COMMS SUPPORT IN INTEL® ETHERNET 800 SERIES

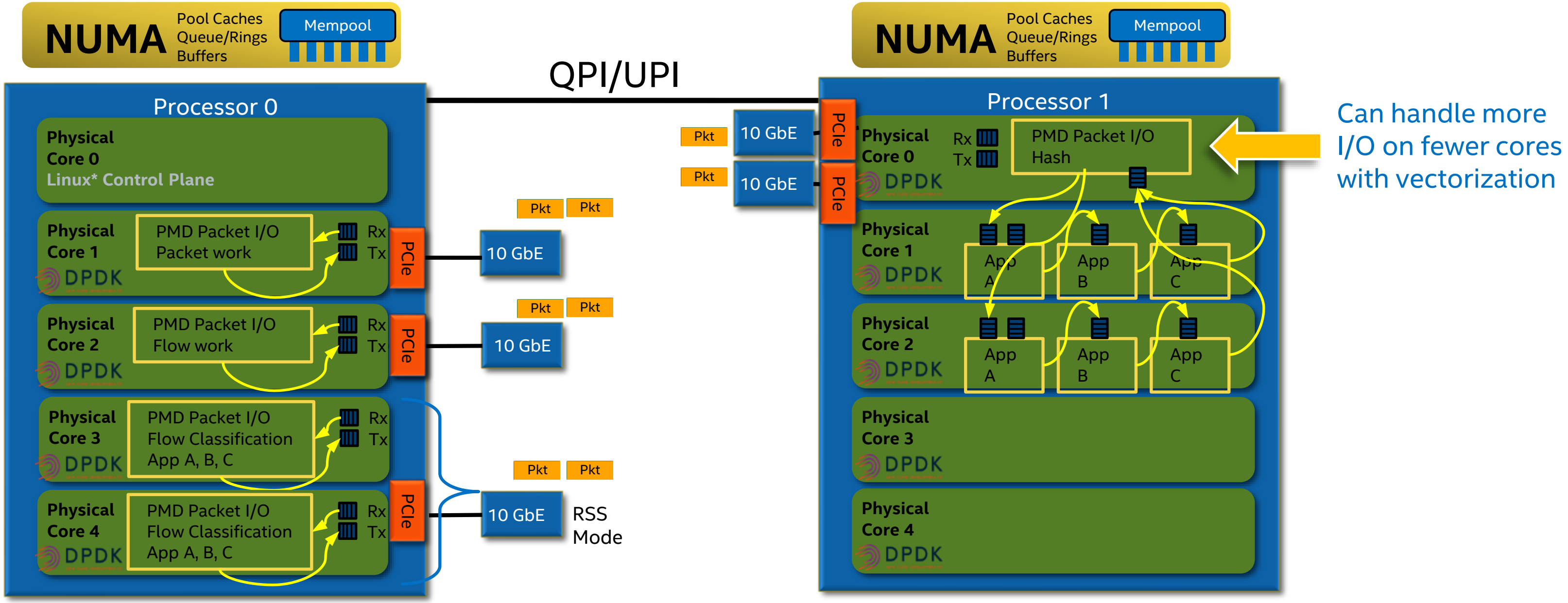
NPG PRC PP SW Team
June 2020

Agenda

- Background
- Programable Pipeline matters in comms
- VNF Use cases

PCIe* Connectivity and Core Usage

Using run-to-completion or pipeline software models



Run to Completion Model

- I/O and Application workload can be handled on a single core
- I/O can be scaled over multiple cores

Pipeline Model

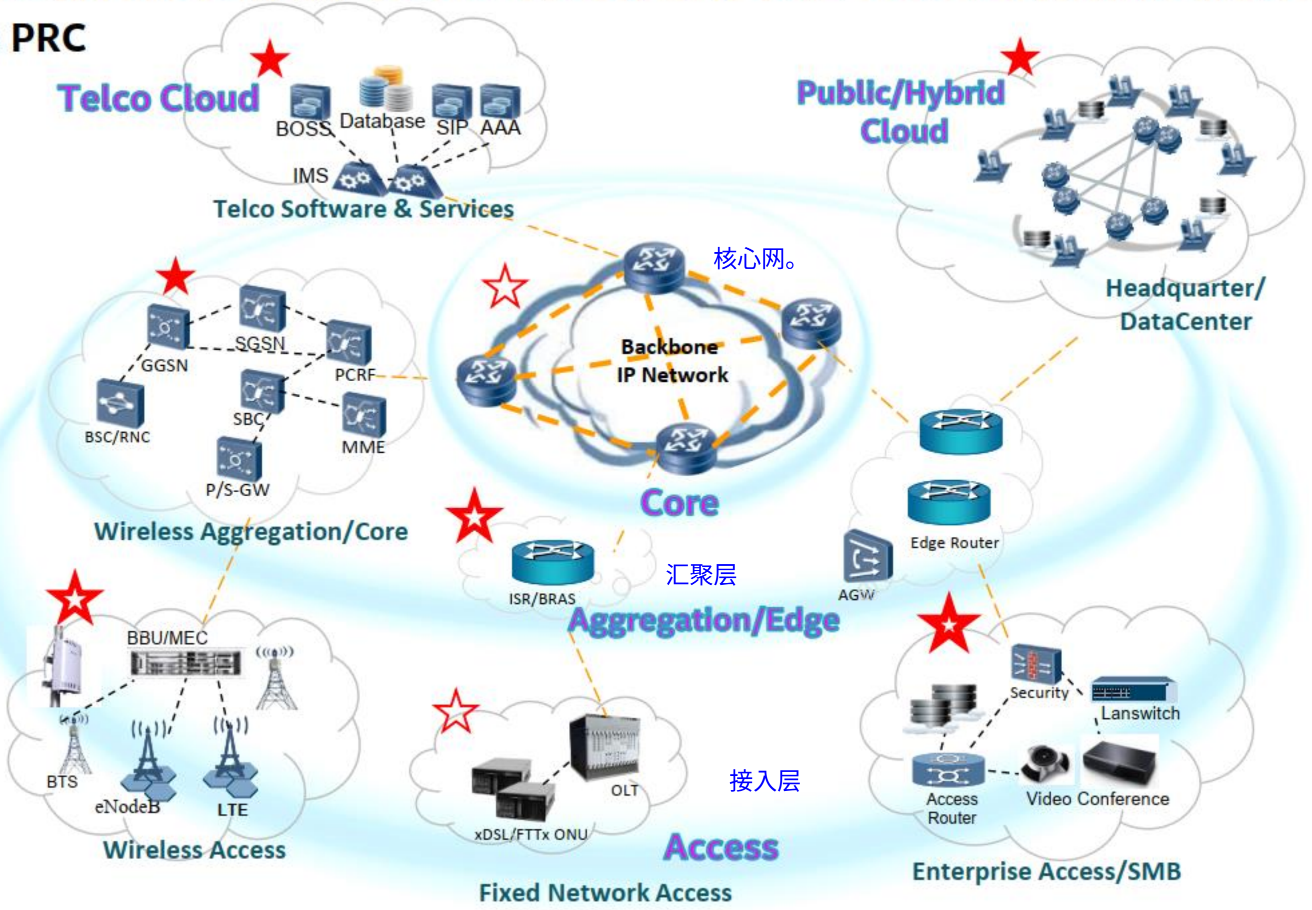
- I/O application disperses packets to other cores
- Application work performed on other cores

```
default_hugepagesz=1G hugepagesz=1G hugepages=16 hugepagesz=2M hugepages=2048 isolcpus=1-11,22-33 nohz_full=1-11,22-33 rcu_nocbs=1-11,22-33
```

