



DPDK

DATA PLANE DEVELOPMENT KIT

Network Interface Controller Drivers

Release 18.02.2

June 15, 2018

1	Overview of Networking Drivers	1
2	Features Overview	4
2.1	Speed capabilities	4
2.2	Link status	4
2.3	Link status event	4
2.4	Removal event	5
2.5	Queue status event	5
2.6	Rx interrupt	5
2.7	Lock-free Tx queue	5
2.8	Fast mbuf free	6
2.9	Free Tx mbuf on demand	6
2.10	Queue start/stop	6
2.11	MTU update	6
2.12	Jumbo frame	6
2.13	Scattered Rx	7
2.14	LRO	7
2.15	TSO	7
2.16	Promiscuous mode	7
2.17	Allmulticast mode	8
2.18	Unicast MAC filter	8
2.19	Multicast MAC filter	8
2.20	RSS hash	8
2.21	RSS key update	8
2.22	RSS reta update	9
2.23	VMDq	9
2.24	SR-IOV	9
2.25	DCB	9
2.26	VLAN filter	9
2.27	Ethertype filter	10
2.28	N-tuple filter	10
2.29	SYN filter	10
2.30	Tunnel filter	10
2.31	Flexible filter	10
2.32	Hash filter	10
2.33	Flow director	11
2.34	Flow control	11
2.35	Flow API	11
2.36	Rate limitation	11

2.37	Traffic mirroring	11
2.38	Inline crypto	11
2.39	CRC offload	12
2.40	VLAN offload	12
2.41	QinQ offload	12
2.42	L3 checksum offload	13
2.43	L4 checksum offload	13
2.44	Timestamp offload	13
2.45	MACsec offload	13
2.46	Inner L3 checksum	14
2.47	Inner L4 checksum	14
2.48	Packet type parsing	14
2.49	Timesync	14
2.50	Rx descriptor status	15
2.51	Tx descriptor status	15
2.52	Basic stats	15
2.53	Extended stats	15
2.54	Stats per queue	15
2.55	FW version	16
2.56	EEPROM dump	16
2.57	Registers dump	16
2.58	LED	16
2.59	Multiprocess aware	16
2.60	BSD nic_uio	16
2.61	Linux UIO	17
2.62	Linux VFIO	17
2.63	Other kdrv	17
2.64	ARMv7	17
2.65	ARMv8	17
2.66	Power8	17
2.67	x86-32	17
2.68	x86-64	17
2.69	Usage doc	18
2.70	Design doc	18
2.71	Perf doc	18
2.72	Other dev ops not represented by a Feature	18
3	Compiling and testing a PMD for a NIC	19
3.1	Driver Compilation	19
3.2	Running testpmd in Linux	20
4	ARK Poll Mode Driver	22
4.1	Overview	22
4.2	Device Parameters	23
4.3	Data Path Interface	23
4.4	Configuration Information	23
4.5	Building DPDK	24
4.6	Supported ARK RTL PCIe Instances	24
4.7	Supported Operating Systems	24
4.8	Supported Features	24
4.9	Unsupported Features	25
4.10	Pre-Requisites	25

4.11	Usage Example	25
5	AVP Poll Mode Driver	26
5.1	Features and Limitations of the AVP PMD	26
5.2	Prerequisites	27
5.3	Launching a VM with an AVP type network attachment	27
6	BNX2X Poll Mode Driver	28
6.1	Supported Features	28
6.2	Non-supported Features	28
6.3	Co-existence considerations	28
6.4	Supported QLogic NICs	29
6.5	Prerequisites	29
6.6	Pre-Installation Configuration	29
6.7	Driver compilation and testing	29
6.8	SR-IOV: Prerequisites and sample Application Notes	29
7	BNXT Poll Mode Driver	32
7.1	Limitations	32
8	CXGBE Poll Mode Driver	33
8.1	Features	33
8.2	Limitations	33
8.3	Supported Chelsio T5 NICs	33
8.4	Supported Chelsio T6 NICs	34
8.5	Prerequisites	34
8.6	Pre-Installation Configuration	34
8.7	Driver compilation and testing	35
8.8	Linux	35
8.9	FreeBSD	36
8.10	Sample Application Notes	39
9	DPAA Poll Mode Driver	40
9.1	NXP DPAA (Data Path Acceleration Architecture - Gen 1)	40
9.2	DPAA DPDK - Poll Mode Driver Overview	41
9.3	Supported DPAA SoCs	42
9.4	Prerequisites	43
9.5	Pre-Installation Configuration	44
9.6	Driver compilation and testing	45
9.7	Limitations	45
10	DPAA2 Poll Mode Driver	46
10.1	NXP DPAA2 (Data Path Acceleration Architecture Gen2)	46
10.2	DPAA2 DPDK - Poll Mode Driver Overview	50
10.3	Supported DPAA2 SoCs	52
10.4	Prerequisites	52
10.5	Pre-Installation Configuration	53
10.6	Driver compilation and testing	53
10.7	Limitations	54
11	Driver for VM Emulated Devices	55
11.1	Validated Hypervisors	55
11.2	Recommended Guest Operating System in Virtual Machine	55

11.3	Setting Up a KVM Virtual Machine	55
11.4	Known Limitations of Emulated Devices	57
12	ENA Poll Mode Driver	58
12.1	Overview	58
12.2	Management Interface	58
12.3	Data Path Interface	59
12.4	Configuration information	59
12.5	Building DPDK	60
12.6	Supported ENA adapters	60
12.7	Supported Operating Systems	60
12.8	Supported features	60
12.9	Unsupported features	60
12.10	Prerequisites	61
12.11	Usage example	61
13	ENIC Poll Mode Driver	62
13.1	How to obtain ENIC PMD integrated DPDK	62
13.2	Configuration information	62
13.3	Flow director support	63
13.4	SR-IOV mode utilization	63
13.5	Generic Flow API support	64
13.6	Limitations	65
13.7	How to build the suite	66
13.8	Supported Cisco VIC adapters	66
13.9	Supported Operating Systems	67
13.10	Supported features	67
13.11	Known bugs and unsupported features in this release	67
13.12	Prerequisites	68
13.13	Additional Reference	68
13.14	Contact Information	69
14	FM10K Poll Mode Driver	70
14.1	FTAG Based Forwarding of FM10K	70
14.2	Vector PMD for FM10K	70
14.3	Limitations	72
15	I40E Poll Mode Driver	74
15.1	Features	74
15.2	Prerequisites	75
15.3	Pre-Installation Configuration	75
15.4	Driver compilation and testing	76
15.5	SR-IOV: Prerequisites and sample Application Notes	76
15.6	Sample Application Notes	77
15.7	Limitations or Known issues	80
15.8	High Performance of Small Packets on 40G NIC	82
15.9	Example of getting best performance with I3fwd example	83
16	IGB Poll Mode Driver	85
16.1	Features	85
16.2	Limitations or Known issues	85
16.3	Supported Chipsets and NICs	85

17 IXGBE Driver	86
17.1 Vector PMD for IXGBE	86
17.2 Application Programming Interface	88
17.3 Sample Application Notes	88
17.4 Limitations or Known issues	88
17.5 Inline crypto processing support	89
17.6 Supported Chipsets and NICs	89
18 Intel Virtual Function Driver	91
18.1 SR-IOV Mode Utilization in a DPDK Environment	91
18.2 Setting Up a KVM Virtual Machine Monitor	97
18.3 DPDK SR-IOV PMD PF/VF Driver Usage Model	101
18.4 SR-IOV (PF/VF) Approach for Inter-VM Communication	101
19 KNI Poll Mode Driver	104
19.1 Usage	104
19.2 Default interface configuration	104
19.3 PMD arguments	105
19.4 PMD log messages	105
19.5 PMD testing	105
20 LiquidIO VF Poll Mode Driver	107
20.1 Supported LiquidIO Adapters	107
20.2 Pre-Installation Configuration	107
20.3 SR-IOV: Prerequisites and Sample Application Notes	108
20.4 Limitations	109
21 MLX4 poll mode driver library	110
21.1 Implementation details	110
21.2 Configuration	111
21.3 Prerequisites	112
21.4 Supported NICs	114
21.5 Quick Start Guide	114
21.6 Performance tuning	114
21.7 Usage example	115
22 MLX5 poll mode driver	117
22.1 Implementation details	117
22.2 Features	117
22.3 Limitations	118
22.4 Statistics	119
22.5 Configuration	119
22.6 Prerequisites	122
22.7 Supported NICs	124
22.8 Quick Start Guide on OFED	124
22.9 Performance tuning	125
22.10 Notes for testpmd	126
22.11 Usage example	126
23 MRVL Poll Mode Driver	129
23.1 Features	129
23.2 Limitations	130

23.3	Prerequisites	130
23.4	Config File Options	130
23.5	QoS Configuration	130
23.6	Building DPDK	132
23.7	Usage Example	132
24	NFP poll mode driver library	133
24.1	Dependencies	133
24.2	Building the software	133
24.3	Driver compilation and testing	134
24.4	Using the PF	134
24.5	PF multiport support	134
24.6	System configuration	134
25	OCTEONTX Poll Mode driver	136
25.1	Features	136
25.2	Supported OCTEONTX SoCs	136
25.3	Unsupported features	136
25.4	Prerequisites	137
25.5	Pre-Installation Configuration	137
25.6	Initialization	138
25.7	Limitations	138
26	QEDE Poll Mode Driver	140
26.1	Supported Features	140
26.2	Non-supported Features	141
26.3	Co-existence considerations	141
26.4	Supported QLogic Adapters	141
26.5	Prerequisites	141
26.6	Driver compilation and testing	142
26.7	SR-IOV: Prerequisites and Sample Application Notes	142
27	Solarflare libefx-based Poll Mode Driver	145
27.1	Features	145
27.2	Non-supported Features	146
27.3	Limitations	146
27.4	Tunnels support	146
27.5	Flow API support	146
27.6	Supported NICs	147
27.7	Prerequisites	148
27.8	Pre-Installation Configuration	148
28	SZEDATA2 poll mode driver library	150
28.1	Prerequisites	150
28.2	Configuration	150
28.3	Using the SZEDATA2 PMD	151
28.4	Example of usage	151
29	Tap Poll Mode Driver	153
29.1	Flow API support	154
29.2	Example	155
29.3	RSS specifics	156

29.4	Systems supporting flow API	156
30	ThunderX NICVF Poll Mode Driver	157
30.1	Features	157
30.2	Supported ThunderX SoCs	157
30.3	Prerequisites	158
30.4	Pre-Installation Configuration	158
30.5	Driver compilation and testing	158
30.6	Linux	158
30.7	Limitations	162
31	VDEV_NETVSC driver	163
31.1	Implementation details	163
31.2	Build options	164
31.3	Run-time parameters	164
32	Poll Mode Driver for Emulated Virtio NIC	165
32.1	Virtio Implementation in DPDK	165
32.2	Features and Limitations of virtio PMD	165
32.3	Prerequisites	166
32.4	Virtio with kni vhost Back End	166
32.5	Virtio with qemu virtio Back End	169
32.6	Virtio PMD Rx/Tx Callbacks	170
32.7	Interrupt mode	170
33	Poll Mode Driver that wraps vhost library	172
33.1	Vhost Implementation in DPDK	172
33.2	Features and Limitations of vhost PMD	172
33.3	Vhost PMD arguments	172
33.4	Vhost PMD event handling	173
33.5	Vhost PMD with testpmd application	173
34	Poll Mode Driver for Paravirtual VMXNET3 NIC	174
34.1	VMXNET3 Implementation in the DPDK	174
34.2	Features and Limitations of VMXNET3 PMD	175
34.3	Prerequisites	175
34.4	VMXNET3 with a Native NIC Connected to a vSwitch	176
34.5	VMXNET3 Chaining VMs Connected to a vSwitch	176
35	Libpcap and Ring Based Poll Mode Drivers	180
35.1	Using the Drivers from the EAL Command Line	180
36	Fail-safe poll mode driver library	185
36.1	Features	185
36.2	Compilation option	185
36.3	Using the Fail-safe PMD from the EAL command line	185
36.4	Using the Fail-safe PMD from an application	187
36.5	Plug-in feature	187
36.6	Plug-out feature	187
36.7	Fail-safe glossary	188

OVERVIEW OF NETWORKING DRIVERS

The networking drivers may be classified in two categories:

- physical for real devices
- virtual for emulated devices

Some physical devices may be shaped through a virtual layer as for SR-IOV. The interface seen in the virtual environment is a VF (Virtual Function).

The ethdev layer exposes an API to use the networking functions of these devices. The bottom half part of ethdev is implemented by the drivers. Thus some features may not be implemented.

There are more differences between drivers regarding some internal properties, portability or even documentation availability. Most of these differences are summarized below.

More details about features can be found in [Features Overview](#).

Feature	a f p a c k e t	a r k	a v f	a v f v e c	a v p	b n x 2 x	b n x 2 x
Speed capabilities		Y	Y	Y		P	P
Link status			Y	Y	Y	Y	Y
Link status event			Y	Y		Y	Y
Removal event							
Queue status event							
Rx interrupt			Y	Y			
Lock-free Tx queue							
Fast mbuf free							
Free Tx mbuf on demand							
Queue start/stop		Y	Y	Y			
MTU update			Y	Y			
Jumbo frame		Y	Y	Y	Y		
Scattered Rx		Y	Y	Y	Y		
LRO							
TSO			Y	Y			
Promiscuous mode			Y	Y	Y	Y	Y
Allmulticast mode			Y	Y			
Unicast MAC filter			Y	Y	Y	Y	Y
Multicast MAC filter			Y	Y		Y	Y
RSS hash			Y	Y			
RSS key update			Y	Y			

Network Interface Controller Drivers, Release 18.02.2

Feature	afpacket	ark	avf	avf.....vec	avp	bnx2x	bnx2x
RSS reta update			Y	Y			
VMDq							
SR-IOV							Y
DCB							
VLAN filter			Y	Y			
Ethertype filter							
N-tuple filter							
SYN filter							
Tunnel filter							
Flexible filter							
Hash filter							
Flow director							
Flow control							
Flow API							
Rate limitation							
Traffic mirroring							
Inline crypto							
CRC offload			Y	Y			
VLAN offload			Y	P	Y		
QinQ offload							
L3 checksum offload			Y	P			
L4 checksum offload			Y	P			
Timestamp offload							
MACsec offload							
Inner L3 checksum							
Inner L4 checksum							
Packet type parsing			Y	Y			
Timesync							
Rx descriptor status			Y	Y			
Tx descriptor status			Y	Y			
Basic stats		Y	Y	Y	Y	Y	Y
Extended stats						Y	Y
Stats per queue		Y			Y		
FW version							
EEPROM dump							
Registers dump							
LED							
Multiprocess aware			Y	Y			
BSD nic_uio			Y	Y			
Linux UIO		Y	Y	Y	Y	Y	Y
Linux VFIO			Y	Y			
Other kdrv							
ARMv7							
ARMv8							
Power8							
x86-32			Y	Y			
x86-64		Y	Y	Y	Y	Y	Y

Feature	afpacket	ark	avf	avf.....vec	avp	bnx2x	bnx2x
Usage doc		Y				Y	Y
Design doc							
Perf doc							

Note: Features marked with “P” are partially supported. Refer to the appropriate NIC guide in the following sections for details.

FEATURES OVERVIEW

This section explains the supported features that are listed in the [Overview of Networking Drivers](#).

As a guide to implementers it also shows the structs where the features are defined and the APIs that can be use to get/set the values.

Following tags used for feature details, these are from driver point of view:

[uses] : Driver uses some kind of input from the application.

[implements] : Driver implements a functionality.

[provides] : Driver provides some kind of data to the application. It is possible to provide data by implementing some function, but “provides” is used for cases where provided data can’t be represented simply by a function.

[related] : Related API with that feature.

2.1 Speed capabilities

Supports getting the speed capabilities that the current device is capable of.

- **[provides] `rte_eth_dev_info`:** `speed_capa:ETH_LINK_SPEED_*`.
- **[related] API:** `rte_eth_dev_info_get()`.

2.2 Link status

Supports getting the link speed, duplex mode and link state (up/down).

- **[implements] `eth_dev_ops`:** `link_update`.
- **[implements] `rte_eth_dev_data`:** `dev_link`.
- **[related] API:** `rte_eth_link_get()`, `rte_eth_link_get_nowait()`.

2.3 Link status event

Supports Link Status Change interrupts.

- **[uses] user config:** `dev_conf.intr_conf.lsc`.

- **[uses] rte_eth_dev_data:** dev_flags: RTE_ETH_DEV_INTR_LSC.
- **[uses] rte_eth_event_type:** RTE_ETH_EVENT_INTR_LSC.
- **[implements] rte_eth_dev_data:** dev_link.
- **[provides] rte_pci_driver.driv_flags:** RTE_PCI_DRV_INTR_LSC.
- **[related] API:** rte_eth_link_get(), rte_eth_link_get_nowait().

2.4 Removal event

Supports device removal interrupts.

- **[uses] user config:** dev_conf.intr_conf.rmv.
- **[uses] rte_eth_dev_data:** dev_flags: RTE_ETH_DEV_INTR_RMV.
- **[uses] rte_eth_event_type:** RTE_ETH_EVENT_INTR_RMV.
- **[provides] rte_pci_driver.driv_flags:** RTE_PCI_DRV_INTR_RMV.

2.5 Queue status event

Supports queue enable/disable events.

- **[uses] rte_eth_event_type:** RTE_ETH_EVENT_QUEUE_STATE.

2.6 Rx interrupt

Supports Rx interrupts.

- **[uses] user config:** dev_conf.intr_conf.rxq.
- **[implements] eth_dev_ops:** rx_queue_intr_enable, rx_queue_intr_disable.
- **[related] API:** rte_eth_dev_rx_intr_enable(), rte_eth_dev_rx_intr_disable().

2.7 Lock-free Tx queue

If a PMD advertises DEV_TX_OFFLOAD_MT_LOCKFREE capable, multiple threads can invoke rte_eth_tx_burst() concurrently on the same Tx queue without SW lock.

- **[uses] rte_eth_txconf, rte_eth_txmode:** offloads: DEV_TX_OFFLOAD_MT_LOCKFREE.
- **[provides] rte_eth_dev_info:** tx_offload_capa, tx_queue_offload_capa: DEV_TX_OFFLOAD_
- **[related] API:** rte_eth_tx_burst().

2.8 Fast mbuf free

Supports optimization for fast release of mbufs following successful Tx. Requires that per queue, all mbufs come from the same mempool and has `refcnt = 1`.

- **[uses]** `rte_eth_txconf, rte_eth_txmode`: `offloads:DEV_TX_OFFLOAD_MBUF_FAST_FREE`.
- **[provides]** `rte_eth_dev_info`: `tx_offload_capa, tx_queue_offload_capa:DEV_TX_OFFLOAD_`

2.9 Free Tx mbuf on demand

Supports freeing consumed buffers on a Tx ring.

- **[implements]** `eth_dev_ops`: `tx_done_cleanup`.
- **[related]** **API**: `rte_eth_tx_done_cleanup()`.

2.10 Queue start/stop

Supports starting/stopping a specific Rx/Tx queue of a port.

- **[implements]** `eth_dev_ops`: `rx_queue_start, rx_queue_stop, tx_queue_start, tx_queue_stop`.
- **[related]** **API**: `rte_eth_dev_rx_queue_start(), rte_eth_dev_rx_queue_stop(), rte_eth_dev_tx_queue_start(), rte_eth_dev_tx_queue_stop()`.

2.11 MTU update

Supports updating port MTU.

- **[implements]** `eth_dev_ops`: `mtu_set`.
- **[implements]** `rte_eth_dev_data`: `mtu`.
- **[provides]** `rte_eth_dev_info`: `max_rx_pktlen`.
- **[related]** **API**: `rte_eth_dev_set_mtu(), rte_eth_dev_get_mtu()`.

2.12 Jumbo frame

Supports Rx jumbo frames.

- **[uses]** `rte_eth_rxconf, rte_eth_rxmode`: `offloads:DEV_RX_OFFLOAD_JUMBO_FRAME, dev_conf.rxmode.max_rx_pkt_len`.
- **[related]** `rte_eth_dev_info`: `max_rx_pktlen`.
- **[related]** **API**: `rte_eth_dev_set_mtu()`.

2.13 Scattered Rx

Supports receiving segmented mbufs.

- **[uses]** `rte_eth_rxconf, rte_eth_rxmode`: `offloads:DEV_RX_OFFLOAD_SCATTER`.
- **[implements]** `datapath`: Scattered Rx function.
- **[implements]** `rte_eth_dev_data`: `scattered_rx`.
- **[provides]** `eth_dev_ops`: `rxq_info_get:scattered_rx`.
- **[related]** `eth_dev_ops`: `rx_pkt_burst`.

2.14 LRO

Supports Large Receive Offload.

- **[uses]** `rte_eth_rxconf, rte_eth_rxmode`: `offloads:DEV_RX_OFFLOAD_TCP_LRO`.
- **[implements]** `datapath`: LRO functionality.
- **[implements]** `rte_eth_dev_data`: `lro`.
- **[provides]** `mbuf`: `mbuf.ol_flags:PKT_RX_LRO, mbuf.tso_segsz`.
- **[provides]** `rte_eth_dev_info`: `rx_offload_capa, rx_queue_offload_capa:DEV_RX_OFFLOAD_`

2.15 TSO

Supports TCP Segmentation Offloading.

- **[uses]** `rte_eth_txconf, rte_eth_txmode`: `offloads:DEV_TX_OFFLOAD_TCP_TSO`.
- **[uses]** `rte_eth_desc_lim`: `nb_seg_max, nb_mtu_seg_max`.
- **[uses]** `mbuf`: `mbuf.ol_flags:PKT_TX_TCP_SEG`.
- **[uses]** `mbuf`: `mbuf.tso_segsz, mbuf.l2_len, mbuf.l3_len, mbuf.l4_len`.
- **[implements]** `datapath`: TSO functionality.
- **[provides]** `rte_eth_dev_info`: `tx_offload_capa, tx_queue_offload_capa:DEV_TX_OFFLOAD_`

2.16 Promiscuous mode

Supports enabling/disabling promiscuous mode for a port.

- **[implements]** `eth_dev_ops`: `promiscuous_enable, promiscuous_disable`.
- **[implements]** `rte_eth_dev_data`: `promiscuous`.
- **[related]** **API**: `rte_eth_promiscuous_enable()`, `rte_eth_promiscuous_disable()`, `rte_eth_promiscuous_get()`.

2.17 Allmulticast mode

Supports enabling/disabling receiving multicast frames.

- **[implements] eth_dev_ops:** `allmulticast_enable`, `allmulticast_disable`.
- **[implements] rte_eth_dev_data:** `all_multicast`.
- **[related] API:** `rte_eth_allmulticast_enable()`, `rte_eth_allmulticast_disable()`, `rte_eth_allmulticast_get()`.

2.18 Unicast MAC filter

Supports adding MAC addresses to enable whitelist filtering to accept packets.

- **[implements] eth_dev_ops:** `mac_addr_set`, `mac_addr_add`, `mac_addr_remove`.
- **[implements] rte_eth_dev_data:** `mac_addrs`.
- **[related] API:** `rte_eth_dev_default_mac_addr_set()`, `rte_eth_dev_mac_addr_add()`, `rte_eth_dev_mac_addr_remove()`, `rte_eth_macaddr_get()`.

2.19 Multicast MAC filter

Supports setting multicast addresses to filter.

- **[implements] eth_dev_ops:** `set_mc_addr_list`.
- **[related] API:** `rte_eth_dev_set_mc_addr_list()`.

2.20 RSS hash

Supports RSS hashing on RX.

- **[uses] user config:** `dev_conf.rxmode.mq_mode = ETH_MQ_RX_RSS_FLAG`.
- **[uses] user config:** `dev_conf.rx_adv_conf.rss_conf`.
- **[provides] rte_eth_dev_info:** `flow_type_rss_offloads`.
- **[provides] mbuf:** `mbuf.ol_flags:PKT_RX_RSS_HASH`, `mbuf.rss`.

2.21 RSS key update

Supports configuration of Receive Side Scaling (RSS) hash computation. Updating Receive Side Scaling (RSS) hash key.

- **[implements] eth_dev_ops:** `rss_hash_update`, `rss_hash_conf_get`.
- **[provides] rte_eth_dev_info:** `hash_key_size`.
- **[related] API:** `rte_eth_dev_rss_hash_update()`, `rte_eth_dev_rss_hash_conf_get()`.

2.22 RSS reta update

Supports updating Redirection Table of the Receive Side Scaling (RSS).

- **[implements] eth_dev_ops:** `reta_update`, `reta_query`.
- **[provides] rte_eth_dev_info:** `reta_size`.
- **[related] API:** `rte_eth_dev_rss_reta_update()`, `rte_eth_dev_rss_reta_query()`.

2.23 VMDq

Supports Virtual Machine Device Queues (VMDq).

- **[uses] user config:** `dev_conf.rxmode.mq_mode = ETH_MQ_RX_VMDQ_FLAG`.
- **[uses] user config:** `dev_conf.rx_adv_conf.vmdq_dcb_conf`.
- **[uses] user config:** `dev_conf.rx_adv_conf.vmdq_rx_conf`.
- **[uses] user config:** `dev_conf.tx_adv_conf.vmdq_dcb_tx_conf`.
- **[uses] user config:** `dev_conf.tx_adv_conf.vmdq_tx_conf`.

2.24 SR-IOV

Driver supports creating Virtual Functions.

- **[implements] rte_eth_dev_data:** `sriov`.

2.25 DCB

Supports Data Center Bridging (DCB).

- **[uses] user config:** `dev_conf.rxmode.mq_mode = ETH_MQ_RX_DCB_FLAG`.
- **[uses] user config:** `dev_conf.rx_adv_conf.vmdq_dcb_conf`.
- **[uses] user config:** `dev_conf.rx_adv_conf.dcb_rx_conf`.
- **[uses] user config:** `dev_conf.tx_adv_conf.vmdq_dcb_tx_conf`.
- **[uses] user config:** `dev_conf.tx_adv_conf.vmdq_tx_conf`.
- **[implements] eth_dev_ops:** `get_dcb_info`.
- **[related] API:** `rte_eth_dev_get_dcb_info()`.

2.26 VLAN filter

Supports filtering of a VLAN Tag identifier.

- **[uses] rte_eth_rxconf, rte_eth_rxmode:** `offloads:DEV_RX_OFFLOAD_VLAN_FILTER`.
- **[implements] eth_dev_ops:** `vlan_filter_set`.

- **[related] API:** `rte_eth_dev_vlan_filter()`.

2.27 Ethertype filter

Supports filtering on Ethernet type.

- **[implements] eth_dev_ops:** `filter_ctrl:RTE_ETH_FILTER_ETHERTYPE`.
- **[related] API:** `rte_eth_dev_filter_ctrl()`, `rte_eth_dev_filter_supported()`.

2.28 N-tuple filter

Supports filtering on N-tuple values.

- **[implements] eth_dev_ops:** `filter_ctrl:RTE_ETH_FILTER_NTUPLE`.
- **[related] API:** `rte_eth_dev_filter_ctrl()`, `rte_eth_dev_filter_supported()`.

2.29 SYN filter

Supports TCP syn filtering.

- **[implements] eth_dev_ops:** `filter_ctrl:RTE_ETH_FILTER_SYN`.
- **[related] API:** `rte_eth_dev_filter_ctrl()`, `rte_eth_dev_filter_supported()`.

2.30 Tunnel filter

Supports tunnel filtering.

- **[implements] eth_dev_ops:** `filter_ctrl:RTE_ETH_FILTER_TUNNEL`.
- **[related] API:** `rte_eth_dev_filter_ctrl()`, `rte_eth_dev_filter_supported()`.

2.31 Flexible filter

Supports a flexible (non-tuple or Ethertype) filter.

- **[implements] eth_dev_ops:** `filter_ctrl:RTE_ETH_FILTER_FLEXIBLE`.
- **[related] API:** `rte_eth_dev_filter_ctrl()`, `rte_eth_dev_filter_supported()`.

2.32 Hash filter

Supports Hash filtering.

- **[implements] eth_dev_ops:** `filter_ctrl:RTE_ETH_FILTER_HASH`.
- **[related] API:** `rte_eth_dev_filter_ctrl()`, `rte_eth_dev_filter_supported()`.

2.33 Flow director

Supports Flow Director style filtering to queues.

- **[implements] eth_dev_ops:** `filter_ctrl:RTE_ETH_FILTER_FDIR`.
- **[provides] mbuf:** `mbuf.ol_flags: PKT_RX_FDIR, PKT_RX_FDIR_ID, PKT_RX_FDIR_FLX`.
- **[related] API:** `rte_eth_dev_filter_ctrl()`, `rte_eth_dev_filter_supported()`.

2.34 Flow control

Supports configuring link flow control.

- **[implements] eth_dev_ops:** `flow_ctrl_get, flow_ctrl_set, priority_flow_ctrl_set`.
- **[related] API:** `rte_eth_dev_flow_ctrl_get()`, `rte_eth_dev_flow_ctrl_set()`, `rte_eth_dev_priority_flow_ctrl_set()`.

2.35 Flow API

Supports the DPDK Flow API for generic filtering.

- **[implements] eth_dev_ops:** `filter_ctrl:RTE_ETH_FILTER_GENERIC`.
- **[implements] rte_flow_ops:** `All`.

2.36 Rate limitation

Supports Tx rate limitation for a queue.

- **[implements] eth_dev_ops:** `set_queue_rate_limit`.
- **[related] API:** `rte_eth_set_queue_rate_limit()`.

2.37 Traffic mirroring

Supports adding traffic mirroring rules.

- **[implements] eth_dev_ops:** `mirror_rule_set, mirror_rule_reset`.
- **[related] API:** `rte_eth_mirror_rule_set()`, `rte_eth_mirror_rule_reset()`.

2.38 Inline crypto

Supports inline crypto processing (eg. inline IPsec). See Security library and PMD documentation for more details.

- **[uses]** `rte_eth_rxconf,rte_eth_rxmode`: `offloads:DEV_RX_OFFLOAD_SECURITY`,
- **[uses]** `rte_eth_txconf,rte_eth_txmode`: `offloads:DEV_TX_OFFLOAD_SECURITY`.
- **[implements]** `rte_security_ops`: `session_create`, `session_update`,
`session_stats_get`, `session_destroy`, `set_pkt_metadata`,
`capabilities_get`.
- **[provides]** `rte_eth_dev_info`: `rx_offload_capa`, `rx_queue_offload_capa:DEV_RX_OFFLOAD_`
`tx_offload_capa`, `tx_queue_offload_capa:DEV_TX_OFFLOAD_SECURITY`.
- **[provides]** `mbuf`: `mbuf.ol_flags:PKT_RX_SEC_OFFLOAD`,
`mbuf.ol_flags:PKT_TX_SEC_OFFLOAD`, `mbuf.ol_flags:PKT_RX_SEC_OFFLOAD_FAILED`.

2.39 CRC offload

Supports CRC stripping by hardware.

- **[uses]** `rte_eth_rxconf,rte_eth_rxmode`: `offloads:DEV_RX_OFFLOAD_CRC_STRIP`.

2.40 VLAN offload

Supports VLAN offload to hardware.

- **[uses]** `rte_eth_rxconf,rte_eth_rxmode`: `offloads:DEV_RX_OFFLOAD_VLAN_STRIP,DEV_RX_OFF`
- **[uses]** `rte_eth_txconf,rte_eth_txmode`: `offloads:DEV_TX_OFFLOAD_VLAN_INSERT`.
- **[implements]** `eth_dev_ops`: `vlan_offload_set`.
- **[provides]** `mbuf`: `mbuf.ol_flags:PKT_RX_VLAN_STRIPPED`, `mbuf.vlan_tci`.
- **[provides]** `rte_eth_dev_info`: `rx_offload_capa`, `rx_queue_offload_capa:DEV_RX_OFFLOAD_`
`tx_offload_capa`, `tx_queue_offload_capa:DEV_TX_OFFLOAD_VLAN_INSERT`.
- **[related]** `API`: `rte_eth_dev_set_vlan_offload()`,
`rte_eth_dev_get_vlan_offload()`.

2.41 QinQ offload

Supports QinQ (queue in queue) offload.

- **[uses]** `rte_eth_rxconf,rte_eth_rxmode`: `offloads:DEV_RX_OFFLOAD_QINQ_STRIP`.
- **[uses]** `rte_eth_txconf,rte_eth_txmode`: `offloads:DEV_TX_OFFLOAD_QINQ_INSERT`.
- **[uses]** `mbuf`: `mbuf.ol_flags:PKT_TX_QINQ_PKT`.
- **[provides]** `mbuf`: `mbuf.ol_flags:PKT_RX_QINQ_STRIPPED`, `mbuf.vlan_tci`,
`mbuf.vlan_tci_outer`.
- **[provides]** `rte_eth_dev_info`: `rx_offload_capa`, `rx_queue_offload_capa:DEV_RX_OFFLOAD_`
`tx_offload_capa`, `tx_queue_offload_capa:DEV_TX_OFFLOAD_QINQ_INSERT`.

2.42 L3 checksum offload

Supports L3 checksum offload.

- **[uses] rte_eth_rxconf,rte_eth_rxmode:** offloads:DEV_RX_OFFLOAD_IPV4_CKSUM.
- **[uses] rte_eth_txconf,rte_eth_txmode:** offloads:DEV_TX_OFFLOAD_IPV4_CKSUM.
- **[uses] mbuf:** mbuf.ol_flags:PKT_TX_IP_CKSUM, mbuf.ol_flags:PKT_TX_IPV4 | PKT_TX_IPV6.
- **[provides] mbuf:** mbuf.ol_flags:PKT_RX_IP_CKSUM_UNKNOWN | PKT_RX_IP_CKSUM_BAD | PKT_RX_IP_CKSUM_GOOD | PKT_RX_IP_CKSUM_NONE.
- **[provides] rte_eth_dev_info:** rx_offload_capa,rx_queue_offload_capa:DEV_RX_OFFLOAD_IP_CKSUM, tx_offload_capa,tx_queue_offload_capa:DEV_TX_OFFLOAD_IPV4_CKSUM.

2.43 L4 checksum offload

Supports L4 checksum offload.

- **[uses] rte_eth_rxconf,rte_eth_rxmode:** offloads:DEV_RX_OFFLOAD_UDP_CKSUM, DEV_RX_OFFLOAD_TCP_CKSUM.
- **[uses] rte_eth_txconf,rte_eth_txmode:** offloads:DEV_TX_OFFLOAD_UDP_CKSUM, DEV_TX_OFFLOAD_TCP_CKSUM.
- **[uses] user config:** dev_conf.rxmode.hw_ip_checksum.
- **[uses] mbuf:** mbuf.ol_flags:PKT_TX_IPV4 | PKT_TX_IPV6, mbuf.ol_flags:PKT_TX_L4_NO_CKSUM | PKT_TX_TCP_CKSUM | PKT_TX_SCTP_CKSUM | PKT_TX_UDP_CKSUM.
- **[provides] mbuf:** mbuf.ol_flags:PKT_RX_L4_CKSUM_UNKNOWN | PKT_RX_L4_CKSUM_BAD | PKT_RX_L4_CKSUM_GOOD | PKT_RX_L4_CKSUM_NONE.
- **[provides] rte_eth_dev_info:** rx_offload_capa,rx_queue_offload_capa:DEV_RX_OFFLOAD_UDP_CKSUM, DEV_RX_OFFLOAD_TCP_CKSUM, tx_offload_capa,tx_queue_offload_capa:DEV_TX_OFFLOAD_UDP_CKSUM, DEV_TX_OFFLOAD_TCP_CKSUM.

2.44 Timestamp offload

Supports Timestamp.

- **[uses] rte_eth_rxconf,rte_eth_rxmode:** offloads:DEV_RX_OFFLOAD_TIMESTAMP.
- **[provides] mbuf:** mbuf.ol_flags:PKT_RX_TIMESTAMP.
- **[provides] mbuf:** mbuf.timestamp.
- **[provides] rte_eth_dev_info:** rx_offload_capa,rx_queue_offload_capa:DEV_RX_OFFLOAD_TIMESTAMP.

2.45 MACsec offload

Supports MACsec.

- **[uses] rte_eth_rxconf,rte_eth_rxmode:** offloads:DEV_RX_OFFLOAD_MACSEC_STRIP.

- **[uses] `rte_eth_txconf,rte_eth_txmode`:** `offloads:DEV_TX_OFFLOAD_MACSEC_INSERT`.
- **[uses] `mbuf`:** `mbuf.ol_flags:PKT_TX_MACSEC`.
- **[provides] `rte_eth_dev_info`:** `rx_offload_capa,rx_queue_offload_capa:DEV_RX_OFFLOAD_`
`tx_offload_capa,tx_queue_offload_capa:DEV_TX_OFFLOAD_MACSEC_INSERT`.

2.46 Inner L3 checksum

Supports inner packet L3 checksum.

- **[uses] `rte_eth_rxconf,rte_eth_rxmode`:** `offloads:DEV_RX_OFFLOAD_OUTER_IPV4_CKSUM`.
- **[uses] `rte_eth_txconf,rte_eth_txmode`:** `offloads:DEV_TX_OFFLOAD_OUTER_IPV4_CKSUM`.
- **[uses] `mbuf`:** `mbuf.ol_flags:PKT_TX_IP_CKSUM,`
`mbuf.ol_flags:PKT_TX_IPV4 | PKT_TX_IPV6,mbuf.ol_flags:PKT_TX_OUTER_IP_CKSUM,`
`mbuf.ol_flags:PKT_TX_OUTER_IPV4 | PKT_TX_OUTER_IPV6`.
- **[uses] `mbuf`:** `mbuf.outer_l2_len,mbuf.outer_l3_len`.
- **[provides] `mbuf`:** `mbuf.ol_flags:PKT_RX_EIP_CKSUM_BAD`.
- **[provides] `rte_eth_dev_info`:** `rx_offload_capa,rx_queue_offload_capa:DEV_RX_OFFLOAD_`
`tx_offload_capa,tx_queue_offload_capa:DEV_TX_OFFLOAD_OUTER_IPV4_CKSUM`.

2.47 Inner L4 checksum

Supports inner packet L4 checksum.

2.48 Packet type parsing

Supports packet type parsing and returns a list of supported types.

- **[implements] `eth_dev_ops`:** `dev_supported_ptypes_get`.
- **[related] API:** `rte_eth_dev_get_supported_ptypes()`.

2.49 Timesync

Supports IEEE1588/802.1AS timestamping.

- **[implements] `eth_dev_ops`:** `timesync_enable, timesync_disable,`
`timesync_read_rx_timestamp, timesync_read_tx_timestamp,`
`timesync_adjust_time,timesync_read_time,timesync_write_time`.
- **[related] API:** `rte_eth_timesync_enable(), rte_eth_timesync_disable(),`
`rte_eth_timesync_read_rx_timestamp(),rte_eth_timesync_read_tx_timestamp,`
`rte_eth_timesync_adjust_time(), rte_eth_timesync_read_time(),`
`rte_eth_timesync_write_time()`.

2.50 Rx descriptor status

Supports check the status of a Rx descriptor. When `rx_descriptor_status` is used, status can be “Available”, “Done” or “Unavailable”. When `rx_descriptor_done` is used, status can be “DD bit is set” or “DD bit is not set”.

- **[implements] eth_dev_ops:** `rx_descriptor_status`.
- **[related] API:** `rte_eth_rx_descriptor_status()`.
- **[implements] eth_dev_ops:** `rx_descriptor_done`.
- **[related] API:** `rte_eth_rx_descriptor_done()`.

2.51 Tx descriptor status

Supports checking the status of a Tx descriptor. Status can be “Full”, “Done” or “Unavailable.”

- **[implements] eth_dev_ops:** `tx_descriptor_status`.
- **[related] API:** `rte_eth_tx_descriptor_status()`.

2.52 Basic stats

Support basic statistics such as: `ipackets`, `opackets`, `ibytes`, `obytes`, `imissed`, `errors`, `oerrors`, `rx_nombuf`.

And per queue stats: `q_ipackets`, `q_opackets`, `q_ibytes`, `q_obytes`, `q_errors`.

These apply to all drivers.

- **[implements] eth_dev_ops:** `stats_get`, `stats_reset`.
- **[related] API:** `rte_eth_stats_get`, `rte_eth_stats_reset()`.

2.53 Extended stats

Supports Extended Statistics, changes from driver to driver.

- **[implements] eth_dev_ops:** `xstats_get`, `xstats_reset`, `xstats_get_names`.
- **[implements] eth_dev_ops:** `xstats_get_by_id`, `xstats_get_names_by_id`.
- **[related] API:** `rte_eth_xstats_get()`, `rte_eth_xstats_reset()`, `rte_eth_xstats_get_names`, `rte_eth_xstats_get_by_id()`, `rte_eth_xstats_get_names_by_id()`, `rte_eth_xstats_get_id_by_name()`.

2.54 Stats per queue

Supports configuring per-queue stat counter mapping.

- **[implements] eth_dev_ops:** `queue_stats_mapping_set`.

- **[related] API:** `rte_eth_dev_set_rx_queue_stats_mapping()`, `rte_eth_dev_set_tx_queue_stats_mapping()`.

2.55 FW version

Supports getting device hardware firmware information.

- **[implements] eth_dev_ops:** `fw_version_get`.
- **[related] API:** `rte_eth_dev_fw_version_get()`.

2.56 EEPROM dump

Supports getting/setting device eeprom data.

- **[implements] eth_dev_ops:** `get_eeprom_length`, `get_eeprom`, `set_eeprom`.
- **[related] API:** `rte_eth_dev_get_eeprom_length()`, `rte_eth_dev_get_eeprom()`, `rte_eth_dev_set_eeprom()`.

2.57 Registers dump

Supports retrieving device registers and registering attributes (number of registers and register size).

- **[implements] eth_dev_ops:** `get_reg`.
- **[related] API:** `rte_eth_dev_get_reg_info()`.

2.58 LED

Supports turning on/off a software controllable LED on a device.

- **[implements] eth_dev_ops:** `dev_led_on`, `dev_led_off`.
- **[related] API:** `rte_eth_led_on()`, `rte_eth_led_off()`.

2.59 Multiprocess aware

Driver can be used for primary-secondary process model.

2.60 BSD nic_uio

BSD `nic_uio` module supported.

2.61 Linux UIO

Works with `igb_uio` kernel module.

- **[provides] RTE_PMD_REGISTER_KMOD_DEP:** `igb_uio`.

2.62 Linux VFIO

Works with `vfio-pci` kernel module.

- **[provides] RTE_PMD_REGISTER_KMOD_DEP:** `vfio-pci`.

2.63 Other kdrv

Kernel module other than above ones supported.

2.64 ARMv7

Support armv7 architecture.

Use `defconfig_arm-armv7a-***`.

2.65 ARMv8

Support armv8a (64bit) architecture.

Use `defconfig_arm64-armv8a-***`

2.66 Power8

Support PowerPC architecture.

Use `defconfig_ppc_64-power8-***`

2.67 x86-32

Support 32bits x86 architecture.

Use `defconfig_x86_x32-native-***` and `defconfig_i686-native-***`.

2.68 x86-64

Support 64bits x86 architecture.

Use `defconfig_x86_64-native-***`.

2.69 Usage doc

Documentation describes usage.

See `doc/guides/nics/*.rst`

2.70 Design doc

Documentation describes design.

See `doc/guides/nics/*.rst`.

2.71 Perf doc

Documentation describes performance values.

See `dptk.org/doc/perf/*`.

2.72 Other dev ops not represented by a Feature

- `rxq_info_get`
- `txq_info_get`
- `vlan_tpid_set`
- `vlan_strip_queue_set`
- `vlan_pvid_set`
- `rx_queue_count`
- `l2_tunnel_offload_set`
- `uc_hash_table_set`
- `uc_all_hash_table_set`
- `udp_tunnel_port_add`
- `udp_tunnel_port_del`
- `l2_tunnel_eth_type_conf`
- `l2_tunnel_offload_set`
- `tx_pkt_prepare`

COMPILING AND TESTING A PMD FOR A NIC

This section demonstrates how to compile and run a Poll Mode Driver (PMD) for the available Network Interface Cards in DPDK using TestPMD.

TestPMD is one of the reference applications distributed with the DPDK. Its main purpose is to forward packets between Ethernet ports on a network interface and as such is the best way to test a PMD.

Refer to the `testpmd` application user guide for detailed information on how to build and run `testpmd`.

3.1 Driver Compilation

To compile a PMD for a platform, run `make` with appropriate target as shown below. Use “`make`” command in Linux and “`gmake`” in FreeBSD. This will also build `testpmd`.

To check available targets:

```
cd <DPDK-source-directory>
make showconfigs
```

Example output:

```
arm-armv7a-linuxapp-gcc
arm64-armv8a-linuxapp-gcc
arm64-dpaa2-linuxapp-gcc
arm64-thunderx-linuxapp-gcc
arm64-xgene1-linuxapp-gcc
i686-native-linuxapp-gcc
i686-native-linuxapp-icc
ppc_64-power8-linuxapp-gcc
x86_64-native-bsdapp-clang
x86_64-native-bsdapp-gcc
x86_64-native-linuxapp-clang
x86_64-native-linuxapp-gcc
x86_64-native-linuxapp-icc
x86_x32-native-linuxapp-gcc
```

To compile a PMD for Linux `x86_64 gcc` target, run the following “`make`” command:

```
make install T=x86_64-native-linuxapp-gcc
```

Use ARM (ThunderX, DPAA, X-Gene) or PowerPC target for respective platform.

