Intel IPU/SmartNIC推动数据中心基础架构演进

Houston Tang, Rosen Xu

intel.

LEGAL DISCLAIMER

- Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration. No computer system can be absolutely secure. Check with your system manufacturer or retailer or learn more at intel.com.

- Tests document performance of components on a particular test, in specific systems. Differences in hardware, software, or configuration will affect actual performance. Consult other sources of information to evaluate performance as you consider your purchase. For more complete information about performance and benchmark results, visit http://www.intel.com/performance.

- Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more complete information visit http://www.intel.com/performance.

- Cost reduction scenarios described are intended as examples of how a given Intel-based product, in the specified circumstances and configurations, may affect future costs and provide cost savings. Circumstances will vary. Intel does not guarantee any costs or cost reduction.

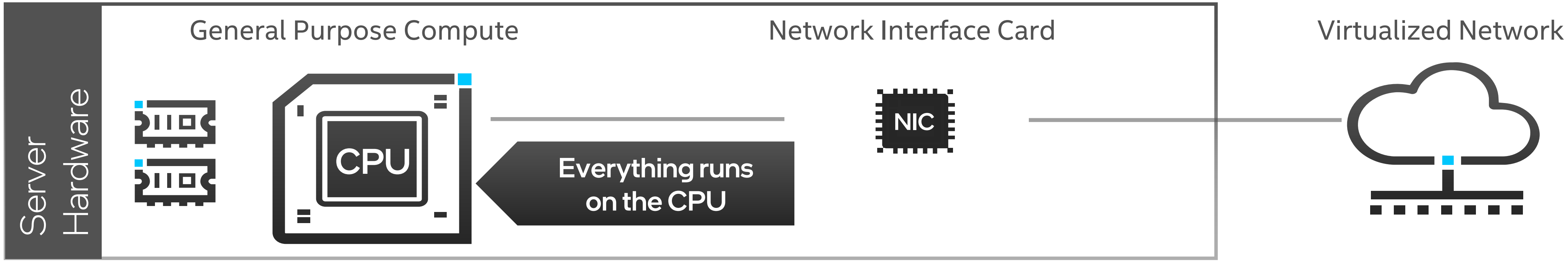
- No license (express or implied, by estoppel or otherwise) to any intellectual property rights is granted by this document.

- Intel does not control or audit third-party benchmark data or the web sites referenced in this document. You should visit the referenced web site and confirm whether referenced data are accurate.

- © 2021 Intel Corporation. Intel, the Intel logo, Agilex Stratix, Arria, and Xeon are trademarks of Intel Corporation in the U.S.
- and/or other countries. *Other names and brands may be claimed as property of others.

IPU/SmartNIC Platforms

Classic Server Architecture



Server Hardware

General Purpose Compute

Network Interface Card

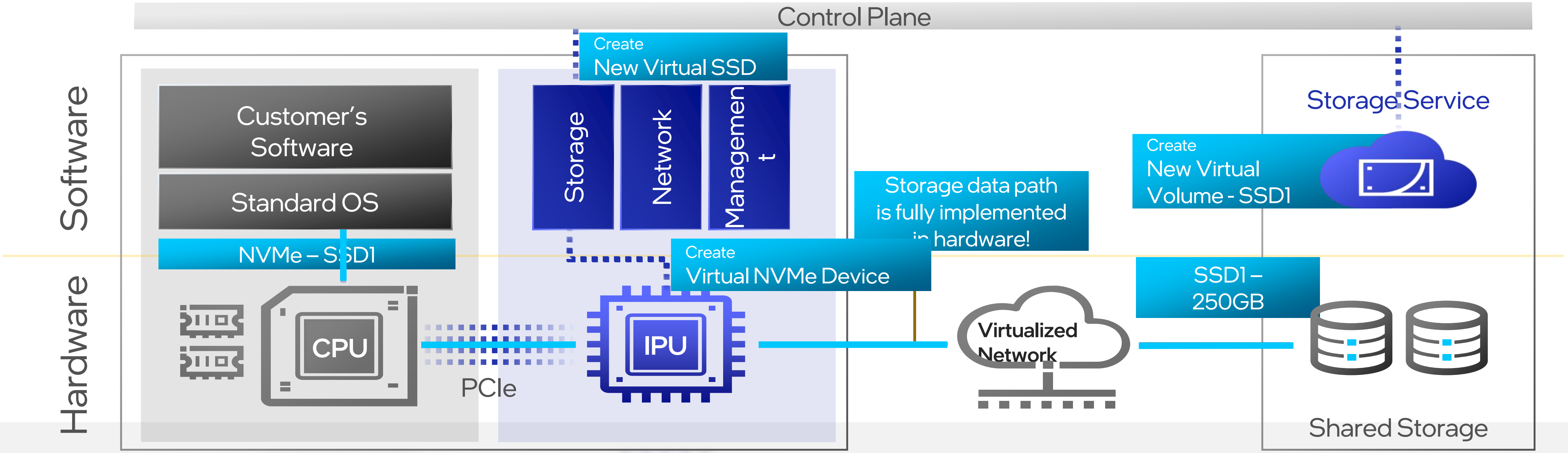
Virtualized Network

CPU

Everything runs on the CPU

NIC

Cloud Server Architecture



Control Plane

Create
New Virtual SSD

Software

Hardware

Customer's Software

Standard OS

NVMe – SSD1

Storage

Network

Managemen t

Storage Service

Create
New Virtual Volume - SSD1

Storage data path is fully implemented in hardware!