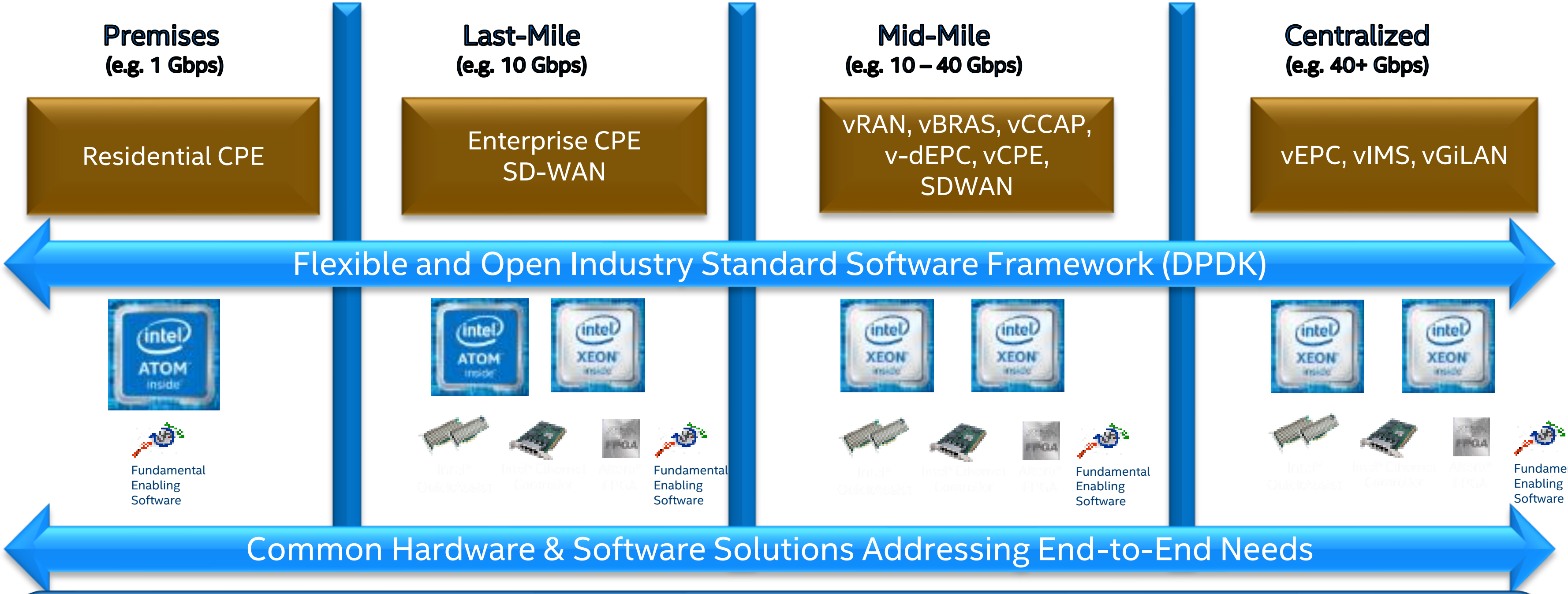
Note: nohz_full and rcu_nocbs is to disable Linux* kernel interrupts, and it’s important for zero-packet loss test. Generally, 1G huge pages are used for performance test.

Architecture Conversion In Telco Network and Cloud Infrastructure

★ IA Adoptions In PRC



Meeting Scalable Data Plane Needs



Premises
(e.g. 1 Gbps)

Residential CPE

Last-Mile
(e.g. 10 Gbps)

Enterprise CPE
SD-WAN

Mid-Mile
(e.g. 10 – 40 Gbps)

vRAN, vBRAS, vCCAP, v-dEPC, vCPE, SDWAN

Centralized
(e.g. 40+ Gbps)

vEPC, vIMS, vGiLAN

Flexible and Open Industry Standard Software Framework (DPDK)

Fundamental Enabling Software

Intel® QuickAssist

Intel® Ethernet Controller

Altera® FPGA

Fundamental Enabling Software

Intel® QuickAssist

Intel® Ethernet Controller

Altera® FPGA

Fundamental Enabling Software

Intel® QuickAssist

Intel® Ethernet Controller

Altera® FPGA

Fundamental Enabling Software

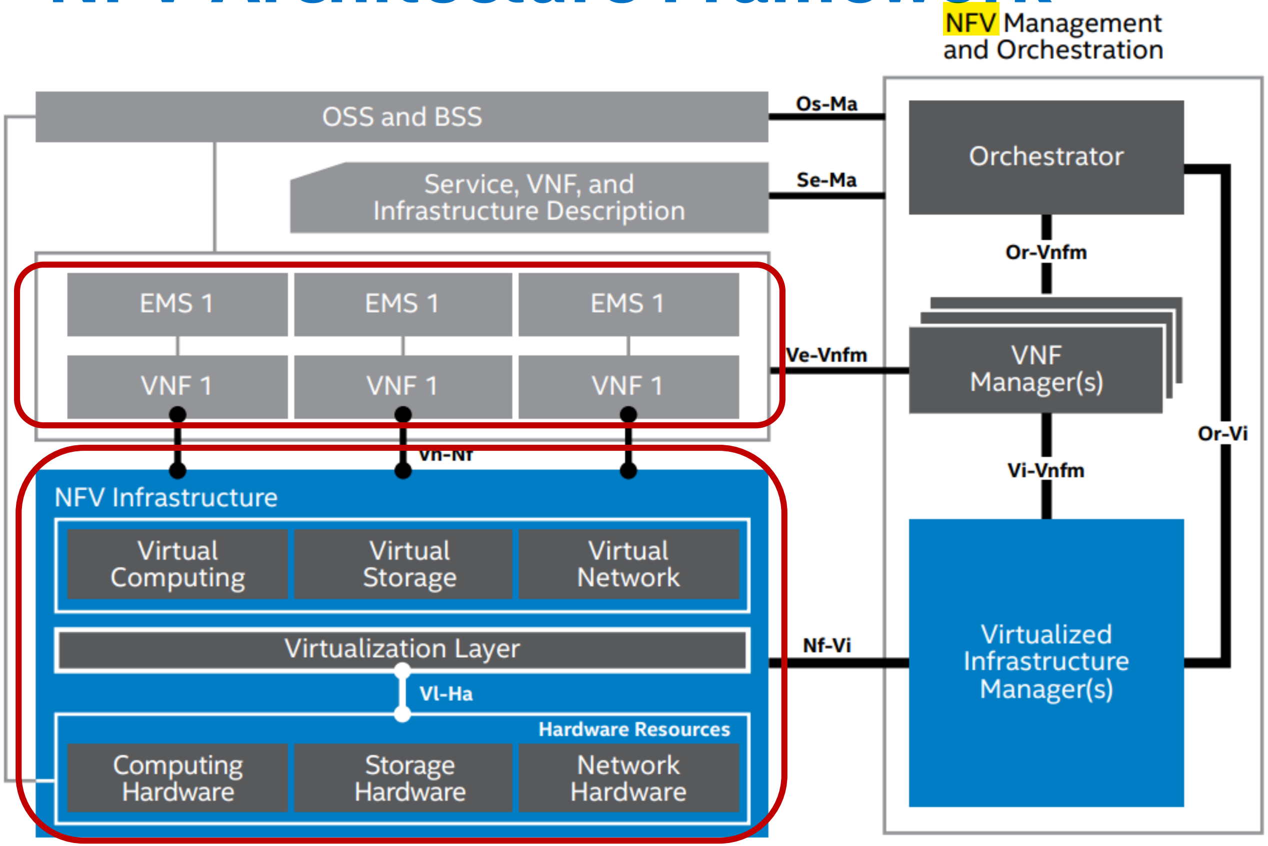
Common Hardware & Software Solutions Addressing End-to-End Needs

- ✓ 1G – 200G+ performance scaling with common architecture
- ✓ Preservation of customer software investment
- ✓ Hardware and software acceleration options to address specific needs

Workloads

Location	Workloads
Access	Base Station, OLT, Router
Aggregation/Edge	BNG, Gateway, UPF
Core	UPF
Cloud	vSwitch, Load Balancer…

NFV Architecture Framework



●━● Execution Reference Points

─ Other Reference Points

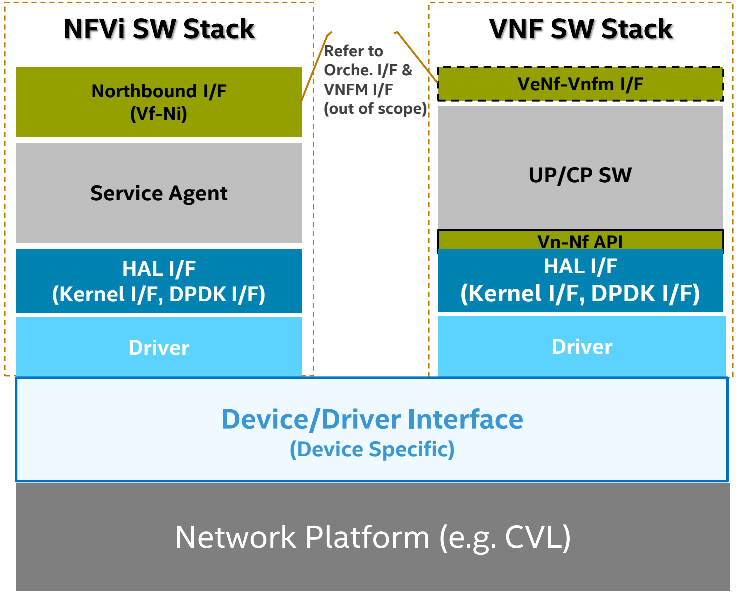
━ Main Network Functions Virtualization (NFV) Reference Points