For more information, refer to the Getting Started Guide for Linux or Getting Started Guide for FreeBSD depending on your platform.

3.2 Running testpmd in Linux

This section demonstrates how to setup and run `testpmd` in Linux.

1. Mount huge pages:

```
mkdir /mnt/huge
mount -t hugetlbfs nodev /mnt/huge
```

2. Request huge pages:

Hugepage memory should be reserved as per application requirement. Check hugepage size configured in the system and calculate the number of pages required.

To reserve 1024 pages of 2MB:

```
echo 1024 > /sys/kernel/mm/hugepages/hugepages-2048kB/nr_hugepages
```

Note: Check `/proc/meminfo` to find system hugepage size:

```
grep "Hugepagesize:" /proc/meminfo
```

Example output:

```
Hugepagesize:      2048 kB
```

3. Load `igb_uio` or `vfio-pci` driver:

```
modprobe uio
insmod ./x86_64-native-linuxapp-gcc/kmod/igb_uio.ko
```

or

```
modprobe vfio-pci
```

4. Setup VFIO permissions for regular users before binding to `vfio-pci`:

```
sudo chmod a+x /dev/vfio
sudo chmod 0666 /dev/vfio/*
```

5. Bind the adapters to `igb_uio` or `vfio-pci` loaded in the previous step:

```
./usertools/dpdk-devbind.py --bind igb_uio DEVICE1 DEVICE2 ...
```

Or setup VFIO permissions for regular users and then bind to `vfio-pci`:

```
./usertools/dpdk-devbind.py --bind vfio-pci DEVICE1 DEVICE2 ...
```

Note: `DEVICE1`, `DEVICE2` are specified via PCI “domain:bus:slot.func” syntax or “bus:slot.func” syntax.

6. Start `testpmd` with basic parameters:

```
./x86_64-native-linuxapp-gcc/app/testpmd -l 0-3 -n 4 -- -i
```

Successful execution will show initialization messages from EAL, PMD and `testpmd` application. A prompt will be displayed at the end for user commands as interactive mode (`-i`) is on.

```
testpmd>
```

Refer to the `testpmd` runtime functions for a list of available commands.

Note: When `testpmd` is built with shared library, use option `-d` to load the dynamic PMD for `rte_eal_init`.

ARK POLL MODE DRIVER

The ARK PMD is a DPDK poll-mode driver for the Atomic Rules Arkville (ARK) family of devices.

More information can be found at the [Atomic Rules website](#).

4.1 Overview

The Atomic Rules Arkville product is DPDK and AXI compliant product that marshals packets across a PCIe conduit between host DPDK mbufs and FPGA AXI streams.

The ARK PMD, and the spirit of the overall Arkville product, has been to take the DPDK API/ABI as a fixed specification; then implement much of the business logic in FPGA RTL circuits. The approach of *working backwards* from the DPDK API/ABI and having the GPP host software *dictate*, while the FPGA hardware *cope*s, results in significant performance gains over a naive implementation.

While this document describes the ARK PMD software, it is helpful to understand what the FPGA hardware is and is not. The Arkville RTL component provides a single PCIe Physical Function (PF) supporting some number of RX/Ingress and TX/Egress Queues. The ARK PMD controls the Arkville core through a dedicated opaque Core BAR (CBAR). To allow users full freedom for their own FPGA application IP, an independent FPGA Application BAR (ABAR) is provided.

One popular way to imagine Arkville's FPGA hardware aspect is as the FPGA PCIe-facing side of a so-called Smart NIC. The Arkville core does not contain any MACs, and is link-speed independent, as well as agnostic to the number of physical ports the application chooses to use. The ARK driver exposes the familiar PMD interface to allow packet movement to and from mbufs across multiple queues.

However FPGA RTL applications could contain a universe of added functionality that an Arkville RTL core does not provide or can not anticipate. To allow for this expectation of user-defined innovation, the ARK PMD provides a dynamic mechanism of adding capabilities without having to modify the ARK PMD.

The ARK PMD is intended to support all instances of the Arkville RTL Core, regardless of configuration, FPGA vendor, or target board. While specific capabilities such as number of physical hardware queue-pairs are negotiated; the driver is designed to remain constant over a broad and extendable feature set.

Intentionally, Arkville by itself DOES NOT provide common NIC capabilities such as offload or receive-side scaling (RSS). These capabilities would be viewed as a gate-level "tax" on

Green-box FPGA applications that do not require such function. Instead, they can be added as needed with essentially no overhead to the FPGA Application.

The ARK PMD also supports optional user extensions, through dynamic linking. The ARK PMD user extensions are a feature of Arkville's DPDK net/ark poll mode driver, allowing users to add their own code to extend the net/ark functionality without having to make source code changes to the driver. One motivation for this capability is that while DPDK provides a rich set of functions to interact with NIC-like capabilities (e.g. MAC addresses and statistics), the Arkville RTL IP does not include a MAC. Users can supply their own MAC or custom FPGA applications, which may require control from the PMD. The user extension is the means providing the control between the user's FPGA application and the existing DPDK features via the PMD.

4.2 Device Parameters

The ARK PMD supports device parameters that are used for packet routing and for internal packet generation and packet checking. This section describes the supported parameters. These features are primarily used for diagnostics, testing, and performance verification under the guidance of an Arkville specialist. The nominal use of Arkville does not require any configuration using these parameters.

“Pkt_dir”

The Packet Director controls connectivity between Arkville's internal hardware components. The features of the Pkt_dir are only used for diagnostics and testing; it is not intended for nominal use. The full set of features are not published at this level.

Format: Pkt_dir=0x00110F10

“Pkt_gen”

The packet generator parameter takes a file as its argument. The file contains configuration parameters used internally for regression testing and are not intended to be published at this level. The packet generator is an internal Arkville hardware component.

Format: Pkt_gen=./config/pg.conf

“Pkt_chkr”

The packet checker parameter takes a file as its argument. The file contains configuration parameters used internally for regression testing and are not intended to be published at this level. The packet checker is an internal Arkville hardware component.

Format: Pkt_chkr=./config/pc.conf

4.3 Data Path Interface

Ingress RX and Egress TX operation is by the nominal DPDK API . The driver supports single-port, multi-queue for both RX and TX.

4.4 Configuration Information

DPDK Configuration Parameters

The following configuration options are available for the ARK PMD:

- **CONFIG_RTE_LIBRTE_ARK_PMD** (default y): Enables or disables inclusion of the ARK PMD driver in the DPDK compilation.
- **CONFIG_RTE_LIBRTE_ARK_PAD_TX** (default y): When enabled TX packets are padded to 60 bytes to support downstream MACS.
- **CONFIG_RTE_LIBRTE_ARK_DEBUG_RX** (default n): Enables or disables debug logging and internal checking of RX ingress logic within the ARK PMD driver.
- **CONFIG_RTE_LIBRTE_ARK_DEBUG_TX** (default n): Enables or disables debug logging and internal checking of TX egress logic within the ARK PMD driver.
- **CONFIG_RTE_LIBRTE_ARK_DEBUG_STATS** (default n): Enables or disables debug logging of detailed packet and performance statistics gathered in the PMD and FPGA.
- **CONFIG_RTE_LIBRTE_ARK_DEBUG_TRACE** (default n): Enables or disables debug logging of detailed PMD events and status.

4.5 Building DPDK

See the DPDK Getting Started Guide for Linux for instructions on how to build DPDK.

By default the ARK PMD library will be built into the DPDK library.

For configuring and using UIO and VFIO frameworks, please also refer the documentation that comes with DPDK suite.

4.6 Supported ARK RTL PCIe Instances

ARK PMD supports the following Arkville RTL PCIe instances including:

- 1d6c:100d - AR-ARKA-FX0 [Arkville 32B DPDK Data Mover]
- 1d6c:100e - AR-ARKA-FX1 [Arkville 64B DPDK Data Mover]

4.7 Supported Operating Systems

Any Linux distribution fulfilling the conditions described in `System Requirements` section of the DPDK documentation or refer to *DPDK Release Notes*. ARM and PowerPC architectures are not supported at this time.

4.8 Supported Features

- Dynamic ARK PMD extensions
- Multiple receive and transmit queues

- Jumbo frames up to 9K
- Hardware Statistics

4.9 Unsupported Features

Features that may be part of, or become part of, the Arkville RTL IP that are not currently supported or exposed by the ARK PMD include:

- PCIe SR-IOV Virtual Functions (VFs)
- Arkville's Packet Generator Control and Status
- Arkville's Packet Director Control and Status
- Arkville's Packet Checker Control and Status
- Arkville's Timebase Management

4.10 Pre-Requisites

1. Prepare the system as recommended by DPDK suite. This includes environment variables, hugepages configuration, tool-chains and configuration
2. Insert `igb_uio` kernel module using the command `'modprobe igb_uio'`
3. Bind the intended ARK device to `igb_uio` module

At this point the system should be ready to run DPDK applications. Once the application runs to completion, the ARK PMD can be detached from `igb_uio` if necessary.

4.11 Usage Example

Follow instructions available in the document [compiling and testing a PMD for a NIC](#) to launch `testpmd` with Atomic Rules ARK devices managed by `librte_pmd_ark`.

Example output:

```
[...]
EAL: PCI device 0000:01:00.0 on NUMA socket -1
EAL:   probe driver: 1d6c:100e rte_ark_pmd
EAL:   PCI memory mapped at 0x7f9b6c400000
PMD: eth_ark_dev_init(): Initializing 0:2:0.1
ARKP PMD CommitID: 378f3a67
Configuring Port 0 (socket 0)
Port 0: DC:3C:F6:00:00:01
Checking link statuses...
Port 0 Link Up - speed 100000 Mbps - full-duplex
Done
testpmd>
```

AVP POLL MODE DRIVER

The Accelerated Virtual Port (AVP) device is a shared memory based device only available on [virtualization platforms](#) from Wind River Systems. The Wind River Systems virtualization platform currently uses QEMU/KVM as its hypervisor and as such provides support for all of the QEMU supported virtual and/or emulated devices (e.g., virtio, e1000, etc.). The platform offers the virtio device type as the default device when launching a virtual machine or creating a virtual machine port. The AVP device is a specialized device available to customers that require increased throughput and decreased latency to meet the demands of their performance focused applications.

The AVP driver binds to any AVP PCI devices that have been exported by the Wind River Systems QEMU/KVM hypervisor. As a user of the DPDK driver API it supports a subset of the full Ethernet device API to enable the application to use the standard device configuration functions and packet receive/transmit functions.

These devices enable optimized packet throughput by bypassing QEMU and delivering packets directly to the virtual switch via a shared memory mechanism. This provides DPDK applications running in virtual machines with significantly improved throughput and latency over other device types.

The AVP device implementation is integrated with the QEMU/KVM live-migration mechanism to allow applications to seamlessly migrate from one hypervisor node to another with minimal packet loss.

5.1 Features and Limitations of the AVP PMD

The AVP PMD driver provides the following functionality.

- Receive and transmit of both simple and chained mbuf packets,
- Chained mbufs may include up to 5 chained segments,
- Up to 8 receive and transmit queues per device,
- Only a single MAC address is supported,
- The MAC address cannot be modified,
- The maximum receive packet length is 9238 bytes,
- VLAN header stripping and inserting,
- Promiscuous mode
- VM live-migration

- PCI hotplug insertion and removal

5.2 Prerequisites

The following prerequisites apply:

- A virtual machine running in a Wind River Systems virtualization environment and configured with at least one neutron port defined with a vif-model set to “avp”.

5.3 Launching a VM with an AVP type network attachment

The following example will launch a VM with three network attachments. The first attachment will have a default vif-model of “virtio”. The next two network attachments will have a vif-model of “avp” and may be used with a DPDK application which is built to include the AVP PMD driver.

```
nova boot --flavor small --image my-image \  
  --nic net-id=${NETWORK1_UUID} \  
  --nic net-id=${NETWORK2_UUID},vif-model=avp \  
  --nic net-id=${NETWORK3_UUID},vif-model=avp \  
  --security-group default my-instance1
```

BNX2X POLL MODE DRIVER

The BNX2X poll mode driver library (**librte_pmd_bnx2x**) implements support for **QLogic 578xx** 10/20 Gbps family of adapters as well as their virtual functions (VF) in SR-IOV context. It is supported on several standard Linux distros like Red Hat 7.x and SLES12 OS. It is compile-tested under FreeBSD OS.

More information can be found at [QLogic Corporation's Official Website](#).

6.1 Supported Features

BNX2X PMD has support for:

- Base L2 features
- Unicast/multicast filtering
- Promiscuous mode
- Port hardware statistics
- SR-IOV VF

6.2 Non-supported Features

The features not yet supported include:

- TSS (Transmit Side Scaling)
- RSS (Receive Side Scaling)
- LRO/TSO offload
- Checksum offload
- SR-IOV PF
- Rx TX scatter gather

6.3 Co-existence considerations

- BCM578xx being a CNA can have both NIC and Storage personalities. However, co-existence with storage protocol drivers (cnic, bnx2fc and bnx2fi) is not supported on the

same adapter. So storage personality has to be disabled on that adapter when used in DPDK applications.

- For SR-IOV case, bnx2x PMD will be used to bind to SR-IOV VF device and Linux native kernel driver (bnx2x) will be attached to SR-IOV PF.

6.4 Supported QLogic NICs

- 578xx

6.5 Prerequisites

- Requires firmware version **7.2.51.0**. It is included in most of the standard Linux distros. If it is not available visit [linux-firmware git repository](#) to get the required firmware.

6.6 Pre-Installation Configuration

6.6.1 Config File Options

The following options can be modified in the `.config` file. Please note that enabling debugging options may affect system performance.

- `CONFIG_RTE_LIBRTE_BNX2X_PMD` (default **n**)
Toggle compilation of bnx2x driver. To use bnx2x PMD set this config parameter to 'y'. Also, in order for firmware binary to load user will need zlib devel package installed.
- `CONFIG_RTE_LIBRTE_BNX2X_DEBUG_TX` (default **n**)
Toggle display of transmit fast path run-time messages.
- `CONFIG_RTE_LIBRTE_BNX2X_DEBUG_RX` (default **n**)
Toggle display of receive fast path run-time messages.
- `CONFIG_RTE_LIBRTE_BNX2X_DEBUG_PERIODIC` (default **n**)
Toggle display of register reads and writes.

6.7 Driver compilation and testing

Refer to the document *compiling and testing a PMD for a NIC* for details.

6.8 SR-IOV: Prerequisites and sample Application Notes

This section provides instructions to configure SR-IOV with Linux OS.

1. Verify SR-IOV and ARI capabilities are enabled on the adapter using `lspci`:

```
lspci -s <slot> -vvv
```

Example output:

```
[...]
Capabilities: [1b8 v1] Alternative Routing-ID Interpretation (ARI)
[...]
Capabilities: [1c0 v1] Single Root I/O Virtualization (SR-IOV)
[...]
Kernel driver in use: igb_uio
```

2. Load the kernel module:

```
modprobe bnx2x
```

Example output:

```
systemd-udevd[4848]: renamed network interface eth0 to ens5f0
systemd-udevd[4848]: renamed network interface eth1 to ens5f1
```

3. Bring up the PF ports:

```
ifconfig ens5f0 up
ifconfig ens5f1 up
```

4. Create VF device(s):

Echo the number of VFs to be created into “sriov_numvfs” sysfs entry of the parent PF.

Example output:

```
echo 2 > /sys/devices/pci0000:00/0000:00:03.0/0000:81:00.0/sriov_numvfs
```

5. Assign VF MAC address:

Assign MAC address to the VF using iproute2 utility. The syntax is: ip link set <PF iface> vf <VF id> mac <macaddr>

Example output:

```
ip link set ens5f0 vf 0 mac 52:54:00:2f:9d:e8
```

6. PCI Passthrough:

The VF devices may be passed through to the guest VM using virt-manager or virsh etc. bnx2x PMD should be used to bind the VF devices in the guest VM using the instructions outlined in the Application notes below.

7. Running testpmd:

Follow instructions available in the document [compiling and testing a PMD for a NIC](#) to run testpmd.

Example output:

```
[...]
EAL: PCI device 0000:84:00.0 on NUMA socket 1
EAL: probe driver: 14e4:168e rte_bnx2x_pmd
EAL: PCI memory mapped at 0x7f14f6fe5000
EAL: PCI memory mapped at 0x7f14f67e5000
EAL: PCI memory mapped at 0x7f15fbd9b000
EAL: PCI device 0000:84:00.1 on NUMA socket 1
EAL: probe driver: 14e4:168e rte_bnx2x_pmd
EAL: PCI memory mapped at 0x7f14f5fe5000
EAL: PCI memory mapped at 0x7f14f57e5000
EAL: PCI memory mapped at 0x7f15fbd4f000
Interactive-mode selected
Configuring Port 0 (socket 0)
PMD: bnx2x_dev_tx_queue_setup(): fp[00] req_bd=512, thresh=512,
```

```
        usable_bd=1020, total_bd=1024,  
            tx_pages=4  
PMD: bnx2x_dev_rx_queue_setup(): fp[00] req_bd=128, thresh=0,  
        usable_bd=510, total_bd=512,  
            rx_pages=1, cq_pages=8  
PMD: bnx2x_print_adapter_info():  
[...]  
Checking link statuses...  
Port 0 Link Up - speed 10000 Mbps - full-duplex  
Port 1 Link Up - speed 10000 Mbps - full-duplex  
Done  
testpmd>
```

BNXT POLL MODE DRIVER

The bnxt poll mode library (`librte_pmd_bnxt`) implements support for:

- **Broadcom NetXtreme-C®/NetXtreme-E® BCM5730X and BCM574XX family of Ethernet Network Controllers**

These adapters support Standards compliant 10/25/50/100Gbps 30MPPS full-duplex throughput.

Information about the NetXtreme family of adapters can be found in the [NetXtreme® Brand](#) section of the [Broadcom website](#).

- **Broadcom StrataGX® BCM5871X Series of Communications Processors**

These ARM based processors target a broad range of networking applications including virtual CPE (vCPE) and NFV appliances, 10G service routers and gateways, control plane processing for Ethernet switches and network attached storage (NAS).

Information about the StrataGX family of adapters can be found in the [StrataGX® BCM5871X Series](#) section of the [Broadcom website](#).

7.1 Limitations

With the current driver, allocated mbufs must be large enough to hold the entire received frame. If the mbufs are not large enough, the packets will be dropped. This is most limiting when jumbo frames are used.

CXGBE POLL MODE DRIVER

The CXGBE PMD (`librte_pmd_cxgbe`) provides poll mode driver support for **Chelsio Terminator** 10/25/40/100 Gbps family of adapters. CXGBE PMD has support for the latest Linux and FreeBSD operating systems.

More information can be found at [Chelsio Communications Official Website](#).

8.1 Features

CXGBE PMD has support for:

- Multiple queues for TX and RX
- Receiver Side Steering (RSS)
- VLAN filtering
- Checksum offload
- Promiscuous mode
- All multicast mode
- Port hardware statistics
- Jumbo frames

8.2 Limitations

The Chelsio Terminator series of devices provide two/four ports but expose a single PCI bus address, thus, `librte_pmd_cxgbe` registers itself as a PCI driver that allocates one Ethernet device per detected port.

For this reason, one cannot whitelist/blacklist a single port without whitelisting/blacklisting the other ports on the same device.

8.3 Supported Chelsio T5 NICs

- 1G NICs: T502-BT
- 10G NICs: T520-BT, T520-CR, T520-LL-CR, T520-SO-CR, T540-CR

- 40G NICs: T580-CR, T580-LP-CR, T580-SO-CR
- Other T5 NICs: T522-CR

8.4 Supported Chelsio T6 NICs

- 25G NICs: T6425-CR, T6225-CR, T6225-LL-CR, T6225-SO-CR
- 100G NICs: T62100-CR, T62100-LP-CR, T62100-SO-CR

8.5 Prerequisites

- Requires firmware version **1.16.43.0** and higher. Visit [Chelsio Download Center](#) to get latest firmware bundled with the latest Chelsio Unified Wire package.

For Linux, installing and loading the latest cxgb4 kernel driver from the Chelsio Unified Wire package should get you the latest firmware. More information can be obtained from the User Guide that is bundled with the Chelsio Unified Wire package.

For FreeBSD, the latest firmware obtained from the Chelsio Unified Wire package must be manually flashed via cxgbetool available in FreeBSD source repository.

Instructions on how to manually flash the firmware are given in section [Linux Installation](#) for Linux and section [FreeBSD Installation](#) for FreeBSD.

8.6 Pre-Installation Configuration

8.6.1 Config File Options

The following options can be modified in the `.config` file. Please note that enabling debugging options may affect system performance.

- `CONFIG_RTE_LIBRTE_CXGBE_PMD` (default **y**)
Toggle compilation of `librte_pmd_cxgbe` driver.
- `CONFIG_RTE_LIBRTE_CXGBE_DEBUG` (default **n**)
Toggle display of generic debugging messages.
- `CONFIG_RTE_LIBRTE_CXGBE_DEBUG_REG` (default **n**)
Toggle display of registers related run-time check messages.
- `CONFIG_RTE_LIBRTE_CXGBE_DEBUG_MBOX` (default **n**)
Toggle display of firmware mailbox related run-time check messages.
- `CONFIG_RTE_LIBRTE_CXGBE_DEBUG_TX` (default **n**)
Toggle display of transmission data path run-time check messages.
- `CONFIG_RTE_LIBRTE_CXGBE_DEBUG_RX` (default **n**)
Toggle display of receiving data path run-time check messages.

- CONFIG_RTE_LIBRTE_CXGBE_TPUT (default y)
Toggle behaviour to prefer Throughput or Latency.

8.7 Driver compilation and testing

Refer to the document *compiling and testing a PMD for a NIC* for details.

8.8 Linux

8.8.1 Linux Installation

Steps to manually install the latest firmware from the downloaded Chelsio Unified Wire package for Linux operating system are as follows:

1. Load the kernel module:

```
modprobe cxgb4
```

2. Use ifconfig to get the interface name assigned to Chelsio card:

```
ifconfig -a | grep "00:07:43"
```

Example output:

```
p1p1      Link encap:Ethernet  HWaddr 00:07:43:2D:EA:C0
p1p2      Link encap:Ethernet  HWaddr 00:07:43:2D:EA:C8
```

3. Install cxgbtool:

```
cd <path_to_uwire>/tools/cxgbtool
make install
```

4. Use cxgbtool to load the firmware config file onto the card:

```
cxgbtool p1p1 loadcfg <path_to_uwire>/src/network/firmware/t5-config.txt
```

5. Use cxgbtool to load the firmware image onto the card:

```
cxgbtool p1p1 loadfw <path_to_uwire>/src/network/firmware/t5fw-*.bin
```

6. Unload and reload the kernel module:

```
modprobe -r cxgb4
modprobe cxgb4
```

7. Verify with ethtool:

```
ethtool -i p1p1 | grep "firmware"
```

Example output:

```
firmware-version: 1.16.43.0, TP 0.1.4.9
```

8.8.2 Running testpmd

This section demonstrates how to launch **testpmd** with Chelsio devices managed by `librte_pmd_cxgbe` in Linux operating system.

1. Load the kernel module:

```
modprobe cxgb4
```

2. Get the PCI bus addresses of the interfaces bound to cxgb4 driver:

```
dmesg | tail -2
```

Example output:

```
cxgb4 0000:02:00.4 p1p1: renamed from eth0
cxgb4 0000:02:00.4 p1p2: renamed from eth1
```

Note: Both the interfaces of a Chelsio 2-port adapter are bound to the same PCI bus address.

3. Unload the kernel module:

```
modprobe -ar cxgb4 csiostor
```

4. Running testpmd

Follow instructions available in the document [compiling and testing a PMD for a NIC](#) to run testpmd.

Note: Currently, CXGBE PMD only supports the binding of PF4 for Chelsio NICs.

Example output:

```
[...]
EAL: PCI device 0000:02:00.4 on NUMA socket -1
EAL:  probe driver: 1425:5401 rte_cxgbe_pmd
EAL:  PCI memory mapped at 0x7fd7c0200000
EAL:  PCI memory mapped at 0x7fd77cdfd000
EAL:  PCI memory mapped at 0x7fd7c10b7000
PMD: rte_cxgbe_pmd: fw: 1.16.43.0, TP: 0.1.4.9
PMD: rte_cxgbe_pmd: Coming up as MASTER: Initializing adapter
Interactive-mode selected
Configuring Port 0 (socket 0)
Port 0: 00:07:43:2D:EA:C0
Configuring Port 1 (socket 0)
Port 1: 00:07:43:2D:EA:C8
Checking link statuses...
PMD: rte_cxgbe_pmd: Port0: passive DA port module inserted
PMD: rte_cxgbe_pmd: Port1: passive DA port module inserted
Port 0 Link Up - speed 10000 Mbps - full-duplex
Port 1 Link Up - speed 10000 Mbps - full-duplex
Done
testpmd>
```

Note: Flow control pause TX/RX is disabled by default and can be enabled via testpmd. Refer section [Enable/Disable Flow Control](#) for more details.

8.9 FreeBSD

8.9.1 FreeBSD Installation

Steps to manually install the latest firmware from the downloaded Chelsio Unified Wire package for FreeBSD operating system are as follows:

1. Load the kernel module:

```
kldload if_cxgbe
```

2. Use dmesg to get the t5nex instance assigned to the Chelsio card:

```
dmesg | grep "t5nex"
```

Example output:

```
t5nex0: <Chelsio T520-CR> irq 16 at device 0.4 on pci2
cx10: <port 0> on t5nex0
cx11: <port 1> on t5nex0
t5nex0: PCIe x8, 2 ports, 14 MSI-X interrupts, 31 eq, 13 iq
```

In the example above, a Chelsio T520-CR card is bound to a t5nex0 instance.

3. Install cxgbetool from FreeBSD source repository:

```
cd <path_to_FreeBSD_source>/tools/tools/cxgbetool/
make && make install
```

4. Use cxgbetool to load the firmware image onto the card:

```
cxgbetool t5nex0 loadfw <path_to_uwire>/src/network/firmware/t5fw-*.bin
```

5. Unload and reload the kernel module:

```
kldunload if_cxgbe
kldload if_cxgbe
```

6. Verify with sysctl:

```
sysctl -a | grep "t5nex" | grep "firmware"
```

Example output:

```
dev.t5nex.0.firmware_version: 1.16.43.0
```

8.9.2 Running testpmd

This section demonstrates how to launch **testpmd** with Chelsio devices managed by `librte_pmd_cxgbe` in FreeBSD operating system.

1. Change to DPDK source directory where the target has been compiled in section [Driver compilation and testing](#):

```
cd <DPDK-source-directory>
```

2. Copy the contigmem kernel module to `/boot/kernel` directory:

```
cp x86_64-native-bsdapp-clang/kmod/contigmem.ko /boot/kernel/
```

3. Add the following lines to `/boot/loader.conf`:

```
# reserve 2 x 1G blocks of contiguous memory using contigmem driver
hw.contigmem.num_buffers=2
hw.contigmem.buffer_size=1073741824
# load contigmem module during boot process
contigmem_load="YES"
```

The above lines load the contigmem kernel module during boot process and allocate 2 x 1G blocks of contiguous memory to be used for DPDK later on. This is to avoid issues with potential memory fragmentation during later system up time, which may result in failure of allocating the contiguous memory required for the contigmem kernel module.

- Restart the system and ensure the contigmem module is loaded successfully:

```
reboot
kldstat | grep "contigmem"
```

Example output:

```
2      1 0xffffffff817f1000 3118      contigmem.ko
```

- Repeat step 1 to ensure that you are in the DPDK source directory.
- Load the cxgbe kernel module:

```
kldload if_cxgbe
```

- Get the PCI bus addresses of the interfaces bound to t5nex driver:

```
pciconf -l | grep "t5nex"
```

Example output:

```
t5nex0@pci0:2:0:4: class=0x020000 card=0x00001425 chip=0x54011425 rev=0x00
```

In the above example, the t5nex0 is bound to 2:0:4 bus address.

Note: Both the interfaces of a Chelsio 2-port adapter are bound to the same PCI bus address.

- Unload the kernel module:

```
kldunload if_cxgbe
```

- Set the PCI bus addresses to hw.nic_uio.bdfs kernel environment parameter:

```
kenv hw.nic_uio.bdfs="2:0:4"
```

This automatically binds 2:0:4 to nic_uio kernel driver when it is loaded in the next step.

Note: Currently, CXGBE PMD only supports the binding of PF4 for Chelsio NICs.

- Load nic_uio kernel driver:

```
kldload ./x86_64-native-bsdapp-clang/kmod/nic_uio.ko
```

- Start testpmd with basic parameters:

```
./x86_64-native-bsdapp-clang/app/testpmd -l 0-3 -n 4 -w 0000:02:00.4 -- -i
```

Example output:

```
[...]
EAL: PCI device 0000:02:00.4 on NUMA socket 0
EAL:  probe driver: 1425:5401 rte_cxgbe_pmd
EAL:  PCI memory mapped at 0x8007ec000
EAL:  PCI memory mapped at 0x842800000
EAL:  PCI memory mapped at 0x80086c000
PMD: rte_cxgbe_pmd: fw: 1.16.43.0, TP: 0.1.4.9
PMD: rte_cxgbe_pmd: Coming up as MASTER: Initializing adapter
Interactive-mode selected
Configuring Port 0 (socket 0)
Port 0: 00:07:43:2D:EA:C0
Configuring Port 1 (socket 0)
Port 1: 00:07:43:2D:EA:C8
Checking link statuses...
PMD: rte_cxgbe_pmd: Port0: passive DA port module inserted
PMD: rte_cxgbe_pmd: Port1: passive DA port module inserted
```

```
Port 0 Link Up - speed 10000 Mbps - full-duplex
Port 1 Link Up - speed 10000 Mbps - full-duplex
Done
testpmd>
```

Note: Flow control pause TX/RX is disabled by default and can be enabled via testpmd. Refer section [Enable/Disable Flow Control](#) for more details.

8.10 Sample Application Notes

8.10.1 Enable/Disable Flow Control

Flow control pause TX/RX is disabled by default and can be enabled via testpmd as follows:

```
testpmd> set flow_ctrl rx on tx on 0 0 0 0 mac_ctrl_frame_fwd off autoneg on 0
testpmd> set flow_ctrl rx on tx on 0 0 0 0 mac_ctrl_frame_fwd off autoneg on 1
```

To disable again, run:

```
testpmd> set flow_ctrl rx off tx off 0 0 0 0 mac_ctrl_frame_fwd off autoneg off 0
testpmd> set flow_ctrl rx off tx off 0 0 0 0 mac_ctrl_frame_fwd off autoneg off 1
```

8.10.2 Jumbo Mode

There are two ways to enable sending and receiving of jumbo frames via testpmd. One method involves using the **mtu** command, which changes the mtu of an individual port without having to stop the selected port. Another method involves stopping all the ports first and then running **max-pkt-len** command to configure the mtu of all the ports with a single command.

- To configure each port individually, run the mtu command as follows:

```
testpmd> port config mtu 0 9000
testpmd> port config mtu 1 9000
```

- To configure all the ports at once, stop all the ports first and run the max-pkt-len command as follows:

```
testpmd> port stop all
testpmd> port config all max-pkt-len 9000
```

DPAA POLL MODE DRIVER

The DPAA NIC PMD (`librte_pmd_dpaa`) provides poll mode driver support for the inbuilt NIC found in the **NXP DPAA** SoC family.

More information can be found at [NXP Official Website](#).

9.1 NXP DPAA (Data Path Acceleration Architecture - Gen 1)

This section provides an overview of the NXP DPAA architecture and how it is integrated into the DPDK.

Contents summary

- DPAA overview
- DPAA driver architecture overview

9.1.1 DPAA Overview

Reference: [FSL DPAA Architecture](#).

The QorIQ Data Path Acceleration Architecture (DPAA) is a set of hardware components on specific QorIQ series multicore processors. This architecture provides the infrastructure to support simplified sharing of networking interfaces and accelerators by multiple CPU cores, and the accelerators themselves.

DPAA includes:

- Cores
- Network and packet I/O
- Hardware offload accelerators
- Infrastructure required to facilitate flow of packets between the components above

Infrastructure components are:

- The Queue Manager (QMan) is a hardware accelerator that manages frame queues. It allows CPUs and other accelerators connected to the SoC datapath to enqueue and dequeue ethernet frames, thus providing the infrastructure for data exchange among CPUs and datapath accelerators.

- The Buffer Manager (BMan) is a hardware buffer pool management block that allows software and accelerators on the datapath to acquire and release buffers in order to build frames.

Hardware accelerators are:

- SEC - Cryptographic accelerator
- PME - Pattern matching engine

The Network and packet I/O component:

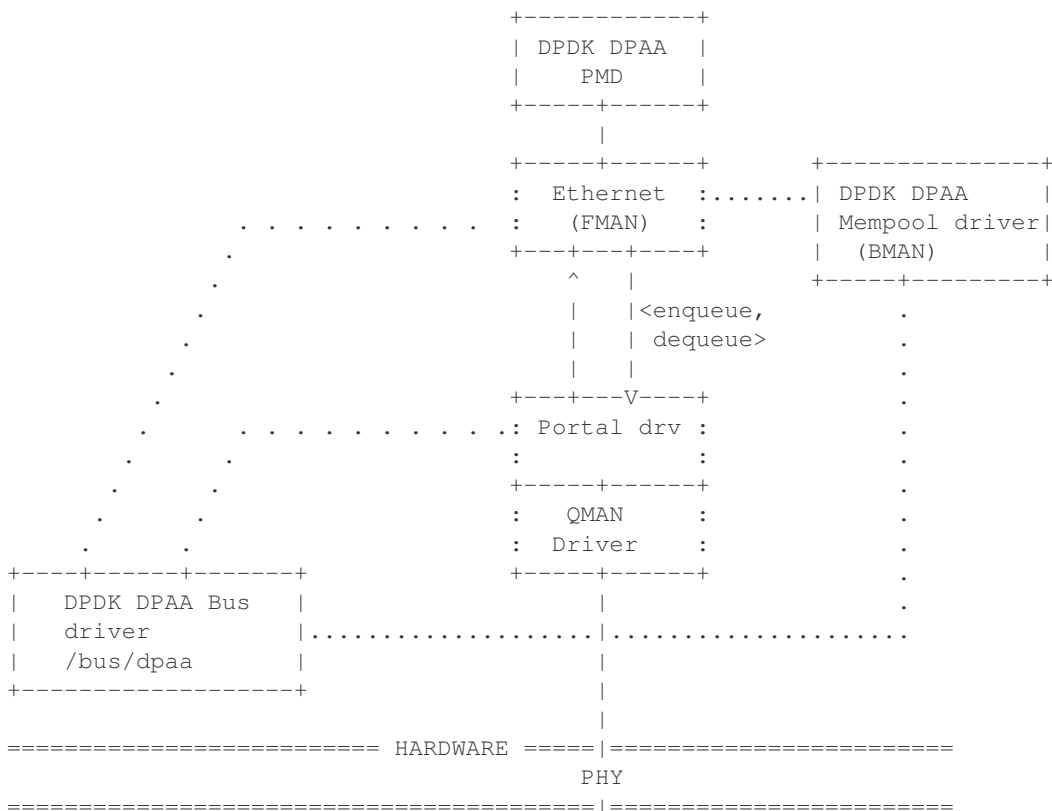
- The Frame Manager (FMan) is a key component in the DPAA and makes use of the DPAA infrastructure (QMan and BMan). FMan is responsible for packet distribution and policing. Each frame can be parsed, classified and results may be attached to the frame. This meta data can be used to select particular QMan queue, which the packet is forwarded to.

9.2 DPAA DPDK - Poll Mode Driver Overview

This section provides an overview of the drivers for DPAA:

- Bus driver and associated “DPAA infrastructure” drivers
- Functional object drivers (such as Ethernet).

Brief description of each driver is provided in layout below as well as in the following sections.



In the above representation, solid lines represent components which interface with DPDK RTE Framework and dotted lines represent DPAA internal components.

9.2.1 DPAA Bus driver

The DPAA bus driver is a `rte_bus` driver which scans the platform like bus. Key functions include:

- Scanning and parsing the various objects and adding them to their respective device list.
- Performing probe for available drivers against each scanned device
- Creating necessary ethernet instance before passing control to the PMD

9.2.2 DPAA NIC Driver (PMD)

DPAA PMD is traditional DPDK PMD which provides necessary interface between RTE framework and DPAA internal components/drivers.

- Once devices have been identified by DPAA Bus, each device is associated with the PMD
- PMD is responsible for implementing necessary glue layer between RTE APIs and lower level QMan and FMan blocks. The Ethernet driver is bound to a FMAN port and implements the interfaces needed to connect the DPAA network interface to the network stack. Each FMAN Port corresponds to a DPDK network interface.

Features

Features of the DPAA PMD are:

- Multiple queues for TX and RX
- Receive Side Scaling (RSS)
- Packet type information
- Checksum offload
- Promiscuous mode

9.2.3 DPAA Mempool Driver

DPAA has a hardware offloaded buffer pool manager, called BMan, or Buffer Manager.

- Using standard Mempools operations RTE API, the mempool driver interfaces with RTE to service each mempool creation, deletion, buffer allocation and deallocation requests.
- Each FMAN instance has a BMan pool attached to it during initialization. Each Tx frame can be automatically released by hardware, if allocated from this pool.

9.3 Supported DPAA SoCs

- LS1043A/LS1023A
- LS1046A/LS1026A

9.4 Prerequisites

There are three main pre-requisites for executing DPAA PMD on a DPAA compatible board:

1. **ARM 64 Tool Chain**

For example, the [*aarch64* Linaro Toolchain](#).

2. **Linux Kernel**

It can be obtained from [NXP's Github hosting](#).

3. **Rootfile system**

Any *aarch64* supporting filesystem can be used. For example, Ubuntu 15.10 (Wily) or 16.04 LTS (Xenial) userland which can be obtained from [here](#).

4. **FMC Tool**

Before any DPDK application can be executed, the Frame Manager Configuration Tool (FMC) need to be executed to set the configurations of the queues. This includes the queue state, RSS and other policies. This tool can be obtained from [NXP \(Freescale\) Public Git Repository](#).

This tool needs configuration files which are available in the [DPDK Extra Scripts](#), described below for DPDK usages.

As an alternative method, DPAA PMD can also be executed using images provided as part of SDK from NXP. The SDK includes all the above prerequisites necessary to bring up a DPAA board.

The following dependencies are not part of DPDK and must be installed separately:

- **NXP Linux SDK**

NXP Linux software development kit (SDK) includes support for family of QorIQ® ARM-Architecture-based system on chip (SoC) processors and corresponding boards.

It includes the Linux board support packages (BSPs) for NXP SoCs, a fully operational tool chain, kernel and board specific modules.

SDK and related information can be obtained from: [NXP QorIQ SDK](#).

- **DPDK Extra Scripts**

DPAA based resources can be configured easily with the help of ready scripts as provided in the DPDK Extra repository.

[DPDK Extras Scripts](#).

Currently supported by DPDK:

- NXP SDK **2.0+**.
- Supported architectures: **arm64 LE**.
- Follow the DPDK Getting Started Guide for Linux to setup the basic DPDK environment.

Note: Some part of dpaa bus code (qbman and fman - library) routines are dual licensed (BSD & GPLv2), however they are used as BSD in DPDK in userspace.

9.5 Pre-Installation Configuration

9.5.1 Config File Options

The following options can be modified in the `config` file. Please note that enabling debugging options may affect system performance.

- `CONFIG_RTE_LIBRTE_DPAA_BUS` (default n)

By default it is enabled only for `defconfig_arm64-dpaa-*` config. Toggle compilation of the `librte_bus_dpaa` driver.
- `CONFIG_RTE_LIBRTE_DPAA_PMD` (default n)

By default it is enabled only for `defconfig_arm64-dpaa-*` config. Toggle compilation of the `librte_pmd_dpaa` driver.
- `CONFIG_RTE_LIBRTE_DPAA_DEBUG_DRIVER` (default n)

Toggles display of bus configurations and enables a debugging queue to fetch error (Rx/Tx) packets to driver. By default, packets with errors (like wrong checksum) are dropped by the hardware.
- `CONFIG_RTE_LIBRTE_DPAA_HWDEBUG` (default n)

Enables debugging of the Queue and Buffer Manager layer which interacts with the DPAA hardware.
- `CONFIG_RTE_MBUF_DEFAULT_MEMPOOL_OPS` (default dpaa)

This is not a DPAA specific configuration - it is a generic RTE config. For optimal performance and hardware utilization, it is expected that DPAA Mempool driver is used for mempools. For that, this configuration needs to be enabled.

9.5.2 Environment Variables

DPAA drivers use the following environment variables to configure its state during application initialization:

- `DPAA_NUM_RX_QUEUES` (default 1)

This defines the number of Rx queues configured for an application, per port. Hardware would distribute across these many number of queues on Rx of packets. In case the application is configured to use lesser number of queues than configured above, it might result in packet loss (because of distribution).
- `DPAA_PUSH_QUEUES_NUMBER` (default 4)

This defines the number of High performance queues to be used for ethdev Rx. These queues use one private HW portal per queue configured, so they are limited in the system. The first configured ethdev queues will be automatically be assigned from these high perf PUSH queues. Any queue configuration beyond that will be standard Rx queues. The application can choose to change their number if HW portals are limited. The valid values are from '0' to '4'. The value shall be set to '0' if the application wants to use eventdev with DPAA device.

9.6 Driver compilation and testing

Refer to the document *compiling and testing a PMD for a NIC* for details.

1. Running testpmd:

Follow instructions available in the document *compiling and testing a PMD for a NIC* to run testpmd.

Example output:

```
./arm64-dpaa-linuxapp-gcc/testpmd -c 0xff -n 1 \  
-- -i --portmask=0x3 --nb-cores=1 --no-flush-rx
```

```
.....  
EAL: Registered [pci] bus.  
EAL: Registered [dpaa] bus.  
EAL: Detected 4 lcore(s)  
.....  
EAL: dpaa: Bus scan completed  
.....  
Configuring Port 0 (socket 0)  
Port 0: 00:00:00:00:00:01  
Configuring Port 1 (socket 0)  
Port 1: 00:00:00:00:00:02  
.....  
Checking link statuses...  
Port 0 Link Up - speed 10000 Mbps - full-duplex  
Port 1 Link Up - speed 10000 Mbps - full-duplex  
Done  
testpmd>
```

9.7 Limitations

9.7.1 Platform Requirement

DPAA drivers for DPDK can only work on NXP SoCs as listed in the Supported DPAA SoCs.

9.7.2 Maximum packet length

The DPAA SoC family support a maximum of a 10240 jumbo frame. The value is fixed and cannot be changed. So, even when the `rxmode.max_rx_pkt_len` member of `struct rte_eth_conf` is set to a value lower than 10240, frames up to 10240 bytes can still reach the host interface.

9.7.3 Multiprocess Support

Current version of DPAA driver doesn't support multi-process applications where I/O is performed using secondary processes. This feature would be implemented in subsequent versions.

DPAA2 POLL MODE DRIVER

The DPAA2 NIC PMD (`librte_pmd_dpaa2`) provides poll mode driver support for the inbuilt NIC found in the **NXP DPAA2** SoC family.

More information can be found at [NXP Official Website](#).

10.1 NXP DPAA2 (Data Path Acceleration Architecture Gen2)

This section provides an overview of the NXP DPAA2 architecture and how it is integrated into the DPDK.

Contents summary

- DPAA2 overview
- Overview of DPAA2 objects
- DPAA2 driver architecture overview

10.1.1 DPAA2 Overview

Reference: [FSL MC BUS in Linux Kernel](#).

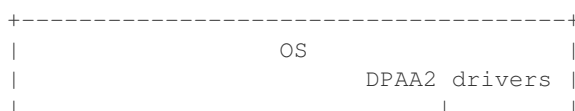
DPAA2 is a hardware architecture designed for high-speed network packet processing. DPAA2 consists of sophisticated mechanisms for processing Ethernet packets, queue management, buffer management, autonomous L2 switching, virtual Ethernet bridging, and accelerator (e.g. crypto) sharing.

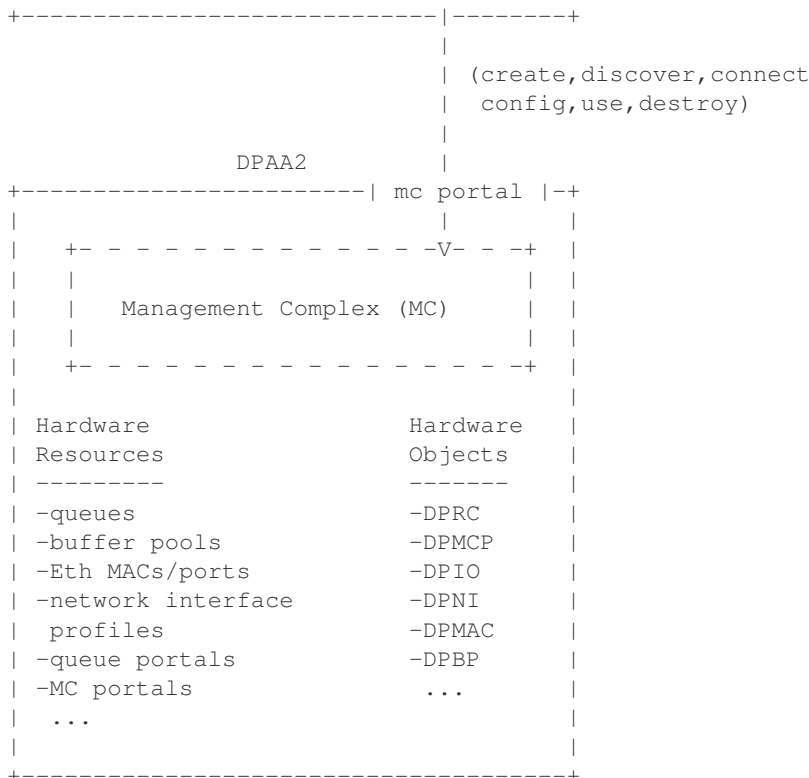
A DPAA2 hardware component called the Management Complex (or MC) manages the DPAA2 hardware resources. The MC provides an object-based abstraction for software drivers to use the DPAA2 hardware.

