DPDK

DATA PLANE DEVELOPMENT KIT

Hyperscale rte_flow toward production is really happening

LI XING, ALIBABA

ALEX ROSENBAUM, MELLANOX

Agenda

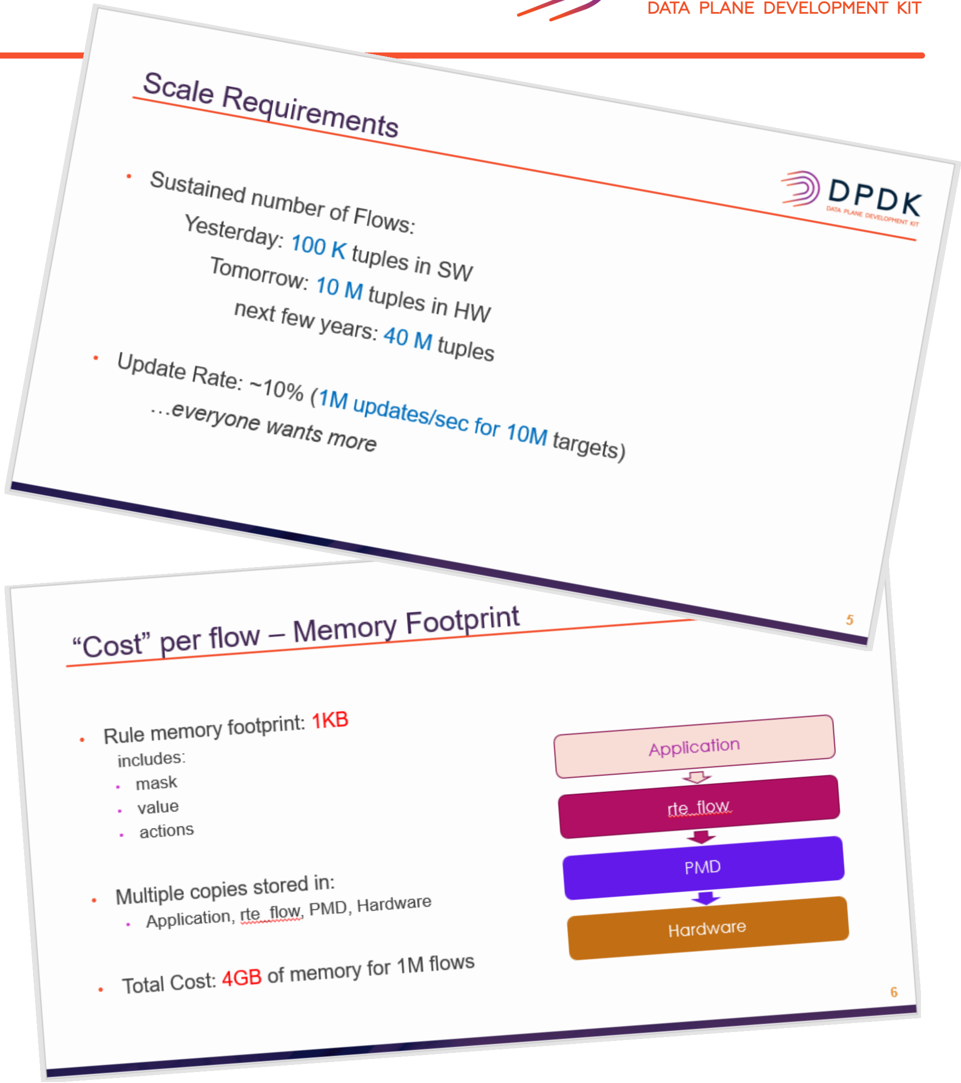
- AliCloud Service Introduction
- Integration work with rte_flow
- Performance Data
- Look into the Future
- Q & A

Quick Recap

Last year [1] we presented a session on the need for handling large scale steering requirements: SW & HW solutions

This year we review integration and feedback of these results

[1] https://youtu.be/8f-6cN-pSXo



Scale Requirements

- Sustained number of Flows:
  - Yesterday: 100 K tuples in SW
  - Tomorrow: 10 M tuples in HW
  - next few years: 40 M tuples
- Update Rate: ~10% (1M updates/sec for 10M targets)
  - …*everyone wants more*

5

"Cost" per flow – Memory Footprint

- Rule memory footprint: 1KB
  - includes:
    - mask
    - value
    - actions
- Multiple copies stored in:
  - Application, rte_flow, PMD, Hardware
- Total Cost: 4GB of memory for 1M flows

