



DPDK

DATA PLANE DEVELOPMENT KIT

DPDK Release Notes

Release 2.2.0

January 16, 2016

1	Description of Release	1
2	DPDK Release 2.2	2
2.1	New Features	2
2.2	Resolved Issues	5
2.3	Known Issues	9
2.4	API Changes	10
2.5	ABI Changes	10
2.6	Shared Library Versions	11
3	DPDK Release 2.1	12
3.1	New Features	12
3.2	Resolved Issues	18
3.3	Known Issues	25
3.4	API Changes	25
3.5	ABI Changes	25
4	DPDK Release 2.0	26
4.1	New Features	26
5	DPDK Release 1.8	28
5.1	New Features	28
6	Supported Operating Systems	29
7	Known Issues and Limitations in Legacy Releases	30
7.1	Unit Test for Link Bonding may fail at test_tlb_tx_burst()	30
7.2	Pause Frame Forwarding does not work properly on igb	30
7.3	In packets provided by the PMD, some flags are missing	30
7.4	The rte_malloc library is not fully implemented	31
7.5	HPET reading is slow	31
7.6	HPET timers do not work on the Osage customer reference platform	31
7.7	Not all variants of supported NIC types have been used in testing	32
7.8	Multi-process sample app requires exact memory mapping	32
7.9	Packets are not sent by the 1 GbE/10 GbE SR-IOV driver when the source MAC is not the MAC assigned to the VF NIC	33
7.10	SR-IOV drivers do not fully implement the rte_ethdev API	33
7.11	PMD does not work with --no-huge EAL command line parameter	34
7.12	Some hardware off-load functions are not supported by the VF Driver	34
7.13	Kernel crash on IGB port unbinding	34

7.14	Twinpond and Ironpond NICs do not report link status correctly	35
7.15	Discrepancies between statistics reported by different NICs	35
7.16	Error reported opening files on DPDK initialization	35
7.17	Intel® QuickAssist Technology sample application does not work on a 32-bit OS on Shumway	35
7.18	Differences in how different Intel NICs handle maximum packet length for jumbo frame	36
7.19	Binding PCI devices to igb_uio fails on Linux kernel 3.9 when more than one device is used	36
7.20	GCC might generate Intel® AVX instructions for processors without Intel® AVX support	36
7.21	Ethertype filter could receive other packets (non-assigned) in Niantic	37
7.22	Cannot set link speed on Intel® 40G Ethernet controller	37
7.23	Stopping the port does not down the link on Intel® 40G Ethernet controller	37
7.24	Devices bound to igb_uio with VT-d enabled do not work on Linux kernel 3.15-3.17	38
7.25	VM power manager may not work on systems with more than 64 cores	38
8	ABI and API Deprecation	39
8.1	Deprecation Notices	39

DESCRIPTION OF RELEASE

This document contains the release notes for Data Plane Development Kit (DPDK) release version 2.2.0 and previous releases.

It lists new features, fixed bugs, API and ABI changes and known issues.

For instructions on compiling and running the release, see the DPDK Getting Started Guide.

DPDK RELEASE 2.2

2.1 New Features

- **Introduce ARMv7 and ARMv8 architectures.**

- It is now possible to build DPDK for the ARMv7 and ARMv8 platforms.
- ARMv7 can be tested with virtual PMD drivers.
- ARMv8 can be tested with virtual and physical PMD drivers.

- **Enabled freeing of ring.**

A new function `rte_ring_free()` has been added to allow the user to free a ring if it was created with `rte_ring_create()`.

- **Added keepalive support to EAL and example application.**

- **Added experimental cryptodev API**

The cryptographic processing of packets is provided as a preview with two drivers for:

- Intel QuickAssist devices
- Intel AES-NI multi-buffer library

Due to its experimental state, the API may change without prior notice.

- **Added ethdev APIs for additional IEEE1588 support.**

Added functions to read, write and adjust system time in the NIC. Added client slave sample application to demonstrate the IEEE1588 functionality.

- **Extended Statistics.**

Defined an extended statistics naming scheme to store metadata in the name string of each statistic. Refer to the Extended Statistics section of the Programmers Guide for more details.

Implemented the extended statistics API for the following PMDs:

- `igb`
- `igbvf`
- `i40e`
- `i40evf`
- `fm10k`

- virtio

- **Added API in ethdev to retrieve RX/TX queue information.**

- Added the ability for the upper layer to query RX/TX queue information.
- Added new fields in `rte_eth_dev_info` to represent information about RX/TX descriptors min/max/align numbers, per queue, for the device.

- **Added RSS dynamic configuration to bonding.**

- **Updated the e1000 base driver.**

The e1000 base driver was updated with several features including the following:

- Added new i218 devices
- Allowed both ULP and EEE in Sx state
- Initialized 88E1543 (Marvell 1543) PHY
- Added flags to set EEE advertisement modes
- Supported inverted format ETrackId
- Added bit to disable packetbuffer read
- Added defaults for i210 RX/TX PBSIZE
- Check more errors for ESB2 init and reset
- Check more NVM read errors
- Return code after setting receive address register
- Removed all NAHUM6LP_HW tags

- **Added e1000 RX interrupt support.**

- **Added igb TSO support for both PF and VF.**

- **Added RSS enhancements to Intel x550 NIC.**

- Added support for 512 entry RSS redirection table.
- Added support for per VF RSS redirection table.

- **Added Flow director enhancements on Intel x550 NIC.**

- Added 2 new flow director modes on x550. One is MAC VLAN mode, the other is tunnel mode.

- **Updated the i40e base driver.**

The i40e base driver was updated with several changes including the following:

- Added promiscuous on VLAN support
- Added a workaround to drop all flow control frames
- Added VF capabilities to virtual channel interface
- Added TX Scheduling related AQ commands
- Added additional PCTYPES supported for FortPark RSS
- Added parsing for CEE DCBX TLVs

- Added FortPark specific registers
- Added AQ functions to handle RSS Key and LUT programming
- Increased PF reset max loop limit
- **Added i40e vector RX/TX.**
- **Added i40e RX interrupt support.**
- **Added i40e flow control support.**
- **Added DCB support to i40e PF driver.**
- **Added RSS/FD input set granularity on Intel X710/XL710.**
- **Added different GRE key length for input set on Intel X710/XL710.**
- **Added flow director support in i40e VF.**
- **Added i40e support of early X722 series.**
 Added early X722 support, for evaluation only, as the hardware is alpha.
- **Added fm10k vector RX/TX.**
- **Added fm10k TSO support for both PF and VF.**
- **Added fm10k VMDQ support.**
- **New NIC Boulder Rapid support.**
 Added support for the Boulder Rapid variant of Intel's fm10k NIC family.
- **Enhanced support for the Chelsio CXGBE driver.**
 - Added support for Jumbo Frames.
 - Optimized forwarding performance for Chelsio T5 40GbE cards.
- **Improved enic TX packet rate.**
 Reduced frequency of TX tail pointer updates to the NIC.
- **Added support for link status interrupts in mlx4.**
- **Added partial support (TX only) for secondary processes in mlx4.**
- **Added support for Mellanox ConnectX-4 adapters (mlx5).**
 The mlx5 poll-mode driver implements support for Mellanox ConnectX-4 EN and Mellanox ConnectX-4 Lx EN families of 10/25/40/50/100 Gb/s adapters.
 Like mlx4, this PMD is only available for Linux and is disabled by default due to external dependencies (libibverbs and libmlx5).
- **Added driver for Netronome nfp-6xxx card.**
 Support for using Netronome nfp-6xxx with PCI VFs.
- **Added virtual szedata2 driver for COMBO cards.**
 Added virtual PMD for COMBO-100G and COMBO-80G cards. PMD is disabled in default configuration.
- **Enhanced support for virtio driver.**

- Virtio ring layout optimization (fixed avail ring)
- Vector RX
- Simple TX
- **Added vhost-user multiple queue support.**
- **Added port hotplug support to vmxnet3.**
- **Added port hotplug support to xenvirt.**
- **Added ethtool shim and sample application.**
- **Added experimental performance thread example application.**

The new sample application demonstrates L3 forwarding with different threading models: pthreads, cgroups, or lightweight threads. The example includes a simple cooperative scheduler.

Due to its experimental state this application may change without notice. The application is supported only for Linux x86_64.

- **Enhancements to the IP pipeline application.**

The following features have been added to the `ip_pipeline` application;

- Added Multiple Producers/Multiple Consumers (MPSC) and fragmentation/reassembly support to software rings.
- Added a dynamic pipeline reconfiguration feature that allows binding a pipeline to other threads at runtime using CLI commands.
- Added enable/disable of `promisc` mode from `ip_pipeline` configuration file.
- Added check on RX queues and TX queues of each link whether they are used correctly in the `ip_pipeline` configuration file.
- Added flow id parameters to the flow-classification table entries.
- Added more functions to the routing pipeline: ARP table enable/disable, Q-in-Q and MPLS encapsulation, add color (traffic-class for QoS) to the MPLS tag.
- Added flow-actions pipeline for traffic metering/marketing (for e.g. Two Rate Three Color Marker (trTCM)), policer etc.
- Modified the pass-through pipeline's actions-handler to implement a generic approach to extract fields from the packet's header and copy them to packet metadata.