The MC uses DPAA2 hardware resources such as queues, buffer pools, and network ports to create functional objects/devices such as network interfaces, an L2 switch, or accelerator instances.

The MC provides memory-mapped I/O command interfaces (MC portals) which DPAA2 software drivers use to operate on DPAA2 objects:

The diagram below shows an overview of the DPAA2 resource management architecture:





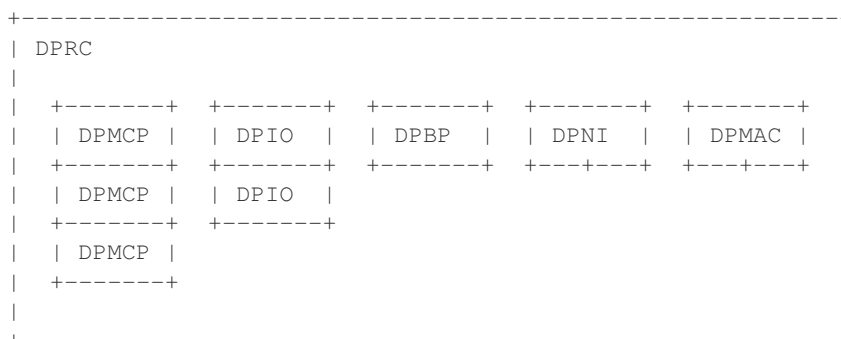
The MC mediates operations such as create, discover, connect, configuration, and destroy. Fast-path operations on data, such as packet transmit/receive, are not mediated by the MC and are done directly using memory mapped regions in DPIO objects.

10.1.2 Overview of DPAA2 Objects

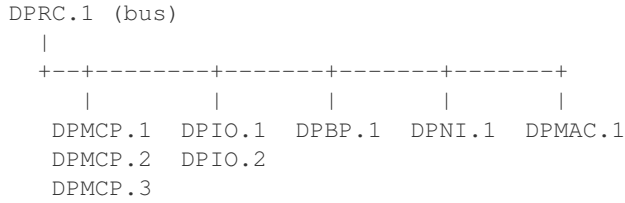
The section provides a brief overview of some key DPAA2 objects. A simple scenario is described illustrating the objects involved in creating a network interfaces.

DPRC (Datapath Resource Container)

A DPRC is a container object that holds all the other types of DPAA2 objects. In the example diagram below there are 8 objects of 5 types (DPMCP, DPIO, DPBP, DPNI, and DPMAC) in the container.



From the point of view of an OS, a DPRC behaves similar to a plug and play bus, like PCI. DPRC commands can be used to enumerate the contents of the DPRC, discover the hardware objects present (including mappable regions and interrupts).



Hardware objects can be created and destroyed dynamically, providing the ability to hot plug/unplug objects in and out of the DPRC.

A DPRC has a mappable MMIO region (an MC portal) that can be used to send MC commands. It has an interrupt for status events (like hotplug).

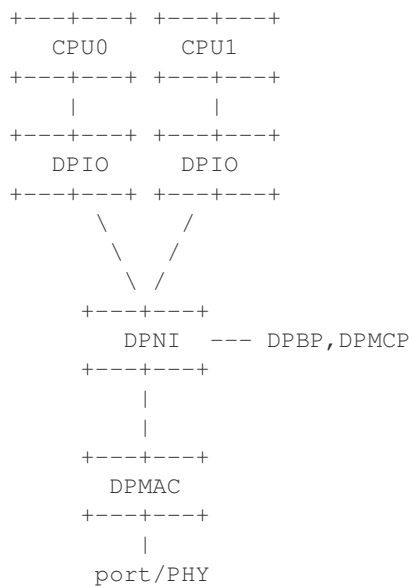
All objects in a container share the same hardware “isolation context”. This means that with respect to an IOMMU the isolation granularity is at the DPRC (container) level, not at the individual object level.

DPRCs can be defined statically and populated with objects via a config file passed to the MC when firmware starts it. There is also a Linux user space tool called “restool” that can be used to create/destroy containers and objects dynamically.

10.1.3 DPAA2 Objects for an Ethernet Network Interface

A typical Ethernet NIC is monolithic– the NIC device contains TX/RX queuing mechanisms, configuration mechanisms, buffer management, physical ports, and interrupts. DPAA2 uses a more granular approach utilizing multiple hardware objects. Each object provides specialized functions. Groups of these objects are used by software to provide Ethernet network interface functionality. This approach provides efficient use of finite hardware resources, flexibility, and performance advantages.

The diagram below shows the objects needed for a simple network interface configuration on a system with 2 CPUs.



Below the objects are described. For each object a brief description is provided along with a summary of the kinds of operations the object supports and a summary of key resources of the object (MMIO regions and IRQs).

DPMAC (Datapath Ethernet MAC): represents an Ethernet MAC, a hardware device that connects to an Ethernet PHY and allows physical transmission and reception of Ethernet frames.

- MMIO regions: none
- IRQs: DPNI link change
- commands: set link up/down, link config, get stats, IRQ config, enable, reset

DPNI (Datapath Network Interface): contains TX/RX queues, network interface configuration, and RX buffer pool configuration mechanisms. The TX/RX queues are in memory and are identified by queue number.

- MMIO regions: none
- IRQs: link state
- commands: port config, offload config, queue config, parse/classify config, IRQ config, enable, reset

DPIO (Datapath I/O): provides interfaces to enqueue and dequeue packets and do hardware buffer pool management operations. The DPAA2 architecture separates the mechanism to access queues (the DPIO object) from the queues themselves. The DPIO provides an MMIO interface to enqueue/dequeue packets. To enqueue something a descriptor is written to the DPIO MMIO region, which includes the target queue number. There will typically be one DPIO assigned to each CPU. This allows all CPUs to simultaneously perform enqueue/dequeue operations. DPIOs are expected to be shared by different DPAA2 drivers.

- MMIO regions: queue operations, buffer management
- IRQs: data availability, congestion notification, buffer pool depletion
- commands: IRQ config, enable, reset

DPBP (Datapath Buffer Pool): represents a hardware buffer pool.

- MMIO regions: none
- IRQs: none
- commands: enable, reset

DPMCP (Datapath MC Portal): provides an MC command portal. Used by drivers to send commands to the MC to manage objects.

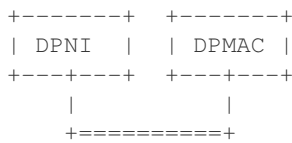
- MMIO regions: MC command portal
- IRQs: command completion
- commands: IRQ config, enable, reset

10.1.4 Object Connections

Some objects have explicit relationships that must be configured:

- DPNI <--> DPMAC
- DPNI <--> DPNI
- DPNI <--> L2-switch-port

A DPNI must be connected to something such as a DPMAC, another DPNI, or L2 switch port. The DPNI connection is made via a DPRC command.



- DPNI <-> DPBP

A network interface requires a ‘buffer pool’ (DPBP object) which provides a list of pointers to memory where received Ethernet data is to be copied. The Ethernet driver configures the DPBPs associated with the network interface.

10.1.5 Interrupts

All interrupts generated by DPAA2 objects are message interrupts. At the hardware level message interrupts generated by devices will normally have 3 components– 1) a non-spoofable ‘device-id’ expressed on the hardware bus, 2) an address, 3) a data value.

In the case of DPAA2 devices/objects, all objects in the same container/DPRC share the same ‘device-id’. For ARM-based SoC this is the same as the stream ID.

10.2 DPAA2 DPDK - Poll Mode Driver Overview

This section provides an overview of the drivers for DPAA2– 1) the bus driver and associated “DPAA2 infrastructure” drivers and 2) functional object drivers (such as Ethernet).

As described previously, a DPRC is a container that holds the other types of DPAA2 objects. It is functionally similar to a plug-and-play bus controller.

Each object in the DPRC is a Linux “device” and is bound to a driver. The diagram below shows the dpaa2 drivers involved in a networking scenario and the objects bound to each driver. A brief description of each driver follows.

A brief description of each driver is provided below.

10.2.1 DPAA2 bus driver

The DPAA2 bus driver is a rte_bus driver which scans the fsl-mc bus. Key functions include:

- Reading the container and setting up vfio group
- Scanning and parsing the various MC objects and adding them to their respective device list.

Additionally, it also provides the object driver for generic MC objects.

10.2.2 DPIO driver

The DPIO driver is bound to DPIO objects and provides services that allow other drivers such as the Ethernet driver to enqueue and dequeue data for their respective objects. Key services include:

- Data availability notifications
- Hardware queuing operations (enqueue and dequeue of data)
- Hardware buffer pool management

To transmit a packet the Ethernet driver puts data on a queue and invokes a DPIO API. For receive, the Ethernet driver registers a data availability notification callback. To dequeue a packet a DPIO API is used.

There is typically one DPIO object per physical CPU for optimum performance, allowing different CPUs to simultaneously enqueue and dequeue data.

The DPIO driver operates on behalf of all DPAA2 drivers active – Ethernet, crypto, compression, etc.

10.2.3 DPBP based Mempool driver

The DPBP driver is bound to a DPBP objects and provides services to create a hardware offloaded packet buffer mempool.

10.2.4 DPAA2 NIC Driver

The Ethernet driver is bound to a DPNI and implements the kernel interfaces needed to connect the DPAA2 network interface to the network stack.

Each DPNI corresponds to a DPDK network interface.

Features

Features of the DPAA2 PMD are:

- Multiple queues for TX and RX
- Receive Side Scaling (RSS)
- MAC/VLAN filtering
- Packet type information
- Checksum offload
- Promiscuous mode
- Multicast mode
- Port hardware statistics
- Jumbo frames
- Link flow control
- Scattered and gather for TX and RX

10.3 Supported DPAA2 SoCs

- LS2080A/LS2040A
- LS2084A/LS2044A
- LS2088A/LS2048A
- LS1088A/LS1048A

10.4 Prerequisites

There are three main pre-requisites for executing DPAA2 PMD on a DPAA2 compatible board:

1. ARM 64 Tool Chain

For example, the [*aarch64* Linaro Toolchain](#).

2. Linux Kernel

It can be obtained from [NXP's Github hosting](#).

3. Rootfile system

Any *aarch64* supporting filesystem can be used. For example, Ubuntu 15.10 (Wily) or 16.04 LTS (Xenial) userland which can be obtained from [here](#).

As an alternative method, DPAA2 PMD can also be executed using images provided as part of SDK from NXP. The SDK includes all the above prerequisites necessary to bring up a DPAA2 board.

The following dependencies are not part of DPDK and must be installed separately:

• NXP Linux SDK

NXP Linux software development kit (SDK) includes support for family of QorIQ® ARM-Architecture-based system on chip (SoC) processors and corresponding boards.

It includes the Linux board support packages (BSPs) for NXP SoCs, a fully operational tool chain, kernel and board specific modules.

SDK and related information can be obtained from: [NXP QorIQ SDK](#).

• DPDK Extra Scripts

DPAA2 based resources can be configured easily with the help of ready scripts as provided in the DPDK Extra repository.

[DPDK Extras Scripts](#).

Currently supported by DPDK:

- NXP SDK **17.08+**.
- MC Firmware version **10.3.1** and higher.
- Supported architectures: **arm64 LE**.
- Follow the DPDK Getting Started Guide for Linux to setup the basic DPDK environment.

Note: Some part of fslmc bus code (mc flib - object library) routines are dual licensed (BSD & GPLv2), however they are used as BSD in DPDK in userspace.

10.5 Pre-Installation Configuration

10.5.1 Config File Options

The following options can be modified in the `config` file. Please note that enabling debugging options may affect system performance.

- `CONFIG_RTE_LIBRTE_FSLMC_BUS` (default n)
By default it is enabled only for `defconfig_arm64-dpaa2-*` config. Toggle compilation of the `librte_bus_fslmc` driver.
- `CONFIG_RTE_LIBRTE_DPAA2_PMD` (default n)
By default it is enabled only for `defconfig_arm64-dpaa2-*` config. Toggle compilation of the `librte_pmd_dpaa2` driver.
- `CONFIG_RTE_LIBRTE_DPAA2_DEBUG_DRIVER` (default n)
Toggle display of generic debugging messages
- `CONFIG_RTE_LIBRTE_DPAA2_USE_PHYS_IOVA` (default y)
Toggle to use physical address vs virtual address for hardware accelerators.
- `CONFIG_RTE_LIBRTE_DPAA2_DEBUG_INIT` (default n)
Toggle display of initialization related messages.
- `CONFIG_RTE_LIBRTE_DPAA2_DEBUG_RX` (default n)
Toggle display of receive fast path run-time message
- `CONFIG_RTE_LIBRTE_DPAA2_DEBUG_TX` (default n)
Toggle display of transmit fast path run-time message
- `CONFIG_RTE_LIBRTE_DPAA2_DEBUG_TX_FREE` (default n)
Toggle display of transmit fast path buffer free run-time message

10.6 Driver compilation and testing

Refer to the document *compiling and testing a PMD for a NIC* for details.

1. Running `testpmd`:

Follow instructions available in the document *compiling and testing a PMD for a NIC* to run `testpmd`.

Example output:

```
./arm64-dpaa2-linuxapp-gcc/testpmd -c 0xff -n 1 \  
-- -i --portmask=0x3 --nb-cores=1 --no-flush-rx  
  
.....  
EAL: Registered [pci] bus.  
EAL: Registered [fslmc] bus.  
EAL: Detected 8 lcore(s)  
EAL: Probing VFIO support...  
EAL: VFIO support initialized  
.....  
PMD: DPAA2: Processing Container = dprc.2  
EAL: fslmc: DPRC contains = 51 devices  
EAL: fslmc: Bus scan completed  
.....  
Configuring Port 0 (socket 0)  
Port 0: 00:00:00:00:00:01  
Configuring Port 1 (socket 0)  
Port 1: 00:00:00:00:00:02  
.....  
Checking link statuses...  
Port 0 Link Up - speed 10000 Mbps - full-duplex  
Port 1 Link Up - speed 10000 Mbps - full-duplex  
Done  
testpmd>
```

10.7 Limitations

10.7.1 Platform Requirement

DPAA2 drivers for DPDK can only work on NXP SoCs as listed in the Supported DPAA2 SoCs.

10.7.2 Maximum packet length

The DPAA2 SoC family support a maximum of a 10240 jumbo frame. The value is fixed and cannot be changed. So, even when the `rxmode.max_rx_pkt_len` member of `struct rte_eth_conf` is set to a value lower than 10240, frames up to 10240 bytes can still reach the host interface.

10.7.3 Other Limitations

- RSS hash key cannot be modified.
- RSS RETA cannot be configured.
- Secondary process packet I/O is not supported.

DRIVER FOR VM EMULATED DEVICES

The DPDK EM poll mode driver supports the following emulated devices:

- qemu-kvm emulated Intel® 82540EM Gigabit Ethernet Controller (qemu e1000 device)
- VMware* emulated Intel® 82545EM Gigabit Ethernet Controller
- VMware emulated Intel® 8274L Gigabit Ethernet Controller.

11.1 Validated Hypervisors

The validated hypervisors are:

- KVM (Kernel Virtual Machine) with Qemu, version 0.14.0
- KVM (Kernel Virtual Machine) with Qemu, version 0.15.1
- VMware ESXi 5.0, Update 1

11.2 Recommended Guest Operating System in Virtual Machine

The recommended guest operating system in a virtualized environment is:

- Fedora* 18 (64-bit)

For supported kernel versions, refer to the *DPDK Release Notes*.

11.3 Setting Up a KVM Virtual Machine

The following describes a target environment:

- Host Operating System: Fedora 14
- Hypervisor: KVM (Kernel Virtual Machine) with Qemu version, 0.14.0
- Guest Operating System: Fedora 14
- Linux Kernel Version: Refer to the DPDK Getting Started Guide
- Target Applications: testpmd

The setup procedure is as follows:

1. Download `qemu-kvm-0.14.0` from <http://sourceforge.net/projects/kvm/files/qemu-kvm/> and install it in the Host OS using the following steps:

When using a recent kernel (2.6.25+) with `kvm` modules included:

```
tar xzf qemu-kvm-release.tar.gz cd qemu-kvm-release
./configure --prefix=/usr/local/kvm
make
sudo make install
sudo /sbin/modprobe kvm-intel
```

When using an older kernel or a kernel from a distribution without the `kvm` modules, you must download (from the same link), compile and install the modules yourself:

```
tar xjf kvm-kmod-release.tar.bz2
cd kvm-kmod-release
./configure
make
sudo make install
sudo /sbin/modprobe kvm-intel
```

Note that `qemu-kvm` installs in the `/usr/local/bin` directory.

For more details about KVM configuration and usage, please refer to: <http://www.linux-kvm.org/page/HOWTO1>.

2. Create a Virtual Machine and install Fedora 14 on the Virtual Machine. This is referred to as the Guest Operating System (Guest OS).
3. Start the Virtual Machine with at least one emulated `e1000` device.

Note: The Qemu provides several choices for the emulated network device backend. Most commonly used is a TAP networking backend that uses a TAP networking device in the host. For more information about Qemu supported networking backends and different options for configuring networking at Qemu, please refer to:

- <http://www.linux-kvm.org/page/Networking>
- <http://wiki.qemu.org/Documentation/Networking>
- <http://qemu.weilnetz.de/qemu-doc.html>

For example, to start a VM with two emulated `e1000` devices, issue the following command:

```
/usr/local/kvm/bin/qemu-system-x86_64 -cpu host -smp 4 -hda qemu1.raw -m 1024
-net nic,model=e1000,vlan=1,macaddr=DE:AD:1E:00:00:01
-net tap,vlan=1,ifname=tapvm01,script=no,downscript=no
-net nic,model=e1000,vlan=2,macaddr=DE:AD:1E:00:00:02
-net tap,vlan=2,ifname=tapvm02,script=no,downscript=no
```

where:

- `-m` = memory to assign
- `-smp` = number of smp cores
- `-hda` = virtual disk image

This command starts a new virtual machine with two emulated `82540EM` devices, backed up with two TAP networking host interfaces, `tapvm01` and `tapvm02`.

```
# ip tuntap show
tapvm01: tap
tapvm02: tap
```


4. Configure your TAP networking interfaces using ip/ifconfig tools.
5. Log in to the guest OS and check that the expected emulated devices exist:

```
# lspci -d 8086:100e
00:04.0 Ethernet controller: Intel Corporation 82540EM Gigabit Ethernet Controller (rev 03)
00:05.0 Ethernet controller: Intel Corporation 82540EM Gigabit Ethernet Controller (rev 03)
```

6. Install the DPDK and run testpmd.

11.4 Known Limitations of Emulated Devices

The following are known limitations:

1. The Qemu e1000 RX path does not support multiple descriptors/buffers per packet. Therefore, `rte_mbuf` should be big enough to hold the whole packet. For example, to allow testpmd to receive jumbo frames, use the following:
`testpmd [options] --mbuf-size=<your-max-packet-size>`
2. Qemu e1000 does not validate the checksum of incoming packets.
3. Qemu e1000 only supports one interrupt source, so link and Rx interrupt should be exclusive.
4. Qemu e1000 does not support interrupt auto-clear, application should disable interrupt immediately when woken up.

ENA POLL MODE DRIVER

The ENA PMD is a DPDK poll-mode driver for the Amazon Elastic Network Adapter (ENA) family.

12.1 Overview

The ENA driver exposes a lightweight management interface with a minimal set of memory mapped registers and an extendable command set through an Admin Queue.

The driver supports a wide range of ENA adapters, is link-speed independent (i.e., the same driver is used for 10GbE, 25GbE, 40GbE, etc.), and it negotiates and supports an extendable feature set.

ENA adapters allow high speed and low overhead Ethernet traffic processing by providing a dedicated Tx/Rx queue pair per CPU core.

The ENA driver supports industry standard TCP/IP offload features such as checksum offload and TCP transmit segmentation offload (TSO).

Receive-side scaling (RSS) is supported for multi-core scaling.

Some of the ENA devices support a working mode called Low-latency Queue (LLQ), which saves several more microseconds.

12.2 Management Interface

ENA management interface is exposed by means of:

- Device Registers
- Admin Queue (AQ) and Admin Completion Queue (ACQ)

ENA device memory-mapped PCIe space for registers (MMIO registers) are accessed only during driver initialization and are not involved in further normal device operation.

AQ is used for submitting management commands, and the results/responses are reported asynchronously through ACQ.

ENA introduces a very small set of management commands with room for vendor-specific extensions. Most of the management operations are framed in a generic Get/Set feature command.

The following admin queue commands are supported:

- Create I/O submission queue
- Create I/O completion queue
- Destroy I/O submission queue
- Destroy I/O completion queue
- Get feature
- Set feature
- Get statistics

Refer to `ena_admin_defs.h` for the list of supported Get/Set Feature properties.

12.3 Data Path Interface

I/O operations are based on Tx and Rx Submission Queues (Tx SQ and Rx SQ correspondingly). Each SQ has a completion queue (CQ) associated with it.

The SQs and CQs are implemented as descriptor rings in contiguous physical memory.

Refer to `ena_eth_io_defs.h` for the detailed structure of the descriptor

The driver supports multi-queue for both Tx and Rx.

12.4 Configuration information

DPDK Configuration Parameters

The following configuration options are available for the ENA PMD:

- **CONFIG_RTE_LIBRTE_ENA_PMD** (default y): Enables or disables inclusion of the ENA PMD driver in the DPDK compilation.
- **CONFIG_RTE_LIBRTE_ENA_DEBUG_INIT** (default y): Enables or disables debug logging of device initialization within the ENA PMD driver.
- **CONFIG_RTE_LIBRTE_ENA_DEBUG_RX** (default n): Enables or disables debug logging of RX logic within the ENA PMD driver.
- **CONFIG_RTE_LIBRTE_ENA_DEBUG_TX** (default n): Enables or disables debug logging of TX logic within the ENA PMD driver.
- **CONFIG_RTE_LIBRTE_ENA_COM_DEBUG** (default n): Enables or disables debug logging of low level tx/rx logic in `ena_com(base)` within the ENA PMD driver.

ENA Configuration Parameters

- **Number of Queues**

This is the requested number of queues upon initialization, however, the actual number of receive and transmit queues to be created will be the minimum between the maximal number supported by the device and number of queues requested.

- **Size of Queues**

This is the requested size of receive/transmit queues, while the actual size will be the minimum between the requested size and the maximal receive/transmit supported by the device.

12.5 Building DPDK

See the DPDK Getting Started Guide for Linux for instructions on how to build DPDK.

By default the ENA PMD library will be built into the DPDK library.

For configuring and using UIO and VFIO frameworks, please also refer the documentation that comes with DPDK suite.

12.6 Supported ENA adapters

Current ENA PMD supports the following ENA adapters including:

- `1d0f:ec20` - ENA VF
- `1d0f:ec21` - ENA VF with LLQ support

12.7 Supported Operating Systems

Any Linux distribution fulfilling the conditions described in `System Requirements` section of the DPDK documentation or refer to *DPDK Release Notes*.

12.8 Supported features

- Jumbo frames up to 9K
- Port Hardware Statistics
- IPv4/TCP/UDP checksum offload
- TSO offload
- Multiple receive and transmit queues
- RSS
- Low Latency Queue for Tx

12.9 Unsupported features

The features supported by the device and not yet supported by this PMD include:

- Asynchronous Event Notification Queue (AENQ)

12.10 Prerequisites

1. Prepare the system as recommended by DPDK suite. This includes environment variables, hugepages configuration, tool-chains and configuration
2. Insert `igb_uio` kernel module using the command `'modprobe igb_uio'`
3. Bind the intended ENA device to `igb_uio` module

At this point the system should be ready to run DPDK applications. Once the application runs to completion, the ENA can be detached from `igb_uio` if necessary.

12.11 Usage example

Follow instructions available in the document *compiling and testing a PMD for a NIC* to launch **testpmd** with Amazon ENA devices managed by `librte_pmd_ena`.

Example output:

```
[...]  
EAL: PCI device 0000:02:00.1 on NUMA socket -1  
EAL: probe driver: 1d0f:ec20 rte_ena_pmd  
EAL: PCI memory mapped at 0x7f9b6c400000  
PMD: eth_ena_dev_init(): Initializing 0:2:0.1  
Interactive-mode selected  
Configuring Port 0 (socket 0)  
Port 0: 00:00:00:11:00:01  
Checking link statuses...  
Port 0 Link Up - speed 10000 Mbps - full-duplex  
Done  
testpmd>
```

ENIC POLL MODE DRIVER

ENIC PMD is the DPDK poll-mode driver for the Cisco System Inc. VIC Ethernet NICs. These adapters are also referred to as vNICs below. If you are running or would like to run DPDK software applications on Cisco UCS servers using Cisco VIC adapters the following documentation is relevant.

13.1 How to obtain ENIC PMD integrated DPDK

ENIC PMD support is integrated into the DPDK suite. `dpdk-<version>.tar.gz` should be downloaded from <http://dpdk.org>

13.2 Configuration information

- **DPDK Configuration Parameters**

The following configuration options are available for the ENIC PMD:

- **CONFIG RTE LIBRTE ENIC PMD** (default y): Enables or disables inclusion of the ENIC PMD driver in the DPDK compilation.

- **vNIC Configuration Parameters**

- **Number of Queues**

The maximum number of receive queues (RQs), work queues (WQs) and completion queues (CQs) are configurable on a per vNIC basis through the Cisco UCS Manager (CIMC or UCSM).

These values should be configured as follows:

- * The number of WQs should be greater or equal to the value of the expected `nb_tx_q` parameter in the call to `rte_eth_dev_configure()`
- * The number of RQs configured in the vNIC should be greater or equal to *twice* the value of the expected `nb_rx_q` parameter in the call to `rte_eth_dev_configure()`. With the addition of Rx scatter, a pair of RQs on the vnic is needed for each receive queue used by DPDK, even if Rx scatter is not being used. Having a vNIC with only 1 RQ is not a valid configuration, and will fail with an error message.
- * The number of CQs should set so that there is one CQ for each WQ, and one CQ for each pair of RQs.

For example: If the application requires 3 Rx queues, and 3 Tx queues, the vNIC should be configured to have at least 3 WQs, 6 RQs (3 pairs), and 6 CQs (3 for use by WQs + 3 for use by the 3 pairs of RQs).

– Size of Queues

Likewise, the number of receive and transmit descriptors are configurable on a per-vNIC basis via the UCS Manager and should be greater than or equal to the `nb_rx_desc` and `nb_tx_desc` parameters expected to be used in the calls to `rte_eth_rx_queue_setup()` and `rte_eth_tx_queue_setup()` respectively. An application requesting more than the set size will be limited to that size.

Unless there is a lack of resources due to creating many vNICs, it is recommended that the WQ and RQ sizes be set to the maximum. This gives the application the greatest amount of flexibility in its queue configuration.

- * *Note:* Since the introduction of Rx scatter, for performance reasons, this PMD uses two RQs on the vNIC per receive queue in DPDK. One RQ holds descriptors for the start of a packet, and the second RQ holds the descriptors for the rest of the fragments of a packet. This means that the `nb_rx_desc` parameter to `rte_eth_rx_queue_setup()` can be a greater than 4096. The exact amount will depend on the size of the mbufs being used for receives, and the MTU size.

For example: If the mbuf size is 2048, and the MTU is 9000, then receiving a full size packet will take 5 descriptors, 1 from the start-of-packet queue, and 4 from the second queue. Assuming that the RQ size was set to the maximum of 4096, then the application can specify up to 1024 + 4096 as the `nb_rx_desc` parameter to `rte_eth_rx_queue_setup()`.

– Interrupts

Only one interrupt per vNIC interface should be configured in the UCS manager regardless of the number receive/transmit queues. The ENIC PMD uses this interrupt to get information about link status and errors in the fast path.

13.3 Flow director support

Advanced filtering support was added to 1300 series VIC firmware starting with version 2.0.13 for C-series UCS servers and version 3.1.2 for UCSM managed blade servers. In order to enable advanced filtering the 'Advanced filter' radio button should be enabled via CIMC or UCSM followed by a reboot of the server.

With advanced filters, perfect matching of all fields of IPv4, IPv6 headers as well as TCP, UDP and SCTP L4 headers is available through flow director. Masking of these fields for partial match is also supported.

Without advanced filter support, the flow director is limited to IPv4 perfect filtering of the 5-tuple with no masking of fields supported.

13.4 SR-IOV mode utilization

UCS blade servers configured with dynamic vNIC connection policies in UCS manager are capable of supporting assigned devices on virtual machines (VMs) through a KVM hypervisor.

Assigned devices, also known as ‘passthrough’ devices, are SR-IOV virtual functions (VFs) on the host which are exposed to VM instances.

The Cisco Virtual Machine Fabric Extender (VM-FEX) gives the VM a dedicated interface on the Fabric Interconnect (FI). Layer 2 switching is done at the FI. This may eliminate the requirement for software switching on the host to route intra-host VM traffic.

Please refer to [Creating a Dynamic vNIC Connection Policy](#) for information on configuring SR-IOV adapter policies using UCS manager.

Once the policies are in place and the host OS is rebooted, VFs should be visible on the host, E.g.:

```
# lspci | grep Cisco | grep Ethernet
0d:00.0 Ethernet controller: Cisco Systems Inc VIC Ethernet NIC (rev a2)
0d:00.1 Ethernet controller: Cisco Systems Inc VIC SR-IOV VF (rev a2)
0d:00.2 Ethernet controller: Cisco Systems Inc VIC SR-IOV VF (rev a2)
0d:00.3 Ethernet controller: Cisco Systems Inc VIC SR-IOV VF (rev a2)
0d:00.4 Ethernet controller: Cisco Systems Inc VIC SR-IOV VF (rev a2)
0d:00.5 Ethernet controller: Cisco Systems Inc VIC SR-IOV VF (rev a2)
0d:00.6 Ethernet controller: Cisco Systems Inc VIC SR-IOV VF (rev a2)
0d:00.7 Ethernet controller: Cisco Systems Inc VIC SR-IOV VF (rev a2)
```

Enable Intel IOMMU on the host and install KVM and libvirt. A VM instance should be created with an assigned device. When using libvirt, this configuration can be done within the domain (i.e. VM) config file. For example this entry maps host VF 0d:00:01 into the VM.

```
<interface type='hostdev' managed='yes'>
  <mac address='52:54:00:ac:ff:b6' />
  <source>
    <address type='pci' domain='0x0000' bus='0x0d' slot='0x00' function='0x1' />
  </source>
```

Alternatively, the configuration can be done in a separate file using the `network` keyword. These methods are described in the libvirt documentation for [Network XML format](#).

When the VM instance is started, the ENIC KVM driver will bind the host VF to `vfiio`, complete provisioning on the FI and bring up the link.

Note: It is not possible to use a VF directly from the host because it is not fully provisioned until the hypervisor brings up the VM that it is assigned to.

In the VM instance, the VF will now be visible. E.g., here the VF 00:04.0 is seen on the VM instance and should be available for binding to a DPDK.

```
# lspci | grep Ether
00:04.0 Ethernet controller: Cisco Systems Inc VIC SR-IOV VF (rev a2)
```

Follow the normal DPDK install procedure, binding the VF to either `igb_uio` or `vfiio` in non-IOMMU mode.

Please see [Limitations](#) for limitations in the use of SR-IOV.

13.5 Generic Flow API support

Generic Flow API is supported. The baseline support is:

- 1200 series VICs

5-tuple exact flow support for 1200 series adapters. This allows:

- Attributes: ingress
- Items: ipv4, ipv6, udp, tcp (must exactly match src/dst IP addresses and ports and all must be specified)
- Actions: queue and void
- Selectors: ‘is’

- **1300 series VICS with advanced filters disabled**

With advanced filters disabled, an IPv4 or IPv6 item must be specified in the pattern.

- Attributes: ingress
- Items: eth, ipv4, ipv6, udp, tcp, vxlan, inner eth, ipv4, ipv6, udp, tcp
- Actions: queue and void
- Selectors: ‘is’, ‘spec’ and ‘mask’. ‘last’ is not supported
- In total, up to 64 bytes of mask is allowed across all headers

- **1300 series VICS with advanced filters enabled**

- Attributes: ingress
- Items: eth, ipv4, ipv6, udp, tcp, vxlan, inner eth, ipv4, ipv6, udp, tcp
- Actions: queue, mark, flag and void
- Selectors: ‘is’, ‘spec’ and ‘mask’. ‘last’ is not supported
- In total, up to 64 bytes of mask is allowed across all headers

More features may be added in future firmware and new versions of the VIC. Please refer to the release notes.

13.6 Limitations

- **VLAN 0 Priority Tagging**

If a vNIC is configured in TRUNK mode by the UCS manager, the adapter will priority tag egress packets according to 802.1Q if they were not already VLAN tagged by software. If the adapter is connected to a properly configured switch, there will be no unexpected behavior.

In test setups where an Ethernet port of a Cisco adapter in TRUNK mode is connected point-to-point to another adapter port or connected through a router instead of a switch, all ingress packets will be VLAN tagged. Programs such as I3fwd which do not account for VLAN tags in packets will misbehave. The solution is to enable VLAN stripping on ingress. The following code fragment is an example of how to accomplish this:

```
vlan_offload = rte_eth_dev_get_vlan_offload(port);
vlan_offload |= ETH_VLAN_STRIP_OFFLOAD;
rte_eth_dev_set_vlan_offload(port, vlan_offload);
```

- Limited flow director support on 1200 series and 1300 series Cisco VIC adapters with old firmware. Please see [Flow director support](#).

- Flow director features are not supported on generation 1 Cisco VIC adapters (M81KR and P81E)
- **SR-IOV**
 - KVM hypervisor support only. VMware has not been tested.
 - Requires VM-FEX, and so is only available on UCS managed servers connected to Fabric Interconnects. It is not on standalone C-Series servers.
 - VF devices are not usable directly from the host. They can only be used as assigned devices on VM instances.
 - Currently, unbind of the ENIC kernel mode driver 'enic.ko' on the VM instance may hang. As a workaround, enic.ko should be blacklisted or removed from the boot process.
 - pci_generic cannot be used as the uio module in the VM. igb_uio or vfio in non-IOMMU mode can be used.
 - The number of RQs in UCSM dynamic vNIC configurations must be at least 2.
 - The number of SR-IOV devices is limited to 256. Components on target system might limit this number to fewer than 256.
- **Flow API**
 - The number of filters that can be specified with the Generic Flow API is dependent on how many header fields are being masked. Use 'flow create' in a loop to determine how many filters your VIC will support (not more than 1000 for 1300 series VICs). Filters are checked for matching in the order they were added. Since there currently is no grouping or priority support, 'catch-all' filters should be added last.

13.7 How to build the suite

The build instructions for the DPDK suite should be followed. By default the ENIC PMD library will be built into the DPDK library.

Refer to the document [compiling and testing a PMD for a NIC](#) for details.

For configuring and using UIO and VFIO frameworks, please refer to the documentation that comes with DPDK suite.

13.8 Supported Cisco VIC adapters

ENIC PMD supports all recent generations of Cisco VIC adapters including:

- VIC 1280
- VIC 1240
- VIC 1225
- VIC 1285
- VIC 1225T
- VIC 1227

- VIC 1227T
- VIC 1380
- VIC 1340
- VIC 1385
- VIC 1387

13.9 Supported Operating Systems

Any Linux distribution fulfilling the conditions described in Dependencies section of DPDK documentation.

13.10 Supported features

- Unicast, multicast and broadcast transmission and reception
- Receive queue polling
- Port Hardware Statistics
- Hardware VLAN acceleration
- IP checksum offload
- Receive side VLAN stripping
- Multiple receive and transmit queues
- Flow Director ADD, UPDATE, DELETE, STATS operation support IPv4 and IPv6
- Promiscuous mode
- Setting RX VLAN (supported via UCSM/CIMC only)
- VLAN filtering (supported via UCSM/CIMC only)
- Execution of application by unprivileged system users
- IPV4, IPV6 and TCP RSS hashing
- Scattered Rx
- MTU update
- SR-IOV on UCS managed servers connected to Fabric Interconnects
- Flow API

13.11 Known bugs and unsupported features in this release

- Signature or flex byte based flow direction
- Drop feature of flow direction
- VLAN based flow direction

- Non-IPV4 flow direction
- Setting of extended VLAN
- UDP RSS hashing
- MTU update only works if Scattered Rx mode is disabled

13.12 Prerequisites

- Prepare the system as recommended by DPDK suite. This includes environment variables, hugepages configuration, tool-chains and configuration.
- Insert vfio-pci kernel module using the command 'modprobe vfio-pci' if the user wants to use VFIO framework.
- Insert uio kernel module using the command 'modprobe uio' if the user wants to use UIO framework.
- DPDK suite should be configured based on the user's decision to use VFIO or UIO framework.
- If the vNIC device(s) to be used is bound to the kernel mode Ethernet driver use 'ip' to bring the interface down. The dpdk-devbind.py tool can then be used to unbind the device's bus id from the ENIC kernel mode driver.
- Bind the intended vNIC to vfio-pci in case the user wants ENIC PMD to use VFIO framework using dpdk-devbind.py.
- Bind the intended vNIC to igb_uio in case the user wants ENIC PMD to use UIO framework using dpdk-devbind.py.

At this point the system should be ready to run DPDK applications. Once the application runs to completion, the vNIC can be detached from vfio-pci or igb_uio if necessary.

Root privilege is required to bind and unbind vNICs to/from VFIO/UIO. VFIO framework helps an unprivileged user to run the applications. For an unprivileged user to run the applications on DPDK and ENIC PMD, it may be necessary to increase the maximum locked memory of the user. The following command could be used to do this.

```
sudo sh -c "ulimit -l <value in Kilo Bytes>"
```

The value depends on the memory configuration of the application, DPDK and PMD. Typically, the limit has to be raised to higher than 2GB. e.g., 2621440

The compilation of any unused drivers can be disabled using the configuration file in config/ directory (e.g., config/common_linuxapp). This would help in bringing down the time taken for building the libraries and the initialization time of the application.

13.13 Additional Reference

- <https://www.cisco.com/c/en/us/products/servers-unified-computing/index.html>
- <https://www.cisco.com/c/en/us/products/interfaces-modules/unified-computing-system-adapters/index.html>

13.14 Contact Information

Any questions or bugs should be reported to DPDK community and to the ENIC PMD maintainers:

- John Daley <johndale@cisco.com>
- Nelson Escobar <neescoba@cisco.com>

FM10K POLL MODE DRIVER

The FM10K poll mode driver library provides support for the Intel FM10000 (FM10K) family of 40GbE/100GbE adapters.

14.1 FTAG Based Forwarding of FM10K

FTAG Based Forwarding is a unique feature of FM10K. The FM10K family of NICs support the addition of a Fabric Tag (FTAG) to carry special information. The FTAG is placed at the beginning of the frame, it contains information such as where the packet comes from and goes, and the vlan tag. In FTAG based forwarding mode, the switch logic forwards packets according to glort (global resource tag) information, rather than the mac and vlan table. Currently this feature works only on PF.

To enable this feature, the user should pass a devargs parameter to the eal like “-w 84:00.0,enable_ftag=1”, and the application should make sure an appropriate FTAG is inserted for every frame on TX side.

14.2 Vector PMD for FM10K

Vector PMD (vPMD) uses Intel® SIMD instructions to optimize packet I/O. It improves load/store bandwidth efficiency of L1 data cache by using a wider SSE/AVX “register (1)”. The wider register gives space to hold multiple packet buffers so as to save on the number of instructions when bulk processing packets.

There is no change to the PMD API. The RX/TX handlers are the only two entries for vPMD packet I/O. They are transparently registered at runtime RX/TX execution if all required conditions are met.

1. To date, only an SSE version of FM10K vPMD is available. To ensure that vPMD is in the binary code, set `CONFIG_RTE_LIBRTE_FM10K_INC_VECTOR=y` in the configure file.

Some constraints apply as pre-conditions for specific optimizations on bulk packet transfers. The following sections explain RX and TX constraints in the vPMD.

14.2.1 RX Constraints

Prerequisites and Pre-conditions

For Vector RX it is assumed that the number of descriptor rings will be a power of 2. With this pre-condition, the ring pointer can easily scroll back to the head after hitting the tail without a conditional check. In addition Vector RX can use this assumption to do a bit mask using `ring_size - 1`.

Features not Supported by Vector RX PMD

Some features are not supported when trying to increase the throughput in vPMD. They are:

- IEEE1588
- Flow director
- Header split
- RX checksum offload

Other features are supported using optional MACRO configuration. They include:

- HW VLAN strip
- L3/L4 packet type

To enable via `RX_OLFLAGS` use `RTE_LIBRTE_FM10K_RX_OLFLAGS_ENABLE=y`.

To guarantee the constraint, the following configuration flags in `dev_conf.rxmode` will be checked:

- `hw_vlan_extend`
- `hw_ip_checksum`
- `header_split`
- `fdir_conf->mode`

RX Burst Size

As vPMD is focused on high throughput, it processes 4 packets at a time. So it assumes that the RX burst should be greater than 4 packets per burst. It returns zero if using `nb_pkt < 4` in the receive handler. If `nb_pkt` is not a multiple of 4, a floor alignment will be applied.

14.2.2 TX Constraint

Features not Supported by TX Vector PMD

TX vPMD only works when `txq_flags` is set to `FM10K_SIMPLE_TX_FLAG`. This means that it does not support TX multi-segment, VLAN offload or TX csum offload. The following MACROS are used for these three features:

- `ETH_TXQ_FLAGS_NOMULTSEGS`
- `ETH_TXQ_FLAGS_NOVLANOFFL`

- `ETH_TXQ_FLAGS_NOXSUMSCTP`
- `ETH_TXQ_FLAGS_NOXSUMUDP`
- `ETH_TXQ_FLAGS_NOXSUMTCP`

14.3 Limitations

14.3.1 Switch manager

The Intel FM10000 family of NICs integrate a hardware switch and multiple host interfaces. The FM10000 PMD driver only manages host interfaces. For the switch component another switch driver has to be loaded prior to the FM10000 PMD driver. The switch driver can be acquired from Intel support. Only Testpoint is validated with DPDK, the latest version that has been validated with DPDK is 4.1.6.

14.3.2 Support for Switch Restart

For FM10000 multi host based design a DPDK app running in the VM or host needs to be aware of the switch's state since it may undergo a quit-restart. When the switch goes down the DPDK app will receive a LSC event indicating link status down, and the app should stop the worker threads that are polling on the Rx/Tx queues. When switch comes up, a LSC event indicating `LINK_UP` is sent to the app, which can then restart the FM10000 port to resume network processing.

14.3.3 CRC striping

The FM10000 family of NICs strip the CRC for every packets coming into the host interface. So, CRC will be stripped even when the `rxmode.hw_strip_crc` member is set to 0 in `struct rte_eth_conf`.

14.3.4 Maximum packet length

The FM10000 family of NICS support a maximum of a 15K jumbo frame. The value is fixed and cannot be changed. So, even when the `rxmode.max_rx_pkt_len` member of `struct rte_eth_conf` is set to a value lower than 15364, frames up to 15364 bytes can still reach the host interface.

14.3.5 Statistic Polling Frequency

The FM10000 NICs expose a set of statistics via the PCI BARs. These statistics are read from the hardware registers when `rte_eth_stats_get()` or `rte_eth_xstats_get()` is called. The packet counting registers are 32 bits while the byte counting registers are 48 bits. As a result, the statistics must be polled regularly in order to ensure the consistency of the returned reads.

Given the PCIe Gen3 x8, about 50Gbps of traffic can occur. With 64 byte packets this gives almost 100 million packets/second, causing 32 bit integer overflow after approx 40 seconds. To ensure these overflows are detected and accounted for in the statistics, it is necessary to

read statistic regularly. It is suggested to read stats every 20 seconds, which will ensure the statistics are accurate.

14.3.6 Interrupt mode

The FM10000 family of NICS need one separate interrupt for mailbox. So only drivers which support multiple interrupt vectors e.g. vfiopci can work for fm10k interrupt mode.

I40E POLL MODE DRIVER

The I40E PMD (`librte_pmd_i40e`) provides poll mode driver support for the Intel X710/XL710/X722 10/40 Gbps family of adapters.

15.1 Features

Features of the I40E PMD are:

- Multiple queues for TX and RX
- Receiver Side Scaling (RSS)
- MAC/VLAN filtering
- Packet type information
- Flow director
- Cloud filter
- Checksum offload
- VLAN/QinQ stripping and inserting
- TSO offload
- Promiscuous mode
- Multicast mode
- Port hardware statistics
- Jumbo frames
- Link state information
- Link flow control
- Mirror on port, VLAN and VSI
- Interrupt mode for RX
- Scattered and gather for TX and RX
- Vector Poll mode driver
- DCB
- VMDQ

- SR-IOV VF
- Hot plug
- IEEE1588/802.1AS timestamping
- VF Daemon (VFD) - EXPERIMENTAL
- Dynamic Device Personalization (DDP)
- Queue region configuration

15.2 Prerequisites

- Identifying your adapter using [Intel Support](#) and get the latest NVM/FW images.
- Follow the DPDK Getting Started Guide for Linux to setup the basic DPDK environment.
- To get better performance on Intel platforms, please follow the “How to get best performance with NICs on Intel platforms” section of the Getting Started Guide for Linux.
- Upgrade the NVM/FW version following the [Intel® Ethernet NVM Update Tool Quick Usage Guide for Linux](#) if needed.