Create
Virtual NVMe Device

SSD1 – 250GB

CPU

PCIe

IPU

Virtualized Network

Shared Storage

Broad Infrastructure Acceleration Portfolio

Dedicated ASIC IPU

Performance and power optimized

Optimized secure networking and storage pipeline

intel IPU

FPGA-based Acceleration

IPU Platforms & Adapters

SmartNICs

Faster time to market for evolving standards

Re-programmable Secure Datapath enables flexible/customizable workload offload (future proof)

Onboard Xeon processor

intel IPU | intel AGILEX | intel STRATIX 10 | intel XEON

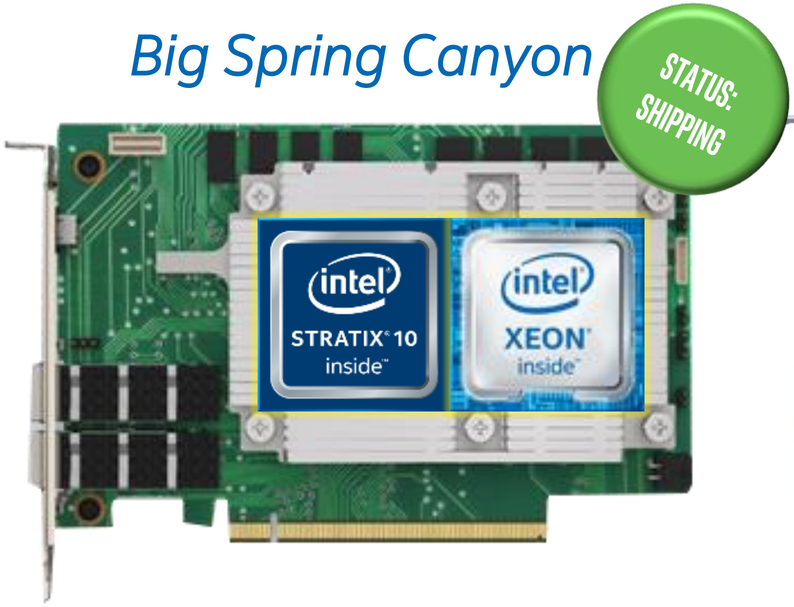
Programmable accelerated infrastructure workloads with customizable packet processing

Intel Ethernet NIC
with DPDK support

intel AGILEX | intel ETHERNET

Intel® FPGA Based IPU Platforms

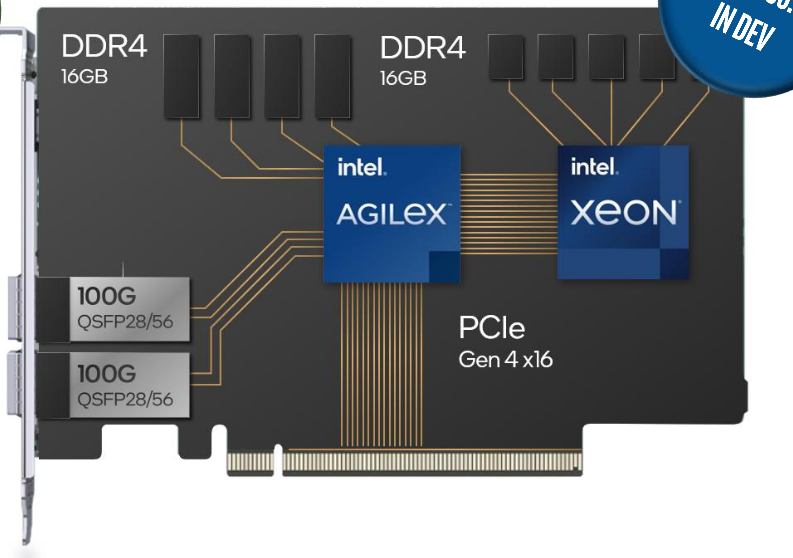
Big Spring Canyon



STATUS: SHIPPING

- 2x25GGbps IPU Platform
- Intel® Xeon® D Hewitt Lake processor
- Intel® Stratix® 10 DX FPGA
- 2 x PCIe Gen 3 x 8