2.2 Resolved Issues

2.2.1 EAL

- **eal/linux: Fixed epoll timeout.**

Fixed issue where the `rte_epoll_wait()` function didn't return when the underlying call to `epoll_wait()` timed out.

2.2.2 Drivers

- **e1000/base: Synchronize PHY interface on non-ME systems.**

On power up, the MAC - PHY interface needs to be set to PCIe, even if the cable is disconnected. In ME systems, the ME handles this on exit from the Sx (Sticky mode) state. In non-ME, the driver handles it. Added a check for non-ME system to the driver code that handles it.

- **e1000/base: Increased timeout of reset check.**

Previously, in `check_reset_block` RSPCIPHY was polled for 100 ms before determining that the ME veto was set. This was not enough and it was increased to 300 ms.

- **e1000/base: Disabled IPv6 extension header parsing on 82575.**

Disabled IPv6 options as per hardware limitation.

- **e1000/base: Prevent ULP flow if cable connected.**

Enabling ULP on link down when the cable is connected caused an infinite loop of link up/down indications in the NDIS driver. The driver now enables ULP only when the cable is disconnected.

- **e1000/base: Support different EEARBC for i210.**

EEARBC has changed on i210. It means EEARBC has a different address on i210 than on other NICs. So, add a new entity named `EEARBC_I210` to the register list and make sure the right one is being used on i210.

- **e1000/base: Fix K1 configuration.**

Added fix for the following updates to the K1 configurations: TX idle period for entering K1 should be 128 ns. Minimum TX idle period in K1 should be 256 ns.

- **e1000/base: Fix link detect flow.**

Fix link detect flow in case where auto-negotiate is not enabled, by calling `e1000_setup_copper_link_generic` instead of `e1000_phy_setup_autoneg`.

- **e1000/base: Fix link check for i354 M88E1112 PHY.**

The `e1000_check_for_link_media_swap()` function is supposed to check PHY page 0 for copper and PHY page 1 for "other" (fiber) links. The driver switched back from page 1 to page 0 too soon, before `e1000_check_for_link_82575()` is executed and was never finding the link on the fiber (other).

If the link is copper, as the M88E1112 page address is set to 1, it should be set back to 0 before checking this link.

- **e1000/base: Fix beacon duration for i217.**

Fix for I217 Packet Loss issue - The Management Engine sets the FEXTNVM4 Beacon Duration incorrectly. This fix ensures that the correct value will always be set. Correct value for this field is 8 usec.

- **e1000/base: Fix TIPG for non 10 half duplex mode.**

TIPG value is increased when setting speed to 10 half duplex to prevent packet loss. However, it was never decreased again when speed changed. This caused performance issues in the NDIS driver. Fix this to restore TIPG to default value on non 10 half duplex.

- **e1000/base: Fix reset of DH89XXCC SGMII.**

For DH89XXCC_SGMII, a write flush leaves registers of this device trashed (0xFFFFFFFF). Add check for this device.

Also, after both Port SW Reset and Device Reset case, the platform should wait at least 3ms before reading any registers. Remove this condition since waiting is conditionally executed only for Device Reset.

- **e1000/base: Fix redundant PHY power down for i210.**

Bit 11 of PHYREG 0 is used to power down PHY. The use of PHYREG 16 is no longer necessary.

- **e1000/base: fix jumbo frame CRC failures.**

Change the value of register 776.20[11:2] for jumbo mode from 0x1A to 0x1F. This is to enlarge the gap between read and write pointers in the TX FIFO.

- **e1000/base: Fix link flap on 82579.**

Several customers have reported a link flap issue on 82579. The symptoms are random and intermittent link losses when 82579 is connected to specific switches. The issue was root caused as an inter-operability problem between the NIC and at least some Broadcom PHYs in the Energy Efficient Ethernet wake mechanism.

To fix the issue, we are disabling the Phase Locked Loop shutdown in 100M Low Power Idle. This solution will cause an increase of power in 100M EEE link. It may cost an additional 28mW in this specific mode.

- **igb: Fixed IEEE1588 frame identification in I210.**

Fixed issue where the flag `PKT_RX_IEEE1588_PTP` was not being set in the Intel I210 NIC, as the EtherType in RX descriptor is in bits 8:10 of Packet Type and not in the default bits 0:2.

- **igb: Fixed VF start with PF stopped.**

VF needs the PF interrupt support initialized even if not started.

- **igb: Fixed VF MAC address when using with DPDK PF.**

Assign a random MAC address in VF when not assigned by PF.

- **igb: Removed CRC bytes from byte counter statistics.**

- **ixgbe: Fixed issue with X550 DCB.**

Fixed a DCB issue with x550 where for 8 TCs (Traffic Classes), if a packet with user priority 6 or 7 was injected to the NIC, then the NIC would only put 3 packets into the queue. There was also a similar issue for 4 TCs.

- **ixgbe: Removed burst size restriction of vector RX.**

Fixed issue where a burst size less than 32 didn't receive anything.

- **ixgbe: Fixed VF start with PF stopped.**

VF needs the PF interrupt support initialized even if not started.

- **ixgbe: Fixed TX hang when RS distance exceeds HW limit.**

Fixed an issue where the TX queue can hang when a lot of highly fragmented packets have to be sent. As part of that fix, `tx_rs_thresh` for ixgbe PMD is not allowed to be greater than 32 to comply with HW restrictions.

- **ixgbe: Fixed rx error statistic counter.**

Fixed an issue that the rx error counter of ixgbe was not accurate. The mac short packet discard count (`mispdc`) was added to the counter. Mac local faults and mac remote faults are removed as they do not count packets but errors, and jabber errors were removed as they are already accounted for by the CRC error counter. Finally the XEC (I3 / I4 checksum error) counter was removed due to errata, see commit 256ff05a9cae for details.

- **ixgbe: Removed CRC bytes from byte counter statistics.**

- **i40e: Fixed base driver allocation when not using first numa node.**

Fixed i40e issue that occurred when a DPDK application didn't initialize ports if memory wasn't available on socket 0.

- **i40e: Fixed maximum of 64 queues per port.**

Fixed an issue in i40e where it would not support more than 64 queues per port, even though the hardware actually supports it. The real number of queues may vary, as long as the total number of queues used in PF, VFs, VMDq and FD does not exceeds the hardware maximum.

- **i40e: Fixed statistics of packets.**

Added discarding packets on VSI to the stats and rectify the old statistics.

- **i40e: Fixed issue of not freeing memzone.**

Fixed an issue of not freeing a memzone in the call to free the memory for adminq DMA.

- **i40e: Removed CRC bytes from byte counter statistics.**

- **mlx: Fixed driver loading.**

The mlx drivers were unable to load when built as a shared library, due to a missing symbol in the mempool library.

- **mlx4: Performance improvements.**

Fixed bugs in TX and RX flows that improves mlx4 performance.

- **mlx4: Fixed TX loss after initialization.**

- **mlx4: Fixed scattered TX with too many segments.**

- **mlx4: Fixed memory registration for indirect mbuf data.**

- **vhost: Fixed Qemu shutdown.**

Fixed issue with libvirt `virsh destroy` not killing the VM.

- **virtio: Fixed crash after changing link state.**

Fixed IO permission in the interrupt handler.

- **virtio: Fixed crash when releasing queue.**

Fixed issue when releasing null control queue.

2.2.3 Libraries

- **hash: Fixed memory allocation of Cuckoo Hash key table.**

Fixed issue where an incorrect Cuckoo Hash key table size could be calculated limiting the size to 4GB.

- **hash: Fixed incorrect lookup if key is all zero.**

Fixed issue in hash library that occurred if an all zero key was not added to the table and the key was looked up, resulting in an incorrect hit.

- **hash: Fixed thread scaling by reducing contention.**

Fixed issue in the hash library where, using multiple cores with hardware transactional memory support, thread scaling did not work, due to the global ring that is shared by all cores.

2.2.4 Examples

- **I3fwd: Fixed crash with IPv6.**
- **vhost_xen: Fixed compile error.**

2.2.5 Other

- This release drops compatibility with Linux kernel 2.6.33. The minimum kernel requirement is now 2.6.34.