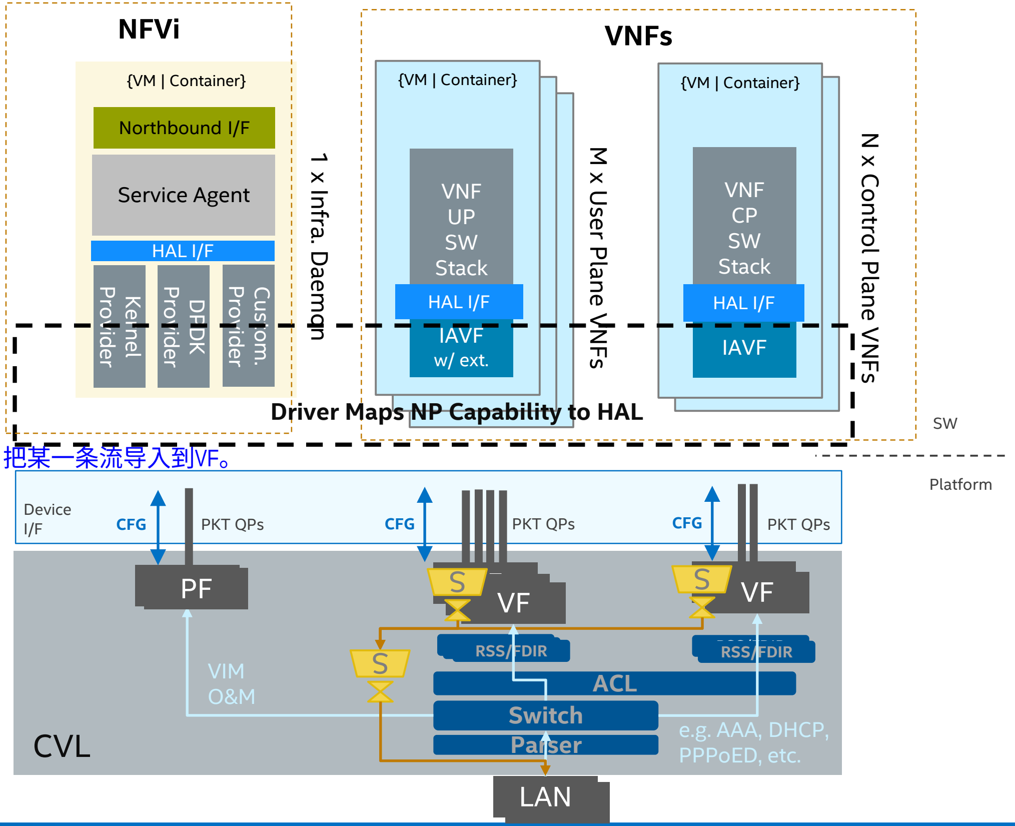
REFERENCE POINTS:

Vl-Ha	Virtualization Layer - Hardware Resources
Vn-Nf	VNF - NFV Infrastructure
Os-Ma	OSS/BSS - NFV Management and Orchestration
Se-Ma	Service, VNF and Infrastructure Description - NFV Management and Orchestration
Ve-Vnfm	VNF/EMS - VNF Manager
Nf-Vi	NFVI - Virtualized Infrastructure Manager
Or-Vnfm	Orchestrator - VNF Manager
Vi-Vnfm	Virtualized Infrastructure Manager - VNF Manager
Or-Vi	Orchestrator - Virtualized Infrastructure Manager

Map Network Platform Capability to SW Stack

把网卡上的不同能力map到上页NFV中





NFVi：把某一条流导入到VF。

云里雾里...问下jingjing

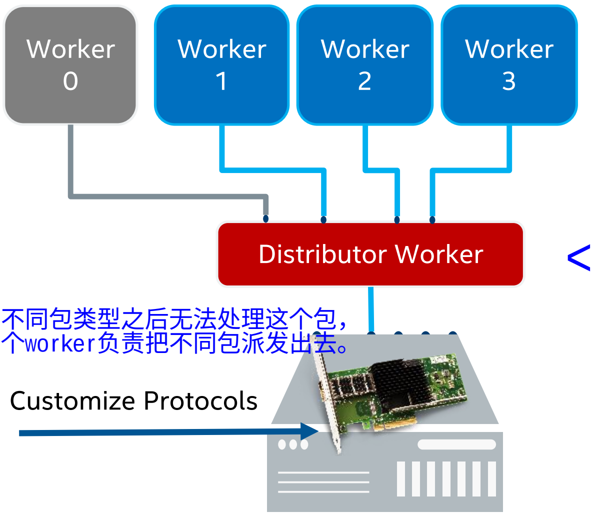
PROGRAMABLE PIPELINE MATTERS IN COMMS

为什么DDP很重要？

OS Protocol vs Comms Protocol

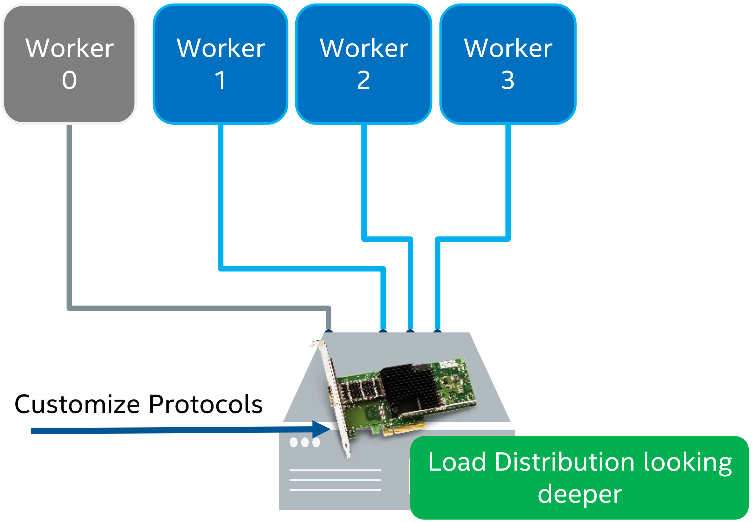
OS Protocols	Comms Protocols	
VLAN	S-VLAN	OSPF
IPv4	C-VLAN	BGP
IPv6	MPLS	GTP
ARP	IPv4	PFCP
ICMP	IPv6	PPPoE
UDP	ARP	L2TP
TCP	ICMP	ESP
VXLAN	UDP	……
	TCP	Customized Protocols

Programmable Pipeline Matters



<-瓶颈

汇聚层：
网卡收到不同包类型之后无法处理这个包，
需要有一个worker负责把不同包派发出去。



parser可以解析的更多。

Default pipeline relies on host to parse undefined protocols

Destination Address	Source Address	Undefined Protocol Start of Payload	?	?

Parsed fields

Payload

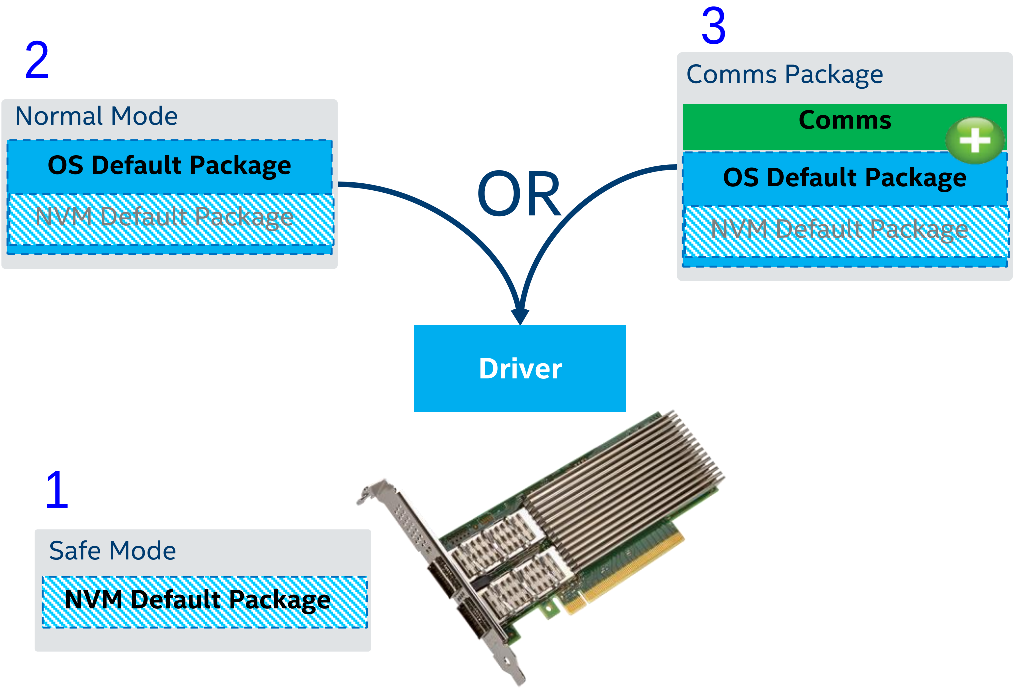
The pipeline parser looks deeper in to the packets

