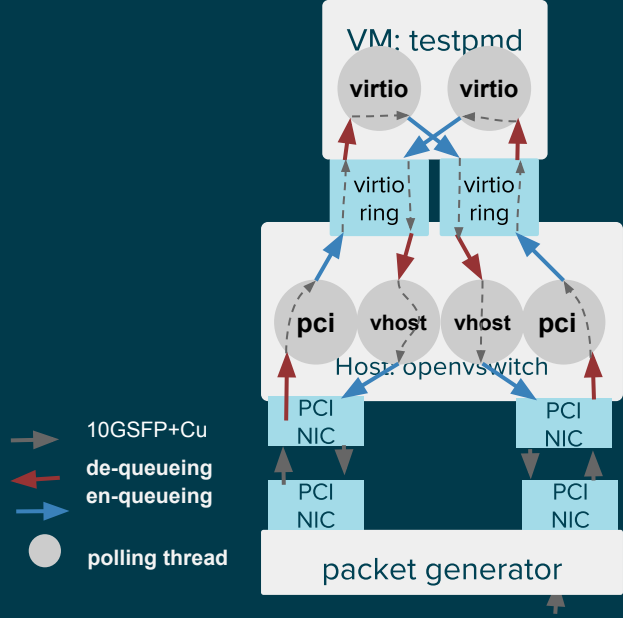


redhat.

Zero Packet loss

VM L2 forwarding, VLAN networks, *intra*-NUMA node, single queue

- 2 x virtio-net interfaces (NUMA node0)
- 2 x 10Gb interfaces (NUMA node0)
- Testpmd DPDK application in VM with 2 x virtio-net
- OVS using 4 PMD threads (2 cores) to process data-plane traffic
- Directly connected packet generator and compute node, no HW switch.
- Bidirectional traffic, 128 flows, *no broadcast or multicast packets*
- Measurement time, zero-loss: 2 hours, non-zero-loss: 5 mins
- Maximum rate while within specified loss:



	Loss: 20 packets-per-million				Loss: 5 packets-per-million				Loss: 1 packet-per-million				Loss: 0 packets-per-million			
Frame size	Mpps	Gbps	Mpps/ core	Gbps/ core	Mpps	Gbps	Mpps/ core	Gbps/ core	Mpps	Gbps	Mpps/ core	Gbps/ core	Mpps	Gbps	Mpps/ core	Gbps/ core
64	**9.38**	**6.30**	**4.69**	**3.15**	**9.15**	**6.15**	**4.57**	**3.07**	**7.39**	**4.97**	**3.69**	**2.48**	**4.34**	**2.92**	**2.17**	**1.46**
256	**7.66**	**16.19**	**3.83**	**8.45**	**7.81**	**17.24**	**3.90**	**8.66**	**7.45**	**6.52**	**3.72**	**3.26**	**2.34**	**5.18**	**1.17**	**2.59**
1024	**2.39**	**19.99**	**1.19**	**9.99**	**2.39**	**19.99**	**1.19**	**9.99**	**2.39**	**19.99**	**1.19**	**9.99**	**2.38**	**19.92**	**1.19**	**9.96**
1500	**1.64**	**19.99**	**0.82**	**9.99**	**1.64**	**19.99**	**0.82**	**9.99**	**1.64**	**19.99**	**0.82**	**9.89**	**1.64**	**19.94**	**0.82**	**9.97**

So whats the big deal?