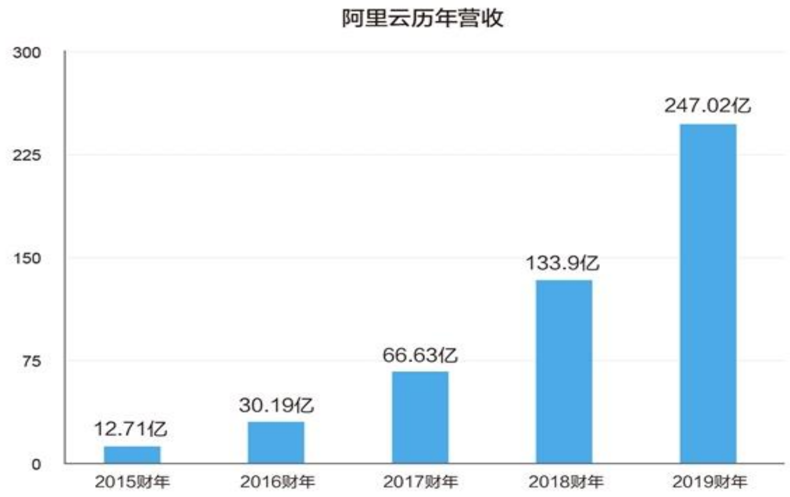
6

AliCloud Service Introduction

Market size: Growth and Key User

- Alibaba cloud services revenues approach $4 billion annual run rate. The segment has shown steady, high growth over the past few years.