15.3 Pre-Installation Configuration

15.3.1 Config File Options

The following options can be modified in the `config` file. Please note that enabling debugging options may affect system performance.

- `CONFIG_RTE_LIBRTE_I40E_PMD` (default `y`)
Toggle compilation of the `librte_pmd_i40e` driver.
- `CONFIG_RTE_LIBRTE_I40E_DEBUG_*` (default `n`)
Toggle display of generic debugging messages.
- `CONFIG_RTE_LIBRTE_I40E_RX_ALLOW_BULK_ALLOC` (default `y`)
Toggle bulk allocation for RX.
- `CONFIG_RTE_LIBRTE_I40E_INC_VECTOR` (default `n`)
Toggle the use of Vector PMD instead of normal RX/TX path. To enable vPMD for RX, bulk allocation for Rx must be allowed.
- `CONFIG_RTE_LIBRTE_I40E_16BYTE_RX_DESC` (default `n`)
Toggle to use a 16-byte RX descriptor, by default the RX descriptor is 32 byte.
- `CONFIG_RTE_LIBRTE_I40E_QUEUE_NUM_PER_PF` (default 64)
Number of queues reserved for PF.
- `CONFIG_RTE_LIBRTE_I40E_QUEUE_NUM_PER_VM` (default 4)
Number of queues reserved for each VMDQ Pool.

- `CONFIG_RTE_LIBRTE_I40E_ITR_INTERVAL` (default -1)
Interrupt Throttling interval.

15.3.2 Runtime Config Options

- Number of Queues per VF (default 4)

The number of queue per VF is determined by its host PF. If the PCI address of an i40e PF is `aaaa:bb.cc`, the number of queues per VF can be configured with EAL parameter like `-w aaaa:bb.cc,queue-num-per-vf=n`. The value `n` can be 1, 2, 4, 8 or 16. If no such parameter is configured, the number of queues per VF is 4 by default.

- Support multiple driver (default disable)

There was a multiple driver support issue during use of 700 series Ethernet Adapter with both Linux kernel and DPDK PMD. To fix this issue, `devargs` parameter `support-multi-driver` is introduced, for example:

```
-w 84:00.0,support-multi-driver=1
```

With the above configuration, DPDK PMD will not change global registers, and will switch PF interrupt from `IntN` to `Int0` to avoid interrupt conflict between DPDK and Linux Kernel.

15.4 Driver compilation and testing

Refer to the document *compiling and testing a PMD for a NIC* for details.

15.5 SR-IOV: Prerequisites and sample Application Notes

1. Load the kernel module:

```
modprobe i40e
```

Check the output in `dmesg`:

```
i40e 0000:83:00.1 ens802f0: renamed from eth0
```

2. Bring up the PF ports:

```
ifconfig ens802f0 up
```

3. Create VF device(s):

Echo the number of VFs to be created into the `sriov_numvfs` sysfs entry of the parent PF.

Example:

```
echo 2 > /sys/devices/pci0000:00/0000:00:03.0/0000:81:00.0/sriov_numvfs
```

4. Assign VF MAC address:

Assign MAC address to the VF using `iproute2` utility. The syntax is:

```
ip link set <PF netdev id> vf <VF id> mac <macaddr>
```

Example:

```
ip link set ens802f0 vf 0 mac a0:b0:c0:d0:e0:f0
```

5. Assign VF to VM, and bring up the VM. Please see the documentation for the *I40E/IXGBE/IGB Virtual Function Driver*.

6. Running testpmd:

Follow instructions available in the document *compiling and testing a PMD for a NIC* to run testpmd.

Example output:

```
...
EAL: PCI device 0000:83:00.0 on NUMA socket 1
EAL: probe driver: 8086:1572 rte_i40e_pmd
EAL: PCI memory mapped at 0x7f7f80000000
EAL: PCI memory mapped at 0x7f7f80800000
PMD: eth_i40e_dev_init(): FW 5.0 API 1.5 NVM 05.00.02 eetrack 8000208a
Interactive-mode selected
Configuring Port 0 (socket 0)
...

PMD: i40e_dev_rx_queue_setup(): Rx Burst Bulk Alloc Preconditions are
satisfied.Rx Burst Bulk Alloc function will be used on port=0, queue=0.

...
Port 0: 68:05:CA:26:85:84
Checking link statuses...
Port 0 Link Up - speed 10000 Mbps - full-duplex
Done

testpmd>
```

15.6 Sample Application Notes

15.6.1 Vlan filter

Vlan filter only works when Promiscuous mode is off.

To start testpmd, and add vlan 10 to port 0:

```
./app/testpmd -l 0-15 -n 4 -- -i --forward-mode=mac
...

testpmd> set promisc 0 off
testpmd> rx_vlan add 10 0
```

15.6.2 Flow Director

The Flow Director works in receive mode to identify specific flows or sets of flows and route them to specific queues. The Flow Director filters can match the different fields for different type of packet: flow type, specific input set per flow type and the flexible payload.

The default input set of each flow type is:

```
ipv4-other : src_ip_address, dst_ip_address
ipv4-frag  : src_ip_address, dst_ip_address
ipv4-tcp   : src_ip_address, dst_ip_address, src_port, dst_port
ipv4-udp   : src_ip_address, dst_ip_address, src_port, dst_port
```

```

ipv4-sctp : src_ip_address, dst_ip_address, src_port, dst_port,
            verification_tag
ipv6-other : src_ip_address, dst_ip_address
ipv6-frag : src_ip_address, dst_ip_address
ipv6-tcp : src_ip_address, dst_ip_address, src_port, dst_port
ipv6-udp : src_ip_address, dst_ip_address, src_port, dst_port
ipv6-sctp : src_ip_address, dst_ip_address, src_port, dst_port,
            verification_tag
l2_payload : ether_type

```

The flex payload is selected from offset 0 to 15 of packet's payload by default, while it is masked out from matching.

Start testpmd with `--disable-rss` and `--pkt-filter-mode=perfect`:

```

./app/testpmd -l 0-15 -n 4 -- -i --disable-rss --pkt-filter-mode=perfect \
              --rxq=8 --txq=8 --nb-cores=8 --nb-ports=1

```

Add a rule to direct ipv4-udp packet whose `dst_ip=2.2.2.5`, `src_ip=2.2.2.3`, `src_port=32`, `dst_port=32` to queue 1:

```

testpmd> flow_director_filter 0 mode IP add flow ipv4-udp \
        src 2.2.2.3 32 dst 2.2.2.5 32 vlan 0 flexbytes () \
        fwd pf queue 1 fd_id 1

```

Check the flow director status:

```

testpmd> show port fdir 0

```

```

##### FDIR infos for port 0 #####
MODE: PERFECT
SUPPORTED FLOW TYPE:  ipv4-frag ipv4-tcp ipv4-udp ipv4-sctp ipv4-other
                    ipv6-frag ipv6-tcp ipv6-udp ipv6-sctp ipv6-other
                    l2_payload

FLEX PAYLOAD INFO:
max_len:      16          payload_limit: 480
payload_unit: 2          payload_seg:   3
bitmask_unit: 2          bitmask_num:  2
MASK:
  vlan_tci: 0x0000,
  src_ipv4: 0x00000000,
  dst_ipv4: 0x00000000,
  src_port: 0x0000,
  dst_port: 0x0000
  src_ipv6: 0x00000000,0x00000000,0x00000000,0x00000000,
  dst_ipv6: 0x00000000,0x00000000,0x00000000,0x00000000
FLEX PAYLOAD SRC OFFSET:
L2_PAYLOAD:  0    1    2    3    4    5    6  ...
L3_PAYLOAD:  0    1    2    3    4    5    6  ...
L4_PAYLOAD:  0    1    2    3    4    5    6  ...
FLEX MASK CFG:
  ipv4-udp:  00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
  ipv4-tcp:  00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
  ipv4-sctp: 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
  ipv4-other:00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
  ipv4-frag: 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
  ipv6-udp:  00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
  ipv6-tcp:  00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
  ipv6-sctp: 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
  ipv6-other:00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
  ipv6-frag: 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
  l2_payload:00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
guarant_count: 1          best_count:   0
guarant_space: 512       best_space:  7168

```

```

collision:    0          free:          0
maxhash:     0          maxlen:       0
add:         0          remove:       0
f_add:       0          f_remove:     0

```

Delete all flow director rules on a port:

```
testpmd> flush_flow_director 0
```

15.6.3 Floating VEB

The Intel® Ethernet Controller X710 and XL710 Family support a feature called “Floating VEB”.

A Virtual Ethernet Bridge (VEB) is an IEEE Edge Virtual Bridging (EVB) term for functionality that allows local switching between virtual endpoints within a physical endpoint and also with an external bridge/network.

A “Floating” VEB doesn’t have an uplink connection to the outside world so all switching is done internally and remains within the host. As such, this feature provides security benefits.

In addition, a Floating VEB overcomes a limitation of normal VEBs where they cannot forward packets when the physical link is down. Floating VEBs don’t need to connect to the NIC port so they can still forward traffic from VF to VF even when the physical link is down.

Therefore, with this feature enabled VFs can be limited to communicating with each other but not an outside network, and they can do so even when there is no physical uplink on the associated NIC port.

To enable this feature, the user should pass a `devargs` parameter to the EAL, for example:

```
-w 84:00.0,enable_floating_veb=1
```

In this configuration the PMD will use the floating VEB feature for all the VFs created by this PF device.

Alternatively, the user can specify which VFs need to connect to this floating VEB using the `floating_veb_list` argument:

```
-w 84:00.0,enable_floating_veb=1,floatig_veb_list=1;3-4
```

In this example VF1, VF3 and VF4 connect to the floating VEB, while other VFs connect to the normal VEB.

The current implementation only supports one floating VEB and one regular VEB. VFs can connect to a floating VEB or a regular VEB according to the configuration passed on the EAL command line.

The floating VEB functionality requires a NIC firmware version of 5.0 or greater.

15.6.4 Dynamic Device Personalization (DDP)

The Intel® Ethernet Controller X*710 support a feature called “Dynamic Device Personalization (DDP)”, which is used to configure hardware by downloading a profile to support protocols/filters which are not supported by default. The DDP functionality requires a NIC firmware version of 6.0 or greater.

Current implementation supports MPLSoUDP/MPLSoGRE/GTP-C/GTP-U/PPPoE/PPPoL2TP, steering can be used with `rte_flow` API.

Load a profile which supports MPLSoUDP/MPLSoGRE and store backup profile:

```
testpmd> ddp add 0 ./mpls.pkgo,./backup.pkgo
```

Delete a MPLS profile and restore backup profile:

```
testpmd> ddp del 0 ./backup.pkgo
```

Get loaded DDP package info list:

```
testpmd> ddp get list 0
```

Display information about a MPLS profile:

```
testpmd> ddp get info ./mpls.pkgo
```

15.6.5 Input set configuration

Input set for any PCTYPE can be configured with user defined configuration, For example, to use only 48bit prefix for IPv6 src address for IPv6 TCP RSS:

```
testpmd> port config 0 pctype 43 hash_inset clear all
testpmd> port config 0 pctype 43 hash_inset set field 13
testpmd> port config 0 pctype 43 hash_inset set field 14
testpmd> port config 0 pctype 43 hash_inset set field 15
```

15.6.6 Queue region configuration

The Ethernet Controller X710/XL710 supports a feature of queue regions configuration for RSS in the PF, so that different traffic classes or different packet classification types can be separated to different queues in different queue regions. There is an API for configuration of queue regions in RSS with a command line. It can parse the parameters of the region index, queue number, queue start index, user priority, traffic classes and so on. Depending on commands from the command line, it will call i40e private APIs and start the process of setting or flushing the queue region configuration. As this feature is specific for i40e only private APIs are used. These new `test_pmd` commands are as shown below. For details please refer to `../testpmd_app_ug/index`.

```
testpmd> set port (port_id) queue-region region_id (value) \
    queue_start_index (value) queue_num (value)
testpmd> set port (port_id) queue-region region_id (value) flowtype (value)
testpmd> set port (port_id) queue-region UP (value) region_id (value)
testpmd> set port (port_id) queue-region flush (on|off)
testpmd> show port (port_id) queue-region
```

15.7 Limitations or Known issues

15.7.1 MPLS packet classification on X710/XL710

For firmware versions prior to 5.0, MPLS packets are not recognized by the NIC. The L2 Payload flow type in flow director can be used to classify MPLS packet by using a command in `testpmd` like:

```
testpmd> flow_director_filter 0 mode IP add flow l2_payload ether 0x8847
flexbytes () fwd pf queue <N> fd_id <M>
```


With the NIC firmware version 5.0 or greater, some limited MPLS support is added: Native MPLS (MPLS in Ethernet) skip is implemented, while no new packet type, no classification or offload are possible. With this change, L2 Payload flow type in flow director cannot be used to classify MPLS packet as with previous firmware versions. Meanwhile, the Ethertype filter can be used to classify MPLS packet by using a command in testpmd like:

```
testpmd> ethertype_filter 0 add mac_ignr 00:00:00:00:00:00 ethertype
0x8847 fwd queue <M>
```

15.7.2 16 Byte RX Descriptor setting on DPDK VF

Currently the VF's RX descriptor mode is decided by PF. There's no PF-VF interface for VF to request the RX descriptor mode, also no interface to notify VF its own RX descriptor mode. For all available versions of the i40e driver, these drivers don't support 16 byte RX descriptor. If the Linux i40e kernel driver is used as host driver, while DPDK i40e PMD is used as the VF driver, DPDK cannot choose 16 byte receive descriptor. The reason is that the RX descriptor is already set to 32 byte by the i40e kernel driver. That is to say, user should keep `CONFIG_RTE_LIBRTE_I40E_16BYTE_RX_DESC=n` in config file. In the future, if the Linux i40e driver supports 16 byte RX descriptor, user should make sure the DPDK VF uses the same RX descriptor mode, 16 byte or 32 byte, as the PF driver.

The same rule for DPDK PF + DPDK VF. The PF and VF should use the same RX descriptor mode. Or the VF RX will not work.

15.7.3 Receive packets with Ethertype 0x88A8

Due to the FW limitation, PF can receive packets with Ethertype 0x88A8 only when floating VEB is disabled.

15.7.4 Incorrect Rx statistics when packet is oversize

When a packet is over maximum frame size, the packet is dropped. However the Rx statistics, when calling `rte_eth_stats_get` incorrectly shows it as received.

15.7.5 VF & TC max bandwidth setting

The per VF max bandwidth and per TC max bandwidth cannot be enabled in parallel. The behavior is different when handling per VF and per TC max bandwidth setting. When enabling per VF max bandwidth, SW will check if per TC max bandwidth is enabled. If so, return failure. When enabling per TC max bandwidth, SW will check if per VF max bandwidth is enabled. If so, disable per VF max bandwidth and continue with per TC max bandwidth setting.

15.7.6 TC TX scheduling mode setting

There're 2 TX scheduling modes for TCs, round robin and strict priority mode. If a TC is set to strict priority mode, it can consume unlimited bandwidth. It means if APP has set the max bandwidth for that TC, it comes to no effect. It's suggested to set the strict priority mode for a TC that is latency sensitive but no consuming much bandwidth.

15.7.7 VF performance is impacted by PCI extended tag setting

To reach maximum NIC performance in the VF the PCI extended tag must be enabled. The DPDK I40E PF driver will set this feature during initialization, but the kernel PF driver does not. So when running traffic on a VF which is managed by the kernel PF driver, a significant NIC performance downgrade has been observed (for 64 byte packets, there is about 25% linerate downgrade for a 25G device and about 35% for a 40G device).

For kernel version ≥ 4.11 , the kernel's PCI driver will enable the extended tag if it detects that the device supports it. So by default, this is not an issue. For kernels ≤ 4.11 or when the PCI extended tag is disabled it can be enabled using the steps below.

1. Get the current value of the PCI configure register:

```
setpci -s <XX:XX.X> a8.w
```

2. Set bit 8:

```
value = value | 0x100
```

3. Set the PCI configure register with new value:

```
setpci -s <XX:XX.X> a8.w=<value>
```

15.7.8 Vlan strip of VF

The VF vlan strip function is only supported in the i40e kernel driver $\geq 2.1.26$.

15.7.9 DCB function

DCB works only when RSS is enabled.

15.7.10 Global configuration warning

I40E PMD will set some global registers to enable some function or set some configure. Then when using different ports of the same NIC with Linux kernel and DPDK, the port with Linux kernel will be impacted by the port with DPDK. For example, register I40E_GL_SWT_L2TAGCTRL is used to control L2 tag, i40e PMD uses I40E_GL_SWT_L2TAGCTRL to set vlan TPID. If setting TPID in port A with DPDK, then the configuration will also impact port B in the NIC with kernel driver, which don't want to use the TPID. So PMD reports warning to clarify what is changed by writing global register.

15.8 High Performance of Small Packets on 40G NIC

As there might be firmware fixes for performance enhancement in latest version of firmware image, the firmware update might be needed for getting high performance. Check with the local Intel's Network Division application engineers for firmware updates. Users should consult the release notes specific to a DPDK release to identify the validated firmware version for a NIC using the i40e driver.

15.8.1 Use 16 Bytes RX Descriptor Size

As i40e PMD supports both 16 and 32 bytes RX descriptor sizes, and 16 bytes size can provide helps to high performance of small packets. Configuration of `CONFIG_RTE_LIBRTE_I40E_16BYTE_RX_DESC` in config files can be changed to use 16 bytes size RX descriptors.

15.8.2 High Performance and per Packet Latency Tradeoff

Due to the hardware design, the interrupt signal inside NIC is needed for per packet descriptor write-back. The minimum interval of interrupts could be set at compile time by `CONFIG_RTE_LIBRTE_I40E_ITR_INTERVAL` in configuration files. Though there is a default configuration, the interval could be tuned by the users with that configuration item depends on what the user cares about more, performance or per packet latency.

15.9 Example of getting best performance with I3fwd example

The following is an example of running the DPDK `l3fwd` sample application to get high performance with an Intel server platform and Intel XL710 NICs.

The example scenario is to get best performance with two Intel XL710 40GbE ports. See Fig. 15.1 for the performance test setup.

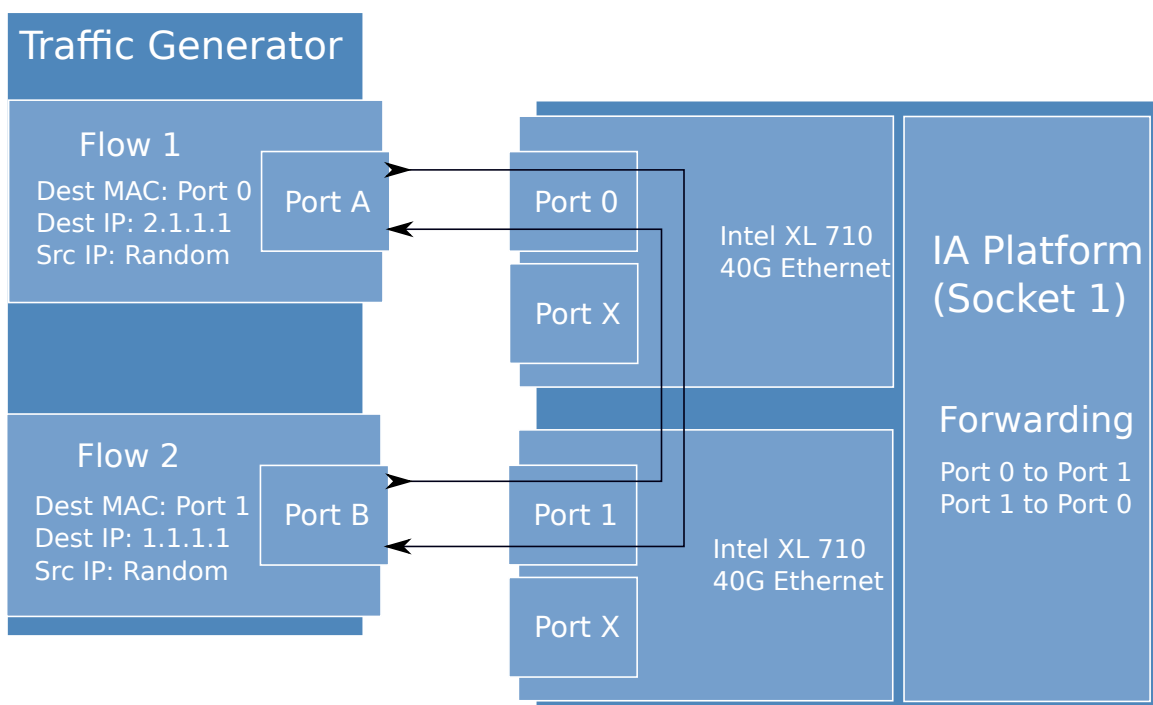


Fig. 15.1: Performance Test Setup

1. Add two Intel XL710 NICs to the platform, and use one port per card to get best performance. The reason for using two NICs is to overcome a PCIe Gen3's limitation since it cannot provide 80G bandwidth for two 40G ports, but two different PCIe Gen3 x8 slot

can. Refer to the sample NICs output above, then we can select 82:00.0 and 85:00.0 as test ports:

```
82:00.0 Ethernet [0200]: Intel XL710 for 40GbE QSFP+ [8086:1583]
85:00.0 Ethernet [0200]: Intel XL710 for 40GbE QSFP+ [8086:1583]
```

2. Connect the ports to the traffic generator. For high speed testing, it's best to use a hardware traffic generator.
3. Check the PCI devices numa node (socket id) and get the cores number on the exact socket id. In this case, 82:00.0 and 85:00.0 are both in socket 1, and the cores on socket 1 in the referenced platform are 18-35 and 54-71. Note: Don't use 2 logical cores on the same core (e.g core18 has 2 logical cores, core18 and core54), instead, use 2 logical cores from different cores (e.g core18 and core19).
4. Bind these two ports to igb_uio.
5. As to XL710 40G port, we need at least two queue pairs to achieve best performance, then two queues per port will be required, and each queue pair will need a dedicated CPU core for receiving/transmitting packets.
6. The DPDK sample application `l3fwd` will be used for performance testing, with using two ports for bi-directional forwarding. Compile the `l3fwd` sample with the default `lpm` mode.
7. The command line of running `l3fwd` would be something like the following:

```
./l3fwd -l 18-21 -n 4 -w 82:00.0 -w 85:00.0 \
-- -p 0x3 --config '(0,0,18), (0,1,19), (1,0,20), (1,1,21)'
```

This means that the application uses core 18 for port 0, queue pair 0 forwarding, core 19 for port 0, queue pair 1 forwarding, core 20 for port 1, queue pair 0 forwarding, and core 21 for port 1, queue pair 1 forwarding.

8. Configure the traffic at a traffic generator.
 - Start creating a stream on packet generator.
 - Set the Ethernet II type to 0x0800.

IGB POLL MODE DRIVER

The IGB PMD (`librte_pmd_e1000`) provides poll mode driver support for Intel 1GbE nics.

16.1 Features

Features of the IGB PMD are:

- Multiple queues for TX and RX
- Receiver Side Scaling (RSS)
- MAC/VLAN filtering
- Packet type information
- Double VLAN
- IEEE 1588
- TSO offload
- Checksum offload
- TCP segmentation offload
- Jumbo frames supported

16.2 Limitations or Known issues

16.3 Supported Chipsets and NICs

- Intel 82576EB 10 Gigabit Ethernet Controller
- Intel 82580EB 10 Gigabit Ethernet Controller
- Intel 82580DB 10 Gigabit Ethernet Controller
- Intel Ethernet Controller I210
- Intel Ethernet Controller I350

IXGBE DRIVER

17.1 Vector PMD for IXGBE

Vector PMD uses Intel® SIMD instructions to optimize packet I/O. It improves load/store bandwidth efficiency of L1 data cache by using a wider SSE/AVX register 1 (1). The wider register gives space to hold multiple packet buffers so as to save instruction number when processing bulk of packets.

There is no change to PMD API. The RX/TX handler are the only two entries for vPMD packet I/O. They are transparently registered at runtime RX/TX execution if all condition checks pass.

1. To date, only an SSE version of IX GBE vPMD is available. To ensure that vPMD is in the binary code, ensure that the option `CONFIG_RTE_IXGBE_INC_VECTOR=y` is in the configure file.

Some constraints apply as pre-conditions for specific optimizations on bulk packet transfers. The following sections explain RX and TX constraints in the vPMD.

17.1.1 RX Constraints

Prerequisites and Pre-conditions

The following prerequisites apply:

- To enable vPMD to work for RX, bulk allocation for Rx must be allowed.

Ensure that the following pre-conditions are satisfied:

- `rxq->rx_free_thresh >= RTE_PMD_IXGBE_RX_MAX_BURST`
- `rxq->rx_free_thresh < rxq->nb_rx_desc`
- `(rxq->nb_rx_desc % rxq->rx_free_thresh) == 0`
- `rxq->nb_rx_desc < (IXGBE_MAX_RING_DESC - RTE_PMD_IXGBE_RX_MAX_BURST)`

These conditions are checked in the code.

Scattered packets are not supported in this mode. If an incoming packet is greater than the maximum acceptable length of one “mbuf” data size (by default, the size is 2 KB), vPMD for RX would be disabled.

By default, `IXGBE_MAX_RING_DESC` is set to 4096 and `RTE_PMD_IXGBE_RX_MAX_BURST` is set to 32.

Feature not Supported by RX Vector PMD

Some features are not supported when trying to increase the throughput in vPMD. They are:

- IEEE1588
- FDIR
- Header split
- RX checksum off load

Other features are supported using optional MACRO configuration. They include:

- HW VLAN strip
- HW extend dual VLAN

To guarantee the constraint, configuration flags in `dev_conf.rxmode` will be checked:

- `hw_vlan_strip`
- `hw_vlan_extend`
- `hw_ip_checksum`
- `header_split`
- `dev_conf`

`fdir_conf->mode` will also be checked.

RX Burst Size

As vPMD is focused on high throughput, it assumes that the RX burst size is equal to or greater than 32 per burst. It returns zero if using `nb_pkt < 32` as the expected packet number in the receive handler.

17.1.2 TX Constraint

Prerequisite

The only prerequisite is related to `tx_rs_thresh`. The `tx_rs_thresh` value must be greater than or equal to `RTE_PMD_IXGBE_TX_MAX_BURST`, but less or equal to `RTE_IXGBE_TX_MAX_FREE_BUF_SZ`. Consequently, by default the `tx_rs_thresh` value is in the range 32 to 64.

Feature not Supported by TX Vector PMD

TX vPMD only works when `txq_flags` is set to `IXGBE_SIMPLE_FLAGS`.

This means that it does not support TX multi-segment, VLAN offload and TX csum offload. The following MACROs are used for these three features:

- `ETH_TXQ_FLAGS_NOMULTSEGS`
- `ETH_TXQ_FLAGS_NOVLANOFFL`

- ETH_TXQ_FLAGS_NOXSUMSCTP
- ETH_TXQ_FLAGS_NOXSUMUDP
- ETH_TXQ_FLAGS_NOXSUMTCP

17.2 Application Programming Interface

In DPDK release v16.11 an API for ixgbe specific functions has been added to the ixgbe PMD. The declarations for the API functions are in the header `rte_pmd_ixgbe.h`.

17.3 Sample Application Notes

17.3.1 l3fwd

When running l3fwd with vPMD, there is one thing to note. In the configuration, ensure that `port_conf.rxmode.hw_ip_checksum=0`. Otherwise, by default, RX vPMD is disabled.

17.3.2 load_balancer

As in the case of l3fwd, set configure `port_conf.rxmode.hw_ip_checksum=0` to enable vPMD. In addition, for improved performance, use `-bsz "(32,32),(64,64),(32,32)"` in `load_balancer` to avoid using the default burst size of 144.

17.4 Limitations or Known issues

17.4.1 Malicious Driver Detection not Supported

The Intel x550 series NICs support a feature called MDD (Malicious Driver Detection) which checks the behavior of the VF driver. If this feature is enabled, the VF must use the advanced context descriptor correctly and set the CC (Check Context) bit. DPDK PF doesn't support MDD, but kernel PF does. We may hit problem in this scenario kernel PF + DPDK VF. If user enables MDD in kernel PF, DPDK VF will not work. Because kernel PF thinks the VF is malicious. But actually it's not. The only reason is the VF doesn't act as MDD required. There's significant performance impact to support MDD. DPDK should check if the advanced context descriptor should be set and set it. And DPDK has to ask the info about the header length from the upper layer, because parsing the packet itself is not acceptable. So, it's too expensive to support MDD. When using kernel PF + DPDK VF on x550, please make sure to use a kernel PF driver that disables MDD or can disable MDD.

Some kernel drivers already disable MDD by default while some kernels can use the command `insmod ixgbe.ko MDD=0,0` to disable MDD. Each "0" in the command refers to a port. For example, if there are 6 ixgbe ports, the command should be changed to `insmod ixgbe.ko MDD=0,0,0,0,0,0`.

17.4.2 Statistics

The statistics of ixgbe hardware must be polled regularly in order for it to remain consistent. Running a DPDK application without polling the statistics will cause registers on hardware to count to the maximum value, and “stick” at that value.

In order to avoid statistic registers every reaching the maximum value, read the statistics from the hardware using `rte_eth_stats_get()` or `rte_eth_xstats_get()`.

The maximum time between statistics polls that ensures consistent results can be calculated as follows:

```
max_read_interval = UINT_MAX / max_packets_per_second
max_read_interval = 4294967295 / 14880952
max_read_interval = 288.6218096127183 (seconds)
max_read_interval = ~4 mins 48 sec.
```

In order to ensure valid results, it is recommended to poll every 4 minutes.

17.4.3 MTU setting

Although the user can set the MTU separately on PF and VF ports, the ixgbe NIC only supports one global MTU per physical port. So when the user sets different MTUs on PF and VF ports in one physical port, the real MTU for all these PF and VF ports is the largest value set. This behavior is based on the kernel driver behavior.

17.4.4 VF MAC address setting

On ixgbe, the concept of “pool” can be used for different things depending on the mode. In VMDq mode, “pool” means a VMDq pool. In IOV mode, “pool” means a VF.

There is no RTE API to add a VF’s MAC address from the PF. On ixgbe, the `rte_eth_dev_mac_addr_add()` function can be used to add a VF’s MAC address, as a workaround.

17.5 Inline crypto processing support

Inline IPsec processing is supported for `RTE_SECURITY_ACTION_TYPE_INLINE_CRYPTO` mode for ESP packets only:

- ESP authentication only: AES-128-GMAC (128-bit key)
- ESP encryption and authentication: AES-128-GCM (128-bit key)

IPsec Security Gateway Sample Application supports inline IPsec processing for ixgbe PMD.

For more details see the IPsec Security Gateway Sample Application and Security library documentation.

17.6 Supported Chipsets and NICs

- Intel 82599EB 10 Gigabit Ethernet Controller

- Intel 82598EB 10 Gigabit Ethernet Controller
- Intel 82599ES 10 Gigabit Ethernet Controller
- Intel 82599EN 10 Gigabit Ethernet Controller
- Intel Ethernet Controller X540-AT2
- Intel Ethernet Controller X550-BT2
- Intel Ethernet Controller X550-AT2
- Intel Ethernet Controller X550-AT
- Intel Ethernet Converged Network Adapter X520-SR1
- Intel Ethernet Converged Network Adapter X520-SR2
- Intel Ethernet Converged Network Adapter X520-LR1
- Intel Ethernet Converged Network Adapter X520-DA1
- Intel Ethernet Converged Network Adapter X520-DA2
- Intel Ethernet Converged Network Adapter X520-DA4
- Intel Ethernet Converged Network Adapter X520-QDA1
- Intel Ethernet Converged Network Adapter X520-T2
- Intel 10 Gigabit AF DA Dual Port Server Adapter
- Intel 10 Gigabit AT Server Adapter
- Intel 10 Gigabit AT2 Server Adapter
- Intel 10 Gigabit CX4 Dual Port Server Adapter
- Intel 10 Gigabit XF LR Server Adapter
- Intel 10 Gigabit XF SR Dual Port Server Adapter
- Intel 10 Gigabit XF SR Server Adapter
- Intel Ethernet Converged Network Adapter X540-T1
- Intel Ethernet Converged Network Adapter X540-T2
- Intel Ethernet Converged Network Adapter X550-T1
- Intel Ethernet Converged Network Adapter X550-T2

INTEL VIRTUAL FUNCTION DRIVER

Supported Intel® Ethernet Controllers (see the *DPDK Release Notes* for details) support the following modes of operation in a virtualized environment:

- **SR-IOV mode:** Involves direct assignment of part of the port resources to different guest operating systems using the PCI-SIG Single Root I/O Virtualization (SR IOV) standard, also known as “native mode” or “pass-through” mode. In this chapter, this mode is referred to as IOV mode.
- **VMDq mode:** Involves central management of the networking resources by an IO Virtual Machine (IOVM) or a Virtual Machine Monitor (VMM), also known as software switch acceleration mode. In this chapter, this mode is referred to as the Next Generation VMDq mode.

18.1 SR-IOV Mode Utilization in a DPDK Environment

The DPDK uses the SR-IOV feature for hardware-based I/O sharing in IOV mode. Therefore, it is possible to partition SR-IOV capability on Ethernet controller NIC resources logically and expose them to a virtual machine as a separate PCI function called a “Virtual Function”. Refer to [Fig. 18.1](#).

Therefore, a NIC is logically distributed among multiple virtual machines (as shown in [Fig. 18.1](#)), while still having global data in common to share with the Physical Function and other Virtual Functions. The DPDK `fm10kvf`, `i40evf`, `igbvf` or `ixgbev` as a Poll Mode Driver (PMD) serves for the Intel® 82576 Gigabit Ethernet Controller, Intel® Ethernet Controller I350 family, Intel® 82599 10 Gigabit Ethernet Controller NIC, Intel® Fortville 10/40 Gigabit Ethernet Controller NIC’s virtual PCI function, or PCIe host-interface of the Intel Ethernet Switch FM10000 Series. Meanwhile the DPDK Poll Mode Driver (PMD) also supports “Physical Function” of such NIC’s on the host.

The DPDK PF/VF Poll Mode Driver (PMD) supports the Layer 2 switch on Intel® 82576 Gigabit Ethernet Controller, Intel® Ethernet Controller I350 family, Intel® 82599 10 Gigabit Ethernet Controller, and Intel® Fortville 10/40 Gigabit Ethernet Controller NICs so that guest can choose it for inter virtual machine traffic in SR-IOV mode.

For more detail on SR-IOV, please refer to the following documents:

- [SR-IOV provides hardware based I/O sharing](#)
- [PCI-SIG-Single Root I/O Virtualization Support on IA](#)
- [Scalable I/O Virtualized Servers](#)

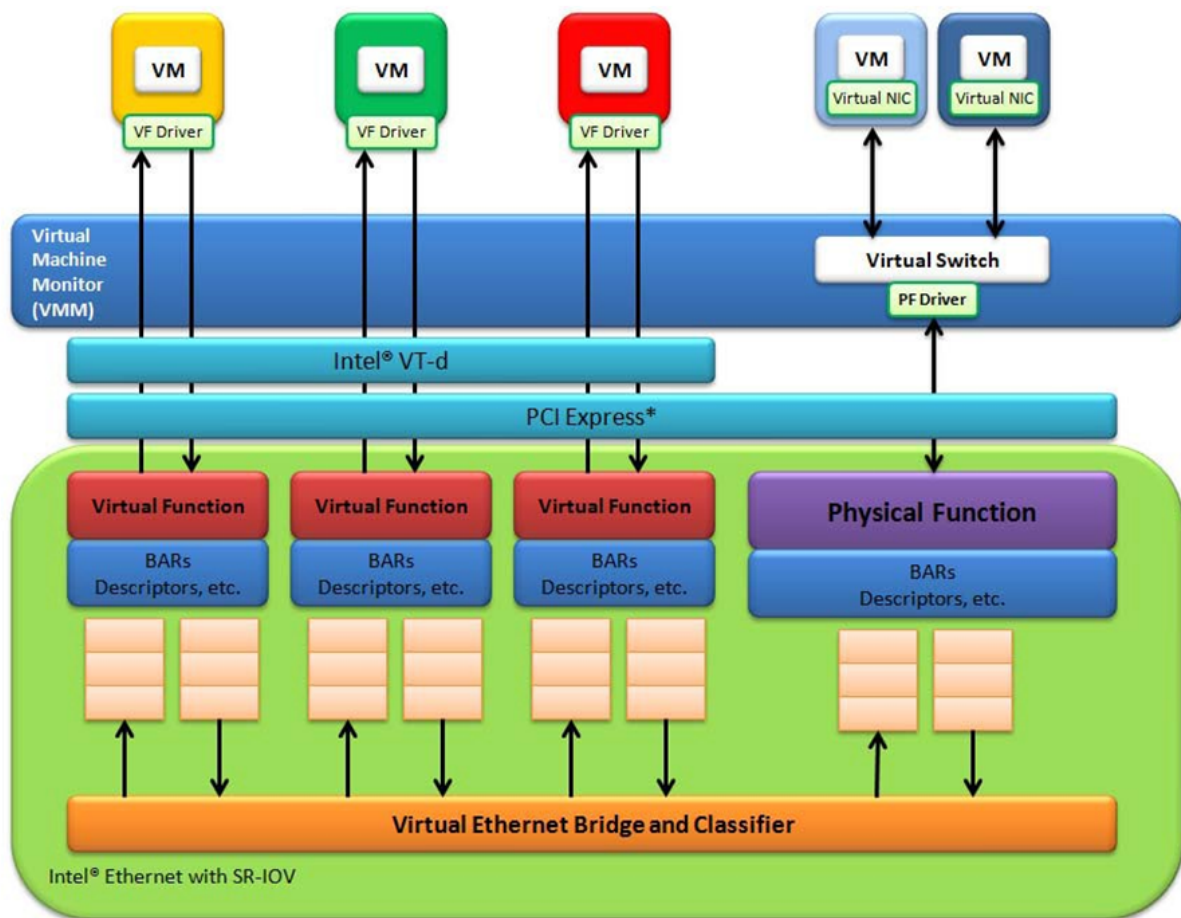


Fig. 18.1: Virtualization for a Single Port NIC in SR-IOV Mode

18.1.1 Physical and Virtual Function Infrastructure

The following describes the Physical Function and Virtual Functions infrastructure for the supported Ethernet Controller NICs.

Virtual Functions operate under the respective Physical Function on the same NIC Port and therefore have no access to the global NIC resources that are shared between other functions for the same NIC port.

A Virtual Function has basic access to the queue resources and control structures of the queues assigned to it. For global resource access, a Virtual Function has to send a request to the Physical Function for that port, and the Physical Function operates on the global resources on behalf of the Virtual Function. For this out-of-band communication, an SR-IOV enabled NIC provides a memory buffer for each Virtual Function, which is called a “Mailbox”.

Intel® Ethernet Adaptive Virtual Function

Adaptive Virtual Function (AVF) is a SR-IOV Virtual Function with the same device id (8086:1889) on different Intel Ethernet Controller. AVF Driver is VF driver which supports for all future Intel devices without requiring a VM update. And since this happens to be an adaptive VF driver, every new drop of the VF driver would add more and more advanced features that can be turned on in the VM if the underlying HW device supports those advanced features based on a device agnostic way without ever compromising on the base functionality. AVF provides generic hardware interface and interface between AVF driver and a compliant PF driver is specified.

Intel products starting Ethernet Controller 700 Series to support Adaptive Virtual Function.

The way to generate Virtual Function is like normal, and the resource of VF assignment depends on the NIC Infrastructure.

For more detail on SR-IOV, please refer to the following documents:

- [Intel® AVF HAS](#)

Note: To use DPDK AVF PMD on Intel® 700 Series Ethernet Controller, the device id (0x1889) need to specified during device assignment in hypervisor. Take qemu for example, the device assignment should carry the AVF device id (0x1889) like `-device vfio-pci,x-pci-device-id=0x1889,host=03:0a.0`.

The PCIE host-interface of Intel Ethernet Switch FM10000 Series VF infrastructure

In a virtualized environment, the programmer can enable a maximum of *64 Virtual Functions (VF)* globally per PCIE host-interface of the Intel Ethernet Switch FM10000 Series device. Each VF can have a maximum of 16 queue pairs. The Physical Function in host could be only configured by the Linux* fm10k driver (in the case of the Linux Kernel-based Virtual Machine [KVM]), DPDK PMD PF driver doesn't support it yet.

For example,

- Using Linux* fm10k driver:

```
rmmod fm10k (To remove the fm10k module)
insmod fm0k.ko max_vfs=2,2 (To enable two Virtual Functions per port)
```

Virtual Function enumeration is performed in the following sequence by the Linux* pci driver for a dual-port NIC. When you enable the four Virtual Functions with the above command, the four enabled functions have a Function# represented by (Bus#, Device#, Function#) in sequence starting from 0 to 3. However:

- Virtual Functions 0 and 2 belong to Physical Function 0
- Virtual Functions 1 and 3 belong to Physical Function 1

Note: The above is an important consideration to take into account when targeting specific packets to a selected port.

Intel® X710/XL710 Gigabit Ethernet Controller VF Infrastructure

In a virtualized environment, the programmer can enable a maximum of *128 Virtual Functions (VF)* globally per Intel® X710/XL710 Gigabit Ethernet Controller NIC device. The number of queue pairs of each VF can be configured by `CONFIG_RTE_LIBRTE_I40E_QUEUE_NUM_PER_VF` in `config` file. The Physical Function in host could be either configured by the Linux* i40e driver (in the case of the Linux Kernel-based Virtual Machine [KVM]) or by DPDK PMD PF driver. When using both DPDK PMD PF/VF drivers, the whole NIC will be taken over by DPDK based application.

For example,

- Using Linux* i40e driver:

```
rmmod i40e (To remove the i40e module)
insmod i40e.ko max_vfs=2,2 (To enable two Virtual Functions per port)
```

- Using the DPDK PMD PF i40e driver:

Kernel Params: `iommu=pt, intel_iommu=on`

```
modprobe uio
insmod igb_uio
./dpdk-devbind.py -b igb_uio bb:ss.f
echo 2 > /sys/bus/pci/devices/0000\:bb\:ss.f/max_vfs (To enable two VFs on a specific PCI
```

Launch the DPDK `testpmd/example` or your own host daemon application using the DPDK PMD library.

Virtual Function enumeration is performed in the following sequence by the Linux* pci driver for a dual-port NIC. When you enable the four Virtual Functions with the above command, the four enabled functions have a Function# represented by (Bus#, Device#, Function#) in sequence starting from 0 to 3. However:

- Virtual Functions 0 and 2 belong to Physical Function 0
- Virtual Functions 1 and 3 belong to Physical Function 1

Note: The above is an important consideration to take into account when targeting specific packets to a selected port.

For Intel® X710/XL710 Gigabit Ethernet Controller, queues are in pairs. One queue pair means one receive queue and one transmit queue. The default number of queue pairs per VF is 4, and can be 16 in maximum.

Intel® 82599 10 Gigabit Ethernet Controller VF Infrastructure

The programmer can enable a maximum of *63 Virtual Functions* and there must be *one Physical Function* per Intel® 82599 10 Gigabit Ethernet Controller NIC port. The reason for this is that the device allows for a maximum of 128 queues per port and a virtual/physical function has to have at least one queue pair (RX/TX). The current implementation of the DPDK ixgbevf driver supports a single queue pair (RX/TX) per Virtual Function. The Physical Function in host could be either configured by the Linux* ixgbe driver (in the case of the Linux Kernel-based Virtual Machine [KVM]) or by DPDK PMD PF driver. When using both DPDK PMD PF/VF drivers, the whole NIC will be taken over by DPDK based application.

For example,

- Using Linux* ixgbe driver:

```
rmmod ixgbe (To remove the ixgbe module)
insmod ixgbe max_vfs=2,2 (To enable two Virtual Functions per port)
```

- Using the DPDK PMD PF ixgbe driver:

Kernel Params: iommu=pt, intel_iommu=on

```
modprobe uio
insmod igb_uio
./dpdk-devbind.py -b igb_uio bb:ss.f
echo 2 > /sys/bus/pci/devices/0000\:bb\:ss.f/max_vfs (To enable two VFs on a specific PCI
```

Launch the DPDK testpmd/example or your own host daemon application using the DPDK PMD library.

- Using the DPDK PMD PF ixgbe driver to enable VF RSS:

Same steps as above to install the modules of uio, igb_uio, specify max_vfs for PCI device, and launch the DPDK testpmd/example or your own host daemon application using the DPDK PMD library.