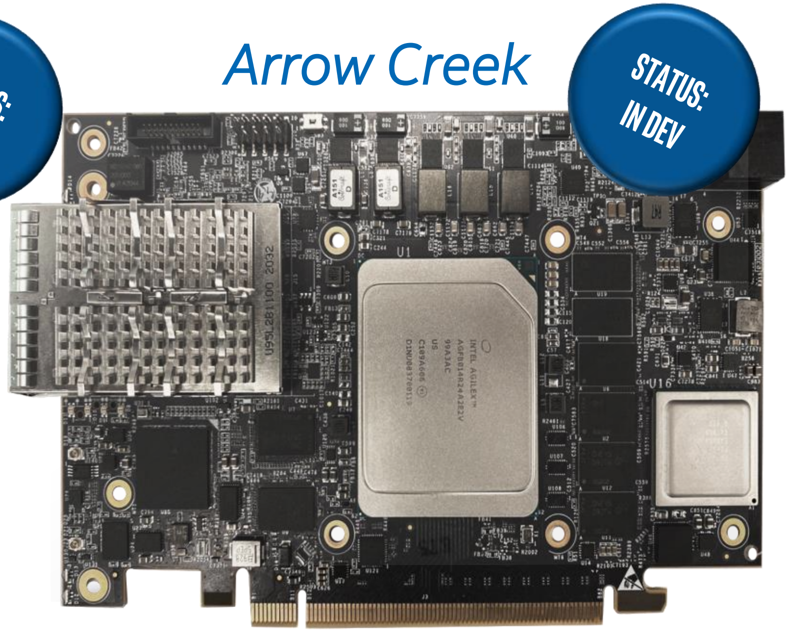
Oak Spring Canyon



STATUS: IN DEV

- 2x100Gbps IPU Platform
- Intel® Xeon® D Ice Lake processor
- Intel® Agilex® FPGA
- 2 x PCIe Gen 4 x16

Arrow Creek



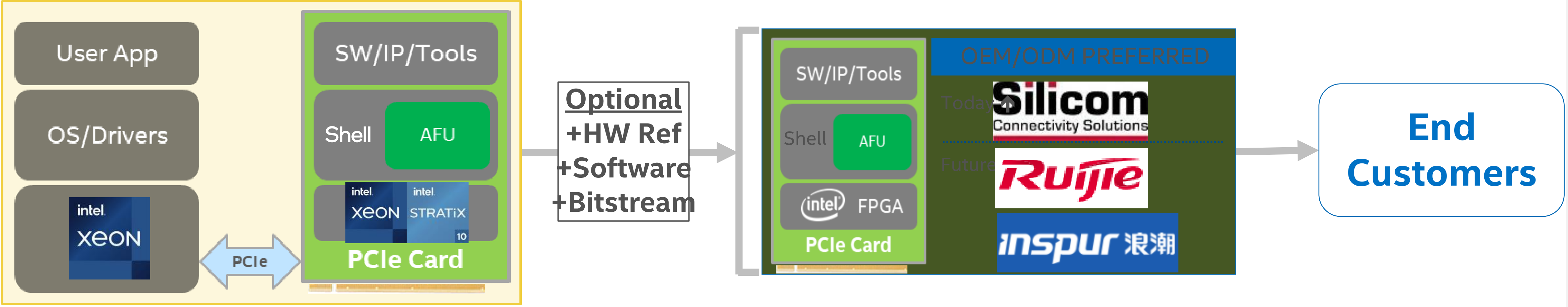
STATUS: IN DEV

- 2x100Gbps SmartNIC Platform
- Intel® Agilex® FPGA
- PCIe Gen 4 x16

Intel FPGA SmartNIC Platform for Cloud Route to Market – ADP

SmartNIC Development Platform From Intel PSG

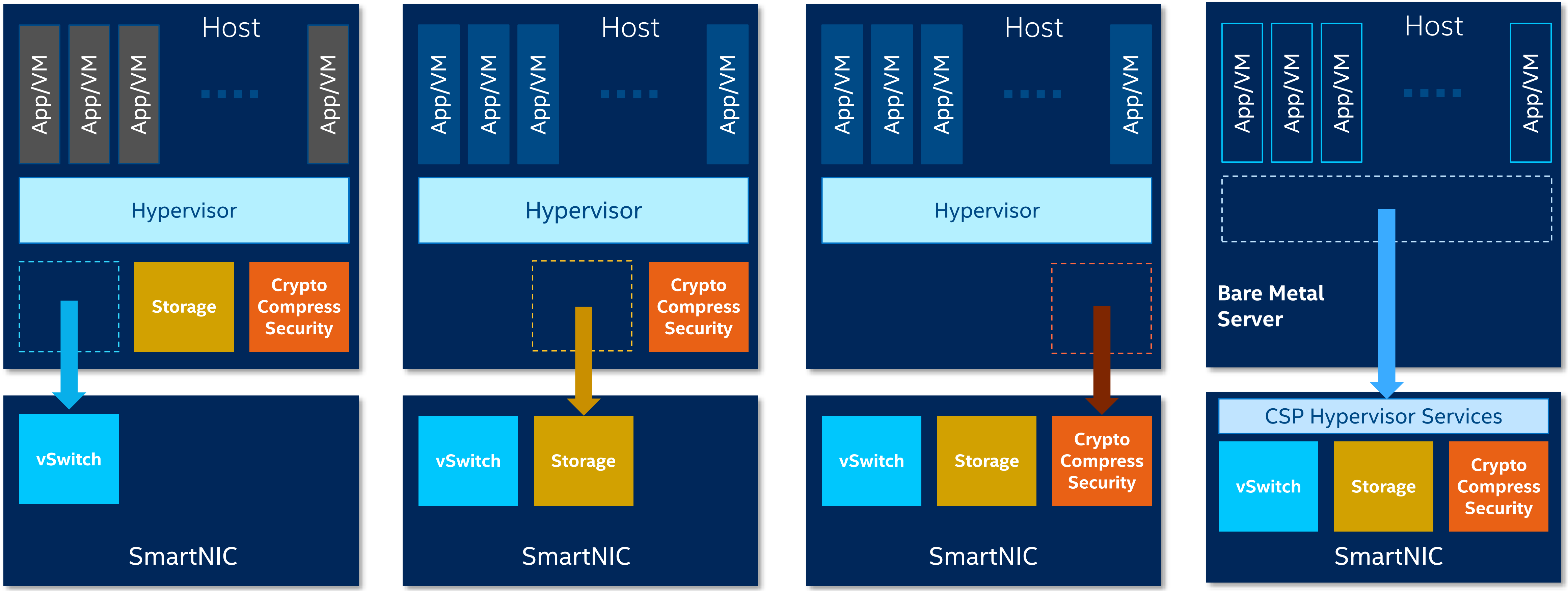
SmartNIC Products From 3rd Parties



Intel's Role
Deliver & Support Reference Platform (H/W & S/W)