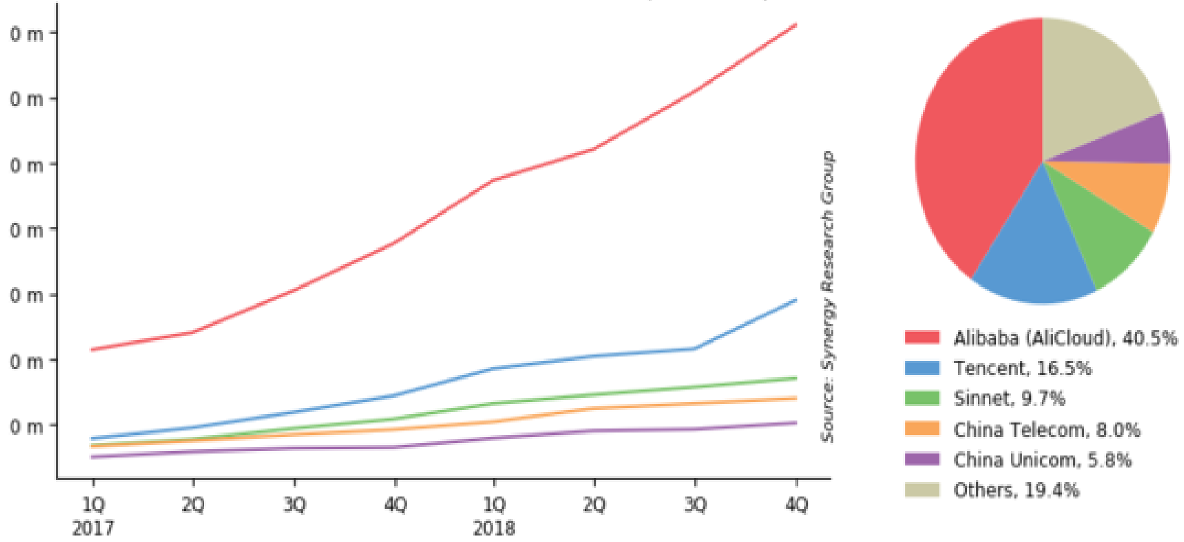
阿里云历年营收



数据来源：阿里巴巴财报

Cloud Infrastructure Services

Table 1. Cloud Infrastructure Services China Revenue for 1Q 2017 to 4Q 2018



Source: Synergy Research Group

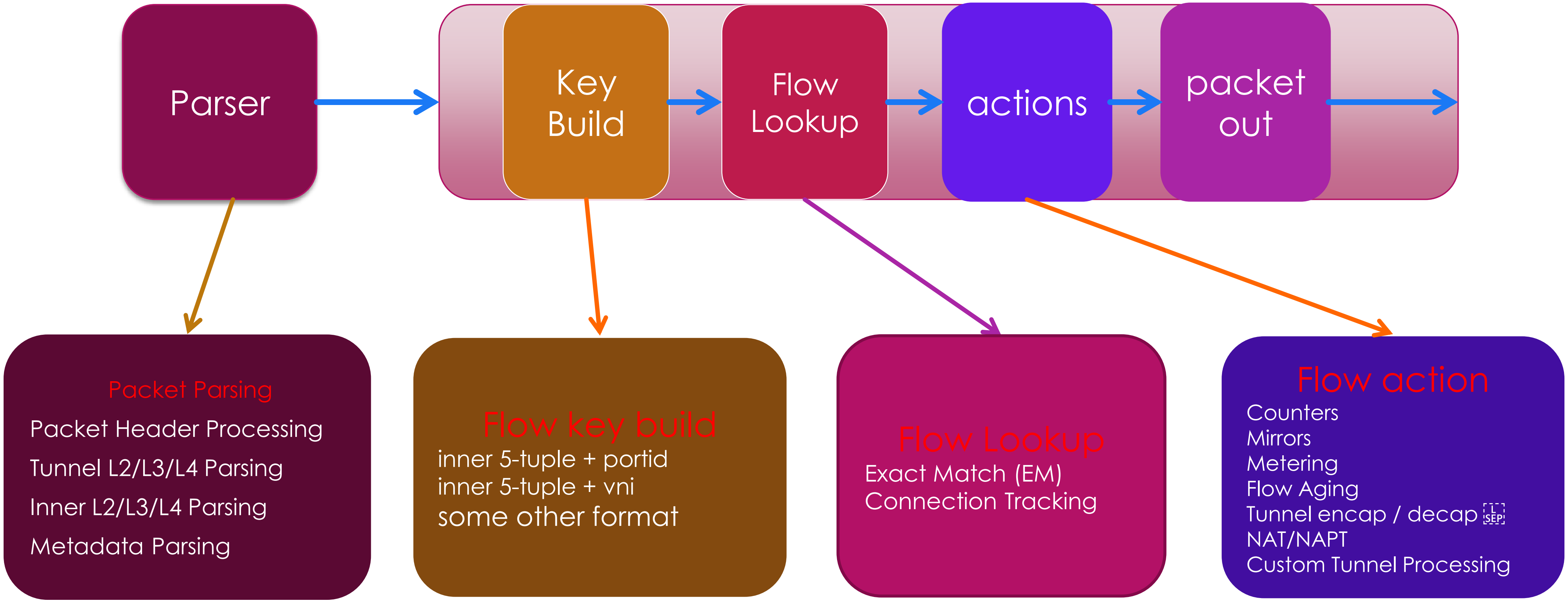
Network Service

- Virtual Private Cloud (VPC)
  - Provides an isolated cloud network to operate resources in a secure environment. Build your isolated network environment based on Alibaba Cloud including customizing the IP address range, network segment, route table, and gateway.
- Server Load balance (SLB)
  - Distributes traffic among multiple instances to improve the service capabilities of your applications. You can use SLB to prevent single point of failures (SPOFs) and improve the availability and the fault tolerance capability of your applications.
- Cloud Enterprise Network (CEN)
  - A dedicated network connection between different cloud environments. This service is used in multiple scenarios such as VPC communication across regions or user accounts, or data transmission between your on-premise data center and the cloud over a leased line.
- Smart Access Gateway (SAG)