Destination Address	Source Address	Defined Protocol Outer Header	Inner Header	Payload

Parsed fields

Payload

Dynamic Device Personalization (DDP) Profile



2

Normal Mode

OS Default Package

NVM Default Package

OR

3

Comms Package

Comms

OS Default Package

NVM Default Package

Driver

1

Safe Mode

NVM Default Package

- **NVM Default Package**

Is a bare minimum package which allows reliable connectivity with basic protocols

- **OS Default Package**

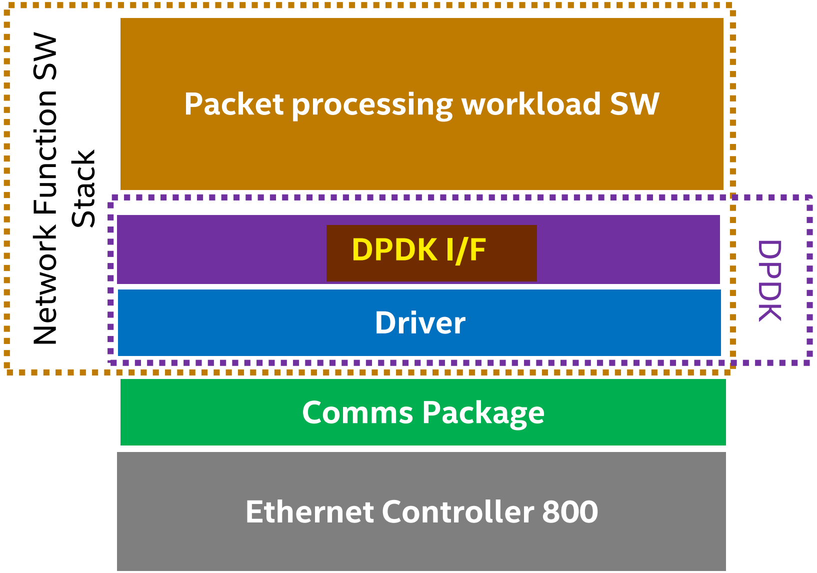
Baseline support for well known protocols and configurations

- **Comms Package**

Comms protocols and configuration built on top of (incremental) existing OS default package

20.05

Bridge SW Stack to Programable Pipeline



DPDK is the bridge between packet processing workload and hardware programable pipeline capability.

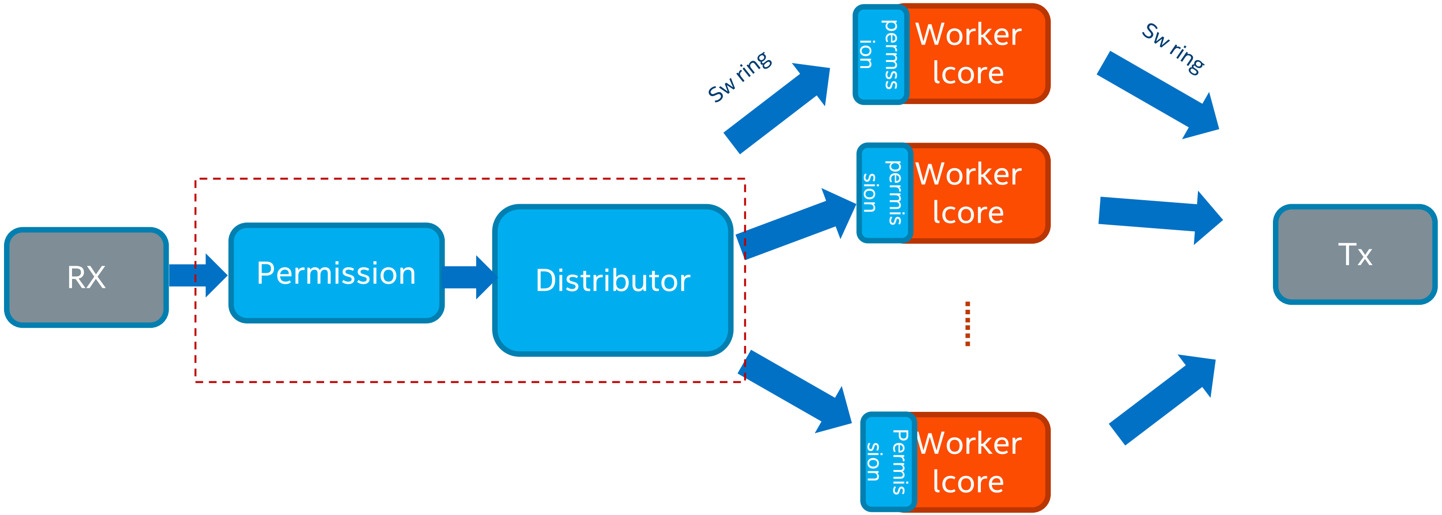
- Download package
- Run time table and rule programing