Partner Role
Production SmartNIC & Solution Support

Leading CSPs are Driving Infrastructure Acceleration



1
- vSwitch widely deployed by CSPs
- First step for most customers
- Infrastructure services accelerated

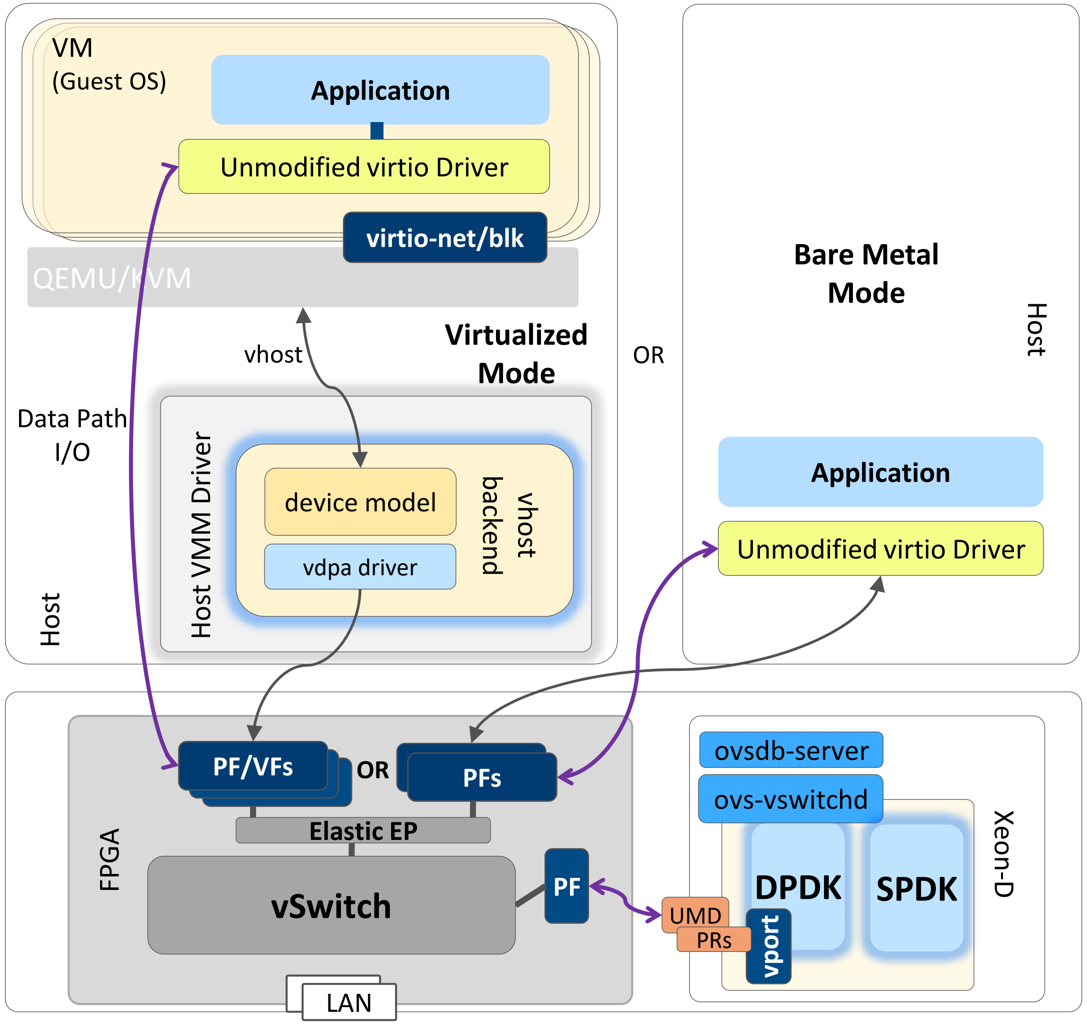
2
- Storage throughput increasing
- Second step for many customers

3
- Inline ops can be accelerated
- Often paired with storage

4
- Tenants rent entire physical server
- Bare metal value
- Improved performance predictability and security for tenant
- Better isolation for landlord