  - Provides an end-to-end cloud deployment solution for connecting hardware and software to Alibaba Cloud. This allows enterprises to connect to the nearest VPC through encrypted connections.
- Thanks to the DPDK community

High Network Performance Scenarios

- VPC Gateway and NFV
- Live broadcast、network accelerate、 online games、online education
- Ali Group Double 11 (E-commerce promotion)
  - Message middleware、 Database、 Big Data

The Data Plane Pipeline for VPC



Parser → Key Build → Flow Lookup → actions → packet out →

Packet Parsing

Packet Header Processing

Tunnel L2/L3/L4 Parsing

Inner L2/L3/L4 Parsing

Metadata Parsing

Flow key build

inner 5-tuple + portid
inner 5-tuple + vni
some other format

Flow Lookup

Exact Match (EM)
Connection Tracking

Flow action

Counters
Mirrors
Metering
Flow Aging
Tunnel encap / decap
NAT/NAPT
Custom Tunnel Processing

The challenges

Trade-off： Functionality VS Performance

Solution： Hardware acceleration + SRIOV

How to satisfy large-scale flow entries?

How to ensure flow timely insertion?

How to aging hardware flow?

How to implement hardware flow state machine？

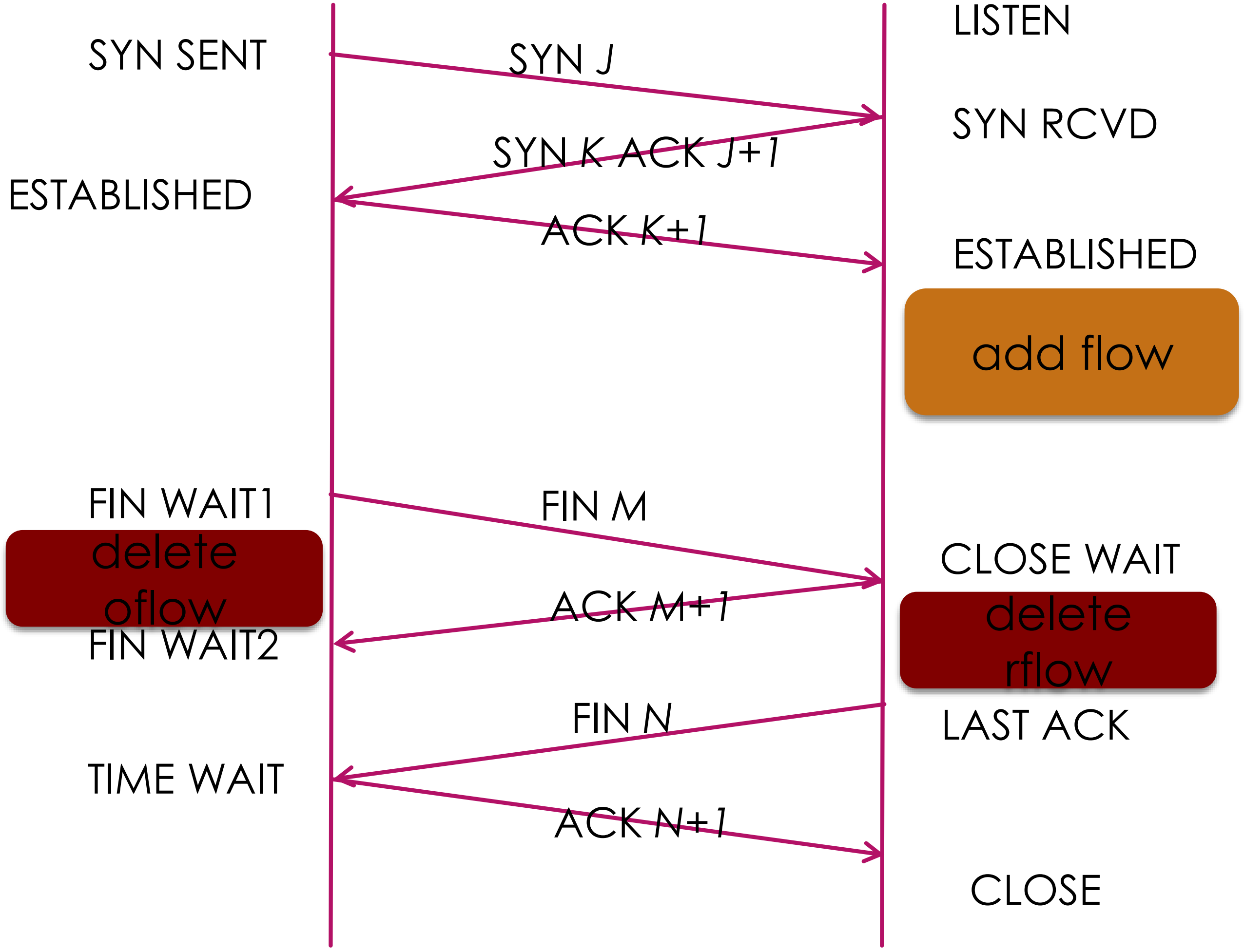
How to be compatible with different hardware？