2.3 Known Issues

- Some drivers do not fill in the packet type when receiving. As the I3fwd example application requires this info, the i40e vector driver must be disabled to benefit of the packet type with i40e.
- Some (possibly all) VF drivers (e.g. i40evf) do not handle any PF reset events/requests in the VF driver. This means that the VF driver may not work after a PF reset in the host side. The workaround is to avoid triggering any PF reset events/requests on the host side.
- 100G link report support is missing.
- **Mellanox PMDs (mlx4 & mlx5):**
 - PMDs do not support CONFIG_RTE_BUILD_COMBINE_LIBS and CONFIG_RTE_BUILD_SHARED_LIB simultaneously.
 - There is performance degradation for small packets when the PMD is compiled with SGE_WR_N = 4 compared to the performance when SGE_WR_N = 1. If scattered packets are not used it is recommended to compile the PMD with SGE_WR_N = 1.
 - When a Multicast or Broadcast packet is sent to the SR-IOV mlx4 VF, it is returned back to the port.
 - PMDs report “bad” L4 checksum when IP packet is received.

- mlx5 PMD reports “bad” checksum although the packet has “good” checksum. Will be fixed in upcoming MLNX_OFED release.

2.4 API Changes

- The deprecated flow director API is removed. It was replaced by `rte_eth_dev_filter_ctrl()`.
- The `dcb_queue` is renamed to `dcb_tc` in following dcb configuration structures: `rte_eth_dcb_rx_conf`, `rte_eth_dcb_tx_conf`, `rte_eth_vmdq_dcb_conf`, `rte_eth_vmdq_dcb_tx_conf`.
- The `rte_eth_rx_queue_count()` function now returns “int” instead of “uint32_t” to allow the use of negative values as error codes on return.
- The function `rte_eal_pci_close_one()` is removed. It was replaced by `rte_eal_pci_detach()`.
- The deprecated ACL API `ipv4vlan` is removed.
- The deprecated hash function `rte_jhash2()` is removed. It was replaced by `rte_jhash_32b()`.
- The deprecated KNI functions are removed: `rte_kni_create()`, `rte_kni_get_port_id()` and `rte_kni_info_get()`.
- The deprecated ring PMD functions are removed: `rte_eth_ring_pair_create()` and `rte_eth_ring_pair_attach()`.
- The devargs union field `virtual` is renamed to `virt` for C++ compatibility.

2.5 ABI Changes

- The EAL and ethdev structures `rte_intr_handle` and `rte_eth_conf` were changed to support RX interrupt. This was already included in 2.1 under the `CONFIG_RTE_NEXT_ABI` #define.
- The ethdev flow director entries for SCTP were changed. This was already included in 2.1 under the `CONFIG_RTE_NEXT_ABI` #define.
- The ethdev flow director structure `rte_eth_fdir_flow_ext` structure was changed. New fields were added to support flow director filtering in VF.
- The size of the ethdev structure `rte_eth_hash_filter_info` is changed by adding a new element `rte_eth_input_set_conf` in a union.
- New fields `rx_desc_lim` and `tx_desc_lim` are added into `rte_eth_dev_info` structure.
- For debug builds, the functions `rte_eth_rx_burst()`, `rte_eth_tx_burst()`, `rte_eth_rx_descriptor_done()` and `rte_eth_rx_queue_count()` will no longer be separate functions in the DPDK libraries. Instead, they will only be present in the `rte_ethdev.h` header file.
- The maximum number of queues per port `CONFIG_RTE_MAX_QUEUES_PER_PORT` is increased to 1024.

- The mbuf structure was changed to support the unified packet type. This was already included in 2.1 under the `CONFIG_RTE_NEXT_ABI` #define.
- The dummy malloc library is removed. The content was moved into EAL in 2.1.
- The LPM structure is changed. The deprecated field `mem_location` is removed.
- `librte_table` LPM: A new parameter to hold the table name will be added to the LPM table parameter structure.
- `librte_table` hash: The key mask parameter is added to the hash table parameter structure for 8-byte key and 16-byte key extendable bucket and LRU tables.
- `librte_port`: Macros to access the packet meta-data stored within the packet buffer has been adjusted to cover the packet mbuf structure.
- `librte_cfgfile`: Allow longer names and values by increasing the constants `CFG_NAME_LEN` and `CFG_VALUE_LEN` to 64 and 256 respectively.
- `vhost`: a new field `enabled` is added to the `vhost_virtqueue` structure.
- `vhost`: a new field `virt_qp_nb` is added to `virtio_net` structure, and the `virtqueue` field is moved to the end of `virtio_net` structure.
- `vhost`: a new operation `vring_state_changed` is added to `virtio_net_device_ops` structure.
- `vhost`: a few spaces are reserved both at `vhost_virtqueue` and `virtio_net` structure for future extension.

2.6 Shared Library Versions

The libraries prepended with a plus sign were incremented in this version.

```

+ librte_dev.so.2
+ librte_acl.so.2
+ librte_cfgfile.so.2
  librte_cmdline.so.1
  librte_distributor.so.1
+ librte_eal.so.2
+ librte_hash.so.2
  librte_ip_frag.so.1
  librte_ivshmem.so.1
  librte_jobstats.so.1
+ librte_kni.so.2
  librte_kvargs.so.1
+ librte_lpm.so.2
+ librte_mbuf.so.2
  librte_mempool.so.1
  librte_meter.so.1
+ librte_pipeline.so.2
  librte_pmd_bond.so.1
+ librte_pmd_ring.so.2
+ librte_port.so.2
  librte_power.so.1
  librte_reorder.so.1
  librte_ring.so.1
  librte_sched.so.1
+ librte_table.so.2
  librte_timer.so.1
+ librte_vhost.so.2

```

DPDK RELEASE 2.1

3.1 New Features

- **Enabled cloning of indirect mbufs.**

This feature removes a limitation of `rte_pktmbuf_attach()` which generated the warning: “mbuf we’re attaching to must be direct”.

Now, when attaching to an indirect mbuf it is possible to:

- Copy all relevant fields (address, length, offload, ...) as before.
- Get the pointer to the mbuf that embeds the data buffer (direct mbuf), and increase the reference counter.

When detaching the mbuf, we can now retrieve this direct mbuf as the pointer is determined from the buffer address.

- **Extended packet type support.**

In previous releases mbuf packet types were indicated by 6 bits in the `ol_flags`. This was not enough for some supported NICs. For example i40e hardware can recognize more than 150 packet types. Not being able to identify these additional packet types limits access to hardware offload capabilities

So an extended “unified” packet type was added to support all possible PMDs. The 16 bit `packet_type` in the mbuf structure was changed to 32 bits and used for this purpose.

To avoid breaking ABI compatibility, the code changes for this feature are enclosed in a `RTE_NEXT_ABI` ifdef. This is enabled by default but can be turned off for ABI compatibility with DPDK R2.0.

- **Reworked memzone to be allocated by malloc and also support freeing.**

In the memory hierarchy, memsegs are groups of physically contiguous hugepages, memzones are slices of memsegs, and malloc slices memzones into smaller memory chunks.

This feature modifies `malloc()` so it partitions memsegs instead of memzones. Now memzones allocate their memory from the malloc heap.

Backward compatibility with API and ABI are maintained.

This allow memzones, and any other structure based on memzones, for example mem-pools, to be freed. Currently only the API from freeing memzones is supported.

- **Interrupt mode PMD.**

This feature introduces a low-latency one-shot RX interrupt into DPDK. It also adds a polling and interrupt mode switch control example.

DPDK userspace interrupt notification and handling mechanism is based on UIO/VFIO with the following limitations:

- Per queue RX interrupt events are only allowed in VFIO which supports multiple MSI-X vectors.
- In UIO, the RX interrupt shares the same vector with other interrupts. When the RX interrupt and LSC interrupt are both enabled, only the former is available.
- RX interrupt is only implemented for the linuxapp target.
- The feature is only currently enabled for tow PMDs: ixgbe and igb.

- **Packet Framework enhancements.**

Several enhancements were made to the Packet Framework:

- A new configuration file syntax has been introduced for IP pipeline applications. Parsing of the configuration file is changed.
- Implementation of the IP pipeline application is modified to make it more structured and user friendly.
- Implementation of the command line interface (CLI) for each pipeline type has been moved to the separate compilation unit. Syntax of pipeline CLI commands has been changed.
- Initialization of IP pipeline is modified to match the new parameters structure.
- New implementation of pass-through pipeline, firewall pipeline, routing pipeline, and flow classification has been added.
- Master pipeline with CLI interface has been added.
- Added extended documentation of the IP Pipeline.

- **Added API for IEEE1588 timestamping.**

This feature adds an ethdev API to enable, disable and read IEEE1588/802.1AS PTP timestamps from devices that support it. The following functions were added:

- `rte_eth_timesync_enable()`
- `rte_eth_timesync_disable()`
- `rte_eth_timesync_read_rx_timestamp()`
- `rte_eth_timesync_read_tx_timestamp()`

The “ieee1588” forwarding mode in testpmd was also refactored to demonstrate the new API.

- **Added multicast address filtering.**

Added multicast address filtering via a new ethdev function `set_mc_addr_list()`.

This overcomes a limitation in previous releases where the receipt of multicast packets on a given port could only be enabled by invoking the `rte_eth_allmulticast_enable()` function. This method did not work for VFs

in SR-IOV architectures when the host PF driver does not allow these operation on VFs. In such cases, joined multicast addresses had to be added individually to the set of multicast addresses that are filtered by the [VF] port.