The available queue number (at most 4) per VF depends on the total number of pool, which is determined by the max number of VF at PF initialization stage and the number of queue specified in config:

- If the max number of VFs (max_vfs) is set in the range of 1 to 32:

If the number of Rx queues is specified as 4 (--rxq=4 in testpmd), then there are totally 32 pools (ETH_32_POOLS), and each VF could have 4 Rx queues;

If the number of Rx queues is specified as 2 (--rxq=2 in testpmd), then there are totally 32 pools (ETH_32_POOLS), and each VF could have 2 Rx queues;

- If the max number of VFs (max_vfs) is in the range of 33 to 64:

If the number of Rx queues is specified as 4 (--rxq=4 in testpmd), then error message is expected as rxq is not correct at this case;

If the number of rxq is 2 (--rxq=2 in testpmd), then there is totally 64 pools (ETH_64_POOLS), and each VF have 2 Rx queues;

On host, to enable VF RSS functionality, rx mq mode should be set as ETH_MQ_RX_VMDQ_RSS or ETH_MQ_RX_RSS mode, and SRIOV mode should be activated (max_vfs >= 1). It also needs config VF RSS information like hash function, RSS key, RSS key length.

Note: The limitation for VF RSS on Intel® 82599 10 Gigabit Ethernet Controller is: The hash and key are shared among PF and all VF, the RETA table with 128 entries is also shared among PF and all VF; So it could not to provide a method to query the hash and reta content per VF on guest, while, if possible, please query them on host for the shared RETA information.

Virtual Function enumeration is performed in the following sequence by the Linux* pci driver for a dual-port NIC. When you enable the four Virtual Functions with the above command, the four enabled functions have a Function# represented by (Bus#, Device#, Function#) in sequence starting from 0 to 3. However:

- Virtual Functions 0 and 2 belong to Physical Function 0
 - Virtual Functions 1 and 3 belong to Physical Function 1
-

Note: The above is an important consideration to take into account when targeting specific packets to a selected port.

Intel® 82576 Gigabit Ethernet Controller and Intel® Ethernet Controller I350 Family VF Infrastructure

In a virtualized environment, an Intel® 82576 Gigabit Ethernet Controller serves up to eight virtual machines (VMs). The controller has 16 TX and 16 RX queues. They are generally referred to (or thought of) as queue pairs (one TX and one RX queue). This gives the controller 16 queue pairs.

A pool is a group of queue pairs for assignment to the same VF, used for transmit and receive operations. The controller has eight pools, with each pool containing two queue pairs, that is, two TX and two RX queues assigned to each VF.

In a virtualized environment, an Intel® Ethernet Controller I350 family device serves up to eight virtual machines (VMs) per port. The eight queues can be accessed by eight different VMs if configured correctly (the i350 has 4x1GbE ports each with 8T X and 8 RX queues), that means, one Transmit and one Receive queue assigned to each VF.

For example,

- Using Linux* igb driver:

```
rmmod igb (To remove the igb module)
insmod igb max_vfs=2,2 (To enable two Virtual Functions per port)
```

- Using DPDK PMD PF igb driver:

Kernel Params: iommu=pt, intel_iommu=on modprobe uio

```
insmod igb_uio
./dpdk-devbind.py -b igb_uio bb:ss.f
echo 2 > /sys/bus/pci/devices/0000\:bb\:ss.f/max_vfs (To enable two VFs on a specific pci
```

Launch DPDK testpmd/example or your own host daemon application using the DPDK PMD library.

Virtual Function enumeration is performed in the following sequence by the Linux* pci driver for a four-port NIC. When you enable the four Virtual Functions with the above command, the four enabled functions have a Function# represented by (Bus#, Device#, Function#) in sequence, starting from 0 to 7. However:

- Virtual Functions 0 and 4 belong to Physical Function 0
- Virtual Functions 1 and 5 belong to Physical Function 1
- Virtual Functions 2 and 6 belong to Physical Function 2
- Virtual Functions 3 and 7 belong to Physical Function 3

Note: The above is an important consideration to take into account when targeting specific packets to a selected port.

18.1.2 Validated Hypervisors

The validated hypervisor is:

- KVM (Kernel Virtual Machine) with Qemu, version 0.14.0

However, the hypervisor is bypassed to configure the Virtual Function devices using the Mailbox interface, the solution is hypervisor-agnostic. Xen* and VMware* (when SR-IOV is supported) will also be able to support the DPDK with Virtual Function driver support.

18.1.3 Expected Guest Operating System in Virtual Machine

The expected guest operating systems in a virtualized environment are:

- Fedora* 14 (64-bit)
- Ubuntu* 10.04 (64-bit)

For supported kernel versions, refer to the *DPDK Release Notes*.

18.2 Setting Up a KVM Virtual Machine Monitor

The following describes a target environment:

- Host Operating System: Fedora 14
- Hypervisor: KVM (Kernel Virtual Machine) with Qemu version 0.14.0
- Guest Operating System: Fedora 14
- Linux Kernel Version: Refer to the *DPDK Getting Started Guide*
- Target Applications: l2fwd, l3fwd-vf

The setup procedure is as follows:

1. Before booting the Host OS, open **BIOS setup** and enable **Intel® VT features**.
2. While booting the Host OS kernel, pass the `intel_iommu=on` kernel command line argument using GRUB. When using DPDK PF driver on host, pass the `iommu=pt` kernel command line argument in GRUB.
3. Download `qemu-kvm-0.14.0` from <http://sourceforge.net/projects/kvm/files/qemu-kvm/> and install it in the Host OS using the following steps:

When using a recent kernel (2.6.25+) with kvm modules included:

```
tar xzf qemu-kvm-release.tar.gz
cd qemu-kvm-release
./configure --prefix=/usr/local/kvm
make
sudo make install
sudo /sbin/modprobe kvm-intel
```

When using an older kernel, or a kernel from a distribution without the kvm modules, you must download (from the same link), compile and install the modules yourself:

```
tar xjf kvm-kmod-release.tar.bz2
cd kvm-kmod-release
./configure
make
sudo make install
sudo /sbin/modprobe kvm-intel
```

qemu-kvm installs in the /usr/local/bin directory.

For more details about KVM configuration and usage, please refer to:

<http://www.linux-kvm.org/page/HOWTO1>.

4. Create a Virtual Machine and install Fedora 14 on the Virtual Machine. This is referred to as the Guest Operating System (Guest OS).
5. Download and install the latest ixgbe driver from:

http://downloadcenter.intel.com/Detail_Desc.aspx?agr=Y&DwnldID=14687

6. In the Host OS

When using Linux kernel ixgbe driver, unload the Linux ixgbe driver and reload it with the `max_vfs=2,2` argument:

```
rmmod ixgbe
modprobe ixgbe max_vfs=2,2
```

When using DPDK PMD PF driver, insert DPDK kernel module `igb_uio` and set the number of VF by `sysfs max_vfs`:

```
modprobe uio
insmod igb_uio
./dpdk-devbind.py -b igb_uio 02:00.0 02:00.1 0e:00.0 0e:00.1
echo 2 > /sys/bus/pci/devices/0000\:02\:00.0/max_vfs
echo 2 > /sys/bus/pci/devices/0000\:02\:00.1/max_vfs
echo 2 > /sys/bus/pci/devices/0000\:0e\:00.0/max_vfs
echo 2 > /sys/bus/pci/devices/0000\:0e\:00.1/max_vfs
```

Note: You need to explicitly specify number of vfs for each port, for example, in the command above, it creates two vfs for the first two ixgbe ports.

Let say we have a machine with four physical ixgbe ports:

```
0000:02:00.0
0000:02:00.1
0000:0e:00.0
0000:0e:00.1
```

The command above creates two vfs for device 0000:02:00.0:

```
ls -alrt /sys/bus/pci/devices/0000\:02\:00.0/virt*
lrwxrwxrwx. 1 root root 0 Apr 13 05:40 /sys/bus/pci/devices/0000:02:00.0/virtfn1 -> ../000
lrwxrwxrwx. 1 root root 0 Apr 13 05:40 /sys/bus/pci/devices/0000:02:00.0/virtfn0 -> ../000
```

It also creates two vfs for device 0000:02:00.1:

```
ls -alrt /sys/bus/pci/devices/0000\:02\:00.1/virt*
lrwxrwxrwx. 1 root root 0 Apr 13 05:51 /sys/bus/pci/devices/0000:02:00.1/virtfn1 -> ../000
lrwxrwxrwx. 1 root root 0 Apr 13 05:51 /sys/bus/pci/devices/0000:02:00.1/virtfn0 -> ../000
```

- List the PCI devices connected and notice that the Host OS shows two Physical Functions (traditional ports) and four Virtual Functions (two for each port). This is the result of the previous step.
- Insert the `pci_stub` module to hold the PCI devices that are freed from the default driver using the following command (see http://www.linux-kvm.org/page/How_to_assign_devices_with_VT-d_in_KVM Section 4 for more information):

```
sudo /sbin/modprobe pci-stub
```

Unbind the default driver from the PCI devices representing the Virtual Functions. A script to perform this action is as follows:

```
echo "8086 10ed" > /sys/bus/pci/drivers/pci-stub/new_id
echo 0000:08:10.0 > /sys/bus/pci/devices/0000:08:10.0/driver/unbind
echo 0000:08:10.0 > /sys/bus/pci/drivers/pci-stub/bind
```

where, 0000:08:10.0 belongs to the Virtual Function visible in the Host OS.

- Now, start the Virtual Machine by running the following command:

```
/usr/local/kvm/bin/qemu-system-x86_64 -m 4096 -smp 4 -boot c -hda lucid.qcow2 -device pci-
```

where:

— `-m` = memory to assign

— `-smp` = number of smp cores

— `-boot` = boot option

— `-hda` = virtual disk image

— `-device` = device to attach

Note: — The `pci-assign,host=08:10.0` value indicates that you want to attach a PCI device to a Virtual Machine and the respective (Bus:Device.Function) numbers should be passed for the Virtual Function to be attached.

— `qemu-kvm-0.14.0` allows a maximum of four PCI devices assigned to a VM, but this is `qemu-kvm` version dependent since `qemu-kvm-0.14.1` allows a maximum of five PCI devices.

— `qemu-system-x86_64` also has a `-cpu` command line option that is used to select the `cpu_model` to emulate in a Virtual Machine. Therefore, it can be used as:

```
/usr/local/kvm/bin/qemu-system-x86_64 -cpu ?
```

(to list all available `cpu_models`)

```
/usr/local/kvm/bin/qemu-system-x86_64 -m 4096 -cpu host -smp 4 -boot c -hda lucid.qcow2 -c
```

(to use the same `cpu_model` equivalent to the host `cpu`)

For more information, please refer to: <http://wiki.qemu.org/Features/CPUModels>.

10. If use vfio-pci to pass through device instead of pci-assign, steps 8 and 9 need to be updated to bind device to vfio-pci and replace pci-assign with vfio-pci when start virtual machine.

```
sudo /sbin/modprobe vfio-pci

echo "8086 10ed" > /sys/bus/pci/drivers/vfio-pci/new_id
echo 0000:08:10.0 > /sys/bus/pci/devices/0000:08:10.0/driver/unbind
echo 0000:08:10.0 > /sys/bus/pci/drivers/vfio-pci/bind

/usr/local/kvm/bin/qemu-system-x86_64 -m 4096 -smp 4 -boot c -hda lucid.qcow2 -device vfio-pci,host=0000:08:10.0
```

11. Install and run DPDK host app to take over the Physical Function. Eg.

```
make install T=x86_64-native-linuxapp-gcc
./x86_64-native-linuxapp-gcc/app/testpmd -l 0-3 -n 4 -- -i
```

12. Finally, access the Guest OS using vncviewer with the localhost:5900 port and check the lspci command output in the Guest OS. The virtual functions will be listed as available for use.

13. Configure and install the DPDK with an x86_64-native-linuxapp-gcc configuration on the Guest OS as normal, that is, there is no change to the normal installation procedure.

```
make config T=x86_64-native-linuxapp-gcc O=x86_64-native-linuxapp-gcc
cd x86_64-native-linuxapp-gcc
make
```

Note: If you are unable to compile the DPDK and you are getting “error: CPU you selected does not support x86-64 instruction set”, power off the Guest OS and start the virtual machine with the correct -cpu option in the qemu- system-x86_64 command as shown in step 9. You must select the best x86_64 cpu_model to emulate or you can select host option if available.

Note: Run the DPDK I2fwd sample application in the Guest OS with Hugepages enabled. For the expected benchmark performance, you must pin the cores from the Guest OS to the Host OS (taskset can be used to do this) and you must also look at the PCI Bus layout on the board to ensure you are not running the traffic over the QPI Interface.

Note:

- The Virtual Machine Manager (the Fedora package name is virt-manager) is a utility for virtual machine management that can also be used to create, start, stop and delete virtual machines. If this option is used, step 2 and 6 in the instructions provided will be different.
 - virsh, a command line utility for virtual machine management, can also be used to bind and unbind devices to a virtual machine in Ubuntu. If this option is used, step 6 in the instructions provided will be different.
 - The Virtual Machine Monitor (see Fig. 18.2) is equivalent to a Host OS with KVM installed as described in the instructions.
-

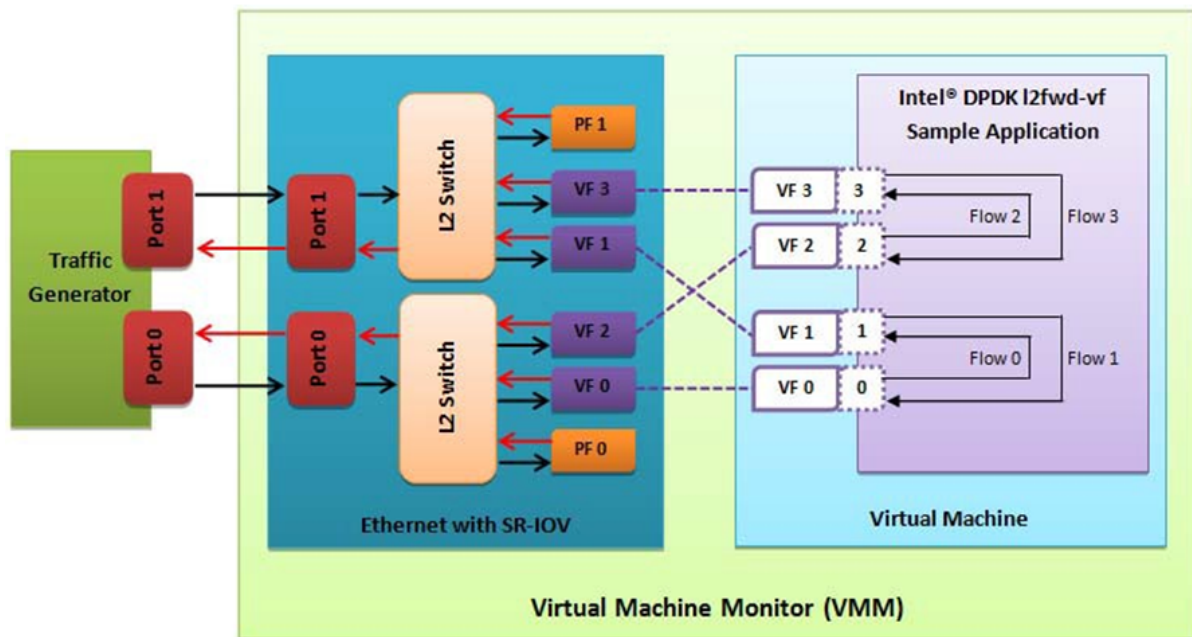


Fig. 18.2: Performance Benchmark Setup

18.3 DPDK SR-IOV PMD PF/VF Driver Usage Model

18.3.1 Fast Host-based Packet Processing

Software Defined Network (SDN) trends are demanding fast host-based packet handling. In a virtualization environment, the DPDK VF PMD driver performs the same throughput result as a non-VT native environment.

With such host instance fast packet processing, lots of services such as filtering, QoS, DPI can be offloaded on the host fast path.

Fig. 18.3 shows the scenario where some VMs directly communicate externally via a VFs, while others connect to a virtual switch and share the same uplink bandwidth.

18.4 SR-IOV (PF/VF) Approach for Inter-VM Communication

Inter-VM data communication is one of the traffic bottle necks in virtualization platforms. SR-IOV device assignment helps a VM to attach the real device, taking advantage of the bridge in the NIC. So VF-to-VF traffic within the same physical port (VM0<->VM1) have hardware acceleration. However, when VF crosses physical ports (VM0<->VM2), there is no such hardware bridge. In this case, the DPDK PMD PF driver provides host forwarding between such VMs.

Fig. 18.4 shows an example. In this case an update of the MAC address lookup tables in both the NIC and host DPDK application is required.

In the NIC, writing the destination of a MAC address belongs to another cross device VM to the PF specific pool. So when a packet comes in, its destination MAC address will match and forward to the host DPDK PMD application.

In the host DPDK application, the behavior is similar to L2 forwarding, that is, the packet is

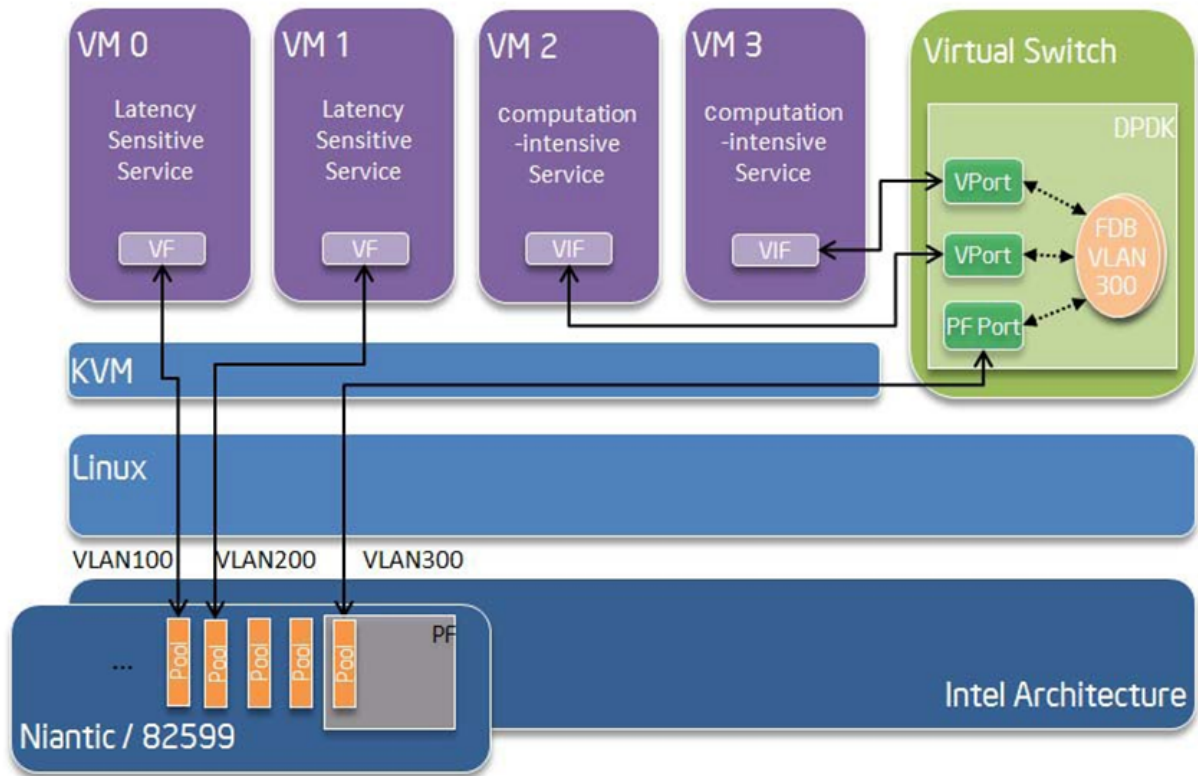


Fig. 18.3: Fast Host-based Packet Processing

forwarded to the correct PF pool. The SR-IOV NIC switch forwards the packet to a specific VM according to the MAC destination address which belongs to the destination VF on the VM.

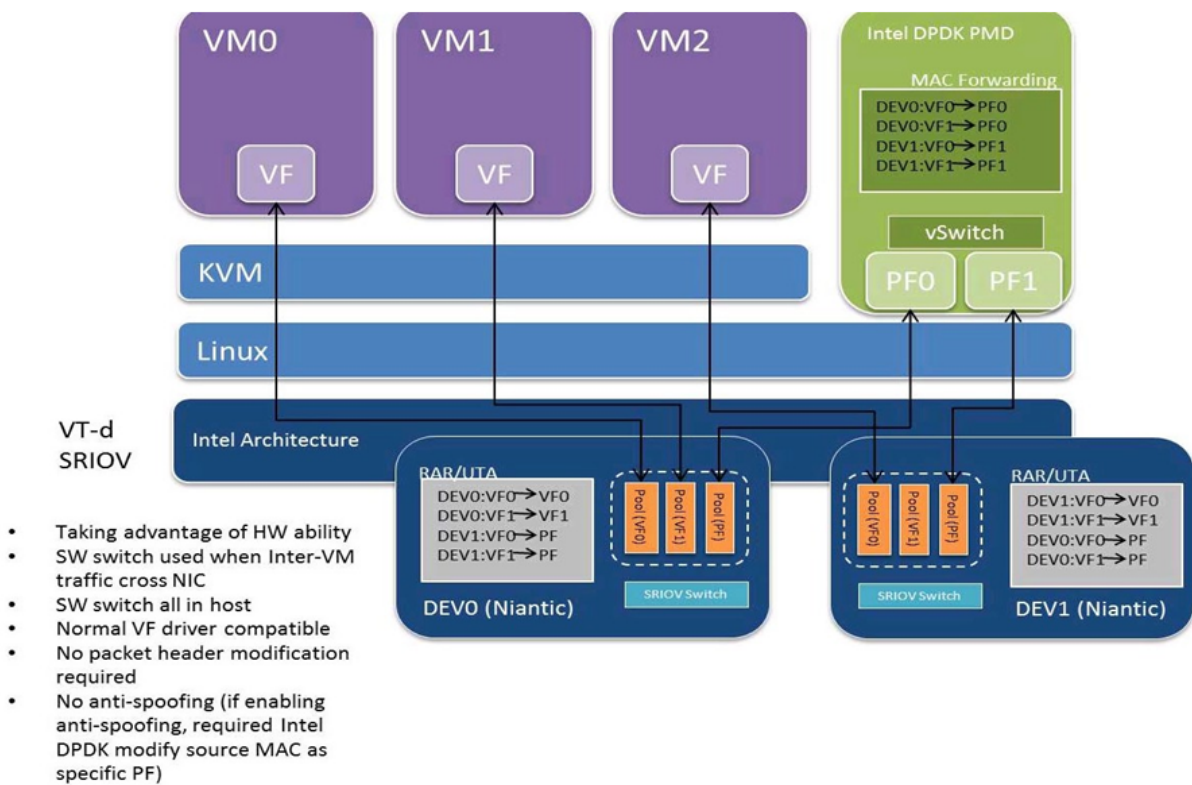


Fig. 18.4: Inter-VM Communication

KNI POLL MODE DRIVER

KNI PMD is wrapper to the `librte_kni` library.

This PMD enables using KNI without having a KNI specific application, any forwarding application can use PMD interface for KNI.

Sending packets to any DPDK controlled interface or sending to the Linux networking stack will be transparent to the DPDK application.

To create a KNI device `net_kni#` device name should be used, and this will create `kni#` Linux virtual network interface.

There is no physical device backend for the virtual KNI device.

Packets sent to the KNI Linux interface will be received by the DPDK application, and DPDK application may forward packets to a physical NIC or to a virtual device (like another KNI interface or PCAP interface).

To forward any traffic from physical NIC to the Linux networking stack, an application should control a physical port and create one virtual KNI port, and forward between two.

Using this PMD requires KNI kernel module be inserted.

19.1 Usage

EAL `--vdev` argument can be used to create KNI device instance, like:

```
testpmd --vdev=net_kni0 --vdev=net_kni1 -- -i
```

Above command will create `kni0` and `kni1` Linux network interfaces, those interfaces can be controlled by standard Linux tools.

When `testpmd` forwarding starts, any packets sent to `kni0` interface forwarded to the `kni1` interface and vice versa.

There is no hard limit on number of interfaces that can be created.

19.2 Default interface configuration

`librte_kni` can create Linux network interfaces with different features, feature set controlled by a configuration struct, and KNI PMD uses a fixed configuration:


```
Interface name: kni#
force bind kernel thread to a core : NO
mbuf size: MAX_PACKET_SZ
```

KNI control path is not supported with the PMD, since there is no physical backend device by default.

19.3 PMD arguments

`no_request_thread`, by default PMD creates a pthread for each KNI interface to handle Linux network interface control commands, like `ifconfig kni0 up`

With `no_request_thread` option, pthread is not created and control commands not handled by PMD.

By default request thread is enabled. And this argument should not be used most of the time, unless this PMD used with customized DPDK application to handle requests itself.

Argument usage:

```
testpmd --vdev "net_kni0,no_request_thread=1" -- -i
```

19.4 PMD log messages

If KNI kernel module (`rte_kni.ko`) not inserted, following error log printed:

```
"KNI: KNI subsystem has not been initialized. Invoke rte_kni_init() first"
```

19.5 PMD testing

It is possible to test PMD quickly using KNI kernel module loopback feature:

- Insert KNI kernel module with loopback support:

```
insmod build/kmod/rte_kni.ko lo_mode=lo_mode_fifo_skb
```

- Start testpmd with no physical device but two KNI virtual devices:

```
./testpmd --vdev net_kni0 --vdev net_kni1 -- -i
...
Configuring Port 0 (socket 0)
KNI: pci: 00:00:00      c580:b8
Port 0: 1A:4A:5B:7C:A2:8C
Configuring Port 1 (socket 0)
KNI: pci: 00:00:00      600:b9
Port 1: AE:95:21:07:93:DD
Checking link statuses...
Port 0 Link Up - speed 10000 Mbps - full-duplex
Port 1 Link Up - speed 10000 Mbps - full-duplex
Done
testpmd>
```

- Observe Linux interfaces

```
$ ifconfig kni0 && ifconfig kni1
kni0: flags=4098<BROADCAST,MULTICAST>  mtu 1500
      ether ae:8e:79:8e:9b:c8  txqueuelen 1000  (Ethernet)
```

```

RX packets 0 bytes 0 (0.0 B)
RX errors 0 dropped 0 overruns 0 frame 0
TX packets 0 bytes 0 (0.0 B)
TX errors 0 dropped 0 overruns 0 carrier 0 collisions 0

```

```

kn1l: flags=4098<BROADCAST,MULTICAST> mtu 1500
ether 9e:76:43:53:3e:9b txqueuelen 1000 (Ethernet)
RX packets 0 bytes 0 (0.0 B)
RX errors 0 dropped 0 overruns 0 frame 0
TX packets 0 bytes 0 (0.0 B)
TX errors 0 dropped 0 overruns 0 carrier 0 collisions 0

```

- Start forwarding with tx_first:

```
testpmd> start tx_first
```

- Quit and check forwarding stats:

```
testpmd> quit
Telling cores to stop...
Waiting for lcores to finish...
```

```

----- Forward statistics for port 0 -----
RX-packets: 35637905      RX-dropped: 0           RX-total: 35637905
TX-packets: 35637947      TX-dropped: 0           TX-total: 35637947
-----

```

```

----- Forward statistics for port 1 -----
RX-packets: 35637915      RX-dropped: 0           RX-total: 35637915
TX-packets: 35637937      TX-dropped: 0           TX-total: 35637937
-----

```

```

+++++++ Accumulated forward statistics for all ports+++++++
RX-packets: 71275820      RX-dropped: 0           RX-total: 71275820
TX-packets: 71275884      TX-dropped: 0           TX-total: 71275884
+++++++

```

LIQUIDIO VF POLL MODE DRIVER

The LiquidIO VF PMD library (`librte_pmd_lio`) provides poll mode driver support for Cavium LiquidIO® II server adapter VFs. PF management and VF creation can be done using kernel driver.

More information can be found at [Cavium Official Website](#).

20.1 Supported LiquidIO Adapters

- LiquidIO II CN2350 210SV/225SV
- LiquidIO II CN2350 210SVPT
- LiquidIO II CN2360 210SV/225SV
- LiquidIO II CN2360 210SVPT

20.2 Pre-Installation Configuration

The following options can be modified in the `config` file. Please note that enabling debugging options may affect system performance.

- `CONFIG_RTE_LIBRTE_LIO_PMD` (default `y`)
Toggle compilation of LiquidIO PMD.
- `CONFIG_RTE_LIBRTE_LIO_DEBUG_DRIVER` (default `n`)
Toggle display of generic debugging messages.
- `CONFIG_RTE_LIBRTE_LIO_DEBUG_INIT` (default `n`)
Toggle display of initialization related messages.
- `CONFIG_RTE_LIBRTE_LIO_DEBUG_RX` (default `n`)
Toggle display of receive fast path run-time messages.
- `CONFIG_RTE_LIBRTE_LIO_DEBUG_TX` (default `n`)
Toggle display of transmit fast path run-time messages.
- `CONFIG_RTE_LIBRTE_LIO_DEBUG_MBOX` (default `n`)
Toggle display of mailbox messages.

- CONFIG_RTE_LIBRTE_LIO_DEBUG_REGS (default n)

Toggle display of register reads and writes.

20.3 SR-IOV: Prerequisites and Sample Application Notes

This section provides instructions to configure SR-IOV with Linux OS.

1. Verify SR-IOV and ARI capabilities are enabled on the adapter using `lspci`:

```
lspci -s <slot> -vvv
```

Example output:

```
[...]
Capabilities: [148 v1] Alternative Routing-ID Interpretation (ARI)
[...]
Capabilities: [178 v1] Single Root I/O Virtualization (SR-IOV)
[...]
Kernel driver in use: LiquidIO
```

2. Load the kernel module:

```
modprobe liquidio
```

3. Bring up the PF ports:

```
ifconfig p4p1 up
ifconfig p4p2 up
```

4. Change PF MTU if required:

```
ifconfig p4p1 mtu 9000
ifconfig p4p2 mtu 9000
```

5. Create VF device(s):

Echo number of VFs to be created into "sriov_numvfs" sysfs entry of the parent PF.

```
echo 1 > /sys/bus/pci/devices/0000:03:00.0/sriov_numvfs
echo 1 > /sys/bus/pci/devices/0000:03:00.1/sriov_numvfs
```

6. Assign VF MAC address:

Assign MAC address to the VF using `iproute2` utility. The syntax is:

```
ip link set <PF iface> vf <VF id> mac <macaddr>
```

Example output:

```
ip link set p4p1 vf 0 mac F2:A8:1B:5E:B4:66
```

7. Assign VF(s) to VM.

The VF devices may be passed through to the guest VM using `qemu` or `virt-manager` or `virsh` etc.

Example `qemu` guest launch command:

```
./qemu-system-x86_64 -name lio-vm -machine accel=kvm \
-cpu host -m 4096 -smp 4 \
-drive file=<disk_file>,if=none,id=disk1,format=<type> \
-device virtio-blk-pci,scsi=off,drive=disk1,id=virtio-disk1,bootindex=1 \
-device vfio-pci,host=03:00.3 -device vfio-pci,host=03:08.3
```

8. Running testpmd

Refer to the document *compiling and testing a PMD for a NIC* to run `testpmd` application.

Note: Use `igb_uio` instead of `vfio-pci` in VM.

Example output:

```
[...]
EAL: PCI device 0000:03:00.3 on NUMA socket 0
EAL: probe driver: 177d:9712 net_liovf
EAL: using IOMMU type 1 (Type 1)
PMD: net_liovf[03:00.3]INFO: DEVICE : CN23XX VF
EAL: PCI device 0000:03:08.3 on NUMA socket 0
EAL: probe driver: 177d:9712 net_liovf
PMD: net_liovf[03:08.3]INFO: DEVICE : CN23XX VF
Interactive-mode selected
USER1: create a new mbuf pool <mbuf_pool_socket_0>: n=171456, size=2176, socket=0
Configuring Port 0 (socket 0)
PMD: net_liovf[03:00.3]INFO: Starting port 0
Port 0: F2:A8:1B:5E:B4:66
Configuring Port 1 (socket 0)
PMD: net_liovf[03:08.3]INFO: Starting port 1
Port 1: 32:76:CC:EE:56:D7
Checking link statuses...
Port 0 Link Up - speed 10000 Mbps - full-duplex
Port 1 Link Up - speed 10000 Mbps - full-duplex
Done
testpmd>
```

9. Enabling VF promiscuous mode

One VF per PF can be marked as trusted for promiscuous mode.

```
ip link set dev <PF iface> vf <VF id> trust on
```

20.4 Limitations

20.4.1 VF MTU

VF MTU is limited by PF MTU. Raise PF value before configuring VF for larger packet size.

20.4.2 VLAN offload

Tx VLAN insertion is not supported and consequently VLAN offload feature is marked partial.

20.4.3 Ring size

Number of descriptors for Rx/Tx ring should be in the range 128 to 512.

20.4.4 CRC striping

LiquidIO adapters strip ethernet FCS of every packet coming to the host interface. So, CRC will be stripped even when the `rxmode.hw_strip_crc` member is set to 0 in `struct rte_eth_conf`.

MLX4 POLL MODE DRIVER LIBRARY

The MLX4 poll mode driver library (`librte_pmd_mlx4`) implements support for **Mellanox ConnectX-3** and **Mellanox ConnectX-3 Pro** 10/40 Gbps adapters as well as their virtual functions (VF) in SR-IOV context.

Information and documentation about this family of adapters can be found on the [Mellanox website](#). Help is also provided by the [Mellanox community](#).

There is also a [section dedicated to this poll mode driver](#).

Note: Due to external dependencies, this driver is disabled by default. It must be enabled manually by setting `CONFIG_RTE_LIBRTE_MLX4_PMD=y` and recompiling DPDK.

21.1 Implementation details

Most Mellanox ConnectX-3 devices provide two ports but expose a single PCI bus address, thus unlike most drivers, `librte_pmd_mlx4` registers itself as a PCI driver that allocates one Ethernet device per detected port.

For this reason, one cannot white/blacklist a single port without also white/blacklisting the others on the same device.

Besides its dependency on `libibverbs` (that implies `libmlx4` and associated kernel support), `librte_pmd_mlx4` relies heavily on system calls for control operations such as querying/updating the MTU and flow control parameters.

For security reasons and robustness, this driver only deals with virtual memory addresses. The way resources allocations are handled by the kernel combined with hardware specifications that allow it to handle virtual memory addresses directly ensure that DPDK applications cannot access random physical memory (or memory that does not belong to the current process).

This capability allows the PMD to coexist with kernel network interfaces which remain functional, although they stop receiving unicast packets as long as they share the same MAC address.

Compiling `librte_pmd_mlx4` causes DPDK to be linked against `libibverbs`.

21.2 Configuration

21.2.1 Compilation options

These options can be modified in the `.config` file.

- `CONFIG_RTE_LIBRTE_MLX4_PMD` (default **n**)

Toggle compilation of `librte_pmd_mlx4` itself.

- `CONFIG_RTE_LIBRTE_MLX4_DLOPEN_DEPS` (default **n**)

Build PMD with additional code to make it loadable without hard dependencies on **libibverbs** nor **libmlx4**, which may not be installed on the target system.

In this mode, their presence is still required for it to run properly, however their absence won't prevent a DPDK application from starting (with `CONFIG_RTE_BUILD_SHARED_LIB` disabled) and they won't show up as missing with `ldd(1)`.

It works by moving these dependencies to a purpose-built rdma-core "glue" plug-in which must either be installed in a directory whose name is based on `CONFIG_RTE_EAL_PMD_PATH` suffixed with `-glue` if set, or in a standard location for the dynamic linker (e.g. `/lib`) if left to the default empty string (`"`).

This option has no performance impact.

- `CONFIG_RTE_LIBRTE_MLX4_DEBUG` (default **n**)

Toggle debugging code and stricter compilation flags. Enabling this option adds additional run-time checks and debugging messages at the cost of lower performance.

- `CONFIG_RTE_LIBRTE_MLX4_TX_MP_CACHE` (default **8**)

Maximum number of cached memory pools (MPs) per TX queue. Each MP from which buffers are to be transmitted must be associated to memory regions (MRs). This is a slow operation that must be cached.

This value is always 1 for RX queues since they use a single MP.

21.2.2 Environment variables

- `MLX4_GLUE_PATH`

A list of directories in which to search for the rdma-core "glue" plug-in, separated by colons or semi-colons.

Only matters when compiled with `CONFIG_RTE_LIBRTE_MLX4_DLOPEN_DEPS` enabled and most useful when `CONFIG_RTE_EAL_PMD_PATH` is also set, since `LD_LIBRARY_PATH` has no effect in this case.

21.2.3 Run-time configuration

- `librte_pmd_mlx4` brings kernel network interfaces up during initialization because it is affected by their state. Forcing them down prevents packets reception.
- **ethtool** operations on related kernel interfaces also affect the PMD.

- `port` parameter [int]

This parameter provides a physical port to probe and can be specified multiple times for additional ports. All ports are probed by default if left unspecified.

21.2.4 Kernel module parameters

The `mlx4_core` kernel module has several parameters that affect the behavior and/or the performance of `librte_pmd_mlx4`. Some of them are described below.

- **num_vfs** (integer or triplet, optionally prefixed by device address strings)

Create the given number of VFs on the specified devices.

- **log_num_mgm_entry_size** (integer)

Device-managed flow steering (DMFS) is required by DPDK applications. It is enabled by using a negative value, the last four bits of which have a special meaning.

- **-1**: force device-managed flow steering (DMFS).
- **-7**: configure optimized steering mode to improve performance with the following limitation: VLAN filtering is not supported with this mode. This is the recommended mode in case VLAN filter is not needed.

21.3 Prerequisites

This driver relies on external libraries and kernel drivers for resources allocations and initialization. The following dependencies are not part of DPDK and must be installed separately:

- **libibverbs** (provided by rdma-core package)

User space verbs framework used by `librte_pmd_mlx4`. This library provides a generic interface between the kernel and low-level user space drivers such as `libmlx4`.

It allows slow and privileged operations (context initialization, hardware resources allocations) to be managed by the kernel and fast operations to never leave user space.

- **libmlx4** (provided by rdma-core package)

Low-level user space driver library for Mellanox ConnectX-3 devices, it is automatically loaded by `libibverbs`.

This library basically implements send/receive calls to the hardware queues.

- **Kernel modules**

They provide the kernel-side verbs API and low level device drivers that manage actual hardware initialization and resources sharing with user space processes.

Unlike most other PMDs, these modules must remain loaded and bound to their devices:

- `mlx4_core`: hardware driver managing Mellanox ConnectX-3 devices.
- `mlx4_en`: Ethernet device driver that provides kernel network interfaces.
- `mlx4_ib`: InfiniBand device driver.
- `ib_uverbs`: user space driver for verbs (entry point for `libibverbs`).

- **Firmware update**

Mellanox OFED releases include firmware updates for ConnectX-3 adapters.

Because each release provides new features, these updates must be applied to match the kernel modules and libraries they come with.

Note: Both libraries are BSD and GPL licensed. Linux kernel modules are GPL licensed.

Depending on system constraints and user preferences either RDMA core library with a recent enough Linux kernel release (recommended) or Mellanox OFED, which provides compatibility with older releases.

21.3.1 Current RDMA core package and Linux kernel (recommended)

- Minimal Linux kernel version: 4.14.
- Minimal RDMA core version: v15 (see [RDMA core installation documentation](#)).

21.3.2 Mellanox OFED as a fallback

- Mellanox OFED version: **4.2, 4.3**.
- firmware version: **2.42.5000** and above.

Note: Several versions of Mellanox OFED are available. Installing the version this DPDK release was developed and tested against is strongly recommended. Please check the [pre-requisites](#).

Installing Mellanox OFED

1. Download latest Mellanox OFED.
2. Install the required libraries and kernel modules either by installing only the required set, or by installing the entire Mellanox OFED:

For bare metal use:

```
./mlnxofedinstall --dpdk --upstream-libs
```

For SR-IOV hypervisors use:

```
./mlnxofedinstall --dpdk --upstream-libs --enable-sriov --hypervisor
```

For SR-IOV virtual machine use:

```
./mlnxofedinstall --dpdk --upstream-libs --guest
```

3. Verify the firmware is the correct one:

```
ibv_devinfo
```

4. Set all ports links to Ethernet, follow instructions on the screen:

```
connectx_port_config
```

5. Continue with [section 2 of the Quick Start Guide](#).

21.4 Supported NICs

- Mellanox(R) ConnectX(R)-3 Pro 40G MCX354A-FCC_Ax (2*40G)

21.5 Quick Start Guide

1. Set all ports links to Ethernet

```
PCI=<NIC PCI address>
echo eth > "/sys/bus/pci/devices/$PCI/mlx4_port0"
echo eth > "/sys/bus/pci/devices/$PCI/mlx4_port1"
```

Note: If using Mellanox OFED one can permanently set the port link to Ethernet using `connectx_port_config` tool provided by it. *Mellanox OFED as a fallback:*

2. In case of bare metal or hypervisor, configure optimized steering mode by adding the following line to `/etc/modprobe.d/mlx4_core.conf`:

```
options mlx4_core log_num_mgm_entry_size=-7
```

Note: If VLAN filtering is used, set `log_num_mgm_entry_size=-1`. Performance degradation can occur on this case.

3. Restart the driver:

```
/etc/init.d/openibd restart
```

or:

```
service openibd restart
```

4. Compile DPDK and you are ready to go. See instructions on Development Kit Build System

21.6 Performance tuning

1. Verify the optimized steering mode is configured:

```
cat /sys/module/mlx4_core/parameters/log_num_mgm_entry_size
```

2. Use the CPU near local NUMA node to which the PCIe adapter is connected, for better performance. For VMs, verify that the right CPU and NUMA node are pinned according to the above. Run:

```
lstopo-no-graphics
```

to identify the NUMA node to which the PCIe adapter is connected.

3. If more than one adapter is used, and root complex capabilities allow to put both adapters on the same NUMA node without PCI bandwidth degradation, it is recommended to locate both adapters on the same NUMA node. This in order to forward packets from one to the other without NUMA performance penalty.

4. Disable pause frames:

```
ethtool -A <netdev> rx off tx off
```

5. Verify IO non-posted prefetch is disabled by default. This can be checked via the BIOS configuration. Please contact your server provider for more information about the settings.

Note: On some machines, depends on the machine integrator, it is beneficial to set the PCI max read request parameter to 1K. This can be done in the following way:

To query the read request size use:

```
setpci -s <NIC PCI address> 68.w
```

If the output is different than 3XXX, set it by:

```
setpci -s <NIC PCI address> 68.w=3XXX
```

The XXX can be different on different systems. Make sure to configure according to the setpci output.

21.7 Usage example

This section demonstrates how to launch **testpmd** with Mellanox ConnectX-3 devices managed by `librte_pmd_mlx4`.

1. Load the kernel modules:

```
modprobe -a ib_uverbs mlx4_en mlx4_core mlx4_ib
```

Alternatively if `MLNX_OFED` is fully installed, the following script can be run:

```
/etc/init.d/openibd restart
```

Note: User space I/O kernel modules (`uio` and `igb_uio`) are not used and do not have to be loaded.

2. Make sure Ethernet interfaces are in working order and linked to kernel verbs. Related `sysfs` entries should be present:

```
ls -d /sys/class/net/*/device/infiniband_verbs/uverbs* | cut -d / -f 5
```

Example output:

```
eth2
eth3
eth4
eth5
```

3. Optionally, retrieve their PCI bus addresses for whitelisting:

```
{
  for intf in eth2 eth3 eth4 eth5;
  do
    (cd "/sys/class/net/${intf}/device/" && pwd -P);
  done;
} |
sed -n 's,.*\/\(.*\),-w \1,p'
```

Example output:

```
-w 0000:83:00.0
-w 0000:83:00.0
-w 0000:84:00.0
-w 0000:84:00.0
```

Note: There are only two distinct PCI bus addresses because the Mellanox ConnectX-3 adapters installed on this system are dual port.

4. Request huge pages:

```
echo 1024 > /sys/kernel/mm/hugepages/hugepages-2048kB/nr_hugepages/nr_hugepages
```

5. Start testpmd with basic parameters:

```
testpmd -l 8-15 -n 4 -w 0000:83:00.0 -w 0000:84:00.0 -- --rxq=2 --txq=2 -i
```

Example output:

```
[...]
EAL: PCI device 0000:83:00.0 on NUMA socket 1
EAL: probe driver: 15b3:1007 librte_pmd_mlx4
PMD: librte_pmd_mlx4: PCI information matches, using device "mlx4_0" (VF: false)
PMD: librte_pmd_mlx4: 2 port(s) detected
PMD: librte_pmd_mlx4: port 1 MAC address is 00:02:c9:b5:b7:50
PMD: librte_pmd_mlx4: port 2 MAC address is 00:02:c9:b5:b7:51
EAL: PCI device 0000:84:00.0 on NUMA socket 1
EAL: probe driver: 15b3:1007 librte_pmd_mlx4
PMD: librte_pmd_mlx4: PCI information matches, using device "mlx4_1" (VF: false)
PMD: librte_pmd_mlx4: 2 port(s) detected
PMD: librte_pmd_mlx4: port 1 MAC address is 00:02:c9:b5:ba:b0
PMD: librte_pmd_mlx4: port 2 MAC address is 00:02:c9:b5:ba:b1
Interactive-mode selected
Configuring Port 0 (socket 0)
PMD: librte_pmd_mlx4: 0x867d60: TX queues number update: 0 -> 2
PMD: librte_pmd_mlx4: 0x867d60: RX queues number update: 0 -> 2
Port 0: 00:02:C9:B5:B7:50
Configuring Port 1 (socket 0)
PMD: librte_pmd_mlx4: 0x867da0: TX queues number update: 0 -> 2
PMD: librte_pmd_mlx4: 0x867da0: RX queues number update: 0 -> 2
Port 1: 00:02:C9:B5:B7:51
Configuring Port 2 (socket 0)
PMD: librte_pmd_mlx4: 0x867de0: TX queues number update: 0 -> 2
PMD: librte_pmd_mlx4: 0x867de0: RX queues number update: 0 -> 2
Port 2: 00:02:C9:B5:BA:B0
Configuring Port 3 (socket 0)
PMD: librte_pmd_mlx4: 0x867e20: TX queues number update: 0 -> 2
PMD: librte_pmd_mlx4: 0x867e20: RX queues number update: 0 -> 2
Port 3: 00:02:C9:B5:BA:B1
Checking link statuses...
Port 0 Link Up - speed 10000 Mbps - full-duplex
Port 1 Link Up - speed 40000 Mbps - full-duplex
Port 2 Link Up - speed 10000 Mbps - full-duplex
Port 3 Link Up - speed 40000 Mbps - full-duplex
Done
testpmd>
```

MLX5 POLL MODE DRIVER

The MLX5 poll mode driver library (`librte_pmd_mlx5`) provides support for **Mellanox ConnectX-4**, **Mellanox ConnectX-4 Lx** and **Mellanox ConnectX-5** families of 10/25/40/50/100 Gb/s adapters as well as their virtual functions (VF) in SR-IOV context.