Just add more CPUs, add more cores

If forwarding 10G of traffic takes ~4 cores

If Storage takes ~2 cores

That is already ¼ of a 24 core chip

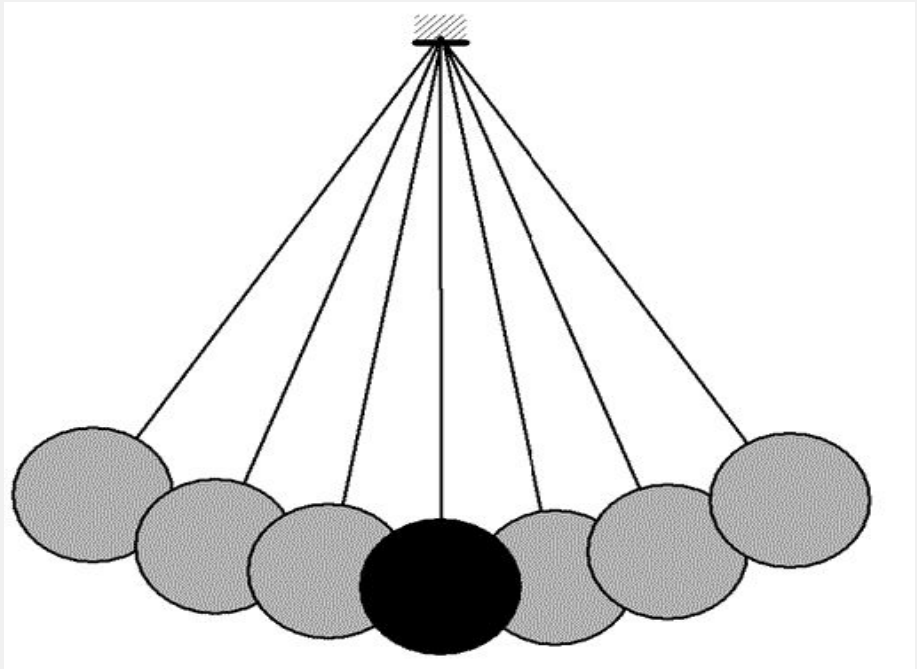
This is “wasted” revenue for the cloud providers

They charge per cycle per second

Swing of the pendulum

All HW

100 - 10 years ago



All SW

9 - 0 years ago

Very near future (some HW, some SW)