- **Added Flow Director extensions.**

Several Flow Director extensions were added such as:

- Support for RSS and Flow Director hashes in vector RX.
- Added Flow Director for L2 payload.

- **Added RSS hash key size query per port.**

This feature supports querying the RSS hash key size of each port. A new field `hash_key_size` has been added in the `rte_eth_dev_info` struct for storing hash key size in bytes.

- **Added userspace ethtool support.**

Added userspace ethtool support to provide a familiar interface for applications that manage devices via kernel-space `ethtool_op` and `net_device_op`.

The initial implementation focuses on operations that can be implemented through existing `netdev` APIs. More operations will be supported in later releases.

- **Updated the ixgbe base driver.**

The ixgbe base driver was updated with several changes including the following:

- Added a new 82599 device id.
- Added new X550 PHY ids.
- Added SFP+ dual-speed support.
- Added wait helper for X550 IOSF accesses.
- Added X550em features.
- Added X557 PHY LEDs support.
- Commands for flow director.
- Issue firmware command when resetting X550em.

See the git log for full details of the ixgbe/base changes.

- **Added additional hotplug support.**

Port hotplug support was added to the following PMDs:

- e1000/igb.
- ixgbe.
- i40e.
- fm10k.
- ring.
- bonding.
- virtio.

Port hotplug support was added to BSD.

- **Added ixgbe LRO support.**

Added LRO support for x540 and 82599 devices.

- **Added extended statistics for ixgbe.**

Implemented `xstats_get()` and `xstats_reset()` in `dev_ops` for ixgbe to expose detailed error statistics to DPDK applications.

These will be implemented for other PMDs in later releases.

- **Added proc_info application.**

Created a new `proc_info` application, by refactoring the existing `dump_cfg` application, to demonstrate the usage of retrieving statistics, and the new extended statistics (see above), for DPDK interfaces.

- **Updated the i40e base driver.**

The i40e base driver was updated with several changes including the following:

- Support for building both PF and VF driver together.
- Support for CEE DCBX on recent firmware versions.
- Replacement of `i40e_debug_read_register()`.
- Rework of `i40e_hmc_get_object_va`.
- Update of shadow RAM read/write functions.
- Enhancement of polling NVM semaphore.
- Enhancements on adminq init and sending asq command.
- Update of get/set LED functions.
- Addition of AOC phy types to case statement in `get_media_type`.
- Support for iSCSI capability.
- Setting of `FLAG_RD` when sending driver version to FW.

See the git log for full details of the i40e/base changes.

- **Added support for port mirroring in i40e.**

Enabled mirror functionality in the i40e driver.

- **Added support for i40e double VLAN, QinQ, stripping and insertion.**

Added support to the i40e driver for offloading double VLAN (QinQ) tags to the mbuf header, and inserting double vlan tags by hardware to the packets to be transmitted. Added a new field `vlan_tci_outer` in the `rte_mbuf` struct, and new flags in `ol_flags` to support this feature.

- **Added fm10k promiscuous mode support.**

Added support for promiscuous/allmulticast enable and disable in the fm10k PF function. VF is not supported yet.

- **Added fm10k jumbo frame support.**

Added support for jumbo frame less than 15K in both VF and PF functions in the fm10k pmd.

- **Added fm10k mac vlan filtering support.**

Added support for the fm10k MAC filter, only available in PF. Updated the VLAN filter to add/delete one static entry in the MAC table for each combination of VLAN and MAC address.

- **Added support for the Broadcom bnx2x driver.**

Added support for the Broadcom NetXtreme II bnx2x driver. It is supported only on Linux 64-bit and disabled by default.

- **Added support for the Chelsio CXGBE driver.**

Added support for the CXGBE Poll Mode Driver for the Chelsio Terminator 5 series of 10G/40G adapters.

- **Enhanced support for Mellanox ConnectX-3 driver (mlx4).**

- Support Mellanox OFED 3.0.
- Improved performance for both RX and TX operations.
- Better link status information.
- Outer L3/L4 checksum offload support.
- Inner L3/L4 checksum offload support for VXLAN.

- **Enabled VMXNET3 vlan filtering.**

Added support for the VLAN filter functionality of the VMXNET3 interface.

- **Added support for vhost live migration.**

Added support to allow live migration of vhost. Without this feature, qemu will report the following error: “migrate: Migration disabled: vhost lacks VHOST_F_LOG_ALL feature”.

- **Added support for pcap jumbo frames.**

Extended the PCAP PMD to support jumbo frames for RX and TX.

- **Added support for the TILE-Gx architecture.**

Added support for the EZchip TILE-Gx family of SoCs.

- **Added hardware memory transactions/lock elision for x86.**

Added the use of hardware memory transactions (HTM) on fast-path for rwlock and spinlock (a.k.a. lock elision). The methods are implemented for x86 using Restricted Transactional Memory instructions (Intel(r) Transactional Synchronization Extensions). The implementation fall-backs to the normal rwlock if HTM is not available or memory transactions fail. This is not a replacement for all rwlock usages since not all critical sections protected by locks are friendly to HTM. For example, an attempt to perform a HW I/O operation inside a hardware memory transaction always aborts the transaction since the CPU is not able to roll-back should the transaction fail. Therefore, hardware transactional locks are not advised to be used around `rte_eth_rx_burst()` and `rte_eth_tx_burst()` calls.

- **Updated Jenkins Hash function**

Updated the version of the Jenkins Hash (jhash) function used in DPDK from the 1996 version to the 2006 version. This gives up to 35% better performance, compared to the original one.

Note, the hashes generated by the updated version differ from the hashes generated by the previous version.

- **Added software implementation of the Toeplitz RSS hash**

Added a software implementation of the Toeplitz hash function used by RSS. It can be used either for packet distribution on a single queue NIC or for simulating RSS computation on a specific NIC (for example after GRE header de-encapsulation).

- **Replaced the existing hash library with a Cuckoo hash implementation.**

Replaced the existing hash library with another approach, using the Cuckoo Hash method to resolve collisions (open addressing). This method pushes items from a full bucket when a new entry must be added to it, storing the evicted entry in an alternative location, using a secondary hash function.

This gives the user the ability to store more entries when a bucket is full, in comparison with the previous implementation.

The API has not been changed, although new fields have been added in the `rte_hash` structure, which has been changed to internal use only.

The main change when creating a new table is that the number of entries per bucket is now fixed, so its parameter is ignored now (it is still there to maintain the same parameters structure).

Also, the maximum burst size in `lookup_burst` function has been increased to 64, to improve performance.

- **Optimized KNI RX burst size computation.**

Optimized KNI RX burst size computation by avoiding checking how many entries are in `kni->rx_q` prior to actually pulling them from the fifo.

- **Added KNI multicast.**

Enabled adding multicast addresses to KNI interfaces by adding an empty callback for `set_rx_mode` (typically used for setting up hardware) so that the `ioctl` succeeds. This is the same thing as the Linux tap interface does.

- **Added cmdline polling mode.**

Added the ability to process console input in the same thread as packet processing by using the `poll()` function.

- **Added VXLAN Tunnel End point sample application.**

Added a Tunnel End point (TEP) sample application that simulates a VXLAN Tunnel Endpoint (VTEP) termination in DPDK. It is used to demonstrate the offload and filtering capabilities of Intel XL710 10/40 GbE NICs for VXLAN packets.

- **Enabled combining of the “-m” and “-no-huge” EAL options.**

Added option to allow combining of the `-m` and `--no-huge` EAL command line options.

This allows user application to run as non-root but with higher memory allocations, and removes a constraint on `--no-huge` mode being limited to 64M.

3.2 Resolved Issues

- **acl: Fix ambiguity between test rules.**

Some test rules had equal priority for the same category. That could cause an ambiguity in building the trie and test results.

- **acl: Fix invalid rule wildness calculation for bitmask field type.**

- **acl: Fix matching rule.**

- **acl: Fix unneeded trie splitting for subset of rules.**

When rebuilding a trie for limited rule-set, don't try to split the rule-set even further.

- **app/testpmd: Fix crash when port id out of bound.**

Fixed issues in testpmd where using a port greater than 32 would cause a seg fault.

Fixes: edab33b1c01d ("app/testpmd: support port hotplug")

- **app/testpmd: Fix reply to a multicast ICMP request.**

Set the IP source and destination addresses in the IP header of the ICMP reply.

- **app/testpmd: fix MAC address in ARP reply.**

Fixed issue where in the `icmpecho` forwarding mode, ARP replies from testpmd contain invalid zero-filled MAC addresses.

Fixes: 31db4d38de72 ("net: change arp header struct declaration")

- **app/testpmd: fix default flow control values.**

Fixes: 422a20a4e62d ("app/testpmd: fix uninitialized flow control variables")

- **bonding: Fix crash when stopping inactive slave.**

- **bonding: Fix device initialization error handling.**

- **bonding: Fix initial link status of slave.**

On Fortville NIC, link status change interrupt callback was not executed when slave in bonding was (re-)started.