Information and documentation about these adapters can be found on the [Mellanox website](#). Help is also provided by the [Mellanox community](#).

There is also a [section dedicated to this poll mode driver](#).

Note: Due to external dependencies, this driver is disabled by default. It must be enabled manually by setting `CONFIG_RTE_LIBRTE_MLX5_PMD=y` and recompiling DPDK.

22.1 Implementation details

Besides its dependency on `libibverbs` (that implies `libmlx5` and associated kernel support), `librte_pmd_mlx5` relies heavily on system calls for control operations such as querying/updating the MTU and flow control parameters.

For security reasons and robustness, this driver only deals with virtual memory addresses. The way resources allocations are handled by the kernel combined with hardware specifications that allow it to handle virtual memory addresses directly ensure that DPDK applications cannot access random physical memory (or memory that does not belong to the current process).

This capability allows the PMD to coexist with kernel network interfaces which remain functional, although they stop receiving unicast packets as long as they share the same MAC address. This means legacy linux control tools (for example: `ethtool`, `ifconfig` and more) can operate on the same network interfaces that owned by the DPDK application.

Enabling `librte_pmd_mlx5` causes DPDK applications to be linked against `libibverbs`.

22.2 Features

- Multi arch support: `x86_64`, `POWER8`, `ARMv8`.
- Multiple TX and RX queues.
- Support for scattered TX and RX frames.
- IPv4, IPv6, TCPv4, TCPv6, UDPv4 and UDPv6 RSS on any number of queues.
- Several RSS hash keys, one for each flow type.

- Configurable RETA table.
- Support for multiple MAC addresses.
- VLAN filtering.
- RX VLAN stripping.
- TX VLAN insertion.
- RX CRC stripping configuration.
- Promiscuous mode.
- Multicast promiscuous mode.
- Hardware checksum offloads.
- Flow director (RTE_FDIR_MODE_PERFECT, RTE_FDIR_MODE_PERFECT_MAC_VLAN and RTE_ETH_FDIR_REJECT).
- Flow API.
- Multiple process.
- KVM and VMware ESX SR-IOV modes are supported.
- RSS hash result is supported.
- Hardware TSO.
- Hardware checksum TX offload for VXLAN and GRE.
- RX interrupts.
- Statistics query including Basic, Extended and per queue.
- Rx HW timestamp.

22.3 Limitations

- Inner RSS for VXLAN frames is not supported yet.
- Hardware checksum RX offloads for VXLAN inner header are not supported yet.
- For secondary process:
 - Forked secondary process not supported.
 - All mempools must be initialized before `rte_eth_dev_start()`.

- Flow pattern without any specific vlan will match for vlan packets as well:

When VLAN spec is not specified in the pattern, the matching rule will be created with VLAN as a wild card. Meaning, the flow rule:

```
flow create 0 ingress pattern eth / vlan vid is 3 / ipv4 / end ...
```

Will only match vlan packets with vid=3. and the flow rules:

```
flow create 0 ingress pattern eth / ipv4 / end ...
```

Or:

```
flow create 0 ingress pattern eth / vlan / ipv4 / end ...
```

Will match any ipv4 packet (VLAN included).

- A multi segment packet must have less than 6 segments in case the Tx burst function is set to multi-packet send or Enhanced multi-packet send. Otherwise it must have less than 50 segments.
- Count action for RTE flow is **only supported in Mellanox OFED**.
- Flows with a VXLAN Network Identifier equal (or ends to be equal) to 0 are not supported.
- VXLAN TSO and checksum offloads are not supported on VM.

22.4 Statistics

MLX5 supports various of methods to report statistics:

Port statistics can be queried using `rte_eth_stats_get()`. The port statistics are through SW only and counts the number of packets received or sent successfully by the PMD.

Extended statistics can be queried using `rte_eth_xstats_get()`. The extended statistics expose a wider set of counters counted by the device. The extended port statistics counts the number of packets received or sent successfully by the port. As Mellanox NICs are using the Bifurcated Linux Driver those counters counts also packet received or sent by the Linux kernel. The counters with `_phy` suffix counts the total events on the physical port, therefore not valid for VF.

Finally per-flow statistics can be queried using `rte_flow_query` when attaching a count action for specific flow. The flow counter counts the number of packets received successfully by the port and match the specific flow.

22.5 Configuration

22.5.1 Compilation options

These options can be modified in the `.config` file.

- `CONFIG_RTE_LIBRTE_MLX5_PMD` (default **n**)

Toggle compilation of `librte_pmd_mlx5` itself.

- `CONFIG_RTE_LIBRTE_MLX5_DLOPEN_DEPS` (default **n**)

Build PMD with additional code to make it loadable without hard dependencies on **libibverbs** nor **libmlx5**, which may not be installed on the target system.

In this mode, their presence is still required for it to run properly, however their absence won't prevent a DPDK application from starting (with `CONFIG_RTE_BUILD_SHARED_LIB` disabled) and they won't show up as missing with `ldd(1)`.

It works by moving these dependencies to a purpose-built rdma-core "glue" plugin which must either be installed in a directory whose name is based on `CONFIG_RTE_EAL_PMD_PATH` suffixed with `-glue` if set, or in a standard location for the dynamic linker (e.g. `/lib`) if left to the default empty string (`"`).

This option has no performance impact.

- `CONFIG_RTE_LIBRTE_MLX5_DEBUG` (default **n**)

Toggle debugging code and stricter compilation flags. Enabling this option adds additional run-time checks and debugging messages at the cost of lower performance.

- `CONFIG_RTE_LIBRTE_MLX5_TX_MP_CACHE` (default **8**)

Maximum number of cached memory pools (MPs) per TX queue. Each MP from which buffers are to be transmitted must be associated to memory regions (MRs). This is a slow operation that must be cached.

This value is always 1 for RX queues since they use a single MP.

22.5.2 Environment variables

- `MLX5_GLUE_PATH`

A list of directories in which to search for the rdma-core “glue” plug-in, separated by colons or semi-colons.

Only matters when compiled with `CONFIG_RTE_LIBRTE_MLX5_DLOPEN_DEPS` enabled and most useful when `CONFIG_RTE_EAL_PMD_PATH` is also set, since `LD_LIBRARY_PATH` has no effect in this case.

- `MLX5_PMD_ENABLE_PADDING`

Enables HW packet padding in PCI bus transactions.

When packet size is cache aligned and CRC stripping is enabled, 4 fewer bytes are written to the PCI bus. Enabling padding makes such packets aligned again.

In cases where PCI bandwidth is the bottleneck, padding can improve performance by 10%.

This is disabled by default since this can also decrease performance for unaligned packet sizes.

- `MLX5_SHUT_UP_BF`

Configures HW Tx doorbell register as IO-mapped.

By default, the HW Tx doorbell is configured as a write-combining register. The register would be flushed to HW usually when the write-combining buffer becomes full, but it depends on CPU design.

Except for vectorized Tx burst routines, a write memory barrier is enforced after updating the register so that the update can be immediately visible to HW.

When vectorized Tx burst is called, the barrier is set only if the burst size is not aligned to `MLX5_VPMD_TX_MAX_BURST`. However, setting this environmental variable will bring better latency even though the maximum throughput can slightly decline.

22.5.3 Run-time configuration

- `librte_pmd_mlx5` brings kernel network interfaces up during initialization because it is affected by their state. Forcing them down prevents packets reception.

- **ethtool** operations on related kernel interfaces also affect the PMD.

- `rxq_cqe_comp_en` parameter [int]

A nonzero value enables the compression of CQE on RX side. This feature allows to save PCI bandwidth and improve performance. Enabled by default.

Supported on:

- x86_64 with ConnectX-4, ConnectX-4 LX and ConnectX-5.
- POWER8 and ARMv8 with ConnectX-4 LX and ConnectX-5.

- `txq_inline` parameter [int]

Amount of data to be inlined during TX operations. Improves latency. Can improve PPS performance when PCI back pressure is detected and may be useful for scenarios involving heavy traffic on many queues.

Because additional software logic is necessary to handle this mode, this option should be used with care, as it can lower performance when back pressure is not expected.

- `txqs_min_inline` parameter [int]

Enable inline send only when the number of TX queues is greater or equal to this value.

This option should be used in combination with `txq_inline` above.

On ConnectX-4, ConnectX-4 LX and ConnectX-5 without Enhanced MPW:

- Disabled by default.
- In case `txq_inline` is set recommendation is 4.

On ConnectX-5 with Enhanced MPW:

- Set to 8 by default.

- `txq_mpw_en` parameter [int]

A nonzero value enables multi-packet send (MPS) for ConnectX-4 Lx and enhanced multi-packet send (Enhanced MPS) for ConnectX-5. MPS allows the TX burst function to pack up multiple packets in a single descriptor session in order to save PCI bandwidth and improve performance at the cost of a slightly higher CPU usage. When `txq_inline` is set along with `txq_mpw_en`, TX burst function tries to copy entire packet data on to TX descriptor instead of including pointer of packet only if there is enough room remained in the descriptor. `txq_inline` sets per-descriptor space for either pointers or inlined packets. In addition, Enhanced MPS supports hybrid mode - mixing inlined packets and pointers in the same descriptor.

This option cannot be used with certain offloads such as `DEV_TX_OFFLOAD_TCP_TSO`, `DEV_TX_OFFLOAD_VXLAN_TNL_TSO`, `DEV_TX_OFFLOAD_GRE_TNL_TSO`, `DEV_TX_OFFLOAD_VLAN_INSERT`. When those offloads are requested the MPS send function will not be used.

It is currently only supported on the ConnectX-4 Lx and ConnectX-5 families of adapters. Enabled by default.

- `txq_mpw_hdr_dseg_en` parameter [int]

A nonzero value enables including two pointers in the first block of TX descriptor. This can be used to lessen CPU load for memory copy.

Effective only when Enhanced MPS is supported. Disabled by default.

- `txq_max_inline_len` parameter [int]

Maximum size of packet to be inlined. This limits the size of packet to be inlined. If the size of a packet is larger than configured value, the packet isn't inlined even though there's enough space remained in the descriptor. Instead, the packet is included with pointer.

Effective only when Enhanced MPS is supported. The default value is 256.

- `tx_vec_en` parameter [int]

A nonzero value enables Tx vector on ConnectX-5 only NIC if the number of global Tx queues on the port is lesser than `MLX5_VPMD_MIN_TXQS`.

This option cannot be used with certain offloads such as `DEV_TX_OFFLOAD_TCP_TSO`, `DEV_TX_OFFLOAD_VXLAN_TNL_TSO`, `DEV_TX_OFFLOAD_GRE_TNL_TSO`, `DEV_TX_OFFLOAD_VLAN_INSERT`. When those offloads are requested the MPS send function will not be used.

Enabled by default on ConnectX-5.

- `rx_vec_en` parameter [int]

A nonzero value enables Rx vector if the port is not configured in multi-segment otherwise this parameter is ignored.

Enabled by default.

22.6 Prerequisites

This driver relies on external libraries and kernel drivers for resources allocations and initialization. The following dependencies are not part of DPDK and must be installed separately:

- **libibverbs**

User space Verbs framework used by `librte_pmd_mlx5`. This library provides a generic interface between the kernel and low-level user space drivers such as `libmlx5`.

It allows slow and privileged operations (context initialization, hardware resources allocations) to be managed by the kernel and fast operations to never leave user space.

- **libmlx5**

Low-level user space driver library for Mellanox ConnectX-4/ConnectX-5 devices, it is automatically loaded by `libibverbs`.

This library basically implements send/receive calls to the hardware queues.

- **Kernel modules**

They provide the kernel-side Verbs API and low level device drivers that manage actual hardware initialization and resources sharing with user space processes.

Unlike most other PMDs, these modules must remain loaded and bound to their devices:

- `mlx5_core`: hardware driver managing Mellanox ConnectX-4/ConnectX-5 devices and related Ethernet kernel network devices.

- mlx5_ib: InfiniBand device driver.
- ib_uverbs: user space driver for Verbs (entry point for libibverbs).

- **Firmware update**

Mellanox OFED releases include firmware updates for ConnectX-4/ConnectX-5 adapters.

Because each release provides new features, these updates must be applied to match the kernel modules and libraries they come with.

Note: Both libraries are BSD and GPL licensed. Linux kernel modules are GPL licensed.

22.6.1 Installation

Either RDMA Core library with a recent enough Linux kernel release (recommended) or Mellanox OFED, which provides compatibility with older releases.

RDMA Core with Linux Kernel

- Minimal kernel version : v4.14 or the most recent 4.14-rc (see [Linux installation documentation](#))
- Minimal rdma-core version: v15+ commit 0c5f5765213a (“Merge pull request #227 from yishaih/tm”) (see [RDMA Core installation documentation](#))

Mellanox OFED

- Mellanox OFED version: **4.2, 4.3**.
- firmware version:
 - ConnectX-4: **12.21.1000** and above.
 - ConnectX-4 Lx: **14.21.1000** and above.
 - ConnectX-5: **16.21.1000** and above.
 - ConnectX-5 Ex: **16.21.1000** and above.

While these libraries and kernel modules are available on OpenFabrics Alliance’s [website](#) and provided by package managers on most distributions, this PMD requires Ethernet extensions that may not be supported at the moment (this is a work in progress).

[Mellanox OFED](#) includes the necessary support and should be used in the meantime. For DPDK, only libibverbs, libmlx5, mlnx-ofed-kernel packages and firmware updates are required from that distribution.

Note: Several versions of Mellanox OFED are available. Installing the version this DPDK release was developed and tested against is strongly recommended. Please check the [pre-requisites](#).

22.7 Supported NICs

- Mellanox(R) ConnectX(R)-4 10G MCX4111A-XCAT (1x10G)
- Mellanox(R) ConnectX(R)-4 10G MCX4121A-XCAT (2x10G)
- Mellanox(R) ConnectX(R)-4 25G MCX4111A-ACAT (1x25G)
- Mellanox(R) ConnectX(R)-4 25G MCX4121A-ACAT (2x25G)
- Mellanox(R) ConnectX(R)-4 40G MCX4131A-BCAT (1x40G)
- Mellanox(R) ConnectX(R)-4 40G MCX413A-BCAT (1x40G)
- Mellanox(R) ConnectX(R)-4 40G MCX415A-BCAT (1x40G)
- Mellanox(R) ConnectX(R)-4 50G MCX4131A-GCAT (1x50G)
- Mellanox(R) ConnectX(R)-4 50G MCX413A-GCAT (1x50G)
- Mellanox(R) ConnectX(R)-4 50G MCX414A-BCAT (2x50G)
- Mellanox(R) ConnectX(R)-4 50G MCX415A-GCAT (2x50G)
- Mellanox(R) ConnectX(R)-4 50G MCX416A-BCAT (2x50G)
- Mellanox(R) ConnectX(R)-4 50G MCX416A-GCAT (2x50G)
- Mellanox(R) ConnectX(R)-4 50G MCX415A-CCAT (1x100G)
- Mellanox(R) ConnectX(R)-4 100G MCX416A-CCAT (2x100G)
- Mellanox(R) ConnectX(R)-4 Lx 10G MCX4121A-XCAT (2x10G)
- Mellanox(R) ConnectX(R)-4 Lx 25G MCX4121A-ACAT (2x25G)
- Mellanox(R) ConnectX(R)-5 100G MCX556A-ECAT (2x100G)
- Mellanox(R) ConnectX(R)-5 Ex EN 100G MCX516A-CDAT (2x100G)

22.8 Quick Start Guide on OFED

1. Download latest Mellanox OFED. For more info check the [prerequisites](#).
2. Install the required libraries and kernel modules either by installing only the required set, or by installing the entire Mellanox OFED:

```
./mlnxofedinstall --upstream-libs --dpdk
```

3. Verify the firmware is the correct one:

```
ibv_devinfo
```

4. Verify all ports links are set to Ethernet:

```
mlxconfig -d <mst device> query | grep LINK_TYPE
LINK_TYPE_P1          ETH (2)
LINK_TYPE_P2          ETH (2)
```

Link types may have to be configured to Ethernet:

```
mlxconfig -d <mst device> set LINK_TYPE_P1/2=1/2/3
```

```
* LINK_TYPE_P1=<1|2|3> , 1=Infiniband 2=Ethernet 3=VPI(auto-sense)
```

For hypervisors verify SR-IOV is enabled on the NIC:

```
mlxconfig -d <mst device> query | grep SRIOV_EN
SRIOV_EN                               True(1)
```

If needed, set enable the set the relevant fields:

```
mlxconfig -d <mst device> set SRIOV_EN=1 NUM_OF_VFS=16
mlxfwreset -d <mst device> reset
```

5. Restart the driver:

```
/etc/init.d/openibd restart
```

or:

```
service openibd restart
```

If link type was changed, firmware must be reset as well:

```
mlxfwreset -d <mst device> reset
```

For hypervisors, after reset write the sysfs number of virtual functions needed for the PF.

To dynamically instantiate a given number of virtual functions (VFs):

```
echo [num_vfs] > /sys/class/infiniband/mlx5_0/device/sriov_numvfs
```

6. Compile DPDK and you are ready to go. See instructions on Development Kit Build System

22.9 Performance tuning

1. Configure aggressive CQE Zipping for maximum performance:

```
mlxconfig -d <mst device> s CQE_COMPRESSION=1
```

To set it back to the default CQE Zipping mode use:

```
mlxconfig -d <mst device> s CQE_COMPRESSION=0
```

2. In case of virtualization:

- Make sure that hypervisor kernel is 3.16 or newer.
- Configure boot with `iommu=pt`.
- Use 1G huge pages.
- Make sure to allocate a VM on huge pages.
- Make sure to set CPU pinning.

3. Use the CPU near local NUMA node to which the PCIe adapter is connected, for better performance. For VMs, verify that the right CPU and NUMA node are pinned according to the above. Run:

```
lstopo-no-graphics
```

to identify the NUMA node to which the PCIe adapter is connected.

4. If more than one adapter is used, and root complex capabilities allow to put both adapters on the same NUMA node without PCI bandwidth degradation, it is recommended to locate both adapters on the same NUMA node. This in order to forward packets from one to the other without NUMA performance penalty.

5. Disable pause frames:

```
ethtool -A <netdev> rx off tx off
```

6. Verify IO non-posted prefetch is disabled by default. This can be checked via the BIOS configuration. Please contact your server provider for more information about the settings.

Note: On some machines, depends on the machine integrator, it is beneficial to set the PCI max read request parameter to 1K. This can be done in the following way:

To query the read request size use:

```
setpci -s <NIC PCI address> 68.w
```

If the output is different than 3XXX, set it by:

```
setpci -s <NIC PCI address> 68.w=3XXX
```

The XXX can be different on different systems. Make sure to configure according to the setpci output.

22.10 Notes for testpmd

Compared to `librte_pmd_mlx4` that implements a single RSS configuration per port, `librte_pmd_mlx5` supports per-protocol RSS configuration.

Since `testpmd` defaults to IP RSS mode and there is currently no command-line parameter to enable additional protocols (UDP and TCP as well as IP), the following commands must be entered from its CLI to get the same behavior as `librte_pmd_mlx4`:

```
> port stop all
> port config all rss all
> port start all
```

22.11 Usage example

This section demonstrates how to launch **testpmd** with Mellanox ConnectX-4/ConnectX-5 devices managed by `librte_pmd_mlx5`.

1. Load the kernel modules:

```
modprobe -a ib_uverbs mlx5_core mlx5_ib
```

Alternatively if `MLNX_OFED` is fully installed, the following script can be run:

```
/etc/init.d/openibd restart
```

Note: User space I/O kernel modules (`uio` and `igb_uio`) are not used and do not have to be loaded.

2. Make sure Ethernet interfaces are in working order and linked to kernel verbs. Related `sysfs` entries should be present:

```
ls -ld /sys/class/net/*/device/infiniband_verbs/uverbs* | cut -d / -f 5
```

Example output:

```
eth30
eth31
eth32
eth33
```

3. Optionally, retrieve their PCI bus addresses for whitelisting:

```
{
  for intf in eth2 eth3 eth4 eth5;
  do
    (cd "/sys/class/net/${intf}/device/" && pwd -P);
  done;
} |
sed -n 's,.*\/\(.*\),-w \1,p'
```

Example output:

```
-w 0000:05:00.1
-w 0000:06:00.0
-w 0000:06:00.1
-w 0000:05:00.0
```

4. Request huge pages:

```
echo 1024 > /sys/kernel/mm/hugepages/hugepages-2048kB/nr_hugepages/nr_hugepages
```

5. Start testpmd with basic parameters:

```
testpmd -l 8-15 -n 4 -w 05:00.0 -w 05:00.1 -w 06:00.0 -w 06:00.1 -- --rxq=2 --txq=2 -i
```

Example output:

```
[...]
EAL: PCI device 0000:05:00.0 on NUMA socket 0
EAL: probe driver: 15b3:1013 librte_pmd_mlx5
PMD: librte_pmd_mlx5: PCI information matches, using device "mlx5_0" (VF: false)
PMD: librte_pmd_mlx5: 1 port(s) detected
PMD: librte_pmd_mlx5: port 1 MAC address is e4:1d:2d:e7:0c:fe
EAL: PCI device 0000:05:00.1 on NUMA socket 0
EAL: probe driver: 15b3:1013 librte_pmd_mlx5
PMD: librte_pmd_mlx5: PCI information matches, using device "mlx5_1" (VF: false)
PMD: librte_pmd_mlx5: 1 port(s) detected
PMD: librte_pmd_mlx5: port 1 MAC address is e4:1d:2d:e7:0c:ff
EAL: PCI device 0000:06:00.0 on NUMA socket 0
EAL: probe driver: 15b3:1013 librte_pmd_mlx5
PMD: librte_pmd_mlx5: PCI information matches, using device "mlx5_2" (VF: false)
PMD: librte_pmd_mlx5: 1 port(s) detected
PMD: librte_pmd_mlx5: port 1 MAC address is e4:1d:2d:e7:0c:fa
EAL: PCI device 0000:06:00.1 on NUMA socket 0
EAL: probe driver: 15b3:1013 librte_pmd_mlx5
PMD: librte_pmd_mlx5: PCI information matches, using device "mlx5_3" (VF: false)
PMD: librte_pmd_mlx5: 1 port(s) detected
PMD: librte_pmd_mlx5: port 1 MAC address is e4:1d:2d:e7:0c:fb
Interactive-mode selected
Configuring Port 0 (socket 0)
PMD: librte_pmd_mlx5: 0x8cba80: TX queues number update: 0 -> 2
PMD: librte_pmd_mlx5: 0x8cba80: RX queues number update: 0 -> 2
Port 0: E4:1D:2D:E7:0C:FE
Configuring Port 1 (socket 0)
PMD: librte_pmd_mlx5: 0x8ccac8: TX queues number update: 0 -> 2
PMD: librte_pmd_mlx5: 0x8ccac8: RX queues number update: 0 -> 2
Port 1: E4:1D:2D:E7:0C:FF
Configuring Port 2 (socket 0)
PMD: librte_pmd_mlx5: 0x8cdb10: TX queues number update: 0 -> 2
PMD: librte_pmd_mlx5: 0x8cdb10: RX queues number update: 0 -> 2
Port 2: E4:1D:2D:E7:0C:FA
```

```
Configuring Port 3 (socket 0)
PMD: librte_pmd_mlx5: 0x8ceb58: TX queues number update: 0 -> 2
PMD: librte_pmd_mlx5: 0x8ceb58: RX queues number update: 0 -> 2
Port 3: E4:1D:2D:E7:0C:FB
Checking link statuses...
Port 0 Link Up - speed 40000 Mbps - full-duplex
Port 1 Link Up - speed 40000 Mbps - full-duplex
Port 2 Link Up - speed 10000 Mbps - full-duplex
Port 3 Link Up - speed 10000 Mbps - full-duplex
Done
testpmd>
```


MRVL POLL MODE DRIVER

The MRVL PMD (`librte_pmd_mrvl`) provides poll mode driver support for the Marvell PPv2 (Packet Processor v2) 1/10 Gbps adapter.

Detailed information about SoCs that use PPv2 can be obtained here:

- <https://www.marvell.com/embedded-processors/armada-70xx/>
- <https://www.marvell.com/embedded-processors/armada-80xx/>

Note: Due to external dependencies, this driver is disabled by default. It must be enabled manually by setting relevant configuration option manually. Please refer to *Config File Options* section for further details.

23.1 Features

Features of the MRVL PMD are:

- Speed capabilities
- Link status
- Queue start/stop
- MTU update
- Jumbo frame
- Promiscuous mode
- Allmulticast mode
- Unicast MAC filter
- Multicast MAC filter
- RSS hash
- VLAN filter
- CRC offload
- L3 checksum offload
- L4 checksum offload
- Packet type parsing

- Basic stats
- QoS

23.2 Limitations

- Number of lcores is limited to 9 by MUSDK internal design. If more lcores need to be allocated, locking will have to be considered. Number of available lcores can be changed via `MRVL_MUSDK_HIFS_RESERVED` define in `mrvl_ethdev.c` source file.
- Flushing vlans added for filtering is not possible due to MUSDK missing functionality. Current workaround is to reset board so that PPv2 has a chance to start in a sane state.

23.3 Prerequisites

- Custom Linux Kernel sources

```
git clone https://github.com/MarvellEmbeddedProcessors/linux-marvell.git -b linux-4.4.52-a
```

- Out of tree `mvpp2x_sysfs` kernel module sources

```
git clone https://github.com/MarvellEmbeddedProcessors/mvpp2x-marvell.git -b mvpp2x-armada-1
```

- MUSDK (Marvell User-Space SDK) sources

```
git clone https://github.com/MarvellEmbeddedProcessors/musdk-marvell.git -b musdk-armada-1
```

MUSDK is a light-weight library that provides direct access to Marvell's PPv2 (Packet Processor v2). Alternatively prebuilt MUSDK library can be requested from [Marvell Extranet](#). Once approval has been granted, library can be found by typing `musdk` in the search box.

MUSDK must be configured with the following features:

```
--enable-bpool-dma=64
```

- DPDK environment

Follow the DPDK Getting Started Guide for Linux to setup DPDK environment.

23.4 Config File Options

The following options can be modified in the `config` file.

- `CONFIG_RTE_LIBRTE_MRVL_PMD` (default n)
Toggle compilation of the `librte_pmd_mrvl` driver.

23.5 QoS Configuration

QoS configuration is done through external configuration file. Path to the file must be given as `cfg` in driver's `vdev` parameter list.

23.5.1 Configuration syntax

```
[port <portnum> default]
default_tc = <default_tc>
mapping_priority = <mapping_priority>
```

```
[port <portnum> tc <traffic_class>]
rxq = <rx_queue_list>
pcp = <pcp_list>
dscp = <dscp_list>
```

```
[port <portnum> tc <traffic_class>]
rxq = <rx_queue_list>
pcp = <pcp_list>
dscp = <dscp_list>
```

Where:

- <portnum>: DPDK Port number (0..n).
- <default_tc>: Default traffic class (e.g. 0)
- <mapping_priority>: QoS priority for mapping (*ip*, *vlan*, *ip/vlan* or *vlan/ip*).
- <traffic_class>: Traffic Class to be configured.
- <rx_queue_list>: List of DPDK RX queues (e.g. 0 1 3-4)
- <pcp_list>: List of PCP values to handle in particular TC (e.g. 0 1 3-4 7).
- <dscp_list>: List of DSCP values to handle in particular TC (e.g. 0-12 32-48 63).

Setting PCP/DSCP values for the default TC is not required. All PCP/DSCP values not assigned explicitly to particular TC will be handled by the default TC.

Configuration file example

```
[port 0 default]
default_tc = 0
qos_mode = ip

[port 0 tc 0]
rxq = 0 1

[port 0 tc 1]
rxq = 2
pcp = 5 6 7
dscp = 26-38

[port 1 default]
default_tc = 0
qos_mode = vlan/ip

[port 1 tc 0]
rxq = 0

[port 1 tc 1]
rxq = 1 2
pcp = 5 6 7
dscp = 26-38
```

Usage example

```
./testpmd --vdev=eth_mrvl,iface=eth0,iface=eth2,cfg=/home/user/mrvl.conf \
-c 7 -- -i -a --rxq=2
```

23.6 Building DPDK

Driver needs precompiled MUSDK library during compilation.

```
export CROSS_COMPILE=<toolchain>/bin/aarch64-linux-gnu-
./bootstrap
./configure --host=aarch64-linux-gnu --enable-bpool-dma=64
make install
```

MUSDK will be installed to *usr/local* under current directory. For the detailed build instructions please consult *doc/musdk_get_started.txt*.

Before the DPDK build process the environmental variable `LIBMUSDK_PATH` with the path to the MUSDK installation directory needs to be exported.

```
export LIBMUSDK_PATH=<musdk>/usr/local
export CROSS=aarch64-linux-gnu-
make config T=arm64-armv8a-linuxapp-gcc
sed -ri 's,(MRVL_PMD=)n,\1y,' build/.config
make
```

23.7 Usage Example

MRVL PMD requires extra out of tree kernel modules to function properly. *musdk_uio* and *mv_pp_uio* sources are part of the MUSDK. Please consult *doc/musdk_get_started.txt* for the detailed build instructions. For *mvpp2x_sysfs* please consult *Documentation/pp22_sysfs.txt* for the detailed build instructions.

```
insmod musdk_uio.ko
insmod mv_pp_uio.ko
insmod mvpp2x_sysfs.ko
```

Additionally interfaces used by DPDK application need to be put up:

```
ip link set eth0 up
ip link set eth2 up
```

In order to run *testpmd* example application following command can be used:

```
./testpmd --vdev=eth_mrvl,iface=eth0,iface=eth2 -c 7 -- \
--burst=128 --txd=2048 --rxd=1024 --rxq=2 --txq=2 --nb-cores=2 \
-i -a --rss-udp
```

NFP POLL MODE DRIVER LIBRARY

Netronome's sixth generation of flow processors pack 216 programmable cores and over 100 hardware accelerators that uniquely combine packet, flow, security and content processing in a single device that scales up to 400-Gb/s.

This document explains how to use DPDK with the Netronome Poll Mode Driver (PMD) supporting Netronome's Network Flow Processor 6xxx (NFP-6xxx) and Netronome's Flow Processor 4xxx (NFP-4xxx).

NFP is a SRIOV capable device and the PMD driver supports the physical function (PF) and the virtual functions (VFs).

24.1 Dependencies

Before using the Netronome's DPDK PMD some NFP configuration, which is not related to DPDK, is required. The system requires installation of **Netronome's BSP (Board Support Package)** along with a specific NFP firmware application. Netronome's NSP ABI version should be 0.20 or higher.

If you have a NFP device you should already have the code and documentation for this configuration. Contact support@netronome.com to obtain the latest available firmware.

The NFP Linux netdev kernel driver for VFs has been a part of the vanilla kernel since kernel version 4.5, and support for the PF since kernel version 4.11. Support for older kernels can be obtained on Github at <https://github.com/Netronome/nfp-driv-kmods> along with the build instructions.

NFP PMD needs to be used along with UIO `igb_uio` or VFIO (`vfio-pci`) Linux kernel driver.

24.2 Building the software

Netronome's PMD code is provided in the `drivers/net/nfp` directory. Although NFP PMD has Netronome's BSP dependencies, it is possible to compile it along with other DPDK PMDs even if no BSP was installed previously. Of course, a DPDK app will require such a BSP installed for using the NFP PMD, along with a specific NFP firmware application.

Default PMD configuration is at the `common_linuxapp configuration` file:

- `CONFIG_RTE_LIBRTE_NFP_PMD=y`

Once the DPDK is built all the DPDK apps and examples include support for the NFP PMD.

24.3 Driver compilation and testing

Refer to the document *compiling and testing a PMD for a NIC* for details.

24.4 Using the PF

NFP PMD supports using the NFP PF as another DPDK port, but it does not have any functionality for controlling VFs. In fact, it is not possible to use the PMD with the VFs if the PF is being used by DPDK, that is, with the NFP PF bound to `igb_uio` or `vfio-pci` kernel drivers. Future DPDK versions will have a PMD able to work with the PF and VFs at the same time and with the PF implementing VF management along with other PF-only functionalities/offloads.

The PMD PF has extra work to do which will delay the DPDK app initialization. This additional effort could be checking if a firmware is already available in the device, uploading the firmware if necessary or configuring the Link state properly when starting or stopping a PF port. Note that firmware upload is not always necessary which is the main delay for NFP PF PMD initialization.

Depending on the Netronome product installed in the system, firmware files should be available under `/lib/firmware/netronome`. DPDK PMD supporting the PF requires a specific link, `/lib/firmware/netronome/nic_dpdk_default.nffw`, which should be created automatically with Netronome's Agilio products installation.

24.5 PF multiport support

Some NFP cards support several physical ports with just one single PCI device. The DPDK core is designed with a 1:1 relationship between PCI devices and DPDK ports, so NFP PMD PF support requires handling the multiport case specifically. During NFP PF initialization, the PMD will extract the information about the number of PF ports from the firmware and will create as many DPDK ports as needed.

Because the unusual relationship between a single PCI device and several DPDK ports, there are some limitations when using more than one PF DPDK port: there is no support for RX interrupts and it is not possible either to use those PF ports with the device hotplug functionality.

24.6 System configuration

1. **Enable SR-IOV on the NFP device:** The current NFP PMD supports the PF and the VFs on a NFP device. However, it is not possible to work with both at the same time because the VFs require the PF being bound to the NFP PF Linux netdev driver. Make sure you are working with a kernel with NFP PF support or get the drivers from the above Github repository and follow the instructions for building and installing it.

VFs need to be enabled before they can be used with the PMD. Before enabling the VFs it is useful to obtain information about the current NFP PCI device detected by the system:

```
lspci -d19ee:
```

Now, for example, configure two virtual functions on a NFP-6xxx device whose PCI system identity is "0000:03:00.0":

```
echo 2 > /sys/bus/pci/devices/0000:03:00.0/sriov_numvfs
```

The result of this command may be shown using lspci again:

```
lspci -d19ee: -k
```

Two new PCI devices should appear in the output of the above command. The -k option shows the device driver, if any, that devices are bound to. Depending on the modules loaded at this point the new PCI devices may be bound to nfp_netvf driver.

OCTEONTX POLL MODE DRIVER

The OCTEONTX ETHDEV PMD (`librte_pmd_octeontx`) provides poll mode ethdev driver support for the inbuilt network device found in the **Cavium OCTEONTX** SoC family as well as their virtual functions (VF) in SR-IOV context.

More information can be found at [Cavium, Inc Official Website](#).

25.1 Features

Features of the OCTEONTX Ethdev PMD are:

- Packet type information
- Promiscuous mode
- Port hardware statistics
- Jumbo frames
- Link state information
- SR-IOV VF
- Multiple queues for TX
- Lock-free Tx queue
- HW offloaded *ethdev Rx queue* to *eventdev event queue* packet injection

25.2 Supported OCTEONTX SoCs

- CN83xx

25.3 Unsupported features

The features supported by the device and not yet supported by this PMD include:

- Receive Side Scaling (RSS)
- Scattered and gather for TX and RX
- Ingress classification support

- Egress hierarchical scheduling, traffic shaping, and marking

25.4 Prerequisites

See `../platform/octeontx` for setup information.

25.5 Pre-Installation Configuration

25.5.1 Config File Options

The following options can be modified in the `config` file. Please note that enabling debugging options may affect system performance.

- `CONFIG_RTE_LIBRTE_OCTEONTX_PMD` (default `y`)
Toggle compilation of the `librte_pmd_octeontx` driver.

25.5.2 Driver compilation and testing

Refer to the document *compiling and testing a PMD for a NIC* for details.

To compile the OCTEONTX PMD for Linux arm64 gcc target, run the following `make` command:

```
cd <DPDK-source-directory>
make config T=arm64-thunderx-linuxapp-gcc install
```

1. Running testpmd:

Follow instructions available in the document *compiling and testing a PMD for a NIC* to run `testpmd`.

Example output:

```
./arm64-thunderx-linuxapp-gcc/app/testpmd -c 700 \
    --base-virtaddr=0x100000000000 \
    --mbuf-pool-ops-name="octeontx_fpavf" \
    --vdev='event_octeontx' \
    --vdev='eth_octeontx,nr_port=2' \
    -- --rxq=1 --txq=1 --nb-core=2 \
    --total-num-mbufs=16384 -i
.....
EAL: Detected 24 lcore(s)
EAL: Probing VFIO support...
EAL: VFIO support initialized
.....
EAL: PCI device 0000:07:00.1 on NUMA socket 0
EAL:   probe driver: 177d:a04b octeontx_ssov
.....
EAL: PCI device 0001:02:00.7 on NUMA socket 0
EAL:   probe driver: 177d:a0dd octeontx_pkivf
.....
EAL: PCI device 0001:03:01.0 on NUMA socket 0
EAL:   probe driver: 177d:a049 octeontx_pkovf
.....
PMD: octeontx_probe(): created ethdev eth_octeontx for port 0
PMD: octeontx_probe(): created ethdev eth_octeontx for port 1
.....
```

```

Configuring Port 0 (socket 0)
Port 0: 00:0F:B7:11:94:46
Configuring Port 1 (socket 0)
Port 1: 00:0F:B7:11:94:47
.....
Checking link statuses...
Port 0 Link Up - speed 40000 Mbps - full-duplex
Port 1 Link Up - speed 40000 Mbps - full-duplex
Done
testpmd>

```

25.6 Initialization

The `octeontx ethdev pmd` is exposed as a vdev device which consists of a set of PKI and PKO PCIe VF devices. On EAL initialization, PKI/PKO PCIe VF devices will be probed and then the vdev device can be created from the application code, or from the EAL command line based on the number of probed/bound PKI/PKO PCIe VF device to DPDK by

- Invoking `rte_vdev_init("eth_octeontx")` from the application
- Using `--vdev="eth_octeontx"` in the EAL options, which will call `rte_vdev_init()` internally

25.6.1 Device arguments

Each ethdev port is mapped to a physical port(LMAC), Application can specify the number of interesting ports with `nr_ports` argument.

25.6.2 Dependency

`eth_octeontx pmd` is depend on `event_octeontx eventdev` device and `octeontx_fpvavf` external mempool handler.

Example:

```

./your_dpdk_application --mbuf-pool-ops-name="octeontx_fpvavf" \
--vdev='event_octeontx' \
--vdev="eth_octeontx,nr_port=2"

```

25.7 Limitations

25.7.1 `octeontx_fpvavf` external mempool handler dependency

The OCTEONTX SoC family NIC has inbuilt HW assisted external mempool manager. This driver will only work with `octeontx_fpvavf` external mempool handler as it is the most performance effective way for packet allocation and Tx buffer recycling on OCTEONTX SoC platform.

25.7.2 CRC striping

The OCTEONTX SoC family NICs strip the CRC for every packets coming into the host interface. So, CRC will be stripped even when the `rxmode.hw_strip_crc` member is set to 0 in `struct rte_eth_conf`.

25.7.3 Maximum packet length

The OCTEONTX SoC family NICs support a maximum of a 32K jumbo frame. The value is fixed and cannot be changed. So, even when the `rxmode.max_rx_pkt_len` member of `struct rte_eth_conf` is set to a value lower than 32k, frames up to 32k bytes can still reach the host interface.

QEDE POLL MODE DRIVER

The QEDE poll mode driver library (`librte_pmd_qede`) implements support for **QLogic FastLinQ QL4xxxx 10G/25G/40G/50G/100G Intelligent Ethernet Adapters (IEA) and Converged Network Adapters (CNA)** family of adapters as well as SR-IOV virtual functions (VF). It is supported on several standard Linux distros like RHEL7.x, SLES12.x and Ubuntu. It is compile-tested under FreeBSD OS.

More information can be found at [QLogic Corporation's Website](#).

26.1 Supported Features

- Unicast/Multicast filtering
- Promiscuous mode
- Allmulti mode
- Port hardware statistics
- Jumbo frames
- Multiple MAC address
- MTU change
- Default pause flow control
- Multiprocess aware
- Scatter-Gather
- Multiple Rx/Tx queues
- RSS (with RETA/hash table/key)
- TSS
- Stateless checksum offloads (IPv4/IPv6/TCP/UDP)
- LRO/TSO
- VLAN offload - Filtering and stripping
- N-tuple filter and flow director (limited support)
- NPAR (NIC Partitioning)
- SR-IOV VF

- VXLAN Tunneling offload
- GENEVE Tunneling offload
- MPLSoUDP Tx Tunneling offload

26.2 Non-supported Features

- SR-IOV PF
- GRE and NVGRE Tunneling offloads

26.3 Co-existence considerations

- QLogic FastLinQ QL4xxxx CNAs can have both NIC and Storage personalities. However, coexistence with storage protocol drivers (qedi and qedf) is not supported on the same adapter. So storage personality has to be disabled on that adapter when used in DPDK applications.
- For SR-IOV case, qede PMD will be used to bind to SR-IOV VF device and Linux native kernel driver (qede) will be attached to SR-IOV PF.

26.4 Supported QLogic Adapters

- QLogic FastLinQ QL4xxxx 10G/25G/40G/50G/100G Intelligent Ethernet Adapters (IEA) and Converged Network Adapters (CNA)

26.5 Prerequisites

- Requires storm firmware version **8.30.12.0**. Firmware may be available in inbox in certain newer Linux distros under the standard directory E.g. `/lib/firmware/qed/qed_init_values-8.30.12.0.bin` If the required firmware files are not available then download it from [QLogic Driver Download Center](#). For downloading firmware file, select adapter category, model and DPDK Poll Mode Driver.
- Requires management firmware (MFW) version **8.30.x.x** or higher to be flashed on to the adapter. If the required management firmware is not available then download from [QLogic Driver Download Center](#). For downloading firmware upgrade utility, select adapter category, model and Linux distro. To flash the management firmware refer to the instructions in the QLogic Firmware Upgrade Utility Readme document.
- SR-IOV requires Linux PF driver version **8.20.x.x** or higher. If the required PF driver is not available then download it from [QLogic Driver Download Center](#). For downloading PF driver, select adapter category, model and Linux distro.

26.5.1 Performance note

- For better performance, it is recommended to use 4K or higher RX/TX rings.

26.5.2 Config File Options

The following options can be modified in the `.config` file. Please note that enabling debugging options may affect system performance.

- `CONFIG_RTE_LIBRTE_QEDE_PMD` (default **y**)
Toggle compilation of QEDE PMD driver.
- `CONFIG_RTE_LIBRTE_QEDE_DEBUG_TX` (default **n**)
Toggle display of transmit fast path run-time messages.
- `CONFIG_RTE_LIBRTE_QEDE_DEBUG_RX` (default **n**)
Toggle display of receive fast path run-time messages.
- `CONFIG_RTE_LIBRTE_QEDE_FW` (default **""**)
Gives absolute path of firmware file. Eg: `"/lib/firmware/qed/qed_init_values-8.30.12.0`
Empty string indicates driver will pick up the firmware file from the default location `/lib/firmware/qed`. CAUTION this option is more for custom firmware, it is not recommended for use under normal condition.

26.6 Driver compilation and testing

Refer to the document *compiling and testing a PMD for a NIC* for details.

26.7 SR-IOV: Prerequisites and Sample Application Notes

This section provides instructions to configure SR-IOV with Linux OS.

Note: `librte_pmd_qede` will be used to bind to SR-IOV VF device and Linux native kernel driver (`qede`) will function as SR-IOV PF driver. Requires PF driver to be 8.10.x.x or higher.

1. Verify SR-IOV and ARI capability is enabled on the adapter using `lspci`:

```
lspci -s <slot> -vvv
```

Example output:

```
[...]
Capabilities: [1b8 v1] Alternative Routing-ID Interpretation (ARI)
[...]
Capabilities: [1c0 v1] Single Root I/O Virtualization (SR-IOV)
[...]
Kernel driver in use: igb_uio
```

2. Load the kernel module:

```
modprobe qede
```

Example output:

```
systemd-udevd[4848]: renamed network interface eth0 to ens5f0
systemd-udevd[4848]: renamed network interface eth1 to ens5f1
```

3. Bring up the PF ports:

```
ifconfig ens5f0 up
ifconfig ens5f1 up
```

4. Create VF device(s):

Echo the number of VFs to be created into "sriov_numvfs" sysfs entry of the parent PF.