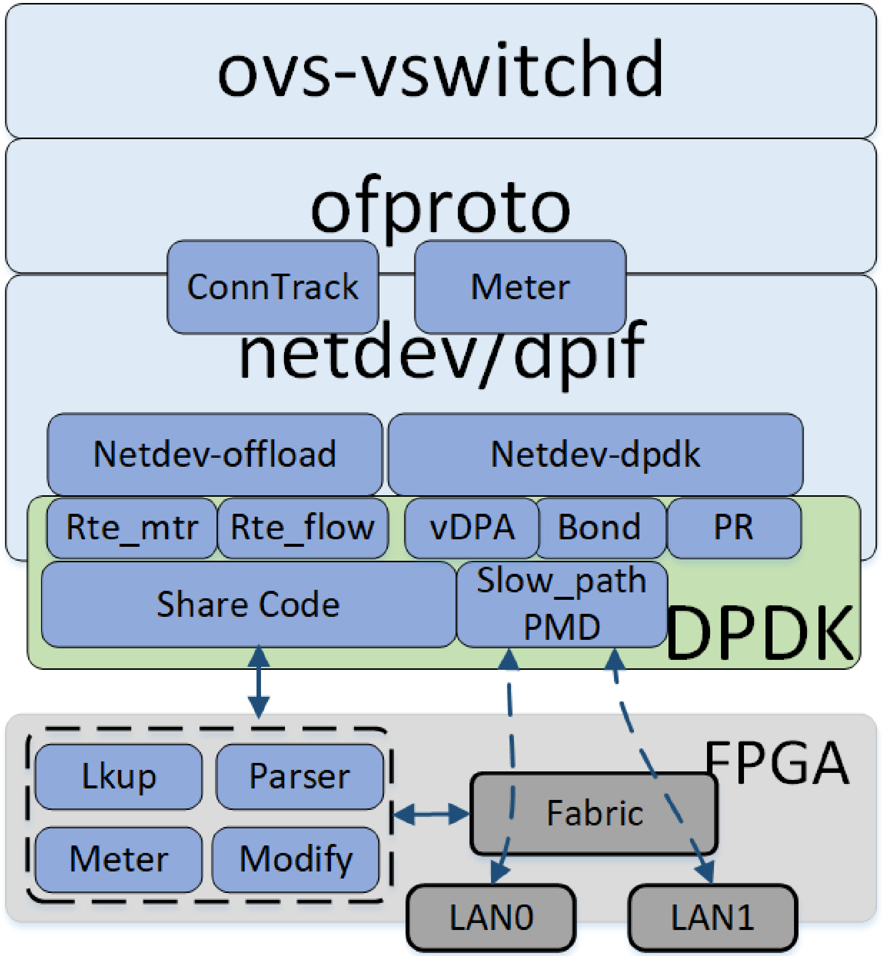
IPU/SmartNIC Software Highlight

Full-Stack Solution



- Virtio SW Ecosystem(net/blk 0.95 and 1.0)
- Virtualization Private Cloud(VPC) is powered by DPDK
- Elastic Block Storage(EBS) is powered by SPDK
- Virtualization and Bare Metal scenarios
- Elastic End Point(PFs/VFs)
- UEFI boot with virtio-blk and PXE virtio-net

Virtualization Private Cloud Acceleration



Data path accelerated by HW

- High throughput
- Committed bandwidth and low latency
- OvS full offload (L2/L3/VxLAN/Geneve/CT/NAT/Meter/GSO)
- High-capacity flow tables

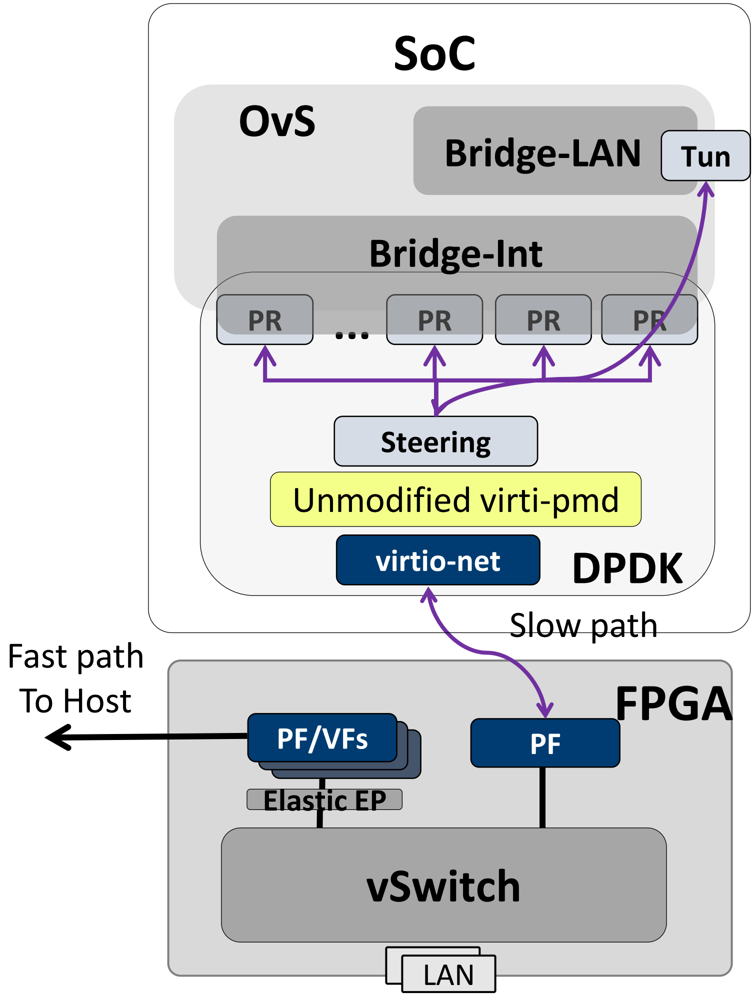
Control plane running in SoC

- OvS DPDK based slow path
- Rte_flow, Rte_mtr API implementation
- Port Representor for host interface
- Rte_eth_bonding for LAG