- **bonding: Fix socket id for LACP slave.**

Fixes: 46fb43683679 ("bond: add mode 4")

- **bonding: Fix device initialization error handling.**

- **cmdline: Fix small memory leak.**

A function in `cmdline.c` had a return that did not free the buf properly.

- **config: Enable same drivers options for Linux and BSD.**

Enabled vector `ixgbe` and `i40e` bulk alloc for BSD as it is already done for Linux.

Fixes: 304caba12643 (“config: fix bsd options”) Fixes: 0ff3324da2eb (“ixgbe: rework vector pmd following mbuf changes”)

- **devargs: Fix crash on failure.**

This problem occurred when passing an invalid PCI id to the blacklist API in devargs.

- **e1000/i40e: Fix descriptor done flag with odd address.**

- **e1000/igb: fix ieee1588 timestamping initialization.**

Fixed issue with e1000 ieee1588 timestamp initialization. On initialization the IEEE1588 functions read the system time to set their timestamp. However, on some 1G NICs, for example, i350, system time is disabled by default and the IEEE1588 timestamp was always 0.

- **eal/bsd: Fix inappropriate header guards.**

- **eal/bsd: Fix virtio on FreeBSD.**

Closing the `/dev/io` fd caused a SIGBUS in inb/outb instructions as the process lost the IOPL privileges once the fd is closed.

Fixes: 8a312224bcde (“eal/bsd: fix fd leak”)

- **eal/linux: Fix comments on vfio MSI.**

- **eal/linux: Fix irq handling with igb_uio.**

Fixed an issue where the the introduction of `uio_pci_generic` broke interrupt handling with `igb_uio`.

Fixes: c112df6875a5 (“eal/linux: toggle interrupt for uio_pci_generic”)

- **eal/linux: Fix numa node detection.**

- **eal/linux: Fix socket value for undetermined numa node.**

Sets zero as the default value of pci device `numa_node` if the socket could not be determined. This provides the same default value as FreeBSD which has no NUMA support, and makes the return value of `rte_eth_dev_socket_id()` be consistent with the API description.

- **eal/ppc: Fix cpu cycle count for little endian.**

On IBM POWER8 PPC64 little endian architecture, the definition of `tsc` union will be different. This fix enables the right output from `rte_rdtsc()`.

- **ethdev: Fix check of threshold for TX freeing.**

Fixed issue where the parameter to `tx_free_thresh` was not consistent between the drivers.

- **ethdev: Fix crash if malloc of user callback fails.**

If `rte_zmalloc()` failed in `rte_eth_dev_callback_register` then the NULL pointer would be dereferenced.

- **ethdev: Fix illegal port access.**

To obtain a detachable flag, `pci_drv` is accessed in `rte_eth_dev_is_detachable()`. However `pci_drv` is only valid if port is enabled. Fixed by checking `rte_eth_dev_is_valid_port()` first.

- **ethdev: Make tables const.**
- **ethdev: Rename and extend the mirror type.**
- **examples/distributor: Fix debug macro.**

The macro to turn on additional debug output when the app was compiled with `-DDEBUG` was broken.

Fixes: 07db4a975094 (“examples/distributor: new sample app”)

- **examples/kni: Fix crash on exit.**
- **examples/vhost: Fix build with debug enabled.**

Fixes: 72ec8d77ac68 (“examples/vhost: rework duplicated code”)

- **fm10k: Fix RETA table initialization.**

The fm10k driver has 128 RETA entries in 32 registers, but it only initialized the first 32 when doing multiple RX queue configurations. This fix initializes all 128 entries.

- **fm10k: Fix RX buffer size.**
- **fm10k: Fix TX multi-segment frame.**
- **fm10k: Fix TX queue cleaning after start error.**
- **fm10k: Fix Tx queue cleaning after start error.**
- **fm10k: Fix default mac/vlan in switch.**
- **fm10k: Fix interrupt fault handling.**
- **fm10k: Fix jumbo frame issue.**
- **fm10k: Fix mac/vlan filtering.**
- **fm10k: Fix maximum VF number.**
- **fm10k: Fix maximum queue number for VF.**

Both PF and VF shared code in function `fm10k_stats_get()`. The function worked with PF, but had problems with VF since it has less queues than PF.

Fixes: a6061d9e7075 (“fm10k: register PF driver”)

- **fm10k: Fix queue disabling.**
- **fm10k: Fix switch synchronization.**
- **i40e/base: Fix error handling of NVM state update.**
- **i40e/base: Fix hardware port number for pass-through.**
- **i40e/base: Rework virtual address retrieval for lan queue.**
- **i40e/base: Update LED blinking.**
- **i40e/base: Workaround for PHY type with firmware < 4.4.**
- **i40e: Disable setting of PHY configuration.**
- **i40e: Fix SCTP flow director.**

- **i40e: Fix check of descriptor done flag.**

Fixes: 4861cde46116 (“i40e: new poll mode driver”) Fixes: 05999aab4ca6 (“i40e: add or delete flow director”)

- **i40e: Fix condition to get VMDQ info.**
- **i40e: Fix registers access from big endian CPU.**
- **i40evf: Clear command when error occurs.**
- **i40evf: Fix RSS with less RX queues than TX queues.**
- **i40evf: Fix crash when setup TX queues.**
- **i40evf: Fix jumbo frame support.**
- **i40evf: Fix offload capability flags.**

Added checksum offload capability flags which have already been supported for a long time.

- **ivshmem: Fix crash in corner case.**

Fixed issues where depending on the configured segments it was possible to hit a segmentation fault as a result of decrementing an unsigned index with value 0.

Fixes: 40b966a211ab (“ivshmem: library changes for mmaping using ivshmem”)

- **ixgbe/base: Fix SFP probing.**
 - **ixgbe/base: Fix TX pending clearing.**
 - **ixgbe/base: Fix X550 CS4227 address.**
 - **ixgbe/base: Fix X550 PCIe master disabling.**
 - **ixgbe/base: Fix X550 check.**
 - **ixgbe/base: Fix X550 init early return.**
 - **ixgbe/base: Fix X550 link speed.**
 - **ixgbe/base: Fix X550em CS4227 speed mode.**
 - **ixgbe/base: Fix X550em SFP+ link stability.**
 - **ixgbe/base: Fix X550em UniPHY link configuration.**
 - **ixgbe/base: Fix X550em flow control for KR backplane.**
 - **ixgbe/base: Fix X550em flow control to be KR only.**
 - **ixgbe/base: Fix X550em link setup without SFP.**
 - **ixgbe/base: Fix X550em mux after MAC reset.**
- Fixes: d2e72774e58c (“ixgbe/base: support X550”)
- **ixgbe/base: Fix bus type overwrite.**
 - **ixgbe/base: Fix init handling of X550em link down.**
 - **ixgbe/base: Fix lan id before first i2c access.**
 - **ixgbe/base: Fix mac type checks.**

- **ixgbe/base: Fix tunneled UDP and TCP frames in flow director.**
- **ixgbe: Check mbuf refcnt when clearing a ring.**

The function to clear the TX ring when a port was being closed, e.g. on exit in testpmd, was not checking the mbuf refcnt before freeing it. Since the function in the vector driver to clear the ring after TX does not setting the pointer to NULL post-free, this caused crashes if mbuf debugging was turned on.

- **ixgbe: Fix RX with buffer address not word aligned.**

Niantic HW expects the Header Buffer Address in the RXD must be word aligned.

- **ixgbe: Fix RX with buffer address not word aligned.**

- **ixgbe: Fix Rx queue reset.**

Fix to reset vector related RX queue fields to their initial values.

Fixes: c95584dc2b18 (“ixgbe: new vectorized functions for Rx/Tx”)

- **ixgbe: Fix TSO in IPv6.**

When TSO was used with IPv6, the generated frames were incorrect. The L4 frame was OK, but the length field of IPv6 header was not populated correctly.

- **ixgbe: Fix X550 flow director check.**

- **ixgbe: Fix check for split packets.**

The check for split packets to be reassembled in the vector ixgbe PMD was incorrectly only checking the first 16 elements of the array instead of all 32.

Fixes: cf4b4708a88a (“ixgbe: improve slow-path perf with vector scattered Rx”)

- **ixgbe: Fix data access on big endian cpu.**

- **ixgbe: Fix flow director flexbytes offset.**

Fixes: d54a9888267c (“ixgbe: support flexpayload configuration of flow director”)

- **ixgbe: Fix number of segments with vector scattered Rx.**

Fixes: cf4b4708a88a (ixgbe: improve slow-path perf with vector scattered Rx)

- **ixgbe: Fix offload config option name.**

The RX_OLFLAGS option was renamed from DISABLE to ENABLE in the driver code and Linux config. It is now renamed also in the BSD config and documentation.