DPDK需要提供api维护各种表。

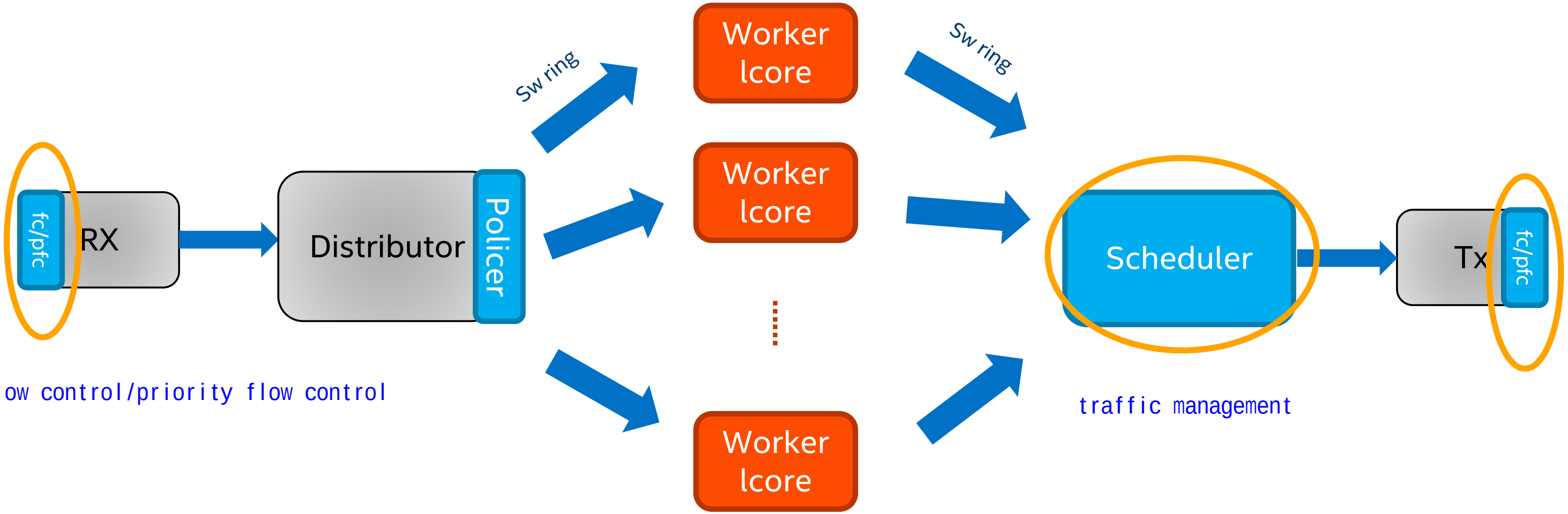
看一下相关的代码。

VNF USE CASES

Simple Pipeline of packet process - Classify



Simple Pipeline of packet process - QoS



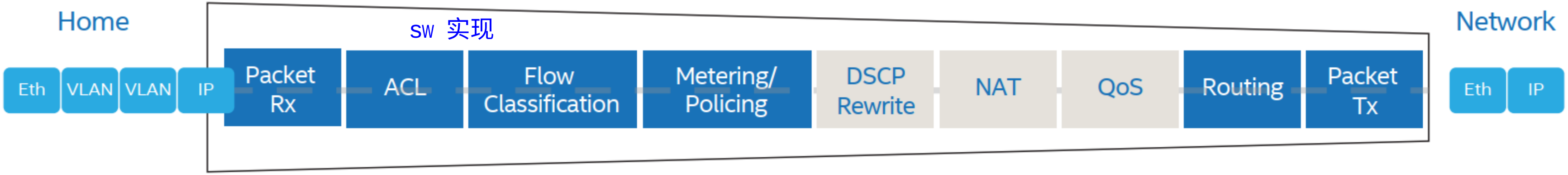
flow control/priority flow control

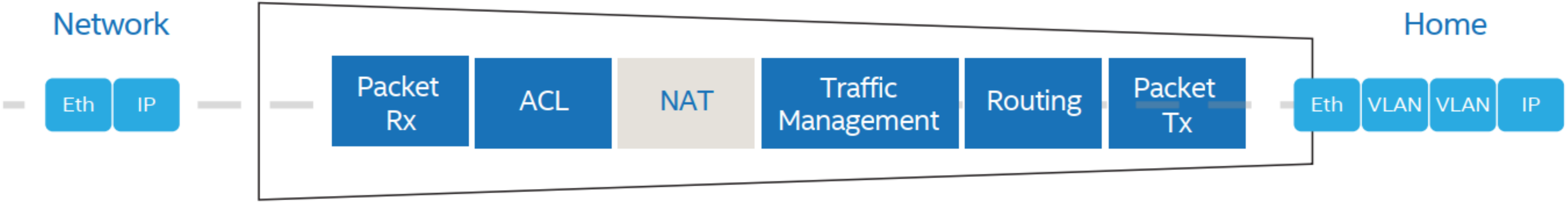
traffic management

vBNG Data Plane Processing Stage Example

汇聚侧，宽带接入网关

上行。meter ing是干啥的？和前面两张图怎么对应？





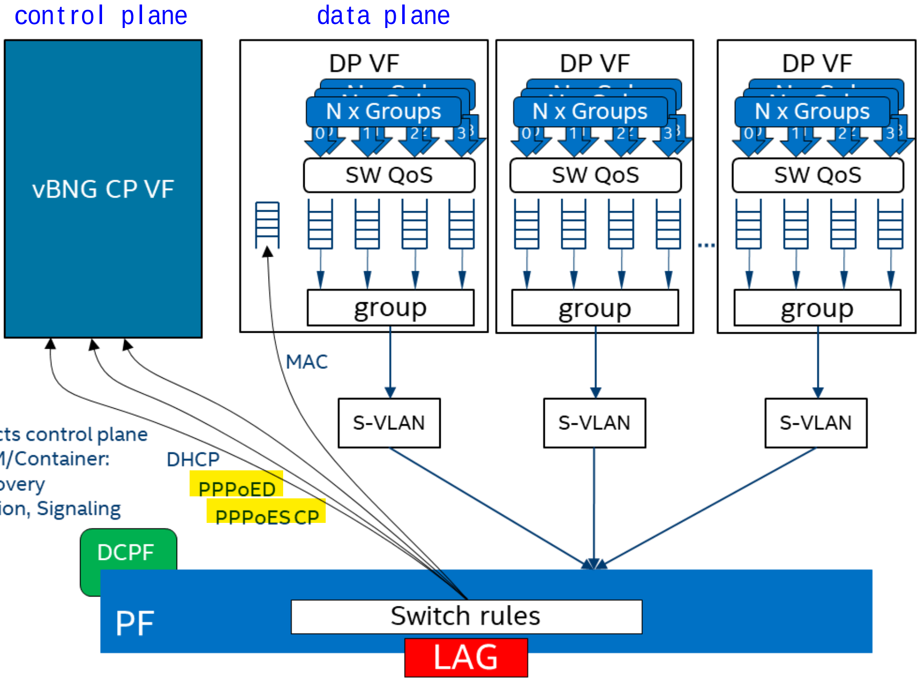
下行。

vBNG requirement to CVL

对QoS的要求很高，k级的用户。

虚拟化。从BNG的角度将控制面和数据面分开。

运行在不同的vm、容器中。



One RX core in run to completion mode.

SW HQOS consolidates subscribers down into predefined groups (maintaining TC)

4 x TX queues per group to enable final QOS in HW

Allows HQOS to scale and enables BW sharing across sub ports/VFs

保证带宽
限流

PF Switch directs control plane traffic to CP VM/Container:

- PPPoE Discovery
- PPPoE Session, Signaling
- DHCP

PPPoE ip flow classification

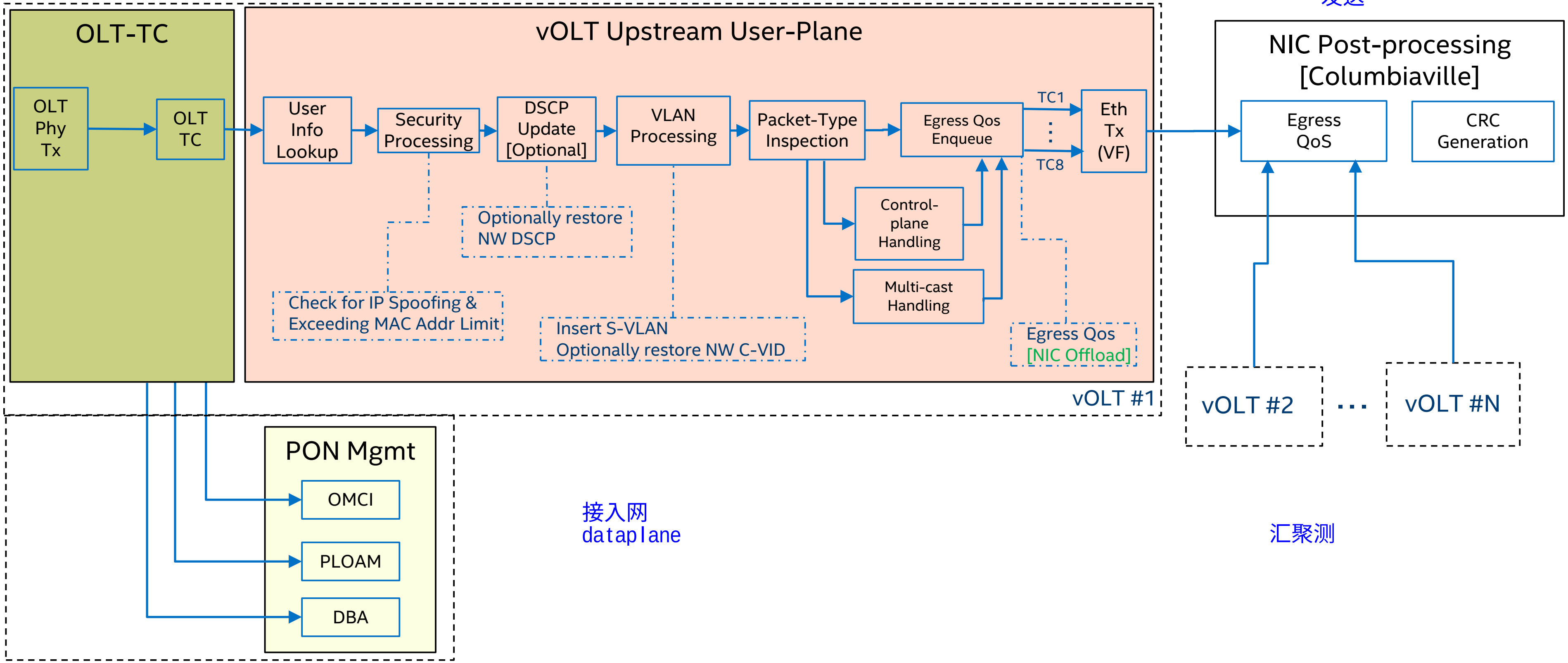
基于MAC的发送到DP VF。基于PPPoE等发送到CP VF。

接入侧

vOLT Data Plane Processing Stage Example

上行流量。包通过网卡发送出去。

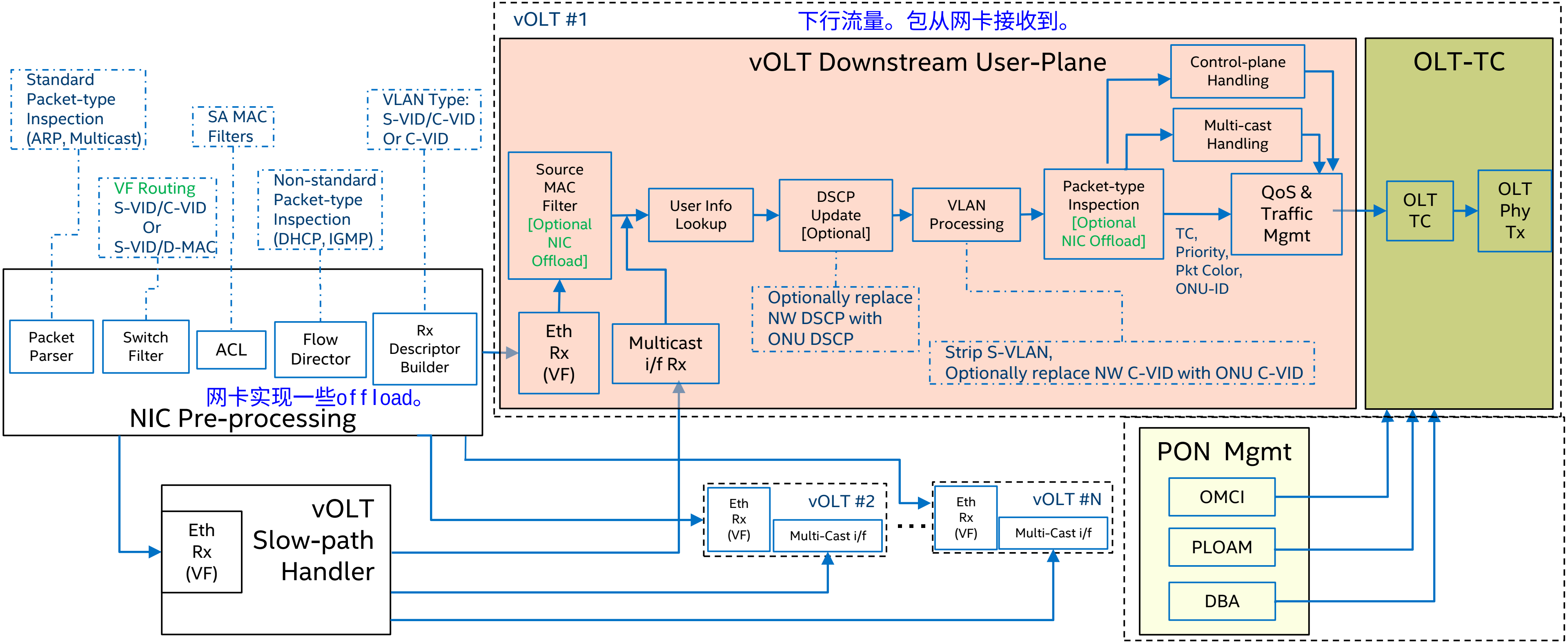
发送



接入网
dataplane

汇聚测

vOLT Data Plane Processing Stage Example



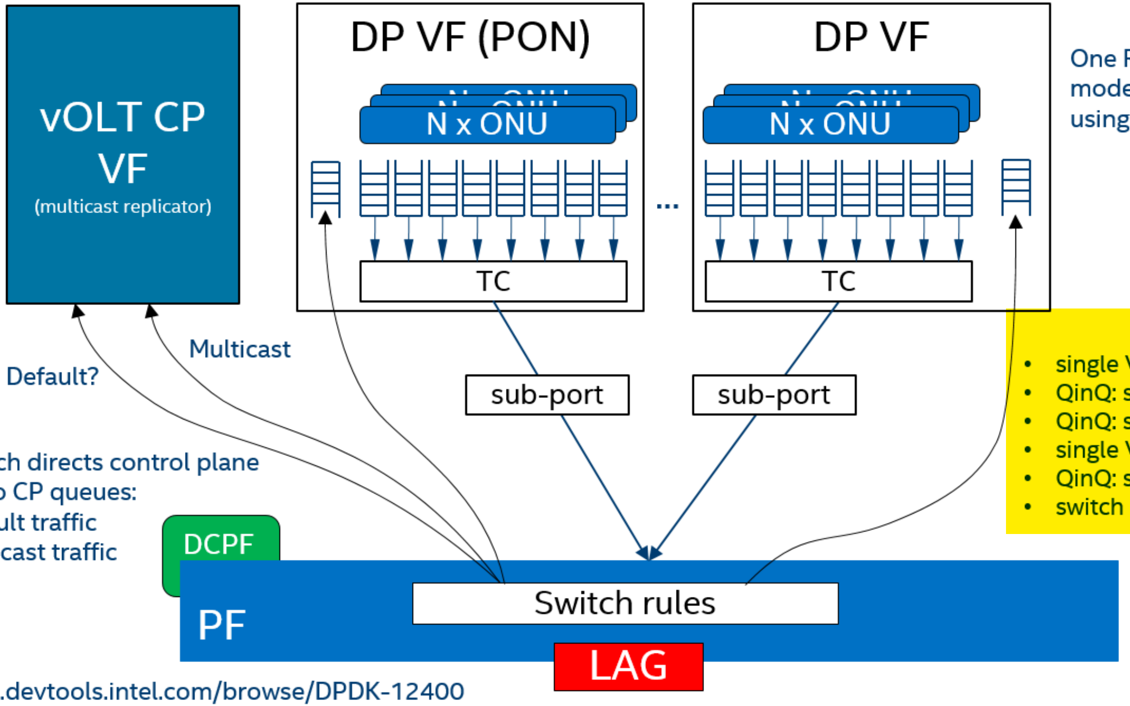
vOLT #1
vOLT Downstream User-Plane
OLT-TC
Standard Packet-type Inspection (ARP, Multicast)
SA MAC Filters
VLAN Type: S-VID/C-VID Or C-VID
VF Routing S-VID/C-VID Or S-VID/D-MAC
Non-standard Packet-type Inspection (DHCP, IGMP)
Packet Parser
Switch Filter
ACL
Flow Director
Rx Descriptor Builder
NIC Pre-processing
Source MAC Filter [Optional NIC Offload]
Eth Rx (VF)
Multicast i/f Rx
User Info Lookup
DSCP Update [Optional]
VLAN Processing
Packet-type Inspection [Optional NIC Offload]
Control-plane Handling
Multi-cast Handling
QoS & Traffic Mgmt
TC, Priority, Pkt Color, ONU-ID
OLT TC
OLT Phy Tx
Optionally replace NW DSCP with ONU DSCP
Strip S-VLAN, Optionally replace NW C-VID with ONU C-VID
vOLT Slow-path Handler
Eth Rx (VF)
vOLT #2
Eth Rx (VF)
Multi-Cast i/f
vOLT #N
Eth Rx (VF)
Multi-Cast i/f
PON Mgmt
OMCI
PLOAM
DBA

下行流量。包从网卡接收到。

网卡实现一些offload。

vOLT requirement to CVL

一个contrl plane的vm/container。多个data plane的vm/container.



One RX core in run to completion mode with TX QoS offloaded to HW using 8 TX queues/8 TC in ETS mode.

- single VLAN: switch on the VLAN (S-VLAN)
- QinQ: switch on S-VLAN + C-VLAN
- QinQ: switch on S-VLAN only
- single VLAN: switch on S-VLAN + MAC DA
- QinQ: switch on S-VLAN + MAC DA
- switch on MAC DA