SW Ecosystem affinity

- Base on OpenSource Community baseline package
- No aware of API change from applications
- Reconfigurability and programmability with SW iteration and evolution

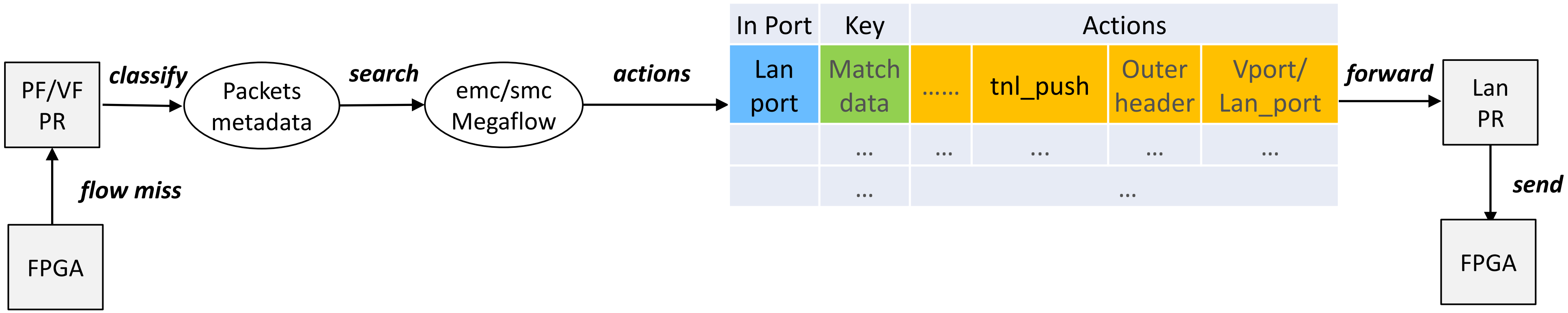
OvS Tunnel Offload Challenge



- There are 2 bridges in OVS instance
  - Br-int: VMs connection and VxLAN/Geneve tunnel access
  - Br-LAN: VxLAN/Geneve peer connection
- Bridges are connected by vport
  - VxLAN/Geneve pop/push
  - Packet recircle in SW pipeline
- Difference between OvS SW pipeline and HW acceleration
  - Encap and Decap are basic operations supported by HW
  - It's not friendly for HW to aware of vport

OvS Tunnel Encap Full Offload

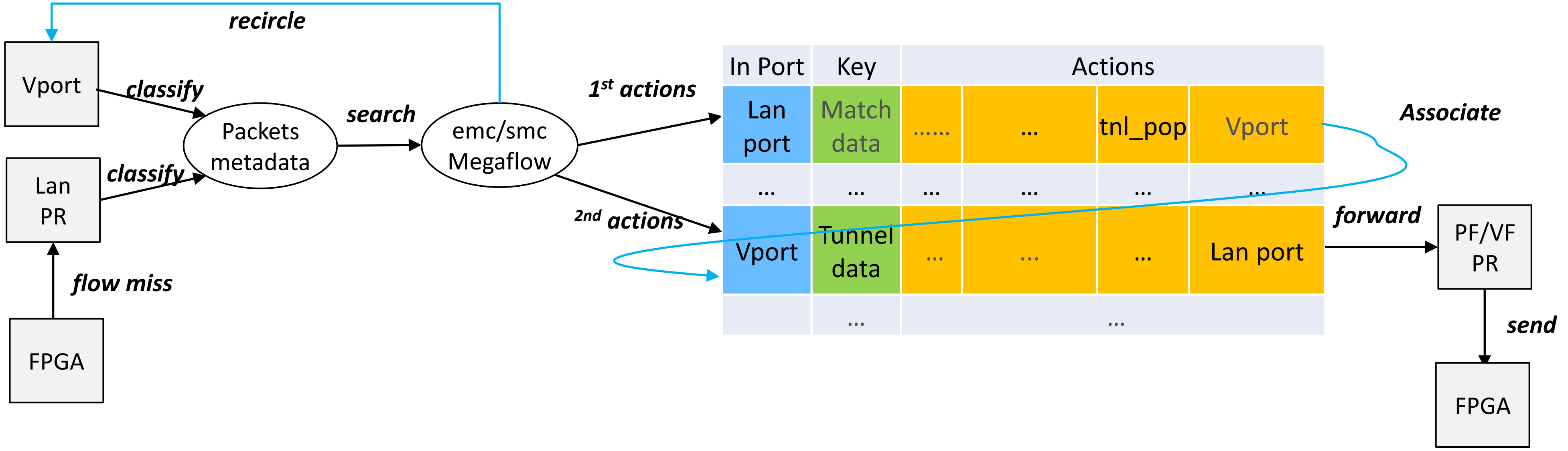
- VxLAN/Geneve Encap – tnl_push
  - Executed by one netdev rte_flow
  - Entire packet header and Encap date in one flow
  - Easily to reuse existing DPDK rte_flow offload process



In Port	Key	Actions			
Lan port	Match data	……	tnl_push	Outer header	Vport/ Lan_port
	…	…	…	…	…
	…	…			