Example output:

```
echo 2 > /sys/devices/pci0000:00/0000:00:03.0/0000:81:00.0/sriov_numvfs
```

5. Assign VF MAC address:

Assign MAC address to the VF using iproute2 utility. The syntax is:

```
ip link set <PF iface> vf <VF id> mac <macaddr>
```

Example output:

```
ip link set ens5f0 vf 0 mac 52:54:00:2f:9d:e8
```

6. PCI Passthrough:

The VF devices may be passed through to the guest VM using `virt-manager` or `virsh`. QEDE PMD should be used to bind the VF devices in the guest VM using the instructions from Driver compilation and testing section above.

7. Running testpmd (Supply `--log-level="pmd.net.qede.driver"`, 7 to view informational messages):

Refer to the document [compiling and testing a PMD for a NIC](#) to run `testpmd` application.

Example output:

```
testpmd -l 0,4-11 -n 4 -- -i --nb-cores=8 --portmask=0xf --rxd=4096 \
--txd=4096 --txfreet=4068 --enable-rx-cksum --rxq=4 --txq=4 \
--rss-ip --rss-udp

[...]

EAL: PCI device 0000:84:00.0 on NUMA socket 1
EAL: probe driver: 1077:1634 rte_qede_pmd
EAL: Not managed by a supported kernel driver, skipped
EAL: PCI device 0000:84:00.1 on NUMA socket 1
EAL: probe driver: 1077:1634 rte_qede_pmd
EAL: Not managed by a supported kernel driver, skipped
EAL: PCI device 0000:88:00.0 on NUMA socket 1
EAL: probe driver: 1077:1656 rte_qede_pmd
EAL: PCI memory mapped at 0x7f738b200000
EAL: PCI memory mapped at 0x7f738b280000
EAL: PCI memory mapped at 0x7f738b300000
PMD: Chip details : BB1
PMD: Driver version : QEDE PMD 8.7.9.0_1.0.0
PMD: Firmware version : 8.7.7.0
PMD: Management firmware version : 8.7.8.0
PMD: Firmware file : /lib/firmware/qed/qed_init_values_zipped-8.7.7.0.bin
[QEDE PMD: (84:00.0:dpgk-port-0)]qede_common_dev_init:macaddr \
00:0e:1e:d2:09:9c

[...]
[QEDE PMD: (84:00.0:dpgk-port-0)]qede_tx_queue_setup:txq 0 num_desc 4096 \
tx_free_thresh 4068 socket 0
[QEDE PMD: (84:00.0:dpgk-port-0)]qede_tx_queue_setup:txq 1 num_desc 4096 \
tx_free_thresh 4068 socket 0
[QEDE PMD: (84:00.0:dpgk-port-0)]qede_tx_queue_setup:txq 2 num_desc 4096 \
```

```

                                tx_free_thresh 4068 socket 0
[QEDE PMD: (84:00.0:dpgk-port-0)]qede_tx_queue_setup:txq 3 num_desc 4096 \
                                tx_free_thresh 4068 socket 0
[QEDE PMD: (84:00.0:dpgk-port-0)]qede_rx_queue_setup:rxq 0 num_desc 4096 \
                                rx_buf_size=2148 socket 0
[QEDE PMD: (84:00.0:dpgk-port-0)]qede_rx_queue_setup:rxq 1 num_desc 4096 \
                                rx_buf_size=2148 socket 0
[QEDE PMD: (84:00.0:dpgk-port-0)]qede_rx_queue_setup:rxq 2 num_desc 4096 \
                                rx_buf_size=2148 socket 0
[QEDE PMD: (84:00.0:dpgk-port-0)]qede_rx_queue_setup:rxq 3 num_desc 4096 \
                                rx_buf_size=2148 socket 0
[QEDE PMD: (84:00.0:dpgk-port-0)]qede_dev_start:port 0
[QEDE PMD: (84:00.0:dpgk-port-0)]qede_dev_start:link status: down
[...]
Checking link statuses...
Port 0 Link Up - speed 25000 Mbps - full-duplex
Port 1 Link Up - speed 25000 Mbps - full-duplex
Port 2 Link Up - speed 25000 Mbps - full-duplex
Port 3 Link Up - speed 25000 Mbps - full-duplex
Done
testpmd>
```


SOLARFLARE LIBEFX-BASED POLL MODE DRIVER

The SFC EFX PMD (`librte_pmd_sfc_efx`) provides poll mode driver support for **Solarflare SFN7xxx and SFN8xxx** family of 10/40 Gbps adapters. SFC EFX PMD has support for the latest Linux and FreeBSD operating systems.

More information can be found at [Solarflare Communications website](#).

27.1 Features

SFC EFX PMD has support for:

- Multiple transmit and receive queues
- Link state information including link status change interrupt
- IPv4/IPv6 TCP/UDP transmit checksum offload
- Inner IPv4/IPv6 TCP/UDP transmit checksum offload
- Port hardware statistics
- Extended statistics (see Solarflare Server Adapter User's Guide for the statistics description)
- Basic flow control
- MTU update
- Jumbo frames up to 9K
- Promiscuous mode
- Allmulticast mode
- TCP segmentation offload (TSO)
- Multicast MAC filter
- IPv4/IPv6 TCP/UDP receive checksum offload
- Inner IPv4/IPv6 TCP/UDP receive checksum offload
- Received packet type information
- Receive side scaling (RSS)
- RSS hash
- Scattered Rx DMA for packet that are larger than a single Rx descriptor

- Deferred receive and transmit queue start
- Transmit VLAN insertion (if running firmware variant supports it)
- Flow API

27.2 Non-supported Features

The features not yet supported include:

- Receive queue interrupts
- Priority-based flow control
- Loopback
- Configurable RX CRC stripping (always stripped)
- Header split on receive
- VLAN filtering
- VLAN stripping
- LRO

27.3 Limitations

Due to requirements on receive buffer alignment and usage of the receive buffer for the auxiliary packet information provided by the NIC up to extra 269 (14 bytes prefix plus up to 255 bytes for end padding) bytes may be required in the receive buffer. It should be taken into account when mbuf pool for receive is created.

27.4 Tunnels support

NVGRE, VXLAN and GENEVE tunnels are supported on SFN8xxx family adapters with full-feature firmware variant running. **sfboot** should be used to configure NIC to run full-feature firmware variant. See Solarflare Server Adapter User's Guide for details.

SFN8xxx family adapters provide either inner or outer packet classes. If adapter firmware advertises support for tunnels then the PMD configures the hardware to report inner classes, and outer classes are not reported in received packets. However, for VXLAN and GENEVE tunnels the PMD does report UDP as the outer layer 4 packet type.

SFN8xxx family adapters report GENEVE packets as VXLAN. If UDP ports are configured for only one tunnel type then it is safe to treat VXLAN packet type indication as the corresponding UDP tunnel type.

27.5 Flow API support

Supported attributes:

- Ingress

Supported pattern items:

- VOID
- ETH (exact match of source/destination addresses, individual/group match of destination address, EtherType)
- VLAN (exact match of VID, double-tagging is supported)
- IPV4 (exact match of source/destination addresses, IP transport protocol)
- IPV6 (exact match of source/destination addresses, IP transport protocol)
- TCP (exact match of source/destination ports)
- UDP (exact match of source/destination ports)

Supported actions:

- VOID
- QUEUE
- RSS

Validating flow rules depends on the firmware variant.

27.5.1 Ethernet destination individual/group match

Ethernet item supports I/G matching, if only the corresponding bit is set in the mask of destination address. If destination address in the spec is multicast, it matches all multicast (and broadcast) packets, otherwise it matches unicast packets that are not filtered by other flow rules.

27.6 Supported NICs

- Solarflare Flareon [Ultra] Server Adapters:
 - Solarflare SFN8522 Dual Port SFP+ Server Adapter
 - Solarflare SFN8522M Dual Port SFP+ Server Adapter
 - Solarflare SFN8042 Dual Port QSFP+ Server Adapter
 - Solarflare SFN8542 Dual Port QSFP+ Server Adapter
 - Solarflare SFN8722 Dual Port SFP+ OCP Server Adapter
 - Solarflare SFN7002F Dual Port SFP+ Server Adapter
 - Solarflare SFN7004F Quad Port SFP+ Server Adapter
 - Solarflare SFN7042Q Dual Port QSFP+ Server Adapter
 - Solarflare SFN7122F Dual Port SFP+ Server Adapter
 - Solarflare SFN7124F Quad Port SFP+ Server Adapter
 - Solarflare SFN7142Q Dual Port QSFP+ Server Adapter

- Solarflare SFN7322F Precision Time Synchronization Server Adapter

27.7 Prerequisites

- Requires firmware version:
 - SFN7xxx: **4.7.1.1001** or higher
 - SFN8xxx: **6.0.2.1004** or higher

Visit [Solarflare Support Downloads](#) to get Solarflare Utilities (either Linux or FreeBSD) with the latest firmware. Follow instructions from Solarflare Server Adapter User's Guide to update firmware and configure the adapter.

27.8 Pre-Installation Configuration

27.8.1 Config File Options

The following options can be modified in the `.config` file. Please note that enabling debugging options may affect system performance.

- `CONFIG_RTE_LIBRTE_SFC_EFX_PMD` (default **y**)
Enable compilation of Solarflare libefx-based poll-mode driver.
- `CONFIG_RTE_LIBRTE_SFC_EFX_DEBUG` (default **n**)
Enable compilation of the extra run-time consistency checks.

27.8.2 Per-Device Parameters

The following per-device parameters can be passed via EAL PCI device whitelist option like “-w 02:00.0,arg1=value1,...”.

Case-insensitive 1/y/yes/on or 0/n/no/off may be used to specify boolean parameters value.

- `rx_datapath` [auto|efx|ef10] (default **auto**)
Choose receive datapath implementation. **auto** allows the driver itself to make a choice based on firmware features available and required by the datapath implementation. **efx** chooses libefx-based datapath which supports Rx scatter. **ef10** chooses EF10 (SFN7xxx, SFN8xxx) native datapath which is more efficient than libefx-based and provides richer packet type classification, but lacks Rx scatter support.
- `tx_datapath` [auto|efx|ef10|ef10_simple] (default **auto**)
Choose transmit datapath implementation. **auto** allows the driver itself to make a choice based on firmware features available and required by the datapath implementation. **efx** chooses libefx-based datapath which supports VLAN insertion (full-feature firmware variant only), TSO and multi-segment mbufs. Mbuf segments may come from different mempools, and mbuf reference counters are treated responsibly. **ef10** chooses EF10 (SFN7xxx, SFN8xxx) native datapath which is more efficient than libefx-based but has no VLAN insertion and TSO support yet. Mbuf segments may come from different mempools, and mbuf reference counters are treated responsibly. **ef10_simple** chooses EF10

(SFN7xxx, SFN8xxx) native datapath which is even more faster than **ef10** but does not support multi-segment mbufs, disallows multiple mempools and neglects mbuf reference counters.

- `perf_profile` [auto|throughput|low-latency] (default **throughput**)

Choose hardware tuning to be optimized for either throughput or low-latency. **auto** allows NIC firmware to make a choice based on installed licences and firmware variant configured using **sfboot**.

- `debug_init` [bool] (default **n**)

Enable extra logging during device initialization and startup.

- `mcdi_logging` [bool] (default **n**)

Enable extra logging of the communication with the NIC's management CPU. The logging is done using `RTE_LOG()` with INFO level and PMD type. The format is consumed by the Solarflare netlogdecode cross-platform tool.

- `stats_update_period_ms` [long] (default **1000**)

Adjust period in milliseconds to update port hardware statistics. The accepted range is 0 to 65535. The value of **0** may be used to disable periodic statistics update. One should note that it's only possible to set an arbitrary value on SFN8xxx provided that firmware version is 6.2.1.1033 or higher, otherwise any positive value will select a fixed update period of **1000** milliseconds

SZEDATA2 POLL MODE DRIVER LIBRARY

The SZEDATA2 poll mode driver library implements support for the Netcope FPGA Boards (**NFB-***), FPGA-based programmable NICs. The SZEDATA2 PMD uses interface provided by the `libsze2` library to communicate with the NFB cards over the `sze2` layer.

More information about the [NFB cards](#) and used technology ([Netcope Development Kit](#)) can be found on the [Netcope Technologies website](#).

Note: This driver has external dependencies. Therefore it is disabled in default configuration files. It can be enabled by setting `CONFIG_RTE_LIBRTE_PMD_SZEDATA2=y` and recompiling.

Note: Currently the driver is supported only on `x86_64` architectures. Only `x86_64` versions of the external libraries are provided.

28.1 Prerequisites

This PMD requires kernel modules which are responsible for initialization and allocation of resources needed for `sze2` layer function. Communication between PMD and kernel modules is mediated by `libsze2` library. These kernel modules and library are not part of DPDK and must be installed separately:

- **libsze2 library**

The library provides API for initialization of `sze2` transfers, receiving and transmitting data segments.

- **Kernel modules**

- `combov3`
- `szedata2_cv3`

Kernel modules manage initialization of hardware, allocation and sharing of resources for user space applications.

Information about getting the dependencies can be found [here](#).

28.2 Configuration

These configuration options can be modified before compilation in the `.config` file:

- CONFIG_RTE_LIBRTE_PMD_SZEDATA2 default value: **n**

Value **y** enables compilation of szedata2 PMD.

- CONFIG_RTE_LIBRTE_PMD_SZEDATA2_AS default value: **0**

This option defines type of firmware address space and must be set according to the used card and mode. Currently supported values are:

- **0** - for cards (modes):
 - * NFB-100G1 (100G1)
- **1** - for cards (modes):
 - * NFB-100G2Q (100G1)
- **2** - for cards (modes):
 - * NFB-40G2 (40G2)
 - * NFB-100G2C (100G2)
 - * NFB-100G2Q (40G2)
- **3** - for cards (modes):
 - * NFB-40G2 (10G8)
 - * NFB-100G2Q (10G8)
- **4** - for cards (modes):
 - * NFB-100G1 (10G10)
- **5** - for experimental firmwares and future use

28.3 Using the SZEDATA2 PMD

From DPDK version 16.04 the type of SZEDATA2 PMD is changed to PMD_PDEV. SZEDATA2 device is automatically recognized during EAL initialization. No special command line options are needed.

Kernel modules have to be loaded before running the DPDK application.

28.4 Example of usage

Read packets from 0. and 1. receive channel and write them to 0. and 1. transmit channel:

```
$RTE_TARGET/app/testpmd -l 0-3 -n 2 \
-- --port-topology=chained --rxq=2 --txq=2 --nb-cores=2 -i -a
```

Example output:

```
[...]
EAL: PCI device 0000:06:00.0 on NUMA socket -1
EAL: probe driver: 1b26:c1c1 rte_szedata2_pmd
PMD: Initializing szedata2 device (0000:06:00.0)
PMD: SZEDATA2 path: /dev/szedataII0
PMD: Available DMA channels RX: 8 TX: 8
PMD: resource0 phys_addr = 0xe8000000 len = 134217728 virt addr = 7f48f8000000
```

```
PMD: szedata2 device (0000:06:00.0) successfully initialized
Interactive-mode selected
Auto-start selected
Configuring Port 0 (socket 0)
Port 0: 00:11:17:00:00:00
Checking link statuses...
Port 0 Link Up - speed 10000 Mbps - full-duplex
Done
Start automatic packet forwarding
  io packet forwarding - CRC stripping disabled - packets/burst=32
  nb forwarding cores=2 - nb forwarding ports=1
  RX queues=2 - RX desc=128 - RX free threshold=0
  RX threshold registers: pthresh=0 hthresh=0 wthresh=0
  TX queues=2 - TX desc=512 - TX free threshold=0
  TX threshold registers: pthresh=0 hthresh=0 wthresh=0
  TX RS bit threshold=0 - TXQ flags=0x0
testpmd>
```


TAP POLL MODE DRIVER

The `rte_eth_tap.c` PMD creates a device using TAP interfaces on the local host. The PMD allows for DPDK and the host to communicate using a raw device interface on the host and in the DPDK application.

The device created is a TAP device, which sends/receives packet in a raw format with a L2 header. The usage for a TAP PMD is for connectivity to the local host using a TAP interface. When the TAP PMD is initialized it will create a number of tap devices in the host accessed via `ifconfig -a` or `ip` command. The commands can be used to assign and query the virtual like device.

These TAP interfaces can be used with Wireshark or `tcpdump` or `Pktgen-DPDK` along with being able to be used as a network connection to the DPDK application. The method enable one or more interfaces is to use the `--vdev=net_tap0` option on the DPDK application command line. Each `--vdev=net_tap1` option given will create an interface named `dtap0`, `dtap1`, and so on.

The interface name can be changed by adding the `iface=foo0`, for example:

```
--vdev=net_tap0,iface=foo0 --vdev=net_tap1,iface=foo1, ...
```

Normally the PMD will generate a random MAC address, but when testing or with a static configuration the developer may need a fixed MAC address style. Using the option `mac=fixed` you can create a fixed known MAC address:

```
--vdev=net_tap0,mac=fixed
```

The MAC address will have a fixed value with the last octet incrementing by one for each interface string containing `mac=fixed`. The MAC address is formatted as `00:d:t:a:p:[00-FF]`. Convert the characters to hex and you get the actual MAC address: `00:64:74:61:70:[00-FF]`.

It is possible to specify a remote netdevice to capture packets from by adding `remote=foo1`, for example:

```
--vdev=net_tap,iface=tap0,remote=foo1
```

If a `remote` is set, the tap MAC address will be set to match the remote one just after netdevice creation. Using TC rules, traffic from the remote netdevice will be redirected to the tap. If the tap is in promiscuous mode, then all packets will be redirected. In `allmulti` mode, all multicast packets will be redirected.

Using the remote feature is especially useful for capturing traffic from a netdevice that has no support in the DPDK. It is possible to add explicit `rte_flow` rules on the tap PMD to capture specific traffic (see next section for examples).

After the DPDK application is started you can send and receive packets on the interface using the standard rx_burst/tx_burst APIs in DPDK. From the host point of view you can use any host tool like tcpdump, Wireshark, ping, Pktgen and others to communicate with the DPDK application. The DPDK application may not understand network protocols like IPv4/6, UDP or TCP unless the application has been written to understand these protocols.

If you need the interface as a real network interface meaning running and has a valid IP address then you can do this with the following commands:

```
sudo ip link set dtap0 up; sudo ip addr add 192.168.0.250/24 dev dtap0
sudo ip link set dtap1 up; sudo ip addr add 192.168.1.250/24 dev dtap1
```

Please change the IP addresses as you see fit.

If routing is enabled on the host you can also communicate with the DPDK App over the internet via a standard socket layer application as long as you account for the protocol handing in the application.

If you have a Network Stack in your DPDK application or something like it you can utilize that stack to handle the network protocols. Plus you would be able to address the interface using an IP address assigned to the internal interface.

29.1 Flow API support

The tap PMD supports major flow API pattern items and actions, when running on linux kernels above 4.2 (“Flower” classifier required). The kernel support can be checked with this command:

```
zcat /proc/config.gz | ( grep 'CLS_FLOWER=' || echo 'not supported' ) |
tee -a /dev/stderr | grep -q '=m' &&
lsmod | ( grep cls_flower || echo 'try modprobe cls_flower' )
```

Supported items:

- eth: src and dst (with variable masks), and eth_type (0xffff mask).
- vlan: vid, pcp, tpid, but not eid. (requires kernel 4.9)
- ipv4/6: src and dst (with variable masks), and ip_proto (0xffff mask).
- udp/tcp: src and dst port (0xffff) mask.

Supported actions:

- DROP
- QUEUE
- PASSTHRU
- RSS (requires kernel 4.9)

It is generally not possible to provide a “last” item. However, if the “last” item, once masked, is identical to the masked spec, then it is supported.

Only IPv4/6 and MAC addresses can use a variable mask. All other items need a full mask (exact match).

As rules are translated to TC, it is possible to show them with something like:

```
tc -s filter show dev tap1 parent 1:
```

29.1.1 Examples of testpmd flow rules

Drop packets for destination IP 192.168.0.1:

```
testpmd> flow create 0 priority 1 ingress pattern eth / ipv4 dst is 1.1.1.1 \
/ end actions drop / end
```

Ensure packets from a given MAC address are received on a queue 2:

```
testpmd> flow create 0 priority 2 ingress pattern eth src is 06:05:04:03:02:01 \
/ end actions queue index 2 / end
```

Drop UDP packets in vlan 3:

```
testpmd> flow create 0 priority 3 ingress pattern eth / vlan vid is 3 / \
ipv4 proto is 17 / end actions drop / end
```

Distribute IPv4 TCP packets using RSS to a given MAC address over queues 0-3:

```
testpmd> flow create 0 priority 4 ingress pattern eth dst is 0a:0b:0c:0d:0e:0f \
/ ipv4 / tcp / end actions rss queues 0 1 2 3 end / end
```

29.2 Example

The following is a simple example of using the TAP PMD with the Pktgen packet generator. It requires that the `socat` utility is installed on the test system.

Build DPDK, then pull down Pktgen and build pktgen using the DPDK SDK/Target used to build the dpdk you pulled down.

Run pktgen from the pktgen directory in a terminal with a commandline like the following:

```
sudo ./app/app/x86_64-native-linuxapp-gcc/app/pktgen -l 1-5 -n 4          \
--proc-type auto --log-level 8 --socket-mem 512,512 --file-prefix pg  \
--vdev=net_tap0 --vdev=net_tap1 -b 05:00.0 -b 05:00.1                \
-b 04:00.0 -b 04:00.1 -b 04:00.2 -b 04:00.3                          \
-b 81:00.0 -b 81:00.1 -b 81:00.2 -b 81:00.3                          \
-b 82:00.0 -b 83:00.0 -- -T -P -m [2:3].0 -m [4:5].1                 \
-f themes/black-yellow.theme
```

Verify with `ifconfig -a` command in a different xterm window, should have a `dtap0` and `dtap1` interfaces created.

Next set the links for the two interfaces to up via the commands below:

```
sudo ip link set dtap0 up; sudo ip addr add 192.168.0.250/24 dev dtap0
sudo ip link set dtap1 up; sudo ip addr add 192.168.1.250/24 dev dtap1
```

Then use `socat` to create a loopback for the two interfaces:

```
sudo socat interface:dtap0 interface:dtap1
```

Then on the Pktgen command line interface you can start sending packets using the commands `start 0` and `start 1` or you can start both at the same time with `start all`. The command `str` is an alias for `start all` and `stp` is an alias for `stop all`.

While running you should see the 64 byte counters increasing to verify the traffic is being looped back. You can use `set all size XXX` to change the size of the packets after you stop the traffic. Use `pktgen help` command to see a list of all commands. You can also use the `-f` option to load commands at startup in command line or Lua script in `pktgen`.

29.3 RSS specifics

Packet distribution in TAP is done by the kernel which has a default distribution. This feature is adding RSS distribution based on eBPF code. The default eBPF code calculates RSS hash based on Toeplitz algorithm for a fixed RSS key. It is calculated on fixed packet offsets. For IPv4 and IPv6 it is calculated over src/dst addresses (8 or 32 bytes for IPv4 or IPv6 respectively) and src/dst TCP/UDP ports (4 bytes).

The RSS algorithm is written in file `tap_bpf_program.c` which does not take part in TAP PMD compilation. Instead this file is compiled in advance to eBPF object file. The eBPF object file is then parsed and translated into eBPF byte code in the format of C arrays of eBPF instructions. The C array of eBPF instructions is part of TAP PMD tree and is taking part in TAP PMD compilation. At run time the C arrays are uploaded to the kernel via BPF system calls and the RSS hash is calculated by the kernel.

It is possible to support different RSS hash algorithms by updating file `tap_bpf_program.c`. In order to add a new RSS hash algorithm follow these steps:

1. Write the new RSS implementation in file `tap_bpf_program.c`

BPF programs which are uploaded to the kernel correspond to C functions under different ELF sections.

2. Install LLVM library and clang compiler versions 3.7 and above
3. Compile `tap_bpf_program.c` via LLVM into an object file:

```
clang -O2 -emit-llvm -c tap_bpf_program.c -o - | llc -march=bpf \
-filetype=obj -o <tap_bpf_program.o>
```

4. Use a tool that receives two parameters: an eBPF object file and a section name, and prints out the section as a C array of eBPF instructions. Embed the C array in your TAP PMD tree.

The C arrays are uploaded to the kernel using BPF system calls.

`tc` (traffic control) is a well known user space utility program used to configure the Linux kernel packet scheduler. It is usually packaged as part of the `iproute2` package. Since commit 11c39b5e9 (“tc: add eBPF support to `f_bpf`”) `tc` can be used to uploads eBPF code to the kernel and can be patched in order to print the C arrays of eBPF instructions just before calling the BPF system call. Please refer to `iproute2` package file `lib/bpf.c` function `bpf_prog_load()`.

An example utility for eBPF instruction generation in the format of C arrays will be added in next releases

29.4 Systems supporting flow API

- “tc flower” classifier requires linux kernel above 4.2
- eBPF/RSS requires linux kernel above 4.9

RH7.3	No flow rule support
RH7.4	No RSS action support
RH7.5	No RSS action support
SLES 15, kernel 4.12	No limitation
Azure Ubuntu 16.04, kernel 4.13	No limitation

THUNDERX NICVF POLL MODE DRIVER

The ThunderX NICVF PMD (`librte_pmd_thunderx_nicvf`) provides poll mode driver support for the inbuilt NIC found in the **Cavium ThunderX** SoC family as well as their virtual functions (VF) in SR-IOV context.

More information can be found at [Cavium, Inc Official Website](#).

30.1 Features

Features of the ThunderX PMD are:

- Multiple queues for TX and RX
- Receive Side Scaling (RSS)
- Packet type information
- Checksum offload
- Promiscuous mode
- Multicast mode
- Port hardware statistics
- Jumbo frames
- Link state information
- Scattered and gather for TX and RX
- VLAN stripping
- SR-IOV VF
- NUMA support
- Multi queue set support (up to 96 queues (12 queue sets)) per port

30.2 Supported ThunderX SoCs

- CN88xx
- CN81xx
- CN83xx

30.3 Prerequisites

- Follow the DPDK Getting Started Guide for Linux to setup the basic DPDK environment.

30.4 Pre-Installation Configuration

30.4.1 Config File Options

The following options can be modified in the `config` file. Please note that enabling debugging options may affect system performance.

- `CONFIG_RTE_LIBRTE_THUNDERX_NICVF_PMD` (default `y`)
Toggle compilation of the `librte_pmd_thunderx_nicvf` driver.
- `CONFIG_RTE_LIBRTE_THUNDERX_NICVF_DEBUG_RX` (default `n`)
Toggle asserts of receive fast path.
- `CONFIG_RTE_LIBRTE_THUNDERX_NICVF_DEBUG_TX` (default `n`)
Toggle asserts of transmit fast path.

30.5 Driver compilation and testing

Refer to the document *compiling and testing a PMD for a NIC* for details.

To compile the ThunderX NICVF PMD for Linux arm64 gcc, use `arm64-thunderx-linuxapp-gcc` as target.

30.6 Linux

30.6.1 SR-IOV: Prerequisites and sample Application Notes

Current ThunderX NIC PF/VF kernel modules maps each physical Ethernet port automatically to virtual function (VF) and presented them as PCIe-like SR-IOV device. This section provides instructions to configure SR-IOV with Linux OS.

1. Verify PF devices capabilities using `lspci`:

```
lspci -vvv
```

Example output:

```
0002:01:00.0 Ethernet controller: Cavium Networks Device a01e (rev 01)
...
Capabilities: [100 v1] Alternative Routing-ID Interpretation (ARI)
...
Capabilities: [180 v1] Single Root I/O Virtualization (SR-IOV)
...
Kernel driver in use: thunder-nic
...
```

Note: Unless `thunder-nic` driver is in use make sure your kernel config includes `CONFIG_THUNDER_NIC_PF` setting.

2. Verify VF devices capabilities and drivers using `lspci`:

```
lspci -vvv
```

Example output:

```
0002:01:00.1 Ethernet controller: Cavium Networks Device 0011 (rev 01)
...
Capabilities: [100 v1] Alternative Routing-ID Interpretation (ARI)
...
Kernel driver in use: thunder-nicvf
...

0002:01:00.2 Ethernet controller: Cavium Networks Device 0011 (rev 01)
...
Capabilities: [100 v1] Alternative Routing-ID Interpretation (ARI)
...
Kernel driver in use: thunder-nicvf
...
```

Note: Unless `thunder-nicvf` driver is in use make sure your kernel config includes `CONFIG_THUNDER_NIC_VF` setting.

3. Pass VF device to VM context (PCIe Passthrough):

The VF devices may be passed through to the guest VM using `qemu` or `virt-manager` or `virsh` etc.

Example `qemu` guest launch command:

```
sudo qemu-system-aarch64 -name vm1 \
-machine virt,gic_version=3,accel=kvm,usb=off \
-cpu host -m 4096 \
-smp 4,sockets=1,cores=8,threads=1 \
-nographic -nodefaults \
-kernel <kernel image> \
-append "root=/dev/vda console=ttyAMA0 rw hugepagesz=512M hugepages=3" \
-device vfio-pci,host=0002:01:00.1 \
-drive file=<rootfs.ext3>,if=none,id=disk1,format=raw \
-device virtio-blk-device,scsi=off,drive=disk1,id=virtio-disk1,bootindex=1 \
-netdev tap,id=net0,ifname=tap0,script=/etc/qemu-ifup_thunder \
-device virtio-net-device,netdev=net0 \
-serial stdio \
-mem-path /dev/huge
```

4. Enable **VFIO-NOIOMMU** mode (optional):

```
echo 1 > /sys/module/vfio/parameters/enable_unsafe_noiommu_mode
```

Note: **VFIO-NOIOMMU** is required only when running in VM context and should not be enabled otherwise.

5. Running `testpmd`:

Follow instructions available in the document [compiling and testing a PMD for a NIC](#) to run `testpmd`.

Example output:

```
./arm64-thunderx-linuxapp-gcc/app/testpmd -l 0-3 -n 4 -w 0002:01:00.2 \
-- -i --no-flush-rx \
--port-topology=loop
...
PMD: rte_nicvf_pmd_init(): librte_pmd_thunderx nicvf version 1.0
...
EAL: probe driver: 177d:11 rte_nicvf_pmd
EAL: using IOMMU type 1 (Type 1)
EAL: PCI memory mapped at 0x3ffade50000
EAL: Trying to map BAR 4 that contains the MSI-X table.
    Trying offsets: 0x40000000000:0x0000, 0x10000:0x1f0000
EAL: PCI memory mapped at 0x3ffadc60000
PMD: nicvf_eth_dev_init(): nicvf: device (177d:11) 2:1:0:2
PMD: nicvf_eth_dev_init(): node=0 vf=1 mode=tns-bypass sqs=false
    loopback_supported=true
PMD: nicvf_eth_dev_init(): Port 0 (177d:11) mac=a6:c6:d9:17:78:01
Interactive-mode selected
Configuring Port 0 (socket 0)
...
PMD: nicvf_dev_configure(): Configured ethdev port0 hwcap=0x0
Port 0: A6:C6:D9:17:78:01
Checking link statuses...
Port 0 Link Up - speed 10000 Mbps - full-duplex
Done
testpmd>
```

30.6.2 Multiple Queue Set per DPDK port configuration

There are two types of VFs:

- Primary VF
- Secondary VF

Each port consists of a primary VF and n secondary VF(s). Each VF provides 8 Tx/Rx queues to a port. When a given port is configured to use more than 8 queues, it requires one (or more) secondary VF. Each secondary VF adds 8 additional queues to the queue set.

During PMD driver initialization, the primary VF's are enumerated by checking the specific flag (see sqs message in DPDK boot log - sqs indicates secondary queue set). They are at the beginning of VF list (the remain ones are secondary VF's).

The primary VFs are used as master queue sets. Secondary VFs provide additional queue sets for primary ones. If a port is configured for more than 8 queues than it will request for additional queues from secondary VFs.

Secondary VFs cannot be shared between primary VFs.

Primary VFs are present on the beginning of the 'Network devices using kernel driver' list, secondary VFs are on the remaining on the remaining part of the list.

Note: The VNIC driver in the multiqueue setup works differently than other drivers like *ixgbe*. We need to bind separately each specific queue set device with the `usertools/dpdk-devbind.py` utility.

Note: Depending on the hardware used, the kernel driver sets a threshold `vf_id`. VFs that try to attached with an id below or equal to this boundary are considered primary VFs. VFs that try to attach with an id above this boundary are considered secondary VFs.

30.6.3 Example device binding

If a system has three interfaces, a total of 18 VF devices will be created on a non-NUMA machine.

Note: NUMA systems have 12 VFs per port and non-NUMA 6 VFs per port.

```
# usertools/dpdk-devbind.py --status

Network devices using DPDK-compatible driver
=====
<none>

Network devices using kernel driver
=====
0000:01:10.0 'Device a026' if= drv=thunder-BGX unused=vfio-pci,uio_pci_generic
0000:01:10.1 'Device a026' if= drv=thunder-BGX unused=vfio-pci,uio_pci_generic
0002:01:00.0 'Device a01e' if= drv=thunder-nic unused=vfio-pci,uio_pci_generic
0002:01:00.1 'Device 0011' if=eth0 drv=thunder-nicvf unused=vfio-pci,uio_pci_generic
0002:01:00.2 'Device 0011' if=eth1 drv=thunder-nicvf unused=vfio-pci,uio_pci_generic
0002:01:00.3 'Device 0011' if=eth2 drv=thunder-nicvf unused=vfio-pci,uio_pci_generic
0002:01:00.4 'Device 0011' if= drv=thunder-nicvf unused=vfio-pci,uio_pci_generic
0002:01:00.5 'Device 0011' if= drv=thunder-nicvf unused=vfio-pci,uio_pci_generic
0002:01:00.6 'Device 0011' if= drv=thunder-nicvf unused=vfio-pci,uio_pci_generic
0002:01:00.7 'Device 0011' if= drv=thunder-nicvf unused=vfio-pci,uio_pci_generic
0002:01:01.0 'Device 0011' if= drv=thunder-nicvf unused=vfio-pci,uio_pci_generic
0002:01:01.1 'Device 0011' if= drv=thunder-nicvf unused=vfio-pci,uio_pci_generic
0002:01:01.2 'Device 0011' if= drv=thunder-nicvf unused=vfio-pci,uio_pci_generic
0002:01:01.3 'Device 0011' if= drv=thunder-nicvf unused=vfio-pci,uio_pci_generic
0002:01:01.4 'Device 0011' if= drv=thunder-nicvf unused=vfio-pci,uio_pci_generic
0002:01:01.5 'Device 0011' if= drv=thunder-nicvf unused=vfio-pci,uio_pci_generic
0002:01:01.6 'Device 0011' if= drv=thunder-nicvf unused=vfio-pci,uio_pci_generic
0002:01:01.7 'Device 0011' if= drv=thunder-nicvf unused=vfio-pci,uio_pci_generic
0002:01:02.0 'Device 0011' if= drv=thunder-nicvf unused=vfio-pci,uio_pci_generic
0002:01:02.1 'Device 0011' if= drv=thunder-nicvf unused=vfio-pci,uio_pci_generic
0002:01:02.2 'Device 0011' if= drv=thunder-nicvf unused=vfio-pci,uio_pci_generic

Other network devices
=====
0002:00:03.0 'Device a01f' unused=vfio-pci,uio_pci_generic
```

We want to bind two physical interfaces with 24 queues each device, we attach two primary VFs and four secondary queues. In our example we choose two 10G interfaces eth1 (0002:01:00.2) and eth2 (0002:01:00.3). We will choose four secondary queue sets from the ending of the list (0002:01:01.7-0002:01:02.2).

1. Bind two primary VFs to the `vfio-pci` driver:

```
usertools/dpdk-devbind.py -b vfio-pci 0002:01:00.2
usertools/dpdk-devbind.py -b vfio-pci 0002:01:00.3
```

2. Bind four primary VFs to the `vfio-pci` driver:

```
usertools/dpdk-devbind.py -b vfio-pci 0002:01:01.7
usertools/dpdk-devbind.py -b vfio-pci 0002:01:02.0
```

```
usertools/dpdk-devbind.py -b vfio-pci 0002:01:02.1
usertools/dpdk-devbind.py -b vfio-pci 0002:01:02.2
```

The nicvf thunderx driver will make use of attached secondary VFs automatically during the interface configuration stage.

30.7 Limitations

30.7.1 CRC striping

The ThunderX SoC family NICs strip the CRC for every packets coming into the host interface irrespective of the offload configuration.

30.7.2 Maximum packet length

The ThunderX SoC family NICs support a maximum of a 9K jumbo frame. The value is fixed and cannot be changed. So, even when the `rxmode.max_rx_pkt_len` member of `struct rte_eth_conf` is set to a value lower than 9200, frames up to 9200 bytes can still reach the host interface.

30.7.3 Maximum packet segments

The ThunderX SoC family NICs support up to 12 segments per packet when working in scatter/gather mode. So, setting MTU will result with `EINVAL` when the frame size does not fit in the maximum number of segments.

VDEV_NETVSC DRIVER

The VDEV_NETVSC driver (`librte_pmd_vdev_netvsc`) provides support for NetVSC interfaces and associated SR-IOV virtual function (VF) devices found in Linux virtual machines running on Microsoft [Hyper-V](#) (including Azure) platforms.

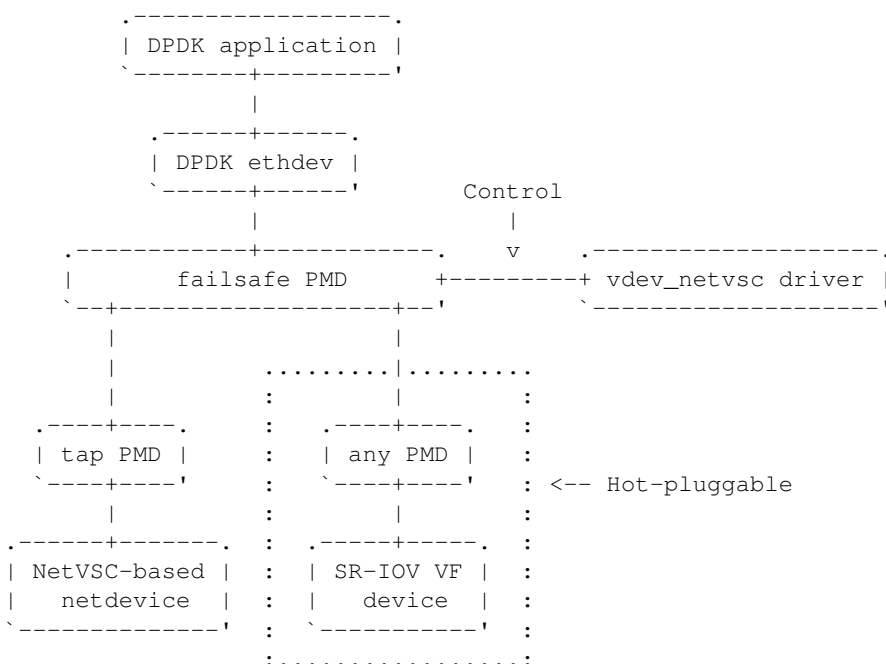
31.1 Implementation details

Each instance of this driver effectively needs to drive two devices: the NetVSC interface proper and its SR-IOV VF (referred to as “physical” from this point on) counterpart sharing the same MAC address.

Physical devices are part of the host system and cannot be maintained during VM migration. From a VM standpoint they appear as hot-plug devices that come and go without prior notice.

When the physical device is present, egress and most of the ingress traffic flows through it; only multicasts and other hypervisor control still flow through NetVSC. Otherwise, NetVSC acts as a fallback for all traffic.

To avoid unnecessary code duplication and ensure maximum performance, handling of physical devices is left to their original PMDs; this virtual device driver (also known as *vdev*) manages other PMDs as summarized by the following block diagram:



This driver implementation may be temporary and should be improved or removed either when hot-plug will be fully supported in EAL and bus drivers or when a new NetVSC driver will be integrated.

31.2 Build options

- `CONFIG_RTE_LIBRTE_VDEV_NETVSC_PMD` (default `y`)

Toggle compilation of this driver.

31.3 Run-time parameters

This driver is invoked automatically in Hyper-V VM systems unless the user invoked it by command line using `--vdev=net_vdev_netvsc` EAL option.

The following device parameters are supported:

- `iface` [string]
Provide a specific NetVSC interface (netdevice) name to attach this driver to. Can be provided multiple times for additional instances.
- `mac` [string]
Same as `iface` except a suitable NetVSC interface is located using its MAC address.
- `force` [int]
If nonzero, forces the use of specified interfaces even if not detected as NetVSC.
- `ignore` [int]
If nonzero, ignores the driver running (actually used to disable the auto-detection in Hyper-V VM).

Note: Not specifying either `iface` or `mac` makes this driver attach itself to all unrouted NetVSC interfaces found on the system. Specifying the device makes this driver attach itself to the device regardless the device routes.

POLL MODE DRIVER FOR EMULATED VIRTIO NIC

Virtio is a para-virtualization framework initiated by IBM, and supported by KVM hypervisor. In the Data Plane Development Kit (DPDK), we provide a virtio Poll Mode Driver (PMD) as a software solution, comparing to SRIOV hardware solution,

for fast guest VM to guest VM communication and guest VM to host communication.

Vhost is a kernel acceleration module for virtio qemu backend. The DPDK extends kni to support vhost raw socket interface, which enables vhost to directly read/ write packets from/to a physical port. With this enhancement, virtio could achieve quite promising performance.

For basic qemu-KVM installation and other Intel EM poll mode driver in guest VM, please refer to Chapter “Driver for VM Emulated Devices”.

In this chapter, we will demonstrate usage of virtio PMD driver with two backends, standard qemu vhost back end and vhost kni back end.

32.1 Virtio Implementation in DPDK

For details about the virtio spec, refer to Virtio PCI Card Specification written by Rusty Russell.

As a PMD, virtio provides packet reception and transmission callbacks `virtio_recv_pkts` and `virtio_xmit_pkts`.

In `virtio_recv_pkts`, index in range [`vq->vq_used_cons_idx` , `vq->vq_ring.used->idx`) in `vring` is available for virtio to burst out.

In `virtio_xmit_pkts`, same index range in `vring` is available for virtio to clean. Virtio will enqueue to be transmitted packets into `vring`, advance the `vq->vq_ring.avail->idx`, and then notify the host back end if necessary.

32.2 Features and Limitations of virtio PMD

In this release, the virtio PMD driver provides the basic functionality of packet reception and transmission.

- It supports merge-able buffers per packet when receiving packets and scattered buffer per packet when transmitting packets. The packet size supported is from 64 to 1518.
- It supports multicast packets and promiscuous mode.

- The descriptor number for the Rx/Tx queue is hard-coded to be 256 by qemu 2.7 and below. If given a different descriptor number by the upper application, the virtio PMD generates a warning and fall back to the hard-coded value. Rx queue size can be configurable and up to 1024 since qemu 2.8 and above. Rx queue size is 256 by default. Tx queue size is still hard-coded to be 256.
- Features of mac/vlan filter are supported, negotiation with vhost/backend are needed to support them. When backend can't support vlan filter, virtio app on guest should not enable vlan filter in order to make sure the virtio port is configured correctly. E.g. do not specify '-enable-hw-vlan' in testpmd command line.
- "RTE_PKTMBUF_HEADROOM" should be defined no less than "sizeof(struct virtio_net_hdr_mrg_rxbuf)", which is 12 bytes when mergeable or "VIRTIO_F_VERSION_1" is set. no less than "sizeof(struct virtio_net_hdr)", which is 10 bytes, when using non-mergeable.
- Virtio does not support runtime configuration.
- Virtio supports Link State interrupt.
- Virtio supports Rx interrupt (so far, only support 1:1 mapping for queue/interrupt).
- Virtio supports software vlan stripping and inserting.
- Virtio supports using port IO to get PCI resource when uio/igb_uio module is not available.

32.3 Prerequisites

The following prerequisites apply:

- In the BIOS, turn VT-x and VT-d on
- Linux kernel with KVM module; vhost module loaded and ioeventfd supported. Qemu standard backend without vhost support isn't tested, and probably isn't supported.

32.4 Virtio with kni vhost Back End

This section demonstrates kni vhost back end example setup for Phy-VM Communication.