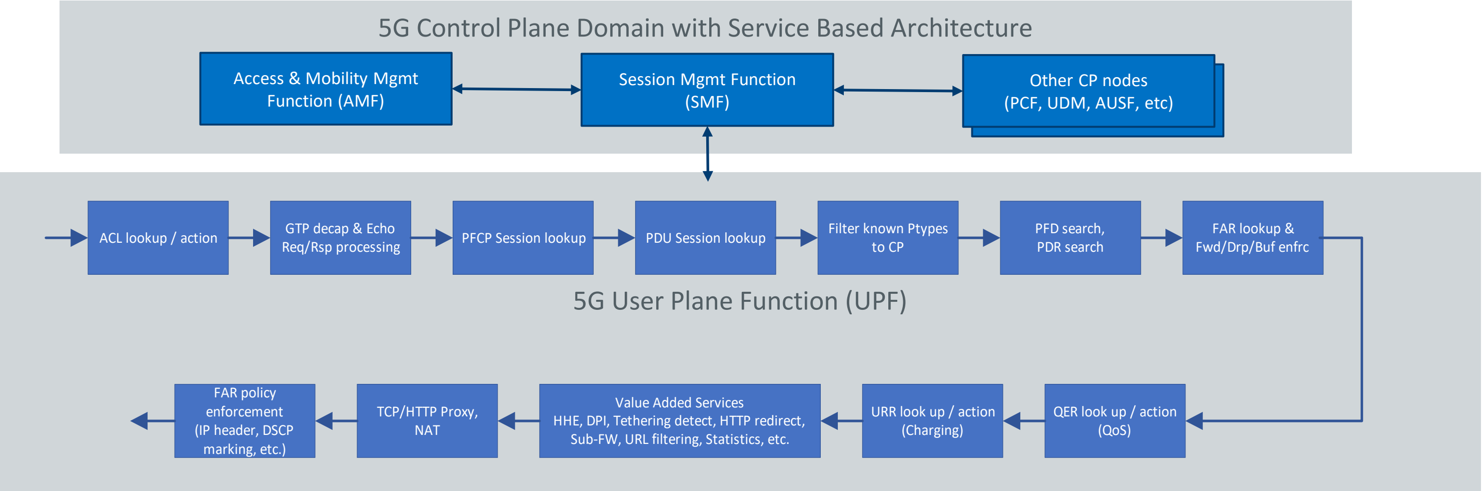
PF Switch directs control plane traffic to CP queues:

- Default traffic
- Multicast traffic

https://jira.devtools.intel.com/browse/DPDK-12400 vBNG和vOLT都是比较成熟的。

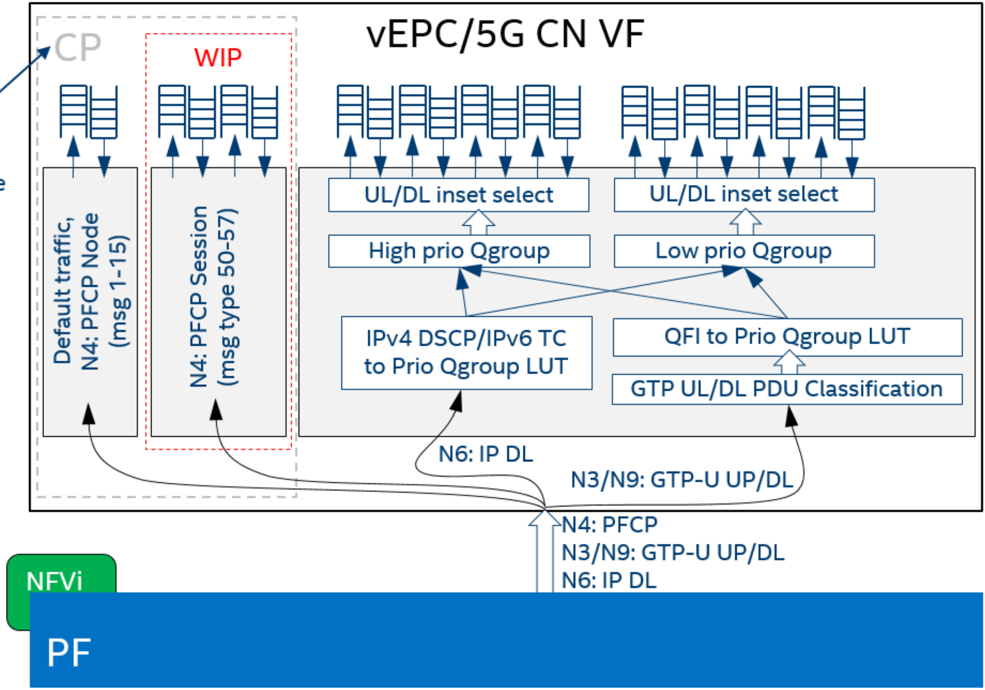
5G UPF – software-based pipeline

新型workload


5G Control Plane Domain with Service Based Architecture
Access & Mobility Mgmt Function (AMF)
Session Mgmt Function (SMF)
Other CP nodes (PCF, UDM, AUSF, etc)
ACL lookup / action
GTP decap & Echo Req/Rsp processing
PFCP Session lookup
PDU Session lookup
Filter known Ptypes to CP
PFD search, PDR search
FAR lookup & Fwd/Drp/Buf enfrc
5G User Plane Function (UPF)
QER look up / action (QoS)
URR look up / action (Charging)
Value Added Services HHE, DPI, Tethering detect, HTTP redirect, Sub-FW, URL filtering, Statistics, etc.
TCP/HTTP Proxy, NAT
FAR policy enforcement (IP header, DSCP marking, etc.)

vEPC/UPF

Control plane runs
on separate queues
in PF-only deployment
Will move to VF in future



不同的Qgroup根据inset做RSS。

根据priority导入不同的Qgroup。

在VF中，识别GTPU包

COMMS PROTOCOL ENABLE IN DPDK

Driver Enabling Software

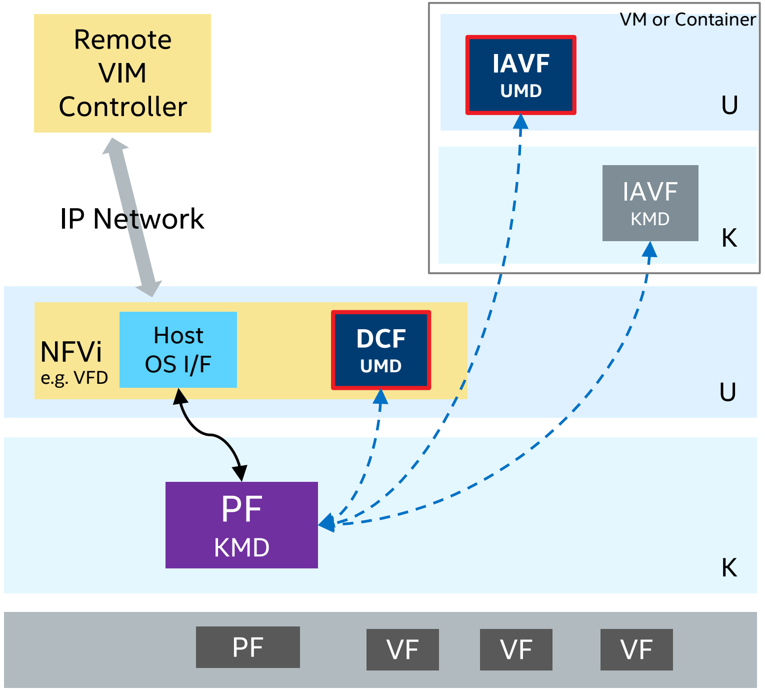
Mode	Network Function	Host Driver	VNF HAL I/F	NFVi HAL I/F	Problems
✓ non-IOV 没有virtualization的情况	DPDK PF	N/A	DPDK I/F (e.g. rte_flow, rte_tm & etc.)	N/A	
✓ IOV	DPDK AVF AVF-Intel VF标准	Kernel LAN	DPDK I/F (e.g. rte_flow, rte_tm & etc.)	Kernel I/F (e.g. ethtools, iprouter, tc & etc.)	VNF Accl. requires extra device advanced features against kernel mainstream readiness. 需要vf配置导流。
✗ IOV	DPDK AVF	DPDK PF	DPDK I/F (e.g. rte_flow, rte_tm & etc.)	DPDK I/F (e.g. rte_flow, rte_tm & etc.)	Hard to sustain with incremental complexity of PF/VF co-existence by Host PF PMD kernel mainstream user space driver develop framework does not support SRIOV management.