Many HW vendors have OVS Offload solutions

NETRONOME **MELLANOX** **CAVIUM**

CHELSIO **BROADCOM**

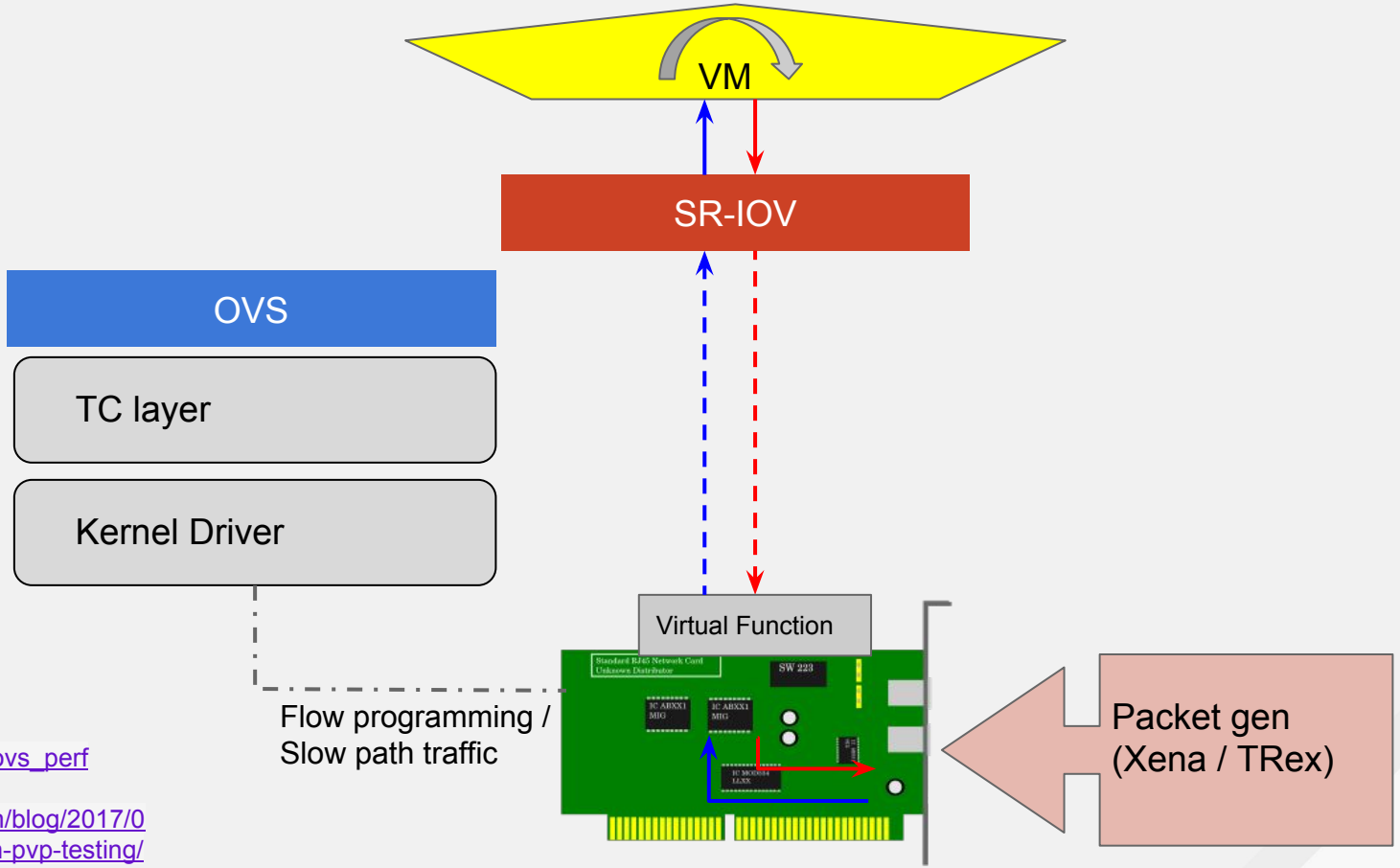
Others

Offloading method

Netronome	TC (kernel)
Mellanox	TC (kernel)
Broadcom	TC (kernel)
Chelsio	TC (kernel)
Cavium	OVS runs in the NIC firmware, offloading is transparent from CPU PoV