Host2VM communication example

1. Load the kni kernel module:

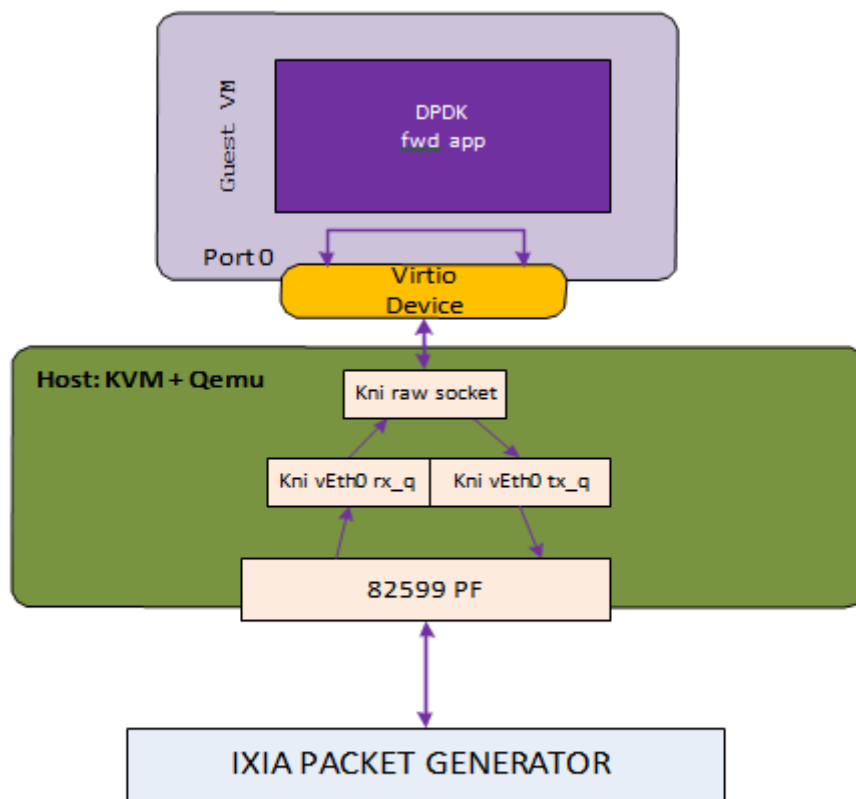
```
insmod rte_kni.ko
```

Other basic DPDK preparations like hugepage enabling, uio port binding are not listed here. Please refer to the *DPDK Getting Started Guide* for detailed instructions.

2. Launch the kni user application:

```
examples/kni/build/app/kni -l 0-3 -n 4 -- -p 0x1 -P --config="(0,1,3) "
```

This command generates one network device vEth0 for physical port. If specify more physical ports, the generated network device will be vEth1, vEth2, and so on.



Host2VM communication example

Fig. 32.1: Host2VM Communication Example Using kni vhost Back End

For each physical port, kni creates two user threads. One thread loops to fetch packets from the physical NIC port into the kni receive queue. The other user thread loops to send packets in the kni transmit queue.

For each physical port, kni also creates a kernel thread that retrieves packets from the kni receive queue, place them onto kni's raw socket's queue and wake up the vhost kernel thread to exchange packets with the virtio virt queue.

For more details about kni, please refer to kni.

3. Enable the kni raw socket functionality for the specified physical NIC port, get the generated file descriptor and set it in the qemu command line parameter. Always remember to set `ioeventfd_on` and `vhost_on`.

Example:

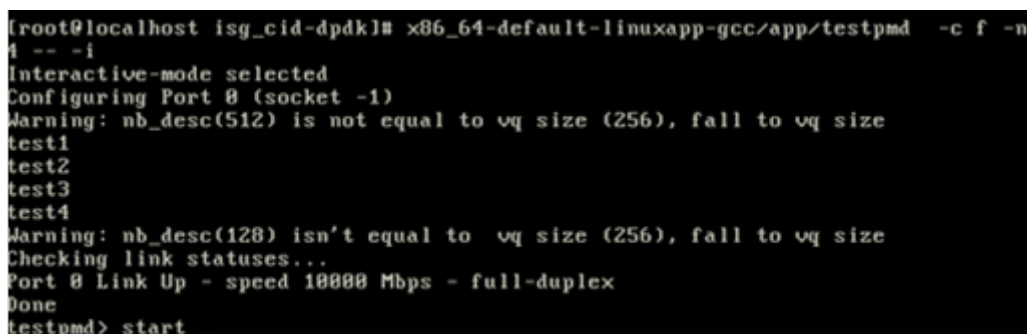
```
echo 1 > /sys/class/net/vEth0/sock_en
fd=`cat /sys/class/net/vEth0/sock_fd`
exec qemu-system-x86_64 -enable-kvm -cpu host \
-m 2048 -smp 4 -name dpdk-test1-vm1 \
-drive file=/data/DPDKVMS/dpdk-vm.img \
-netdev tap, fd=$fd,id=mynet_kni, script=no,vhost=on \
-device virtio-net-pci,netdev=mynet_kni,bus=pci.0,addr=0x3,ioeventfd=on \
-vnc:1 -daemonize
```

In the above example, virtio port 0 in the guest VM will be associated with vEth0, which in turns corresponds to a physical port, which means received packets come from vEth0, and transmitted packets is sent to vEth0.

4. In the guest, bind the virtio device to the `uio_pci_generic` kernel module and start the forwarding application. When the virtio port in guest bursts Rx, it is getting packets from the raw socket's receive queue. When the virtio port bursts Tx, it is sending packet to the `tx_q`.

```
modprobe uio
echo 512 > /sys/devices/system/node/node0/hugepages/hugepages-2048kB/nr_hugepages
modprobe uio_pci_generic
python usertools/dpdk-devbind.py -b uio_pci_generic 00:03.0
```

We use `testpmd` as the forwarding application in this example.



```
(root@localhost isg_cid-dpdk) x86_64-default-linuxapp-gcc/app/testpmd -c f -n
1 -- -i
Interactive-mode selected
Configuring Port 0 (socket -1)
Warning: nb_desc(512) is not equal to vq size (256), fall to vq size
test1
test2
test3
test4
Warning: nb_desc(128) isn't equal to vq size (256), fall to vq size
Checking link statuses...
Port 0 Link Up - speed 10000 Mbps - full-duplex
Done
testpmd> start _
```

Fig. 32.2: Running testpmd

5. Use IXIA packet generator to inject a packet stream into the KNI physical port.

The packet reception and transmission flow path is:

IXIA packet generator->82599 PF->KNI Rx queue->KNI raw socket queue->Guest VM virtio port 0 Rx burst->Guest VM virtio port 0 Tx burst-> KNI Tx queue ->82599 PF->

IXIA packet generator

32.5 Virtio with qemu virtio Back End

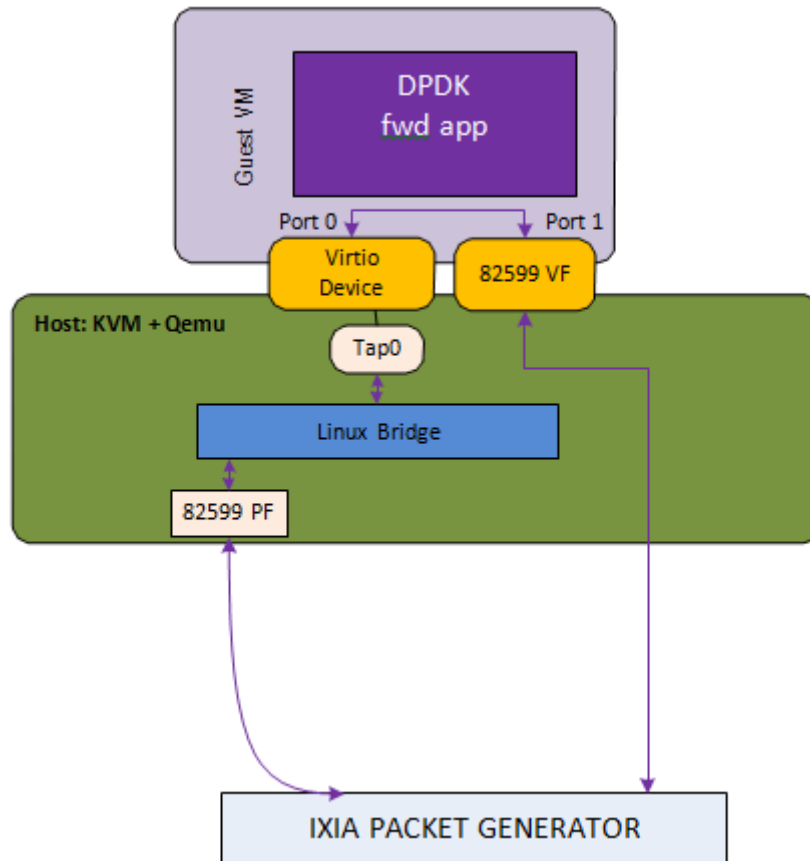


Fig. 32.3: Host2VM Communication Example Using qemu vhost Back End

```
qemu-system-x86_64 -enable-kvm -cpu host -m 2048 -smp 2 -mem-path /dev/
hugepages -mem-prealloc
-drive file=/data/DPDKVMS/dpdk-vm1
-netdev tap,id=vm1_p1,ifname=tap0,script=no,vhost=on
-device virtio-net-pci,netdev=vm1_p1,bus=pci.0,addr=0x3,ioeventfd=on
-device pci-assign,host=04:10.1 \
```

In this example, the packet reception flow path is:

IXIA packet generator->82599 PF->Linux Bridge->TAP0's socket queue-> Guest VM virtio port 0 Rx burst-> Guest VM 82599 VF port1 Tx burst-> IXIA packet generator

The packet transmission flow is:

IXIA packet generator-> Guest VM 82599 VF port1 Rx burst-> Guest VM virtio port 0 Tx burst-> tap -> Linux Bridge->82599 PF-> IXIA packet generator

32.6 Virtio PMD Rx/Tx Callbacks

Virtio driver has 3 Rx callbacks and 2 Tx callbacks.

Rx callbacks:

1. `virtio_recv_pkts`: Regular version without mergeable Rx buffer support.
2. `virtio_recv_mergeable_pkts`: Regular version with mergeable Rx buffer support.
3. `virtio_recv_pkts_vec`: Vector version without mergeable Rx buffer support, also fixes the available ring indexes and uses vector instructions to optimize performance.

Tx callbacks:

1. `virtio_xmit_pkts`: Regular version.
2. `virtio_xmit_pkts_simple`: Vector version fixes the available ring indexes to optimize performance.

By default, the non-vector callbacks are used:

- For Rx: If mergeable Rx buffers is disabled then `virtio_recv_pkts` is used; otherwise `virtio_recv_mergeable_pkts`.
- For Tx: `virtio_xmit_pkts`.

Vector callbacks will be used when:

- `txq_flags` is set to `VIRTIO_SIMPLE_FLAGS` (0xF01), which implies:
 - Single segment is specified.
 - No offload support is needed.
- Mergeable Rx buffers is disabled.

The corresponding callbacks are:

- For Rx: `virtio_recv_pkts_vec`.
- For Tx: `virtio_xmit_pkts_simple`.

Example of using the vector version of the virtio poll mode driver in `testpmd`:

```
testpmd -l 0-2 -n 4 -- -i --txqflags=0xF01 --rxq=1 --txq=1 --nb-cores=1
```

32.7 Interrupt mode

There are three kinds of interrupts from a virtio device over PCI bus: config interrupt, Rx interrupts, and Tx interrupts. Config interrupt is used for notification of device configuration changes, especially link status (lsc). Interrupt mode is translated into Rx interrupts in the context of DPDK.

Note: Virtio PMD already has support for receiving lsc from qemu when the link status changes, especially when vhost user disconnects. However, it fails to do that if the VM is created by qemu 2.6.2 or below, since the capability to detect vhost user disconnection is introduced in qemu 2.7.0.

32.7.1 Prerequisites for Rx interrupts

To support Rx interrupts, #. Check if guest kernel supports VFIO-NOIOMMU:

Linux started to support VFIO-NOIOMMU since 4.8.0. Make sure the guest kernel is compiled with:

```
CONFIG_VFIO_NOIOMMU=y
```

1. Properly set msix vectors when starting VM:

Enable multi-queue when starting VM, and specify msix vectors in qemu cmd-line. (N+1) is the minimum, and (2N+2) is mostly recommended.

```
$ (QEMU) ... -device virtio-net-pci,mq=on,vectors=2N+2 ...
```

2. In VM, insert vfio module in NOIOMMU mode:

```
modprobe vfio enable_unsafe_noiommu_mode=1
modprobe vfio-pci
```

3. In VM, bind the virtio device with vfio-pci:

```
python usertools/dpdk-devbind.py -b vfio-pci 00:03.0
```

32.7.2 Example

Here we use l3fwd-power as an example to show how to get started.

Example:

```
$ l3fwd-power -l 0-1 -- -p 1 -P --config="(0,0,1)" \
--no-numa --parse-ptype
```

POLL MODE DRIVER THAT WRAPS VHOST LIBRARY

This PMD is a thin wrapper of the DPDK vhost library. The user can handle virtqueues as one of normal DPDK port.

33.1 Vhost Implementation in DPDK

Please refer to Chapter “Vhost Library” of *DPDK Programmer’s Guide* to know detail of vhost.

33.2 Features and Limitations of vhost PMD

Currently, the vhost PMD provides the basic functionality of packet reception, transmission and event handling.

- It has multiple queues support.
- It supports `RTE_ETH_EVENT_INTR_LSC` and `RTE_ETH_EVENT_QUEUE_STATE` events.
- It supports Port Hotplug functionality.
- Don’t need to stop RX/TX, when the user wants to stop a guest or a virtio-net driver on guest.

33.3 Vhost PMD arguments

The user can specify below arguments in `-vdev` option.

1. `iface:`

It is used to specify a path to connect to a QEMU virtio-net device.

2. `queues:`

It is used to specify the number of queues virtio-net device has. (Default: 1)

3. `iommu-support:`

It is used to enable iommu support in vhost library. (Default: 0 (disabled))

33.4 Vhost PMD event handling

This section describes how to handle vhost PMD events.

The user can register an event callback handler with `rte_eth_dev_callback_register()`. The registered callback handler will be invoked with one of below event types.

1. `RTE_ETH_EVENT_INTR_LSC`:

It means link status of the port was changed.

2. `RTE_ETH_EVENT_QUEUE_STATE`:

It means some of queue statuses were changed. Call `rte_eth_vhost_get_queue_event()` in the callback handler. Because changing multiple statuses may occur only one event, call the function repeatedly as long as it doesn't return negative value.

33.5 Vhost PMD with testpmd application

This section demonstrates vhost PMD with testpmd DPDK sample application.

1. Launch the testpmd with vhost PMD:

```
./testpmd -l 0-3 -n 4 --vdev 'net_vhost0,iface=/tmp/sock0,queues=1' -- -i
```

Other basic DPDK preparations like hugepage enabling here. Please refer to the *DPDK Getting Started Guide* for detailed instructions.

2. Launch the QEMU:

```
qemu-system-x86_64 <snip>
    -chardev socket,id=chr0,path=/tmp/sock0 \
    -netdev vhost-user,id=net0,chardev=chr0,vhostforce,queues=1 \
    -device virtio-net-pci,netdev=net0
```

This command attaches one virtio-net device to QEMU guest. After initialization processes between QEMU and DPDK vhost library are done, status of the port will be linked up.

POLL MODE DRIVER FOR PARAVIRTUAL VMXNET3 NIC

The VMXNET3 adapter is the next generation of a paravirtualized NIC, introduced by VMware* ESXi. It is designed for performance, offers all the features available in VMXNET2, and adds several new features such as, multi-queue support (also known as Receive Side Scaling, RSS), IPv6 offloads, and MSI/MSI-X interrupt delivery. One can use the same device in a DPDK application with VMXNET3 PMD introduced in DPDK API.

In this chapter, two setups with the use of the VMXNET3 PMD are demonstrated:

1. Vmxnet3 with a native NIC connected to a vSwitch
2. Vmxnet3 chaining VMs connected to a vSwitch

34.1 VMXNET3 Implementation in the DPDK

For details on the VMXNET3 device, refer to the VMXNET3 driver's vmxnet3 directory and support manual from VMware*.

For performance details, refer to the following link from VMware:

http://www.vmware.com/pdf/vsp_4_vmxnet3_perf.pdf

As a PMD, the VMXNET3 driver provides the packet reception and transmission callbacks, `vmxnet3_recv_pkts` and `vmxnet3_xmit_pkts`.

The VMXNET3 PMD handles all the packet buffer memory allocation and resides in guest address space and it is solely responsible to free that memory when not needed. The packet buffers and features to be supported are made available to hypervisor via VMXNET3 PCI configuration space BARs. During RX/TX, the packet buffers are exchanged by their GPAs, and the hypervisor loads the buffers with packets in the RX case and sends packets to vSwitch in the TX case.

The VMXNET3 PMD is compiled with `vmxnet3` device headers. The interface is similar to that of the other PMDs available in the DPDK API. The driver pre-allocates the packet buffers and loads the command ring descriptors in advance. The hypervisor fills those packet buffers on packet arrival and write completion ring descriptors, which are eventually pulled by the PMD. After reception, the DPDK application frees the descriptors and loads new packet buffers for the coming packets. The interrupts are disabled and there is no notification required. This keeps performance up on the RX side, even though the device provides a notification feature.

In the transmit routine, the DPDK application fills packet buffer pointers in the descriptors of the command ring and notifies the hypervisor. In response the hypervisor takes packets and passes them to the vSwitch, It writes into the completion descriptors ring. The rings are read

by the PMD in the next transmit routine call and the buffers and descriptors are freed from memory.

34.2 Features and Limitations of VMXNET3 PMD

In release 1.6.0, the VMXNET3 PMD provides the basic functionality of packet reception and transmission. There are several options available for filtering packets at VMXNET3 device level including:

1. MAC Address based filtering:
 - Unicast, Broadcast, All Multicast modes - SUPPORTED BY DEFAULT
 - Multicast with Multicast Filter table - NOT SUPPORTED
 - Promiscuous mode - SUPPORTED
 - RSS based load balancing between queues - SUPPORTED
2. VLAN filtering:
 - VLAN tag based filtering without load balancing - SUPPORTED

Note:

- Release 1.6.0 does not support separate headers and body receive cmd_ring and hence, multiple segment buffers are not supported. Only cmd_ring_0 is used for packet buffers, one for each descriptor.
- Receive and transmit of scattered packets is not supported.
- Multicast with Multicast Filter table is not supported.

34.3 Prerequisites

The following prerequisites apply:

- Before starting a VM, a VMXNET3 interface to a VM through VMware vSphere Client must be assigned. This is shown in the figure below.

Note: Depending on the Virtual Machine type, the VMware vSphere Client shows Ethernet adaptors while adding an Ethernet device. Ensure that the VM type used offers a VMXNET3 device. Refer to the VMware documentation for a listed of VMs.

Note: Follow the *DPDK Getting Started Guide* to setup the basic DPDK environment.

Note: Follow the *DPDK Sample Application's User Guide*, L2 Forwarding/L3 Forwarding and TestPMD for instructions on how to run a DPDK application using an assigned VMXNET3 device.

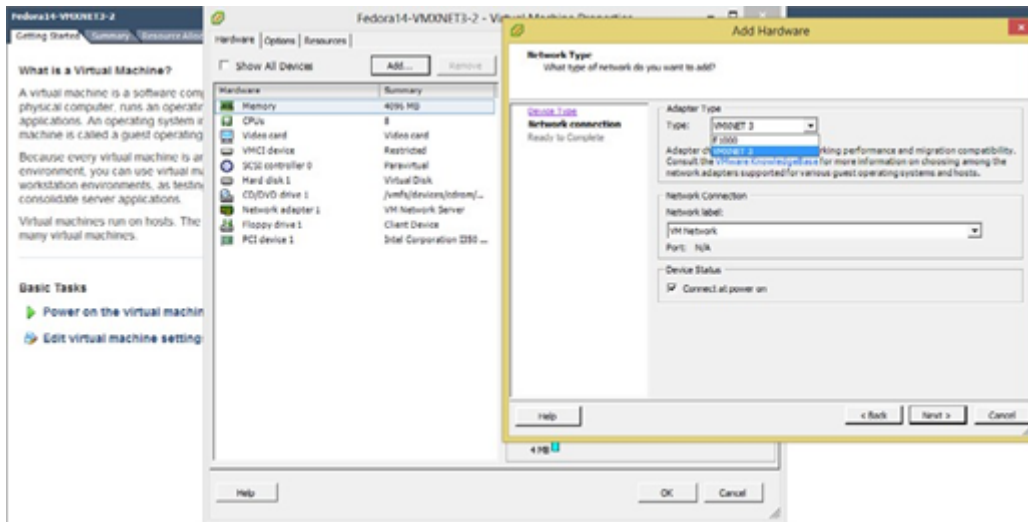


Fig. 34.1: Assigning a VMXNET3 interface to a VM using VMware vSphere Client

34.4 VMXNET3 with a Native NIC Connected to a vSwitch

This section describes an example setup for Phy-vSwitch-VM-Phy communication.

Note: Other instructions on preparing to use DPDK such as, hugepage enabling, uio port binding are not listed here. Please refer to *DPDK Getting Started Guide and DPDK Sample Application's User Guide* for detailed instructions.

The packet reception and transmission flow path is:

```
Packet generator -> 82576
                  -> VMware ESXi vSwitch
                  -> VMXNET3 device
                  -> Guest VM VMXNET3 port 0 rx burst
                  -> Guest VM 82599 VF port 0 tx burst
                  -> 82599 VF
                  -> Packet generator
```

34.5 VMXNET3 Chaining VMs Connected to a vSwitch

The following figure shows an example VM-to-VM communication over a Phy-VM-vSwitch-VM-Phy communication channel.

Note: When using the L2 Forwarding or L3 Forwarding applications, a destination MAC address needs to be written in packets to hit the other VM's VMXNET3 interface.

In this example, the packet flow path is:

```
Packet generator -> 82599 VF
                  -> Guest VM 82599 port 0 rx burst
                  -> Guest VM VMXNET3 port 1 tx burst
                  -> VMXNET3 device
                  -> VMware ESXi vSwitch
                  -> VMXNET3 device
                  -> Guest VM VMXNET3 port 0 rx burst
```

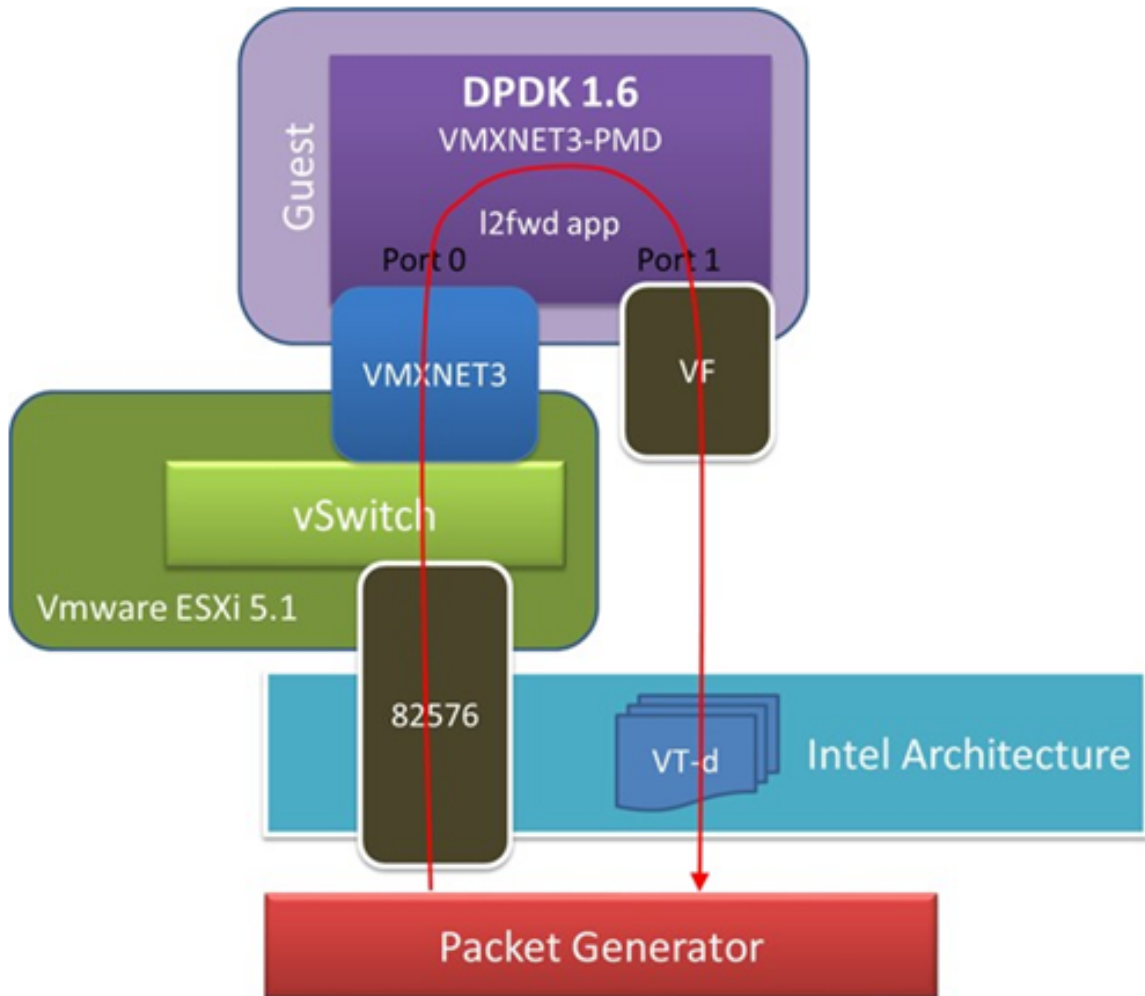



Fig. 34.2: VMXNET3 with a Native NIC Connected to a vSwitch

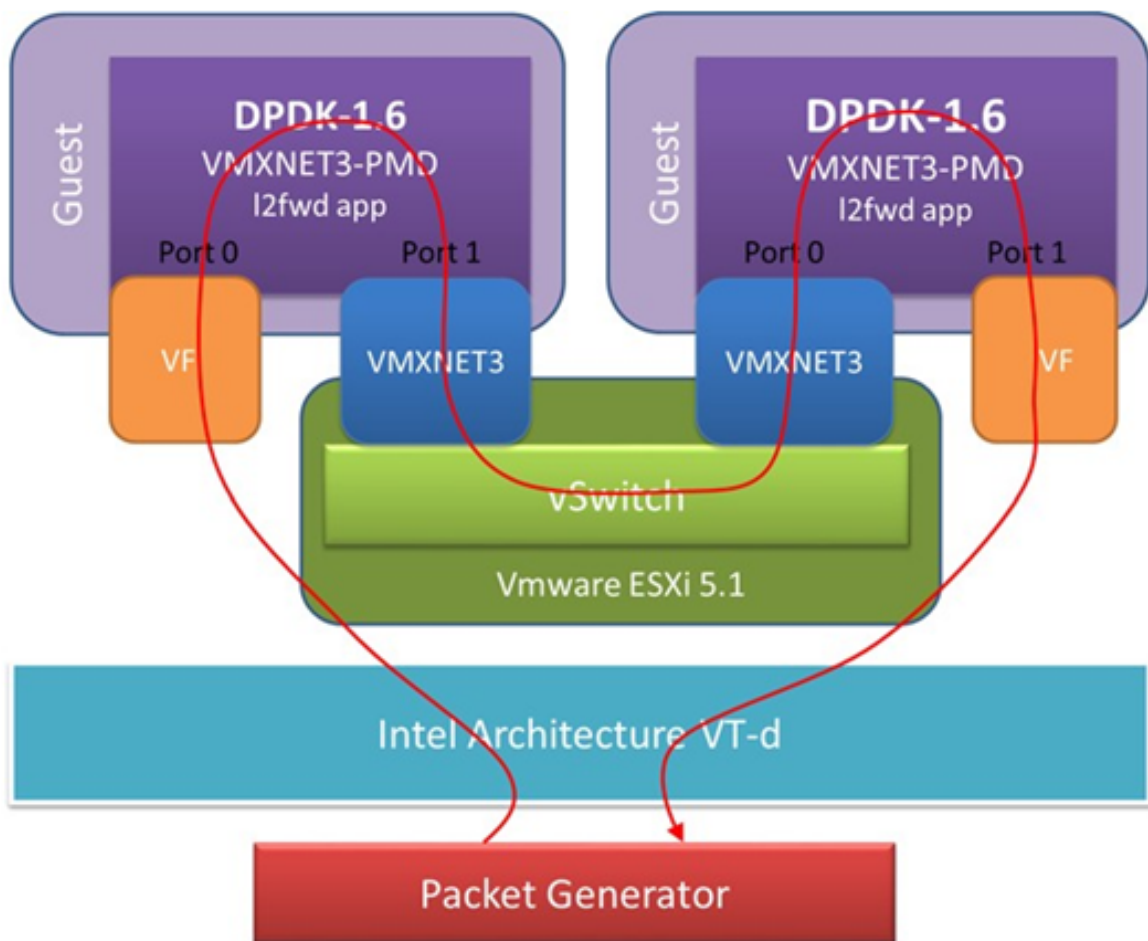


Fig. 34.3: VMXNET3 Chaining VMs Connected to a vSwitch

```
-> Guest VM 82599 VF port 1 tx burst  
-> 82599 VF  
-> Packet generator
```

LIBPCAP AND RING BASED POLL MODE DRIVERS

In addition to Poll Mode Drivers (PMDs) for physical and virtual hardware, the DPDK also includes pure-software PMDs, two of these drivers are:

- A libpcap -based PMD (`librte_pmd_pcap`) that reads and writes packets using libpcap, - both from files on disk, as well as from physical NIC devices using standard Linux kernel drivers.
- A ring-based PMD (`librte_pmd_ring`) that allows a set of software FIFOs (that is, `rte_ring`) to be accessed using the PMD APIs, as though they were physical NICs.

Note: The libpcap -based PMD is disabled by default in the build configuration files, owing to an external dependency on the libpcap development files which must be installed on the board. Once the libpcap development files are installed, the library can be enabled by setting `CONFIG_RTE_LIBRTE_PMD_PCAP=y` and recompiling the DPDK.

35.1 Using the Drivers from the EAL Command Line

For ease of use, the DPDK EAL also has been extended to allow pseudo-Ethernet devices, using one or more of these drivers, to be created at application startup time during EAL initialization.

To do so, the `-vdev=` parameter must be passed to the EAL. This takes take options to allow ring and pcap-based Ethernet to be allocated and used transparently by the application. This can be used, for example, for testing on a virtual machine where there are no Ethernet ports.

35.1.1 Libpcap-based PMD

Pcap-based devices can be created using the virtual device `-vdev` option. The device name must start with the `net_pcap` prefix followed by numbers or letters. The name is unique for each device. Each device can have multiple stream options and multiple devices can be used. Multiple device definitions can be arranged using multiple `-vdev`. Device name and stream options must be separated by commas as shown below:

```
$RTE_TARGET/app/testpmd -l 0-3 -n 4 \  
  --vdev 'net_pcap0,stream_opt0=..,stream_opt1=..' \  
  --vdev='net_pcap1,stream_opt0=..'
```

Device Streams

Multiple ways of stream definitions can be assessed and combined as long as the following two rules are respected:

- A device is provided with two different streams - reception and transmission.
- A device is provided with one network interface name used for reading and writing packets.

The different stream types are:

- `rx_pcap`: Defines a reception stream based on a pcap file. The driver reads each packet within the given pcap file as if it was receiving it from the wire. The value is a path to a valid pcap file.

```
rx_pcap=/path/to/file.pcap
```

- `tx_pcap`: Defines a transmission stream based on a pcap file. The driver writes each received packet to the given pcap file. The value is a path to a pcap file. The file is overwritten if it already exists and it is created if it does not.

```
tx_pcap=/path/to/file.pcap
```

- `rx_iface`: Defines a reception stream based on a network interface name. The driver reads packets coming from the given interface using the Linux kernel driver for that interface. The value is an interface name.

```
rx_iface=eth0
```

- `tx_iface`: Defines a transmission stream based on a network interface name. The driver sends packets to the given interface using the Linux kernel driver for that interface. The value is an interface name.

```
tx_iface=eth0
```

- `iface`: Defines a device mapping a network interface. The driver both reads and writes packets from and to the given interface. The value is an interface name.

```
iface=eth0
```

Examples of Usage

Read packets from one pcap file and write them to another:

```
$RTE_TARGET/app/testpmd -l 0-3 -n 4 \
  --vdev 'net_pcap0,rx_pcap=file_rx.pcap,tx_pcap=file_tx.pcap' \
  -- --port-topology=chained
```

Read packets from a network interface and write them to a pcap file:

```
$RTE_TARGET/app/testpmd -l 0-3 -n 4 \
  --vdev 'net_pcap0,rx_iface=eth0,tx_pcap=file_tx.pcap' \
  -- --port-topology=chained
```

Read packets from a pcap file and write them to a network interface:

```
$RTE_TARGET/app/testpmd -l 0-3 -n 4 \
  --vdev 'net_pcap0,rx_pcap=file_rx.pcap,tx_iface=eth1' \
  -- --port-topology=chained
```

Forward packets through two network interfaces:

```
$RTE_TARGET/app/testpmd -l 0-3 -n 4 \
  --vdev 'net_pcap0,iface=eth0' --vdev='net_pcap1;iface=eth1'
```

Using libpcap-based PMD with the testpmd Application

One of the first things that testpmd does before starting to forward packets is to flush the RX streams by reading the first 512 packets on every RX stream and discarding them. When using a libpcap-based PMD this behavior can be turned off using the following command line option:

```
--no-flush-rx
```

It is also available in the runtime command line:

```
set flush_rx on/off
```

It is useful for the case where the rx_pcap is being used and no packets are meant to be discarded. Otherwise, the first 512 packets from the input pcap file will be discarded by the RX flushing operation.

```
$RTE_TARGET/app/testpmd -l 0-3 -n 4 \
  --vdev 'net_pcap0,rx_pcap=file_rx.pcap,tx_pcap=file_tx.pcap' \
  -- --port-topology=chained --no-flush-rx
```

Note: The network interface provided to the PMD should be up. The PMD will return an error if interface is down, and the PMD itself won't change the status of the external network interface.

35.1.2 Rings-based PMD

To run a DPDK application on a machine without any Ethernet devices, a pair of ring-based rte_ethdevs can be used as below. The device names passed to the `--vdev` option must start with `net_ring` and take no additional parameters. Multiple devices may be specified, separated by commas.

```
./testpmd -l 1-3 -n 4 --vdev=net_ring0 --vdev=net_ring1 -- -i
EAL: Detected lcore 1 as core 1 on socket 0
...

Interactive-mode selected
Configuring Port 0 (socket 0)
Configuring Port 1 (socket 0)
Checking link statuses...
Port 0 Link Up - speed 10000 Mbps - full-duplex
Port 1 Link Up - speed 10000 Mbps - full-duplex
Done

testpmd> start tx_first
io packet forwarding - CRC stripping disabled - packets/burst=16
nb forwarding cores=1 - nb forwarding ports=2
RX queues=1 - RX desc=128 - RX free threshold=0
RX threshold registers: pthresh=8 hthresh=8 wthresh=4
TX queues=1 - TX desc=512 - TX free threshold=0
TX threshold registers: pthresh=36 hthresh=0 wthresh=0
TX RS bit threshold=0 - TXQ flags=0x0

testpmd> stop
Telling cores to stop...
Waiting for lcores to finish...
```

```

----- Forward statistics for port 0 -----
RX-packets: 231192368      RX-dropped: 0      RX-total: 231192368
TX-packets: 231192384      TX-dropped: 0      TX-total: 231192384
-----

----- Forward statistics for port 1 -----
RX-packets: 231192368      RX-dropped: 0      RX-total: 231192368
TX-packets: 231192384      TX-dropped: 0      TX-total: 231192384
-----

+++++++ Accumulated forward statistics for allports+++++++
RX-packets: 462384736  RX-dropped: 0  RX-total: 462384736
TX-packets: 462384768  TX-dropped: 0  TX-total: 462384768
+++++++

Done.

```

35.1.3 Using the Poll Mode Driver from an Application

Both drivers can provide similar APIs to allow the user to create a PMD, that is, `rte_ethdev` structure, instances at run-time in the end-application, for example, using `rte_eth_from_rings()` or `rte_eth_from_pcaps()` APIs. For the rings-based PMD, this functionality could be used, for example, to allow data exchange between cores using rings to be done in exactly the same way as sending or receiving packets from an Ethernet device. For the libpcap-based PMD, it allows an application to open one or more pcap files and use these as a source of packet input to the application.

Usage Examples

To create two pseudo-Ethernet ports where all traffic sent to a port is looped back for reception on the same port (error handling omitted for clarity):

```

#define RING_SIZE 256
#define NUM_RINGS 2
#define SOCKET0 0

struct rte_ring *ring[NUM_RINGS];
int port0, port1;

ring[0] = rte_ring_create("R0", RING_SIZE, SOCKET0, RING_F_SP_ENQ|RING_F_SC_DEQ);
ring[1] = rte_ring_create("R1", RING_SIZE, SOCKET0, RING_F_SP_ENQ|RING_F_SC_DEQ);

/* create two ethdev's */

port0 = rte_eth_from_rings("net_ring0", ring, NUM_RINGS, ring, NUM_RINGS, SOCKET0);
port1 = rte_eth_from_rings("net_ring1", ring, NUM_RINGS, ring, NUM_RINGS, SOCKET0);

```

To create two pseudo-Ethernet ports where the traffic is switched between them, that is, traffic sent to port 0 is read back from port 1 and vice-versa, the final two lines could be changed as below:

```

port0 = rte_eth_from_rings("net_ring0", &ring[0], 1, &ring[1], 1, SOCKET0);
port1 = rte_eth_from_rings("net_ring1", &ring[1], 1, &ring[0], 1, SOCKET0);

```

This type of configuration could be useful in a pipeline model, for example, where one may want to have inter-core communication using pseudo Ethernet devices rather than raw rings, for reasons of API consistency.

Enqueuing and dequeuing items from an `rte_ring` using the rings-based PMD may be slower than using the native rings API. This is because DPDK Ethernet drivers make use of function pointers to call the appropriate enqueue or dequeue functions, while the `rte_ring` specific functions are direct function calls in the code and are often inlined by the compiler.

Once an `ethdev` has been created, for either a ring or a pcap-based PMD, it should be configured and started in the same way as a regular Ethernet device, that is, by calling `rte_eth_dev_configure()` to set the number of receive and transmit queues, then calling `rte_eth_rx_queue_setup()` / `tx_queue_setup()` for each of those queues and finally calling `rte_eth_dev_start()` to allow transmission and reception of packets to begin.

FAIL-SAFE POLL MODE DRIVER LIBRARY

The Fail-safe poll mode driver library (**librte_pmd_failsafe**) is a virtual device that allows using any device supporting hotplug (sudden device removal and plugging on its bus), without modifying other components relying on such device (application, other PMDs).

Additionally to the Seamless Hotplug feature, the Fail-safe PMD offers the ability to redirect operations to secondary devices when the primary has been removed from the system.

Note: The library is enabled by default. You can enable it or disable it manually by setting the `CONFIG_RTE_LIBRTE_PMD_FAILSAFE` configuration option.

36.1 Features

The Fail-safe PMD only supports a limited set of features. If you plan to use a device underneath the Fail-safe PMD with a specific feature, this feature must be supported by the Fail-safe PMD to avoid throwing any error.

A notable exception is the device removal feature. The fail-safe PMD being a virtual device, it cannot currently be removed in the sense of a specific bus hotplug, like for PCI for example. It will however enable this feature for its sub-device automatically, detecting those that are capable and register the relevant callback for such event.

Check the feature matrix for the complete set of supported features.

36.2 Compilation option

This option can be modified in the `$RTE_TARGET/build/.config` file.

- `CONFIG_RTE_LIBRTE_PMD_FAILSAFE` (default **y**)
Toggle compiling `librte_pmd_failsafe`.

36.3 Using the Fail-safe PMD from the EAL command line

The Fail-safe PMD can be used like most other DPDK virtual devices, by passing a `--vdev` parameter to the EAL when starting the application. The device name must start with the `net_failsafe` prefix, followed by numbers or letters. This name must be unique for each device. Each fail-safe instance must have at least one sub-device, up to `RTE_MAX_ETHPORTS-1`.

A sub-device can be any legal DPDK device, including possibly another fail-safe instance.

36.3.1 Fail-safe command line parameters

- **dev(<iface>)** parameter

This parameter allows the user to define a sub-device. The <iface> part of this parameter must be a valid device definition. It could be the argument provided to any `-w` device specification or the argument that would be given to a `--vdev` parameter (including a fail-safe). Enclosing the device definition within parenthesis here allows using additional sub-device parameters if need be. They will be passed on to the sub-device.

Note: In case of whitelist sub-device probed by EAL, fail-safe PMD will take the device as is, which means that EAL device options are taken in this case. When trying to use a PCI device automatically probed in blacklist mode, the syntax for the fail-safe must be with the full PCI id: Domain:Bus:Device.Function. See the usage example section.

- **exec(<shell command>)** parameter

This parameter allows the user to provide a command to the fail-safe PMD to execute and define a sub-device. It is done within a regular shell context. The first line of its output is read by the fail-safe PMD and otherwise interpreted as if passed by the regular **dev** parameter. Any other line is discarded. If the command fail or output an incorrect string, the sub-device is not initialized. All commas within the `shell command` are replaced by spaces before executing the command. This helps using scripts to specify devices.

- **fd(<file descriptor number>)** parameter

This parameter reads a device definition from an arbitrary file descriptor number in <iface> format as described above.

The file descriptor is read in non-blocking mode and is never closed in order to take only the last line into account (unlike `exec()`) at every probe attempt.

- **mac** parameter [MAC address]

This parameter allows the user to set a default MAC address to the fail-safe and all of its sub-devices. If no default mac address is provided, the fail-safe PMD will read the MAC address of the first of its sub-device to be successfully probed and use it as its default MAC address, trying to set it to all of its other sub-devices. If no sub-device was successfully probed at initialization, then a random MAC address is generated, that will be subsequently applied to all sub-device once they are probed.

- **hotplug_poll** parameter [UINT64] (default **2000**)

This parameter allows the user to configure the amount of time in milliseconds between two slave upkeep round.

36.3.2 Usage example

This section shows some example of using **testpmd** with a fail-safe PMD.

1. To build a PMD and configure DPDK, refer to the document *compiling and testing a PMD for a NIC*.

2. Start testpmd. The slave device should be blacklisted from normal EAL operations to avoid probing it twice when in PCI blacklist mode.

```
$RTE_TARGET/build/app/testpmd -c 0xff -n 4 \
  --vdev 'net_failsafe0,mac=de:ad:be:ef:01:02,dev(84:00.0),dev(net_ring0)' \
  -b 84:00.0 -b 00:04.0 -- -i
```

If the slave device being used is not blacklisted, it will be probed by the EAL first. When the fail-safe then tries to initialize it the probe operation fails.

Note that PCI blacklist mode is the default PCI operating mode.

3. Alternatively, it can be used alongside any other device in whitelist mode.

```
$RTE_TARGET/build/app/testpmd -c 0xff -n 4 \
  --vdev 'net_failsafe0,mac=de:ad:be:ef:01:02,dev(84:00.0),dev(net_ring0)' \
  -w 81:00.0 -- -i
```

4. Start testpmd using a flexible device definition

```
$RTE_TARGET/build/app/testpmd -c 0xff -n 4 --no-pci \
  --vdev='net_failsafe0,exec(echo 84:00.0)' -- -i
```

5. Start testpmd, automatically probing the device 84:00.0 and using it with the fail-safe.

```
$RTE_TARGET/build/app/testpmd -c 0xff -n 4 \
  --vdev 'net_failsafe0,dev(0000:84:00.0),dev(net_ring0)' -- -i
```

36.4 Using the Fail-safe PMD from an application

This driver strives to be as seamless as possible to existing applications, in order to propose the hotplug functionality in the easiest way possible.

Care must be taken, however, to respect the **ether** API concerning device access, and in particular, using the `RTE_ETH_FOREACH_DEV` macro to iterate over ethernet devices, instead of directly accessing them or by writing one's own device iterator.

36.5 Plug-in feature

A sub-device can be defined without existing on the system when the fail-safe PMD is initialized. Upon probing this device, the fail-safe PMD will detect its absence and postpone its use. It will then register for a periodic check on any missing sub-device.

During this time, the fail-safe PMD can be used normally, configured and told to emit and receive packets. It will store any applied configuration, and try to apply it upon the probing of its missing sub-device. After this configuration pass, the new sub-device will be synchronized with other sub-devices, i.e. be started if the fail-safe PMD has been started by the user before.

36.6 Plug-out feature

A sub-device supporting the device removal event can be removed from its bus at any time. The fail-safe PMD will register a callback for such event and react accordingly. It will try to safely stop, close and uninit the sub-device having emitted this event, allowing it to free its eventual resources.

36.7 Fail-safe glossary

Fallback device [Secondary device] The fail-safe will fail-over onto this device when the preferred device is absent.

Preferred device [Primary device] The first declared sub-device in the fail-safe parameters. When this device is plugged, it is always used as emitting device. It is the main sub-device and is used as target for configuration operations if there is any ambiguity.

Upkeep round Periodical process when slaves are serviced. Each devices having a state different to that of the fail-safe device itself, is synchronized with it. Additionally, each slave having the remove flag set are cleaned-up.

Slave In the context of the fail-safe PMD, synonymous to sub-device.

Sub-device A device being utilized by the fail-safe PMD. This is another PMD running underneath the fail-safe PMD. Any sub-device can disappear at any time. The fail-safe will ensure that the device removal happens gracefully.

Figures

Fig. 18.1 *Virtualization for a Single Port NIC in SR-IOV Mode*

Fig. 18.2 *Performance Benchmark Setup*

Fig. 18.3 *Fast Host-based Packet Processing*

Fig. 18.4 *Inter-VM Communication*

Fig. 32.1 *Host2VM Communication Example Using kni vhost Back End*

Fig. 32.3 *Host2VM Communication Example Using qemu vhost Back End*

Fig. 34.1 *Assigning a VMXNET3 interface to a VM using VMware vSphere Client*

Fig. 34.2 *VMXNET3 with a Native NIC Connected to a vSwitch*

Fig. 34.3 *VMXNET3 Chaining VMs Connected to a vSwitch*