Solve the problems

- Huge flow table size：At least 4 million flows
  a) use host memory to expand the flow table.
  b) flow rebalance(offload elephant flow – high BW flows).
  c) batch deletion
- Flow aging & Flow insertion rate
  a) Hardware flow aging
  b) Flow insertion rate is greater than 200k/s (sw base line)

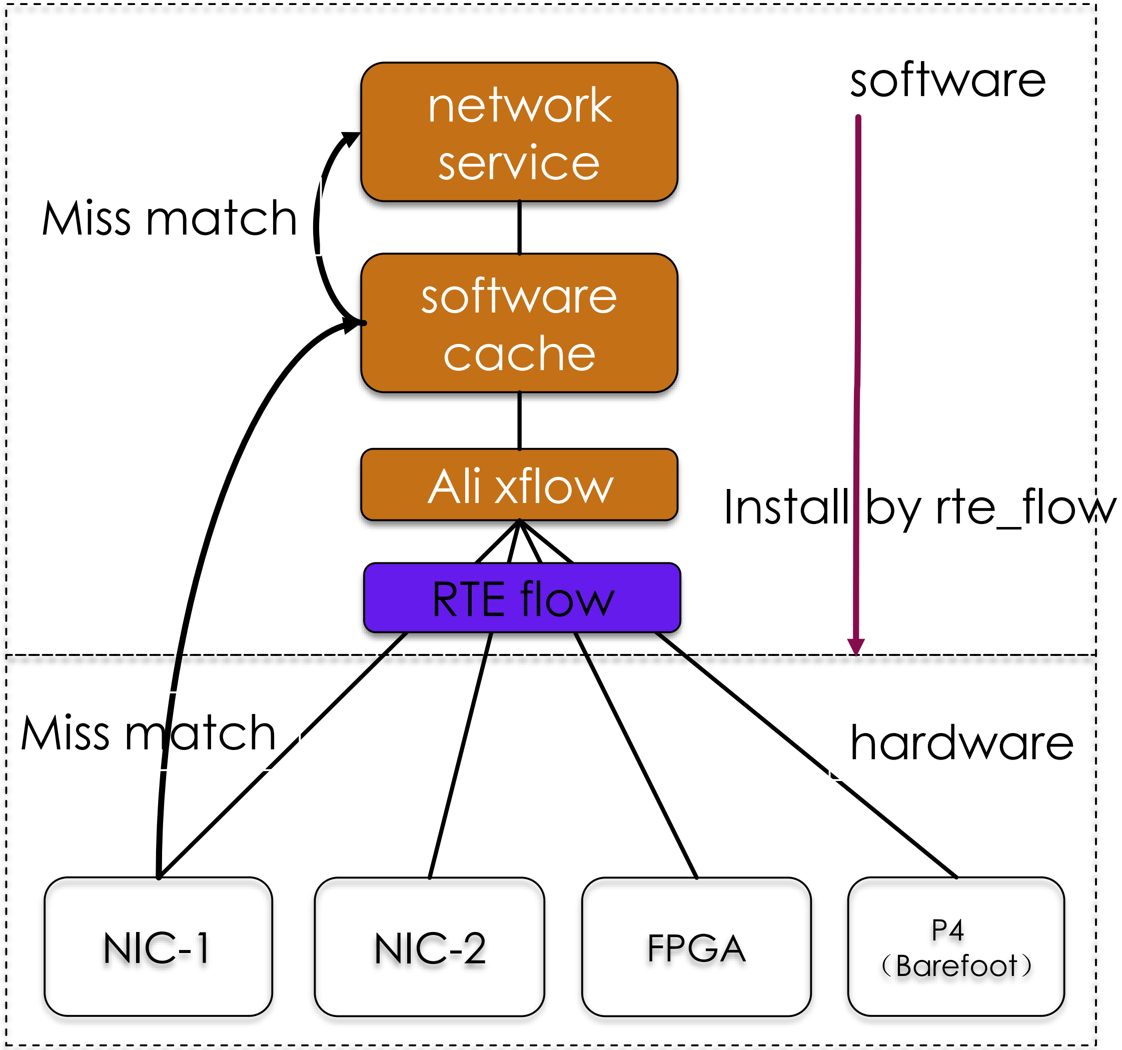
Solve the problems

- Hardware state machines
  a) Software management state machine.
  b) Send all packets with SYN, RST, FIN flag to host for processing.
  c) Update hardware flow when status changes.



Integration work with rte_flow

- Integrated with Different hardware
  a) software flow cache （states）
  b) abstraction layer： Ali xflow
    a) different vendor API
    b) different feature list
    c) Bugfix and workaround before the community



Integration work with rte_flow

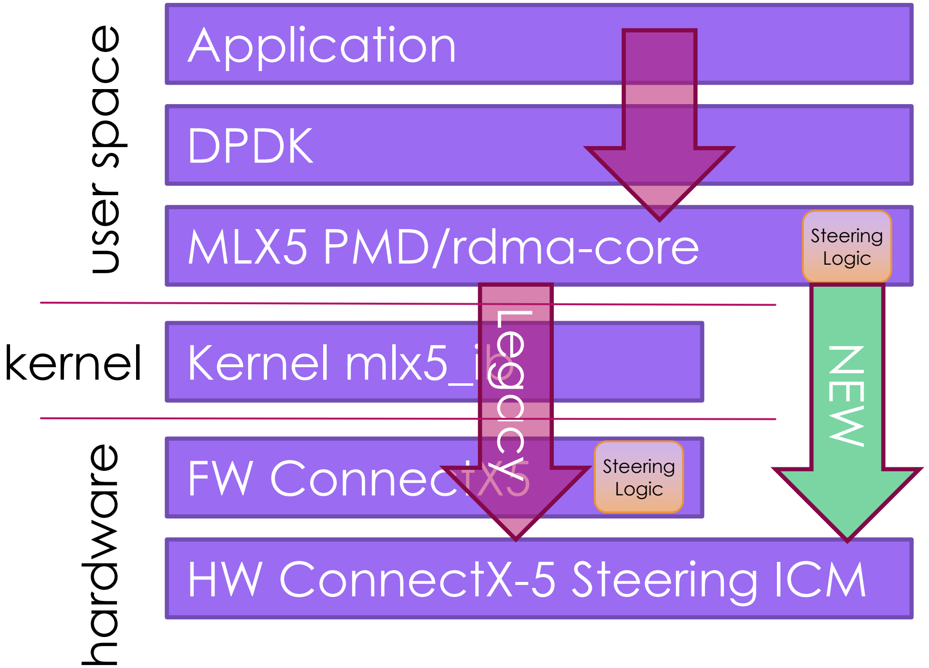
Mellanox rte_vswitch Design

- rte_vswitch, private software library layer, bridges gaps between upstream DPDK/PMD and Alibaba vSwitch requirements
  - Handle ConnectX-5 NIC Rx + TX steering as single FDB steering domain
    - ➢ Rule is mapped twice! in Rx and Tx, depending on context…
      DECAP+wire, ENCAP+vport
  - Steering Rule Ageing
    - ➢ track HW activity timers, flush HW rule
  - Flush full flow tables (vport’s)
    - ➢ Remove 1000’nd of rules quickly when VM is migrated

Mellanox High Rate Rule Insertion Design

- Mellanox driver is bifurcated
  - Has kernel driver (+queues) in parallel to user space queues
  - Exposed to user space via ibverbs/rdma-core