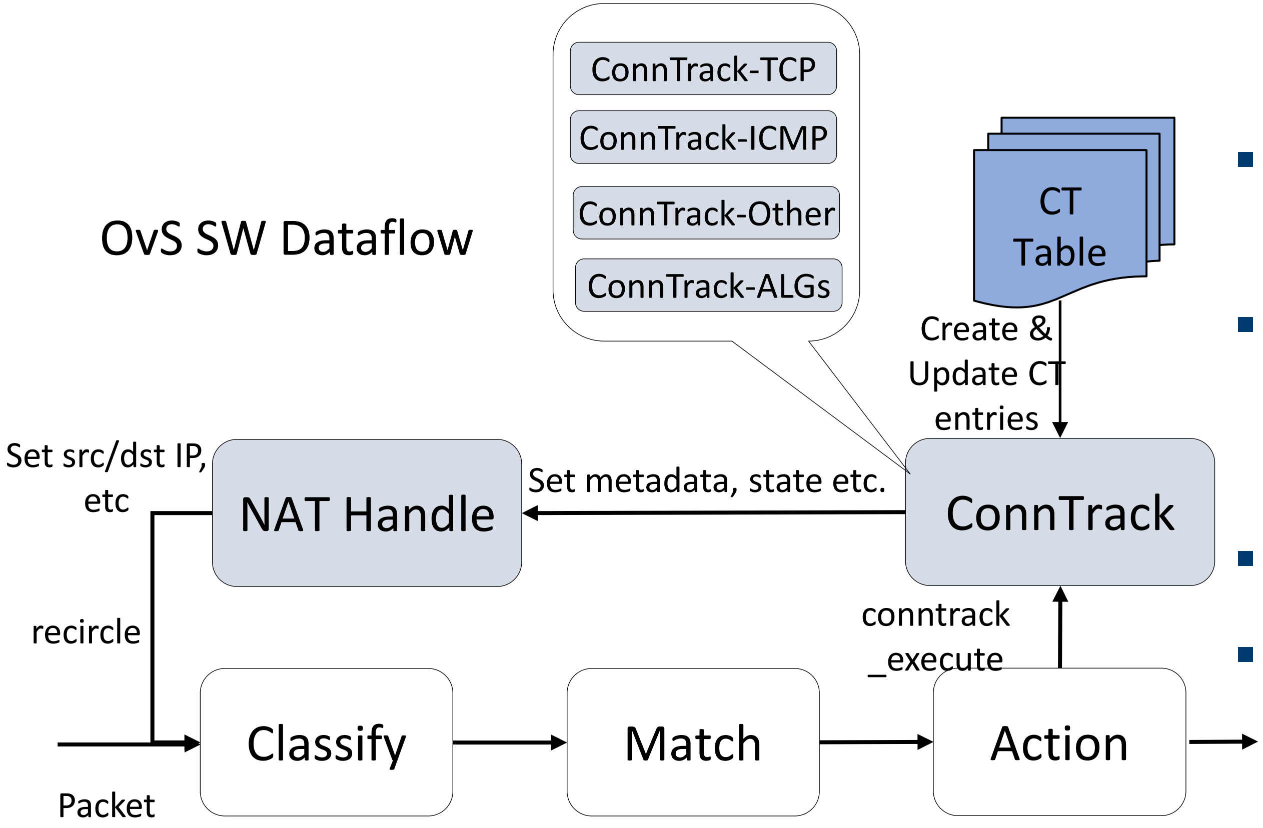
OvS Tunnel Decap Full Offload

- VxLAN/Geneve Decap – tnl_pop
  - Executed by two netdev rte_flow
  - Decap tunnel and forward to vport in 1st netdev rte_flow
  - Packets recircled in OvS SW pipeline and execute 2nd netdev rte_flow
  - Focusing on association for two netdev rte_flow



In Port	Key	Actions			
Lan port	Match data	……	…	tnl_pop	Vport
…	…	…	…	…	…
Vport	Tunnel data	…	…	…	Lan port
	…	…			