Driver Enabling SW Strategy

- Minimize replicated effort on device specific function enabling
- Incremental building re-useful SW for multi-gen NIC coverage
- NPG/ND co-design, KMD/UMD co-existence

DCF + iAVF advanced Feature is the way

基于VF：把DPDK的接口暴露到control plane，让客户配置。



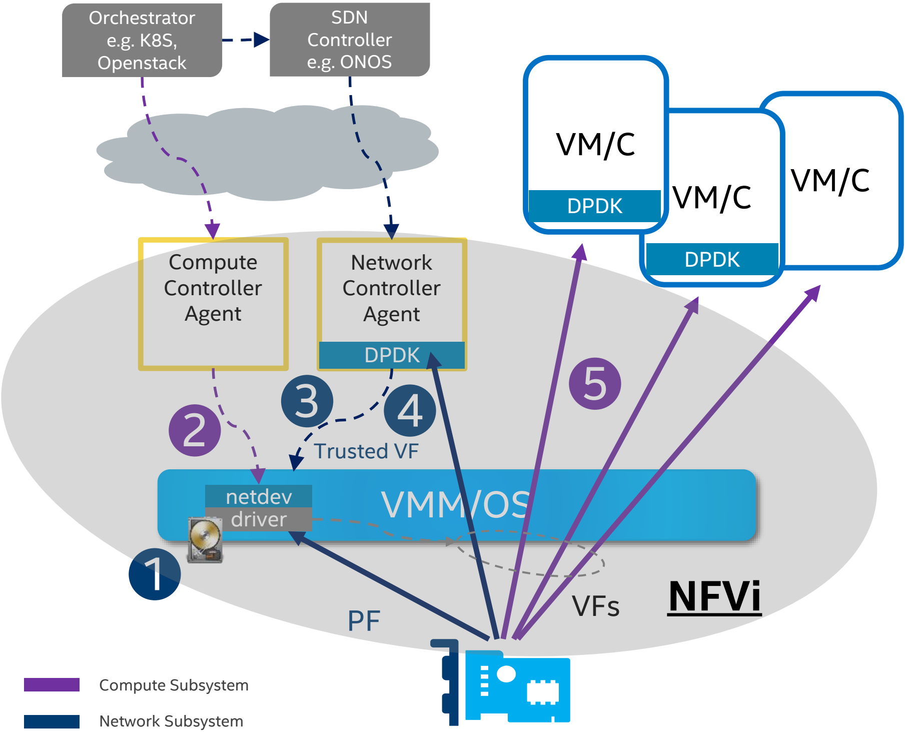
Configuration Flow

- DCF
  - Device Config Function
  - Advanced Cap over trust VF
  - Functional at named VF
  - Single entity per port
  - AQ-CMD over virtnnl
- iAVF
  - Extend virtnnl to support advanced feature
  - More protocols
  - RSS input set change
  - FD supporting

View of Deployment

后面haiyue会讲。

DCF，了解一下腾讯的use case

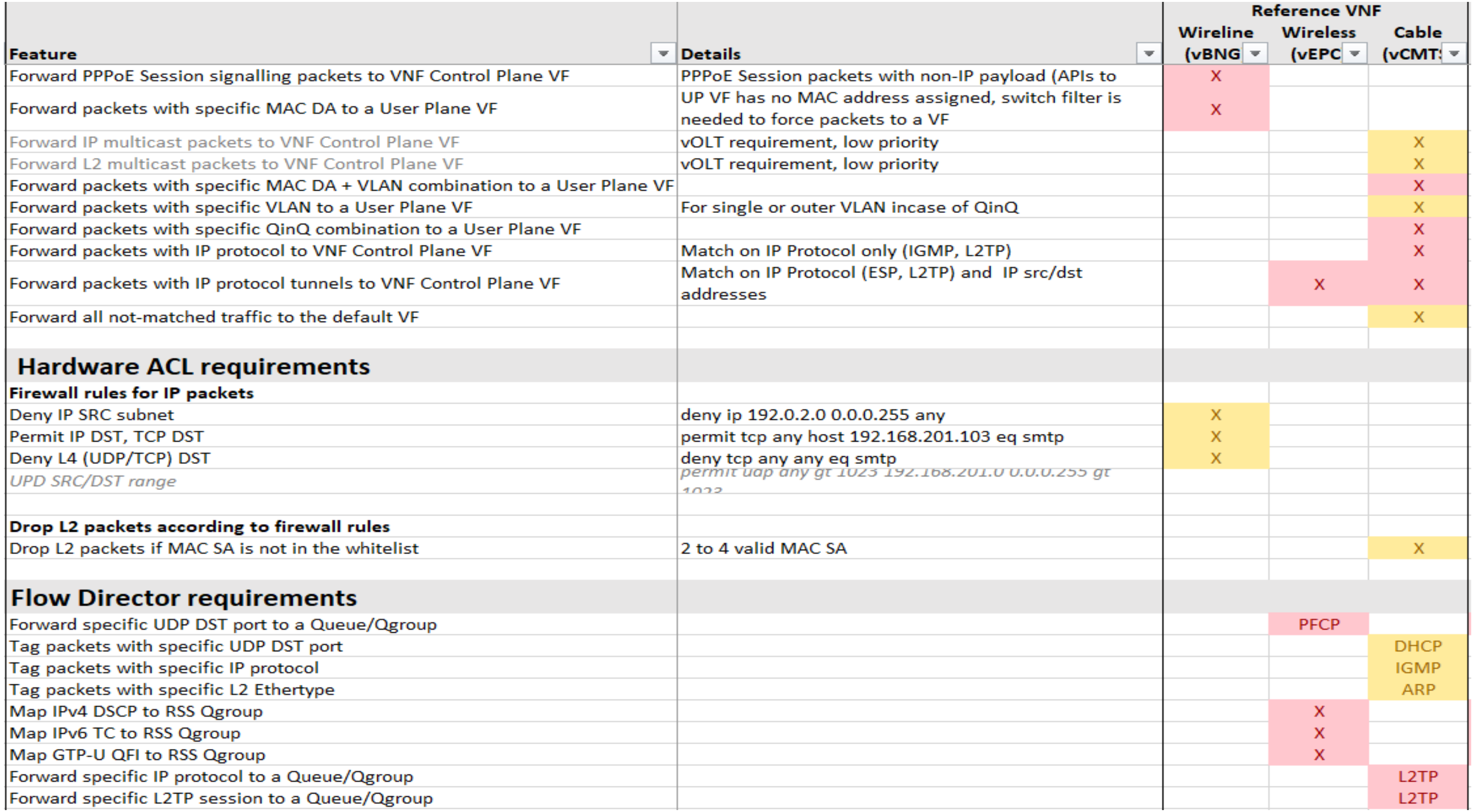


🟪 Compute Subsystem

🟦 Network Subsystem

1. Load DEF/COMM Package
2. PF driver generates SR-IOV VFs
   echo 4 > /sys/class/net/…/sriov_numvfs
3. Turn-on trust mode on a dedicated VF
   ip link set dev $eth $*vf*_id *trust* on
4. Assign trusted VF (VF0) to network controller agent
   … options : dpdk-devargs=$BDF,cap=dcf,representor=[x]
5. Assign other VFs to VMs/Containers

Requirements for Comms NFV



		Reference VNF		
Feature	**Details**	**Wireline (vBNG**	**Wireless (vEPC**	**Cable (vCMT:**
Forward PPPoE Session signalling packets to VNF Control Plane VF	PPPoE Session packets with non-IP payload (APIs to	X		
Forward packets with specific MAC DA to a User Plane VF	UP VF has no MAC address assigned, switch filter is needed to force packets to a VF	X		
Forward IP multicast packets to VNF Control Plane VF	vOLT requirement, low priority			X
Forward L2 multicast packets to VNF Control Plane VF	vOLT requirement, low priority			X
Forward packets with specific MAC DA + VLAN combination to a User Plane VF				X
Forward packets with specific VLAN to a User Plane VF	For single or outer VLAN incase of QinQ			X
Forward packets with specific QinQ combination to a User Plane VF				X
Forward packets with IP protocol to VNF Control Plane VF	Match on IP Protocol only (IGMP, L2TP)			X
Forward packets with IP protocol tunnels to VNF Control Plane VF	Match on IP Protocol (ESP, L2TP) and IP src/dst addresses		X	X
Forward all not-matched traffic to the default VF				X
Hardware ACL requirements				
Firewall rules for IP packets				
Deny IP SRC subnet	deny ip 192.0.2.0 0.0.0.255 any	X		
Permit IP DST, TCP DST	permit tcp any host 192.168.201.103 eq smtp	X		
Deny L4 (UDP/TCP) DST	deny tcp any any eq smtp	X		
UPD SRC/DST range	*permit udp any gt 1023 192.168.201.0 0.0.0.255 gt 1023*			
Drop L2 packets according to firewall rules				
Drop L2 packets if MAC SA is not in the whitelist	2 to 4 valid MAC SA			X
Flow Director requirements				
Forward specific UDP DST port to a Queue/Qgroup			PFCP	
Tag packets with specific UDP DST port				DHCP
Tag packets with specific IP protocol				IGMP
Tag packets with specific L2 Ethertype				ARP
Map IPv4 DSCP to RSS Qgroup			X	
Map IPv6 TC to RSS Qgroup			X	
Map GTP-U QFI to RSS Qgroup			X	
Forward specific IP protocol to a Queue/Qgroup				L2TP
Forward specific L2TP session to a Queue/Qgroup				L2TP

intel®