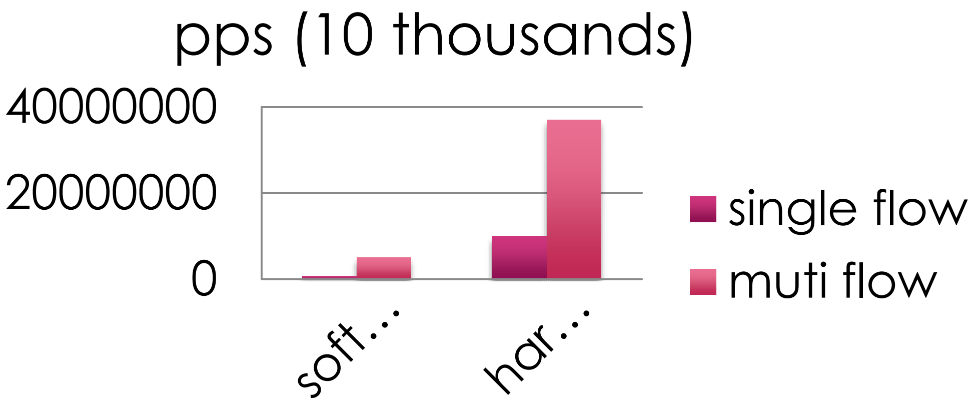
- All legacy steering commands pass via Kernel driver to FW which managed the steering logic
- New user space steering logic can manage ‘islands’ of steering in HW
  - Increases user space memory footprint (replacing FW/Kernel pages)
  - Added multiple flow tables (rte flow groups)
  - Designed with dynamically resizable flow table size

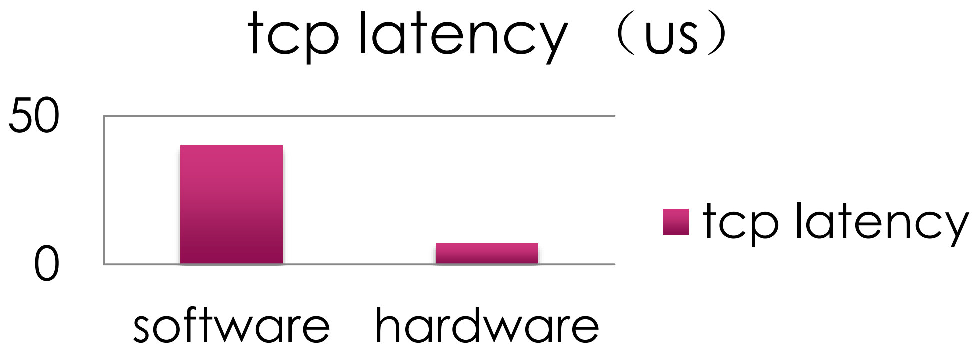


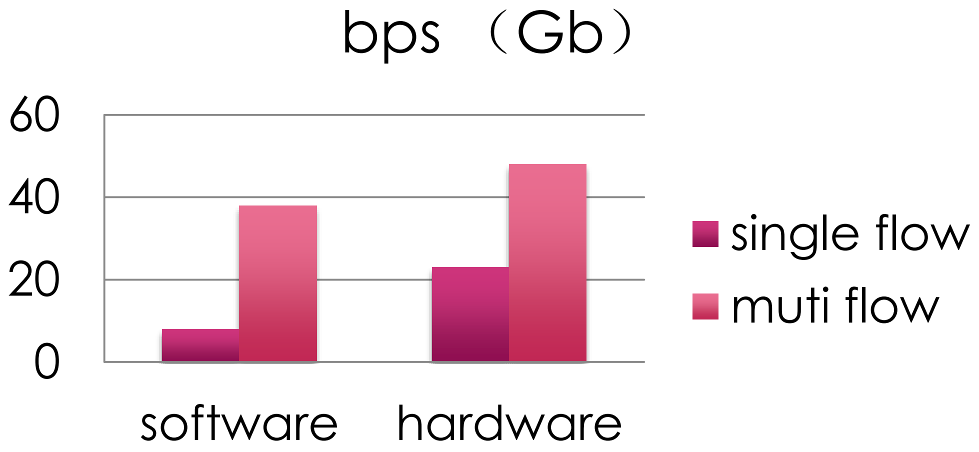
Performance Data

Benchmarks

- Hardware accelerate VS software forwarding





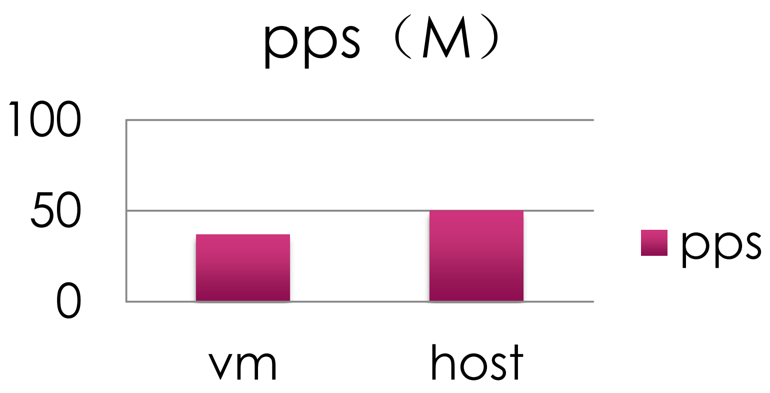


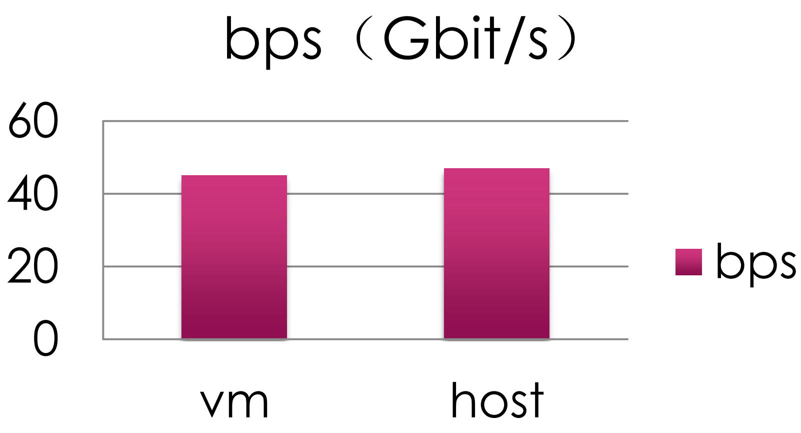
benchmark information:

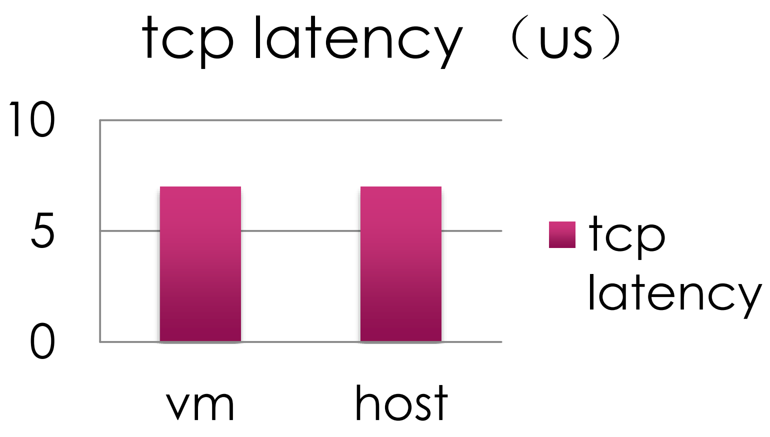
- CPU info: Intel(R) Xeon(R) Platinum 8269C
- CPU MHz: 3232.355
- Num of cpus used by pktgen：4
- NIC mode：SRIOV
- NIC ports： 2 * 25g

Benchmarks

- Virtual machine VS Physical Host





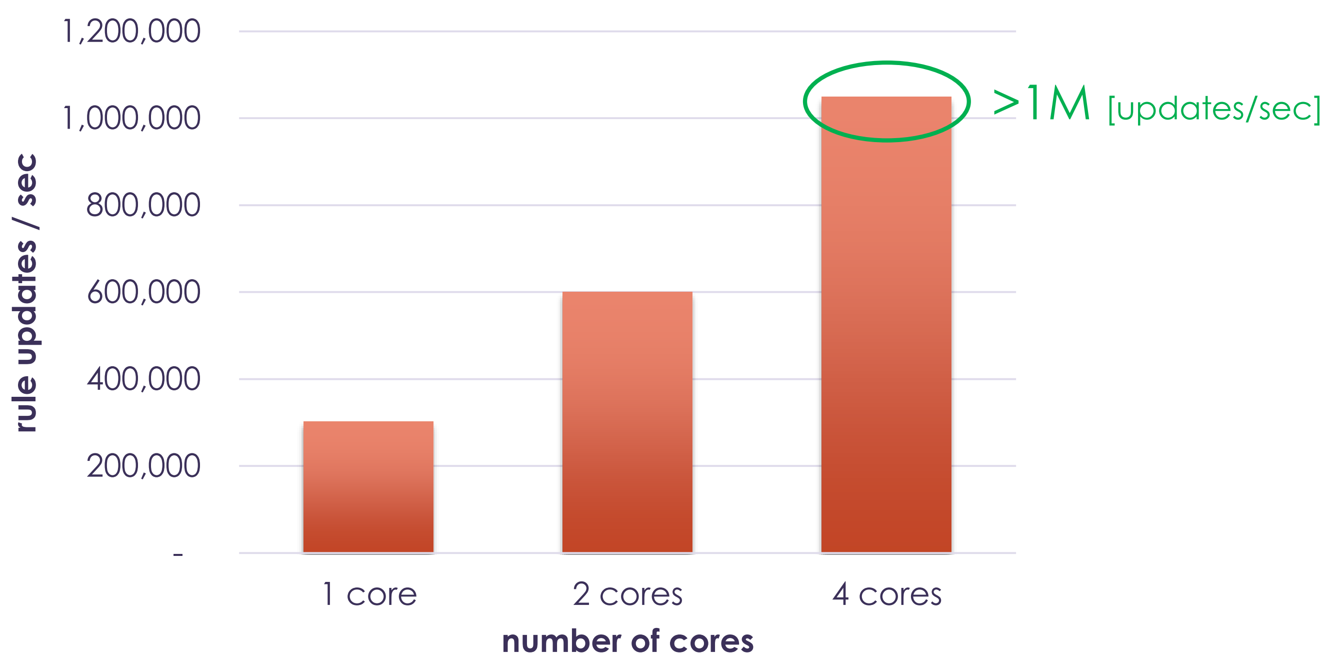


benchmark information:

- CPU info: Intel(R) Xeon(R) Platinum 8269C
- CPU MHz: 3232.355
- Num of cpus used by pktgen：4
- NIC mode：SRIOV
- NIC ports： 2 * 25g