Fixes: 359f106a69a9 (“ixgbe: prefer enabling olflags rather than not disabling”)

- **ixgbe: Fix release queue mbufs.**

The calculations of what mbufs were valid in the RX and TX queues were incorrect when freeing the mbufs for the vector PMD. This led to crashes due to invalid reference counts when mbuf debugging was turned on, and possibly other more subtle problems (such as mbufs being freed when in use) in other cases.

Fixes: c95584dc2b18 (“ixgbe: new vectorized functions for Rx/Tx”)

- **ixgbe: Move PMD specific fields out of base driver.**

Move `rx_bulk_alloc_allowed` and `rx_vec_allowed` from `ixgbe_hw` to `ixgbe_adapter`.

Fixes: 01fa1d6215fa (“ixgbe: unify Rx setup”)

- **ixgbe: Rename TX queue release function.**
- **ixgbev: Fix RX function selection.**

The logic to select ixgbe the VF RX function is different than the PF.

- **ixgbev: Fix link status for PF up/down events.**
- **kni: Fix RX loop limit.**

Loop processing packets dequeued from rx_q was using the number of packets requested, not how many it actually received.

- **kni: Fix ioctl in containers, like Docker.**
- **kni: Fix multicast ioctl handling.**
- **log: Fix crash after log_history dump.**
- **lpm: Fix big endian support.**
- **lpm: Fix depth small entry add.**

- **mbuf: Fix cloning with private mbuf data.**

Added a new `priv_size` field in mbuf structure that should be initialized at mbuf pool creation. This field contains the size of the application private data in mbufs.

Introduced new static inline functions `rte_mbuf_from_indirect()` and `rte_mbuf_to_baddr()` to replace the existing macros, which take the private size into account when attaching and detaching mbufs.

- **mbuf: Fix data room size calculation in pool init.**

Deduct the mbuf data room size from `mempool->elt_size` and `priv_size`, instead of using an hardcoded value that is not related to the real buffer size.

To use `rte_pktmbuf_pool_init()`, the user can either:

- Give a NULL parameter to `rte_pktmbuf_pool_init()`: in this case, the private size is assumed to be 0, and the room size is `mp->elt_size - sizeof(struct rte_mbuf)`.
- Give the `rte_pktmbuf_pool_private` filled with appropriate `data_room_size` and `priv_size` values.

- **mbuf: Fix init when private size is not zero.**

Allow the user to use the default `rte_pktmbuf_init()` function even if the mbuf private size is not 0.

- **mempool: Add structure for object headers.**

Each object stored in mempools are prefixed by a header, allowing for instance to retrieve the mempool pointer from the object. When debug is enabled, a cookie is also added in this header that helps to detect corruptions and double-frees.

Introduced a structure that materializes the content of this header, and will simplify future patches adding things in this header.

- **mempool: Fix pages computation to determine number of objects.**

- **mempool: Fix returned value after counting objects.**
Fixes: 148f963fb532 (“xen: core library changes”)
- **mlx4: Avoid requesting TX completion events to improve performance.**
Instead of requesting a completion event for each TX burst, request it on a fixed schedule once every MLX4_PMD_TX_PER_COMP_REQ (currently 64) packets to improve performance.
- **mlx4: Fix compilation as a shared library and on 32 bit platforms.**
- **mlx4: Fix possible crash on scattered mbuf allocation failure.**
Fixes issue where failing to allocate a segment, `mlx4_rx_burst_sp()` could call `rte_pktmbuf_free()` on an incomplete scattered mbuf whose next pointer in the last segment is not set.
- **mlx4: Fix support for multiple vlan filters.**
This fixes the “Multiple RX VLAN filters can be configured, but only the first one works” bug.
- **pcap: Fix storage of name and type in queues.**
`pcap_rx_queue/pcap_tx_queue` should store it’s own copy of name/type values, not the pointer to temporary allocated space.
- **pci: Fix memory leaks and needless increment of map address.**
- **pci: Fix uio mapping differences between linux and bsd.**
- **port: Fix unaligned access to metadata.**
Fix `RTE_MBUF_METADATA` macros to allow for unaligned accesses to meta-data fields.
- **ring: Fix return of new port id on creation.**
- **timer: Fix race condition.**
Eliminate problematic race condition in `rte_timer_manage()` that can lead to corruption of per-icore pending-lists (implemented as skip-lists).
- **vfio: Fix overflow of BAR region offset and size.**
Fixes: 90a1633b2347 (“eal/Linux: allow to map BARs with MSI-X tables”)
- **vhost: Fix enqueue/dequeue to handle chained vring descriptors.**
- **vhost: Fix race for connection fd.**
- **vhost: Fix virtio freeze due to missed interrupt.**
- **virtio: Fix crash if CQ is not negotiated.**
Fix NULL dereference if virtio control queue is not negotiated.
- **virtio: Fix ring size negotiation.**
Negotiate the virtio ring size. The host may allow for very large rings but application may only want a smaller ring. Conversely, if the number of descriptors requested exceeds the virtio host queue size, then just silently use the smaller host size.

This fixes issues with virtio in non-QEMU environments. For example Google Compute Engine allows up to 16K elements in ring.

- **vmxnet3: Fix link state handling.**

3.3 Known Issues

- When running the `vmdq` sample or `vhost` sample applications with the Intel(R) XL710 (i40e) NIC, the configuration option `CONFIG_RTE_MAX_QUEUES_PER_PORT` should be increased from 256 to 1024.
- VM power manager may not work on systems with more than 64 cores.

3.4 API Changes

- The order that user supplied RX and TX callbacks are called in has been changed to the order that they were added (fifo) in line with end-user expectations. The previous calling order was the reverse of this (lifo) and was counter intuitive for users. The actual API is unchanged.

3.5 ABI Changes

- The `rte_hash` structure has been changed to internal use only.

DPDK RELEASE 2.0

4.1 New Features

- Poll-mode driver support for an early release of the PCIE host interface of the Intel(R) Ethernet Switch FM10000.
 - Basic Rx/Tx functions for PF/VF
 - Interrupt handling support for PF/VF
 - Per queue start/stop functions for PF/VF
 - Support Mailbox handling between PF/VF and PF/Switch Manager
 - Receive Side Scaling (RSS) for PF/VF
 - Scatter receive function for PF/VF
 - Reta update/query for PF/VF
 - VLAN filter set for PF
 - Link status query for PF/VF

Note: The software is intended to run on pre-release hardware and may contain unknown or unresolved defects or issues related to functionality and performance. The poll mode driver is also pre-release and will be updated to a released version post hardware and base driver release. Should the official hardware release be made between DPDK releases an updated poll-mode driver will be made available.

- Link Bonding
 - Support for adaptive load balancing (mode 6) to the link bonding library.
 - Support for registration of link status change callbacks with link bonding devices.
 - Support for slaves devices which do not support link status change interrupts in the link bonding library via a link status polling mechanism.
- PCI Hotplug with NULL PMD sample application
- ABI versioning
- x32 ABI
- Non-EAL Thread Support
- Multi-pthread Support

- Re-order Library
- ACL for AVX2
- Architecture Independent CRC Hash
- uio_pci_generic Support
- KNI Optimizations
- Vhost-user support
- Virtio (link, vlan, mac, port IO, perf)
- IXGBE-VF RSS
- RX/TX Callbacks
- Unified Flow Types
- Indirect Attached MBUF Flag
- Use default port configuration in TestPMD
- Tunnel offloading in TestPMD
- Poll Mode Driver - 40 GbE Controllers (librte_pmd_i40e)
 - Support for Flow Director
 - Support for ethertype filter
 - Support RSS in VF
 - Support configuring redirection table with different size from 1GbE and 10 GbE
 - 128/512 entries of 40GbE PF
 - 64 entries of 40GbE VF
 - Support configuring hash functions
 - Support for VXLAN packet on Intel® 40GbE Controllers
- Poll Mode Driver for Mellanox ConnectX-3 EN adapters (mlx4)

Note: This PMD is only available for Linux and is disabled by default due to external dependencies (libibverbs and libmlx4). Please refer to the NIC drivers guide for more information.

- Packet Distributor Sample Application
- Job Stats library and Sample Application.
- Enhanced Jenkins hash (jhash) library

Note: The hash values returned by the new jhash library are different from the ones returned by the previous library.

5.1 New Features

- Link Bonding
 - Support for 802.3ad link aggregation (mode 4) and transmit load balancing (mode 5) to the link bonding library.
 - Support for registration of link status change callbacks with link bonding devices.
 - Support for slaves devices which do not support link status change interrupts in the link bonding library via a link status polling mechanism.
- Poll Mode Driver - 40 GbE Controllers (librte_pmd_i40e)
 - Support for Flow Director
 - Support for ethertype filter
 - Support RSS in VF
 - Support configuring redirection table with different size from 1GbE and 10 GbE
 - 128/512 entries of 40GbE PF
 - 64 entries of 40GbE VF
 - Support configuring hash functions
 - Support for VXLAN packet on Intel 40GbE Controllers
- Packet Distributor Sample Application

SUPPORTED OPERATING SYSTEMS

The following Linux distributions were successfully used to compile or run DPDK.