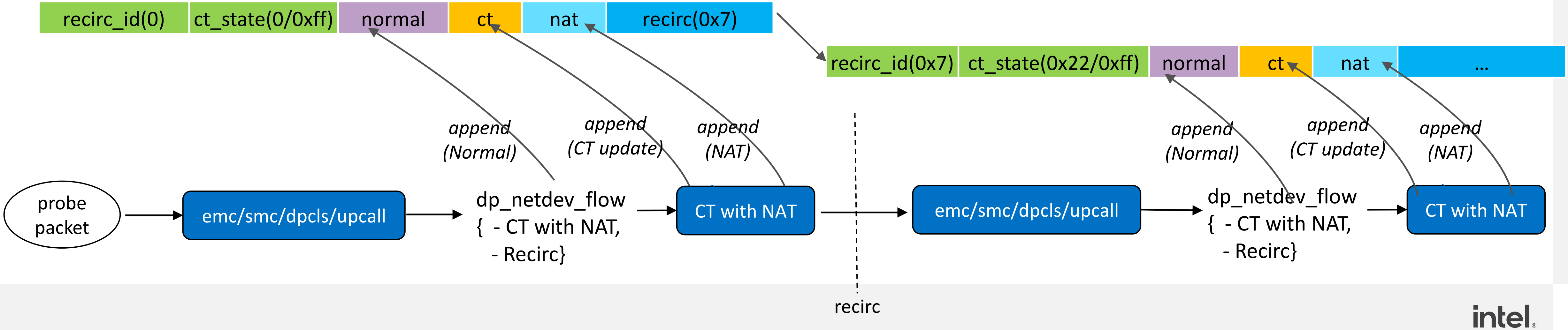
OvS ConnTrack Offload Challenge



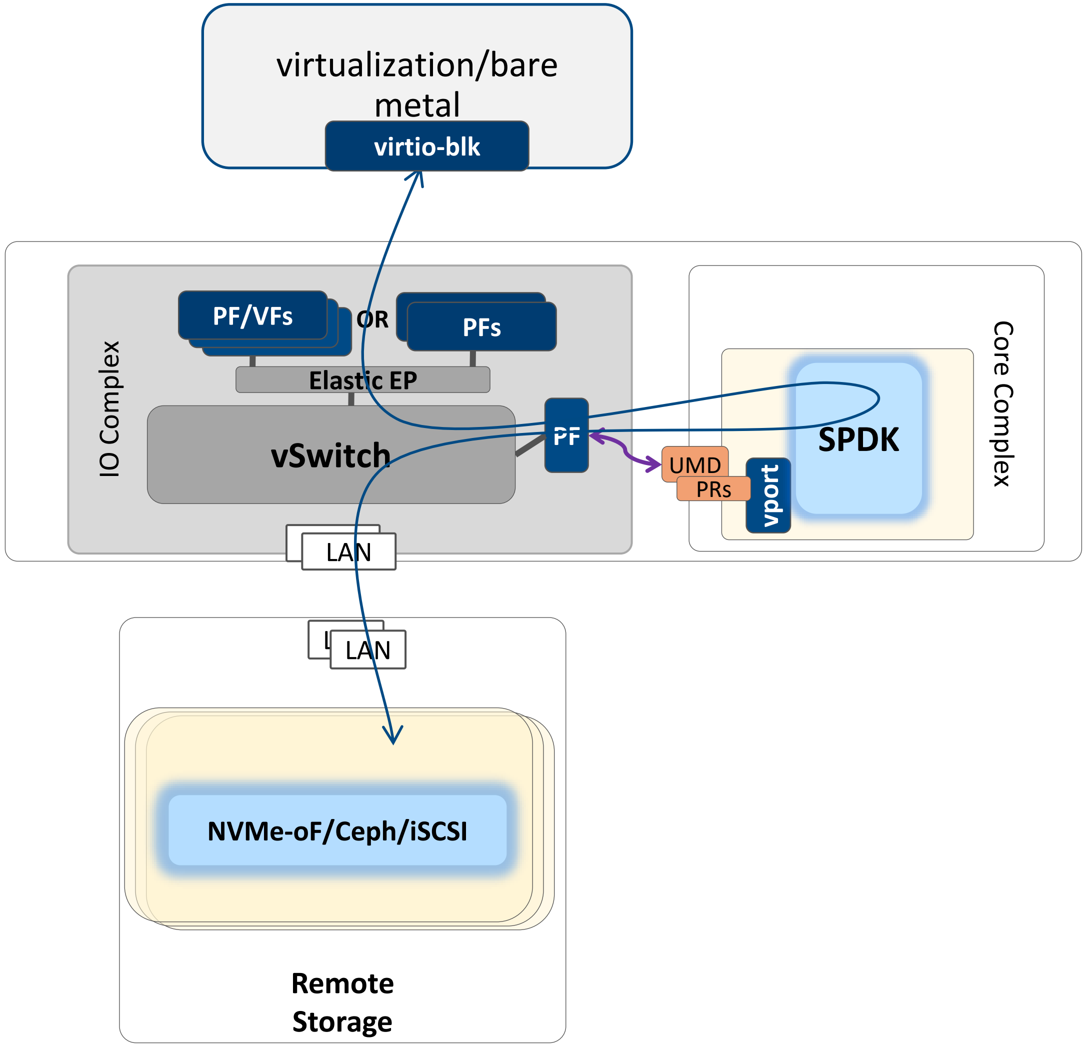
- Tracking both stateless and stateful protocols(TCP/ICMP/ALG etc)
- CT state for every packets(New/Established/Related/Invalid)
- Recircle with new CT state
- Supporting NAT features

OvS ConnTrack DPDK Flow Chain

- DPDK is enhanced to support flow recirc
  - OvS recirc_id is represented by rte_flow_attr->group
  - Implement RTE_FLOW_ACTION_TYPE_JUMP action
  - For HW supports recircle, it’s easy to mapping rte_flow chain to HW pipeline
  - For HW doesn’t support recircle, rte_flow chain merging is necessary
- ConnTrack state offload
  - For HW implemented ConnTrack FSM – State of every packet remains consistent between SW and HW
  - For HW doesn’t implemented ConnTrack FSM – Making the decision to offload in packets in ‘est’ state



Elastic Block Storage Acceleration



- Storage offload
  - SPDK based acceleration
  - IO pass to SoC SPDK
- Remote storage
  - NVMe-over Fabric target
  - Ceph RBD target
  - iSCSI target
  - Storage Volume Resize
  - QoS and Bond
- Remote boot
  - Virtualization scenario, it’s what it is
  - BM scenario, UEFI OptionROM virtio driver probe and boot
  - Cloud remote boot, legacy PXE boot

intel®