Benchmarks: Multi thread scalability

- Test: simple 5 tuple, with simple action
  - Match: Ether(type=0x800) / IP(srcip,dstip) / UDP(dport,sport)
  - Action: queue



Mellanox Lab Test Cases
Benchmark Information:

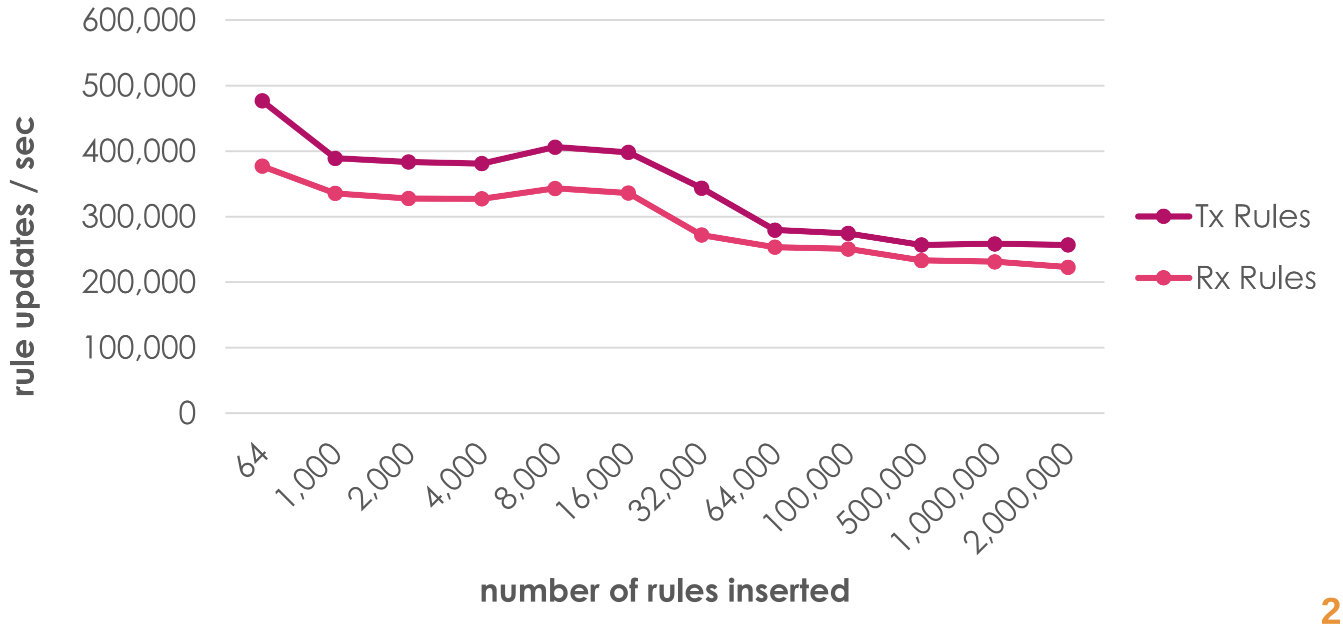
- CPU info: Intel(R) Xeon(R) E5-2699 v4
- CPU MHz: 2.20GHz
- NIC: ConnectX-5, 2 * 25g, PCIe3.0 x16

Benchmarks: Overlay VXLAN Networking

- Test 1 - Tx Rule:
  - Match: METADATA + Ether(type=0x800) / IP(srcip,dstip) / UDP(dport,sport)
  - Action: COUNT, MODIFY_HEADER, ENCAP(vxlan)
- Test 2 - Rx Rule:
  - Match: Ether(type=0x800) / IP / UDP(dport=250) / VXLAN(vni=0x20,flags=0x8) / IP(srcip,dstip) / UDP(dport,sport)
  - Action: COUNT, DECAP, MODIFY_HEADER, MARK and RSS

Mellanox Lab Test Cases
Benchmark Information:

- CPU info: Intel(R) Xeon(R) E5-2699 v4
- CPU MHz: 2.20GHz
- Num of cores：1
- NIC: ConnectX-5, 2 * 25g, PCIe3.0 x16



Look into the Future

Debuggability

Due to the complexity of online scenarios, bugs are inevitable. We need the following capabilities for hardware debugging analysis ：

- with >1 mill rte_flow’s need
- See existing device configuration
- See hardware flow table

DPDK Generic

- Flexible allocation for network queues ：Different users have different requirements for network capabilities and resources. We need the ability to flexibly manage the allocation of queues.

- Hairpin forwarding for SLB and NFV：SLB and NFV are mainly used for traffic forwarding.
  - https://patches.dpdk.org/patch/57663/

- Firmware hot upgrade: Firmware‘s new feature extensions and bugfix updates require no impact on user services.

Device Specific

- More counter and more meter resource
  - Mellanox ConnectX-5: 4,000
  - Mellanox ConnectX-6Dx: 4,000,000

- Memory footprint reduction
  - >4K per rule …
  - rule caching per layer: rte_flow, driver, fw, hw

Thanks !