- FreeBSD 10
- Fedora release 20
- Ubuntu 14.04 LTS
- Wind River Linux 6
- Red Hat Enterprise Linux 6.5
- SUSE Enterprise Linux 11 SP3

These distributions may need additional packages that are not installed by default, or a specific kernel. Refer to the Linux guide and FreeBSD guide for details.

KNOWN ISSUES AND LIMITATIONS IN LEGACY RELEASES

This section describes known issues with the DPDK software that aren't covered in the version specific release notes sections.

7.1 Unit Test for Link Bonding may fail at `test_tlb_tx_burst()`

Description: Unit tests will fail in `test_tlb_tx_burst()` function with error for uneven distribution of packets.

Implication: Unit test `link_bonding_autotest` will fail.

Resolution/Workaround: There is no workaround available.

Affected Environment/Platform: Fedora 20.

Driver/Module: Link Bonding.

7.2 Pause Frame Forwarding does not work properly on `igb`

Description: For `igb` devices `rte_eth_flow_ctrl_set` does not work as expected. Pause frames are always forwarded on `igb`, regardless of the `RFCE`, `MPMCF` and `DPF` registers.

Implication: Pause frames will never be rejected by the host on 1G NICs and they will always be forwarded.

Resolution/Workaround: There is no workaround available.

Affected Environment/Platform: All.

Driver/Module: Poll Mode Driver (PMD).

7.3 In packets provided by the PMD, some flags are missing

Description: In packets provided by the PMD, some flags are missing. The application does not have access to information provided by the hardware (packet is broadcast, packet is multicast, packet is IPv4 and so on).

Implication: The `ol_flags` field in the `rte_mbuf` structure is not correct and should not be used.

Resolution/Workaround: The application has to parse the Ethernet header itself to get the information, which is slower.

Affected Environment/Platform: All.

Driver/Module: Poll Mode Driver (PMD).

7.4 The `rte_malloc` library is not fully implemented

Description: The `rte_malloc` library is not fully implemented.

Implication: All debugging features of `rte_malloc` library described in architecture documentation are not yet implemented.

Resolution/Workaround: No workaround available.

Affected Environment/Platform: All.

Driver/Module: `rte_malloc`.

7.5 HPET reading is slow

Description: Reading the HPET chip is slow.

Implication: An application that calls `rte_get_hpet_cycles()` or `rte_timer_manage()` runs slower.

Resolution/Workaround: The application should not call these functions too often in the main loop. An alternative is to use the TSC register through `rte_rdtsc()` which is faster, but specific to an Icore and is a cycle reference, not a time reference.

Affected Environment/Platform: All.

Driver/Module: Environment Abstraction Layer (EAL).

7.6 HPET timers do not work on the Osage customer reference platform

Description: HPET timers do not work on the Osage customer reference platform which includes an Intel® Xeon® processor 5500 series processor) using the released BIOS from Intel.

Implication: On Osage boards, the implementation of the `rte_delay_us()` function must be changed to not use the HPET timer.

Resolution/Workaround: This can be addressed by building the system with the `CONFIG_RTE_LIBEAL_USE_HPET=n` configuration option or by using the `--no-hpet` EAL option.

Affected Environment/Platform: The Osage customer reference platform. Other vendor platforms with Intel® Xeon® processor 5500 series processors should work correctly, provided the BIOS supports HPET.

Driver/Module: `lib/librte_eal/common/include/rte_cycles.h`

7.7 Not all variants of supported NIC types have been used in testing

Description: The supported network interface cards can come in a number of variants with different device ID's. Not all of these variants have been tested with the Intel® DPDK.

The NIC device identifiers used during testing:

- Intel® Ethernet Controller XL710 for 40GbE QSFP+ [8086:1584]
- Intel® Ethernet Controller XL710 for 40GbE QSFP+ [8086:1583]
- Intel® Ethernet Controller X710 for 10GbE SFP+ [8086:1572]
- Intel® 82576 Gigabit Ethernet Controller [8086:10c9]
- Intel® 82576 Quad Copper Gigabit Ethernet Controller [8086:10e8]
- Intel® 82580 Dual Copper Gigabit Ethernet Controller [8086:150e]
- Intel® I350 Quad Copper Gigabit Ethernet Controller [8086:1521]
- Intel® 82599 Dual Fibre 10 Gigabit Ethernet Controller [8086:10fb]
- Intel® Ethernet Server Adapter X520-T2 [8086: 151c]
- Intel® Ethernet Controller X540-T2 [8086:1528]
- Intel® 82574L Gigabit Network Connection [8086:10d3]
- Emulated Intel® 82540EM Gigabit Ethernet Controller [8086:100e]
- Emulated Intel® 82545EM Gigabit Ethernet Controller [8086:100f]
- Intel® Ethernet Server Adapter X520-4 [8086:154a]
- Intel® Ethernet Controller I210 [8086:1533]

Implication: Risk of issues with untested variants.

Resolution/Workaround: Use tested NIC variants. For those supported Ethernet controllers, additional device IDs may be added to the software if required.

Affected Environment/Platform: All.

Driver/Module: Poll-mode drivers

7.8 Multi-process sample app requires exact memory mapping

Description: The multi-process example application assumes that it is possible to map the hugepage memory to the same virtual addresses in client and server applications. Occasionally, very rarely with 64-bit, this does not occur and a client application will fail on startup. The Linux “address-space layout randomization” security feature can sometimes cause this to occur.

Implication: A multi-process client application fails to initialize.

Resolution/Workaround: See the “Multi-process Limitations” section in the Intel® DPDK Programmer’s Guide for more information.

Affected Environment/Platform: All.

Driver/Module: Multi-process example application

7.9 Packets are not sent by the 1 GbE/10 GbE SR-IOV driver when the source MAC is not the MAC assigned to the VF NIC

Description: The 1 GbE/10 GbE SR-IOV driver can only send packets when the Ethernet header's source MAC address is the same as that of the VF NIC. The reason for this is that the Linux `ixgbe` driver module in the host OS has its anti-spoofing feature enabled.

Implication: Packets sent using the 1 GbE/10 GbE SR-IOV driver must have the source MAC address correctly set to that of the VF NIC. Packets with other source address values are dropped by the NIC if the application attempts to transmit them.

Resolution/Workaround: Configure the Ethernet source address in each packet to match that of the VF NIC.

Affected Environment/Platform: All.

Driver/Module: 1 GbE/10 GbE VF Poll Mode Driver (PMD).

7.10 SR-IOV drivers do not fully implement the `rte_ethdev` API

Description: The SR-IOV drivers only supports the following `rte_ethdev` API functions:

- `rte_eth_dev_configure()`
- `rte_eth_tx_queue_setup()`
- `rte_eth_rx_queue_setup()`
- `rte_eth_dev_info_get()`
- `rte_eth_dev_start()`
- `rte_eth_tx_burst()`
- `rte_eth_rx_burst()`
- `rte_eth_dev_stop()`
- `rte_eth_stats_get()`
- `rte_eth_stats_reset()`
- `rte_eth_link_get()`
- `rte_eth_link_get_no_wait()`

Implication: Calling an unsupported function will result in an application error.

Resolution/Workaround: Do not use other `rte_ethdev` API functions in applications that use the SR-IOV drivers.

Affected Environment/Platform: All.

Driver/Module: VF Poll Mode Driver (PMD).

7.11 PMD does not work with `--no-huge` EAL command line parameter

Description: Currently, the DPDK does not store any information about memory allocated by `malloc()` (for example, NUMA node, physical address), hence PMD drivers do not work when the `--no-huge` command line parameter is supplied to EAL.

Implication: Sending and receiving data with PMD will not work.

Resolution/Workaround: Use huge page memory or use VFIO to map devices.

Affected Environment/Platform: Systems running the DPDK on Linux

Driver/Module: Poll Mode Driver (PMD).

7.12 Some hardware off-load functions are not supported by the VF Driver

Description: Currently, configuration of the following items is not supported by the VF driver:

- IP/UDP/TCP checksum offload
- Jumbo Frame Receipt
- HW Strip CRC

Implication: Any configuration for these items in the VF register will be ignored. The behavior is dependent on the current PF setting.

Resolution/Workaround: For the PF (Physical Function) status on which the VF driver depends, there is an option item under PMD in the config file. For others, the VF will keep the same behavior as PF setting.

Affected Environment/Platform: All.

Driver/Module: VF (SR-IOV) Poll Mode Driver (PMD).

7.13 Kernel crash on IGB port unbinding

Description: Kernel crash may occur when unbinding 1G ports from the `igb_uio` driver, on 2.6.3x kernels such as shipped with Fedora 14.

Implication: Kernel crash occurs.

Resolution/Workaround: Use newer kernels or do not unbind ports.

Affected Environment/Platform: 2.6.3x kernels such as shipped with Fedora 14

Driver/Module: IGB Poll Mode Driver (PMD).