Accepted in upstream netdev

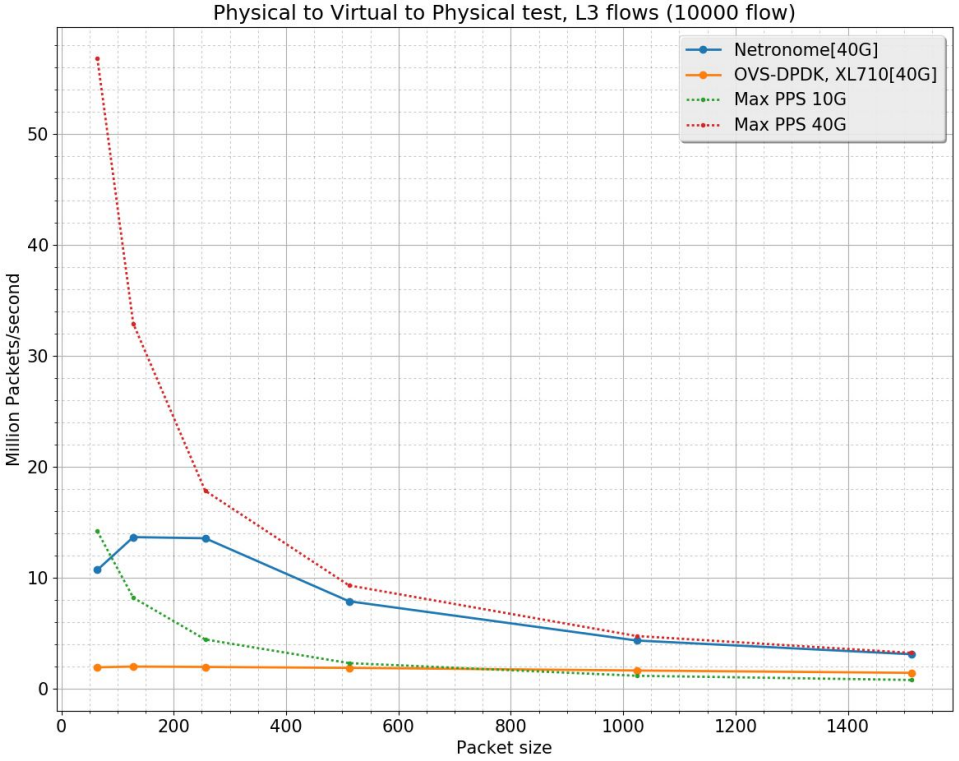
Example PVP test

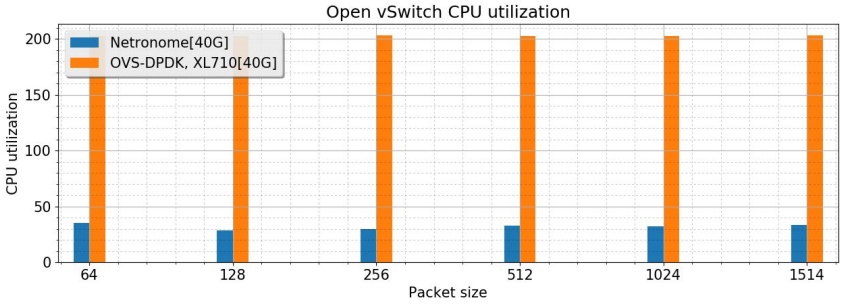


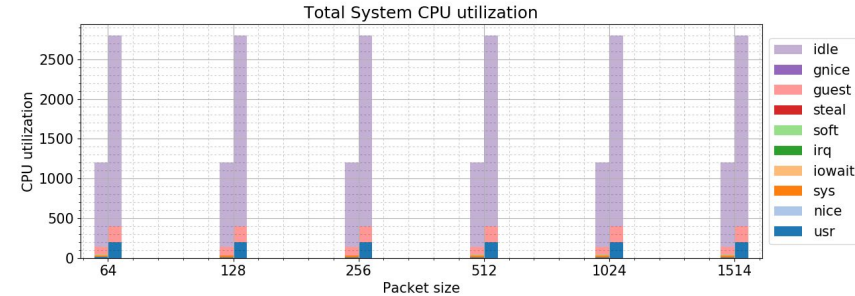
https://github.com/chaudron/ovs_perf

https://developers.redhat.com/blog/2017/09/28/automated-open-vswitch-pvp-testing/