7.14 Twinpond and Ironpond NICs do not report link status correctly

Description: Twin Pond/Iron Pond NICs do not bring the physical link down when shutting down the port.

Implication: The link is reported as up even after issuing `shutdown` command unless the cable is physically disconnected.

Resolution/Workaround: None.

Affected Environment/Platform: Twin Pond and Iron Pond NICs

Driver/Module: Poll Mode Driver (PMD).

7.15 Discrepancies between statistics reported by different NICs

Description: Gigabit Ethernet devices from Intel include CRC bytes when calculating packet reception statistics regardless of hardware CRC stripping state, while 10-Gigabit Ethernet devices from Intel do so only when hardware CRC stripping is disabled.

Implication: There may be a discrepancy in how different NICs display packet reception statistics.

Resolution/Workaround: None

Affected Environment/Platform: All.

Driver/Module: Poll Mode Driver (PMD).

7.16 Error reported opening files on DPDK initialization

Description: On DPDK application startup, errors may be reported when opening files as part of the initialization process. This occurs if a large number, for example, 500 or more, or if hugepages are used, due to the per-process limit on the number of open files.

Implication: The DPDK application may fail to run.

Resolution/Workaround: If using 2 MB hugepages, consider switching to a fewer number of 1 GB pages. Alternatively, use the `ulimit` command to increase the number of files which can be opened by a process.

Affected Environment/Platform: All.

Driver/Module: Environment Abstraction Layer (EAL).

7.17 Intel® QuickAssist Technology sample application does not work on a 32-bit OS on Shumway

Description: The Intel® Communications Chipset 89xx Series device does not fully support NUMA on a 32-bit OS. Consequently, the sample application cannot work properly on Shumway, since it requires NUMA on both nodes.

Implication: The sample application cannot work in 32-bit mode with emulated NUMA, on multi-socket boards.

Resolution/Workaround: There is no workaround available.

Affected Environment/Platform: Shumway

Driver/Module: All.

7.18 Differences in how different Intel NICs handle maximum packet length for jumbo frame

Description: 10 Gigabit Ethernet devices from Intel do not take VLAN tags into account when calculating packet size while Gigabit Ethernet devices do so for jumbo frames.

Implication: When receiving packets with VLAN tags, the actual maximum size of useful payload that Intel Gigabit Ethernet devices are able to receive is 4 bytes (or 8 bytes in the case of packets with extended VLAN tags) less than that of Intel 10 Gigabit Ethernet devices.

Resolution/Workaround: Increase the configured maximum packet size when using Intel Gigabit Ethernet devices.

Affected Environment/Platform: All.

Driver/Module: Poll Mode Driver (PMD).

7.19 Binding PCI devices to igb_uio fails on Linux kernel 3.9 when more than one device is used

Description: A known bug in the uio driver included in Linux kernel version 3.9 prevents more than one PCI device to be bound to the igb_uio driver.

Implication: The Poll Mode Driver (PMD) will crash on initialization.

Resolution/Workaround: Use earlier or later kernel versions, or apply the following [patch](#).

Affected Environment/Platform: Linux systems with kernel version 3.9

Driver/Module: igb_uio module

7.20 GCC might generate Intel® AVX instructions for processors without Intel® AVX support

Description: When compiling Intel® DPDK (and any DPDK app), gcc may generate Intel® AVX instructions, even when the processor does not support Intel® AVX.

Implication: Any DPDK app might crash while starting up.

Resolution/Workaround: Either compile using icc or set `EXTRA_CFLAGS='-O3'` prior to compilation.

Affected Environment/Platform: Platforms which processor does not support Intel® AVX.

Driver/Module: Environment Abstraction Layer (EAL).

7.21 Ethertype filter could receive other packets (non-assigned) in Niantic

Description: On Intel® Ethernet Controller 82599EB When Ethertype filter (priority enable) was set, unmatched packets also could be received on the assigned queue, such as ARP packets without 802.1q tags or with the user priority not equal to set value. Launch the testpmd by disabling RSS and with multiply queues, then add the ethertype filter like the following and then start forwarding:

```
add_ethertype_filter 0 ethertype 0x0806 priority enable 3 queue 2 index 1
```

When sending ARP packets without 802.1q tag and with user priority as non-3 by tester, all the ARP packets can be received on the assigned queue.

Implication: The user priority comparing in Ethertype filter cannot work probably. It is a NIC's issue due to the following: "In fact, ETQF.UP is not functional, and the information will be added in errata of 82599 and X540."

Resolution/Workaround: None

Affected Environment/Platform: All.

Driver/Module: Poll Mode Driver (PMD).

7.22 Cannot set link speed on Intel® 40G Ethernet controller

Description: On Intel® 40G Ethernet Controller you cannot set the link to specific speed.

Implication: The link speed cannot be changed forcibly, though it can be configured by application.

Resolution/Workaround: None

Affected Environment/Platform: All.

Driver/Module: Poll Mode Driver (PMD).

7.23 Stopping the port does not down the link on Intel® 40G Ethernet controller

Description: On Intel® 40G Ethernet Controller stopping the port does not really down the port link.

Implication: The port link will be still up after stopping the port.

Resolution/Workaround: None

Affected Environment/Platform: All.

Driver/Module: Poll Mode Driver (PMD).

7.24 Devices bound to igb_uio with VT-d enabled do not work on Linux kernel 3.15-3.17

Description: When VT-d is enabled (`iommu=pt intel_iommu=on`), devices are 1:1 mapped. In the Linux kernel unbinding devices from drivers removes that mapping which result in IOMMU errors. Introduced in Linux kernel 3.15 commit, solved in Linux kernel 3.18 commit.

Implication: Devices will not be allowed to access memory, resulting in following kernel errors:

```
dmar: DRHD: handling fault status reg 2
dmar: DMAR:[DMA Read] Request device [02:00.0] fault addr a0c58000
DMAR:[fault reason 02] Present bit in context entry is clear
```

Resolution/Workaround: Use earlier or later kernel versions, or avoid driver binding on boot by blacklisting the driver modules. I.e., in the case of `ixgbe`, we can pass the kernel command line option: `modprobe.blacklist=ixgbe`. This way we do not need to unbind the device to bind it to `igb_uio`.

Affected Environment/Platform: Linux systems with kernel versions 3.15 to 3.17.

Driver/Module: `igb_uio` module.

7.25 VM power manager may not work on systems with more than 64 cores

Description: When using VM power manager on a system with more than 64 cores, VM(s) should not use cores 64 or higher.

Implication: VM power manager should not be used with VM(s) that are using cores 64 or above.

Resolution/Workaround: Do not use cores 64 or above.

Affected Environment/Platform: Platforms with more than 64 cores.

Driver/Module: VM power manager application.

ABI AND API DEPRECATION

See the `guidelines` document for details of the ABI policy. API and ABI deprecation notices are to be posted here.

8.1 Deprecation Notices

- The following fields have been deprecated in `rte_eth_stats`: `ibadcrc`, `ibadlen`, `imcasts`, `fdirmatch`, `fdirmiss`, `tx_pause_xon`, `rx_pause_xon`, `tx_pause_xoff`, `rx_pause_xoff`
- The `ethdev` structures `rte_eth_link`, `rte_eth_dev_info` and `rte_eth_conf` must be updated to support 100G link and to have a cleaner link speed API.
- ABI changes is planned for the `reta` field in struct `rte_eth_rss_reta_entry64` which handles at most 256 queues (8 bits) while newer NICs support larger tables (512 queues). It should be integrated in release 2.3.
- ABI changes are planned for struct `rte_eth_fdir_flow` in order to support extend flow director's input set. The release 2.2 does not contain these ABI changes, but release 2.3 will, and no backwards compatibility is planned.
- ABI changes are planned for `rte_eth_ipv4_flow` and `rte_eth_ipv6_flow` to include more fields to be matched against. The release 2.2 does not contain these ABI changes, but release 2.3 will.
- ABI changes are planned for adding four new flow types. This impacts `RTE_ETH_FLOW_MAX`. The release 2.2 does not contain these ABI changes, but release 2.3 will.
- ABI changes are planned for `rte_eth_tunnel_filter_conf`. Change the fields of `outer_mac` and `inner_mac` from pointer to struct in order to keep the code's readability. The release 2.2 does not contain these ABI changes, but release 2.3 will, and no backwards compatibility is planned.
- The scheduler statistics structure will change to allow keeping track of RED actions.
- `librte_pipeline`: The prototype for the pipeline input port, output port and table action handlers will be updated: the pipeline parameter will be added, the packets mask parameter will be either removed (for input port action handler) or made input-only.
- ABI changes are planned in `cmdline` buffer size to allow the use of long commands (such as `RETA` update in `testpmd`). This should impact `CMDLINE_PARSE_RESULT_BUFSIZE`, `STR_TOKEN_SIZE` and `RDLINE_BUF_SIZE`. It should be integrated in release 2.3.