NETRONOME

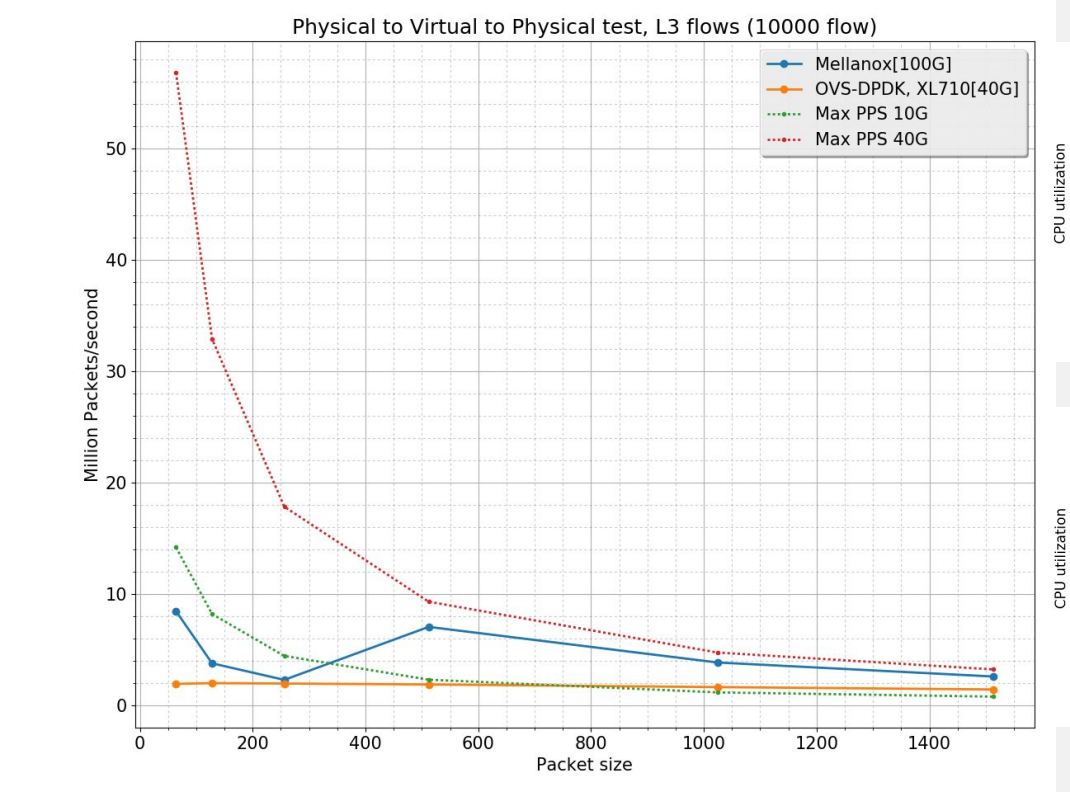


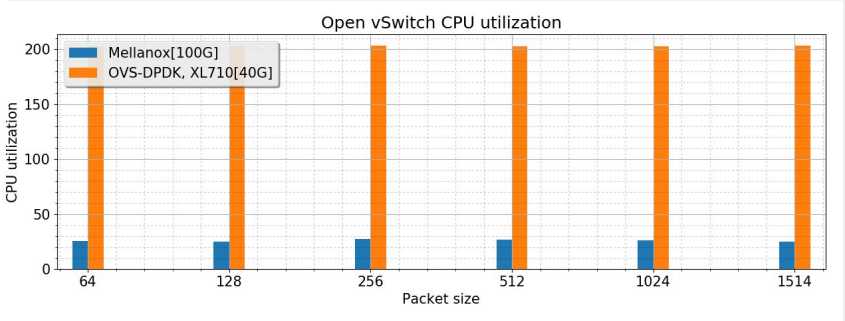


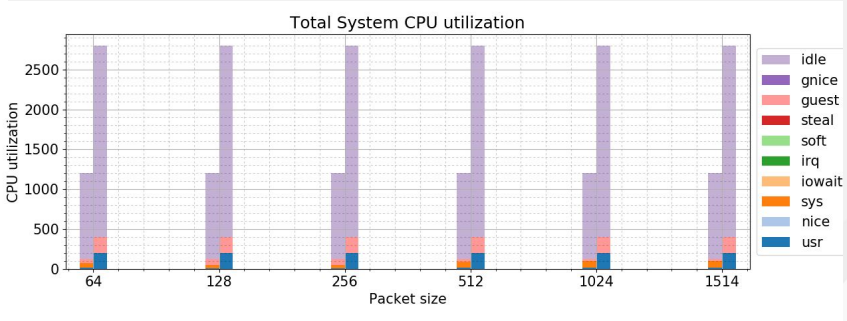


redhat.

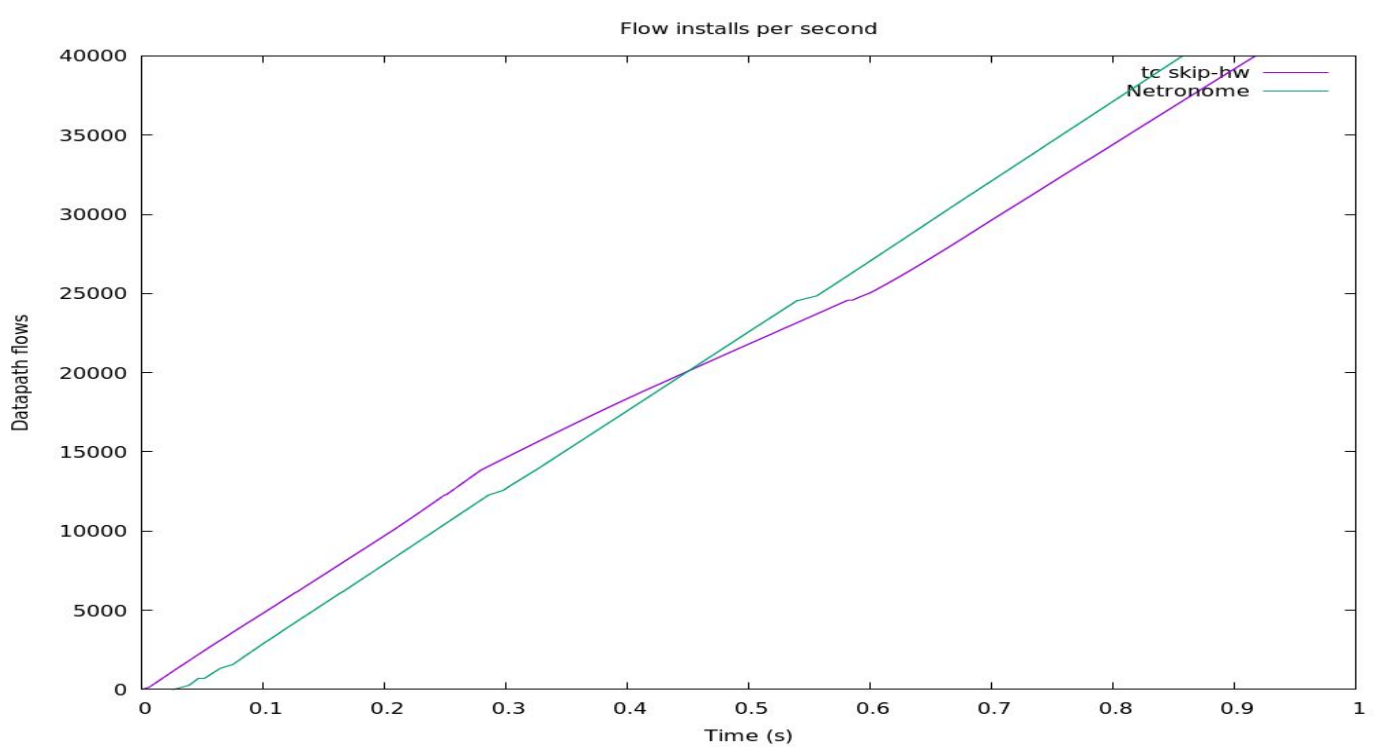
Mellanox





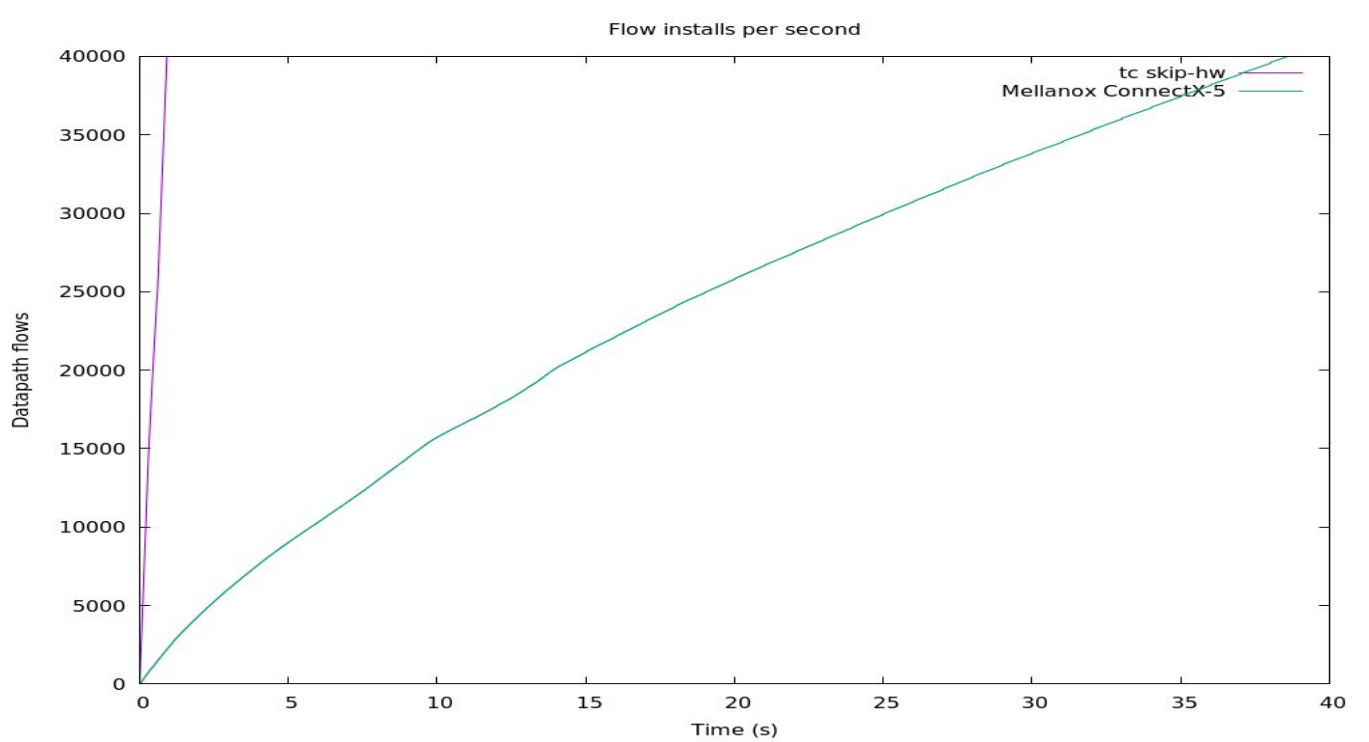


NETRONOME



Purple is sw only

Mellanox



What is WIP / To-do List

Rome wasn't built in a day…

- Connection tracking offload
- Openstack integration
- Flow insertion / deletion rate improvement
- Expand to do additional actions
- Metrics / statistics / billing
- System level logging (supportability)
- Support for sending to multiple ports
- QOS
- Kubernetes integration
- Migration from one card to another

We Are Hiring !!!!

THANK YOU

For further questions / comments:
rkhan@redhat.com


plus.google.com/+RedHat


facebook.com/redhatinc


linkedin.com/company/red-hat


twitter.com/RedHatNews


youtube.com/user/RedHatVideos

More information

SW used for testing

Netronome:
Linux upstream kernel, v4.13 for PVP test results. Linux V4.14rc4 for TC insertion rates.
OVS master branch from October 26th (7b997d4). DPDK/testpmd on VM v16.07

Mellanox:
Linux upstream kernel, net-next commit e1ea2f9856b7.
OVS master branch commit b05af21631ce, DPDK 17.11-rc2 (all git tips from Oct 30th).

OVS-DPDK:
RHEL7.4 latest kernel, OVS master branch from September 26th (97ee6d4), DPDK v17.05.2,
DPDK/testpmd on VM v16.07