DPDK

DATA PLANE DEVELOPMENT KIT

Embracing Externally Allocated Memory

YONGSEOK KOH

MELLANOX

Externally Allocated Memory

- Allocated and managed outside of DPDK
  - Inherently not using hugepages of DPDK for zero-copy
    - Storage buffer
    - GPU device memory
  - VPP has its own memory management system

Case 1 – Private Memory Management

- What if application already has its own memory management and doesn't want to use DPDK memory?
  - rte_mempool_populate_iova() could be used for mempool
    - Still need to register memory for DMA via separate call
      - rte_vfio_dma_map()
    - Not for other data structure
    - May be deprecated

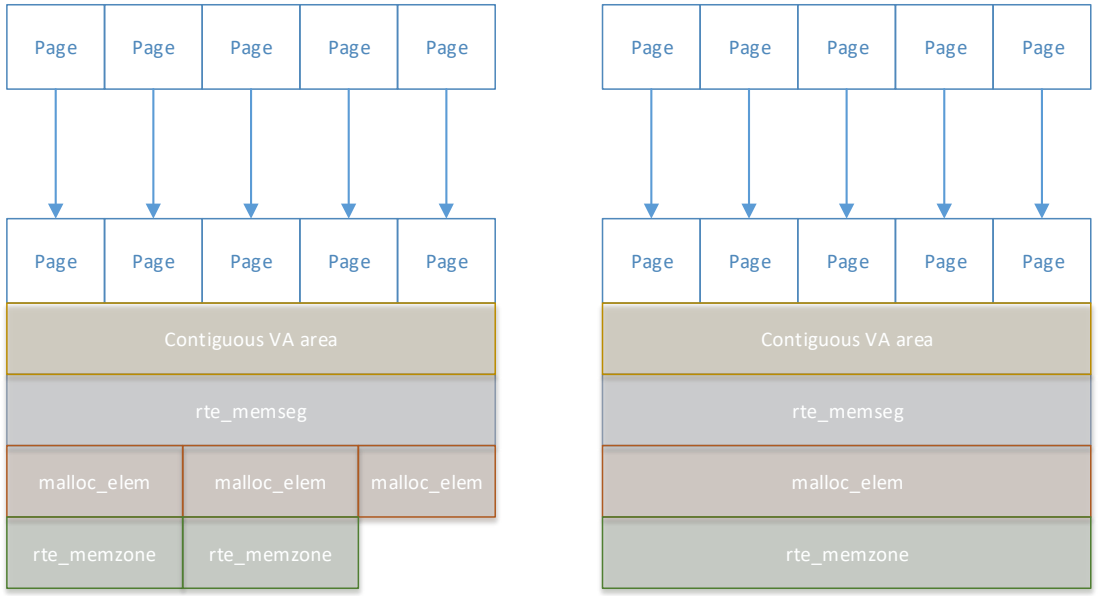
Case 2 – Integrate External Memory

- Can it be integrated with DPDK seamlessly?
  - In v18.11, Anatoly introduced:
    - [PATCH v9 00/21] Support externally allocated memory in DPDK
    - Programmer's Guide – Support for Externally Allocated Memory

From Dublin Summit 2018, by Anatoly (1/2)

Legacy DPDK Memory Architecture

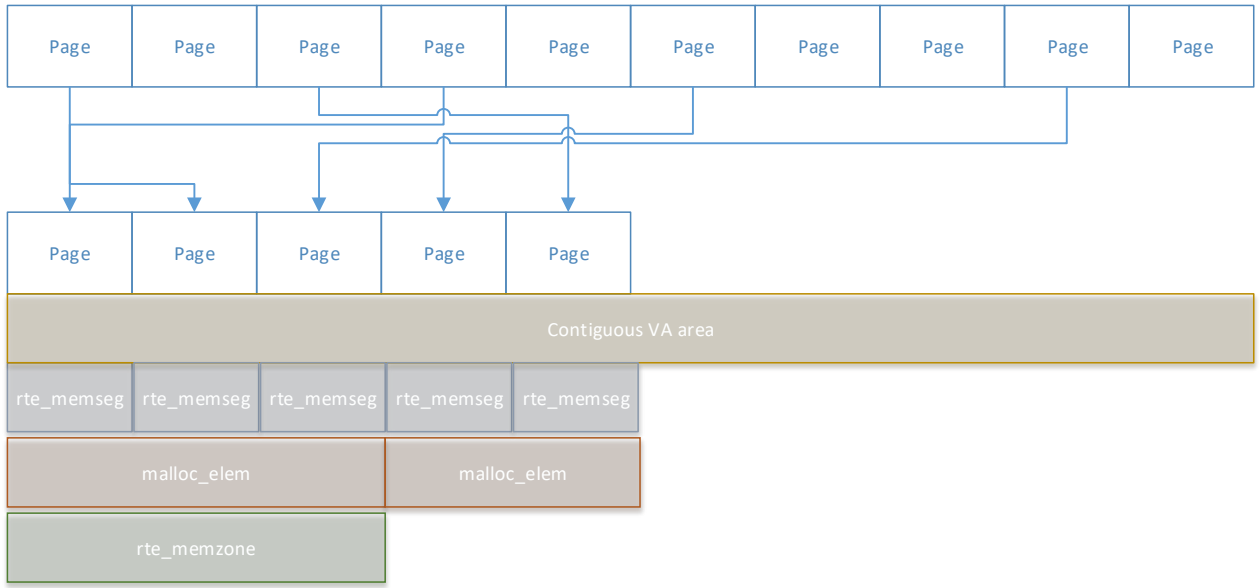
- VA layout follows PA layout
- VA and PA layout is fixed



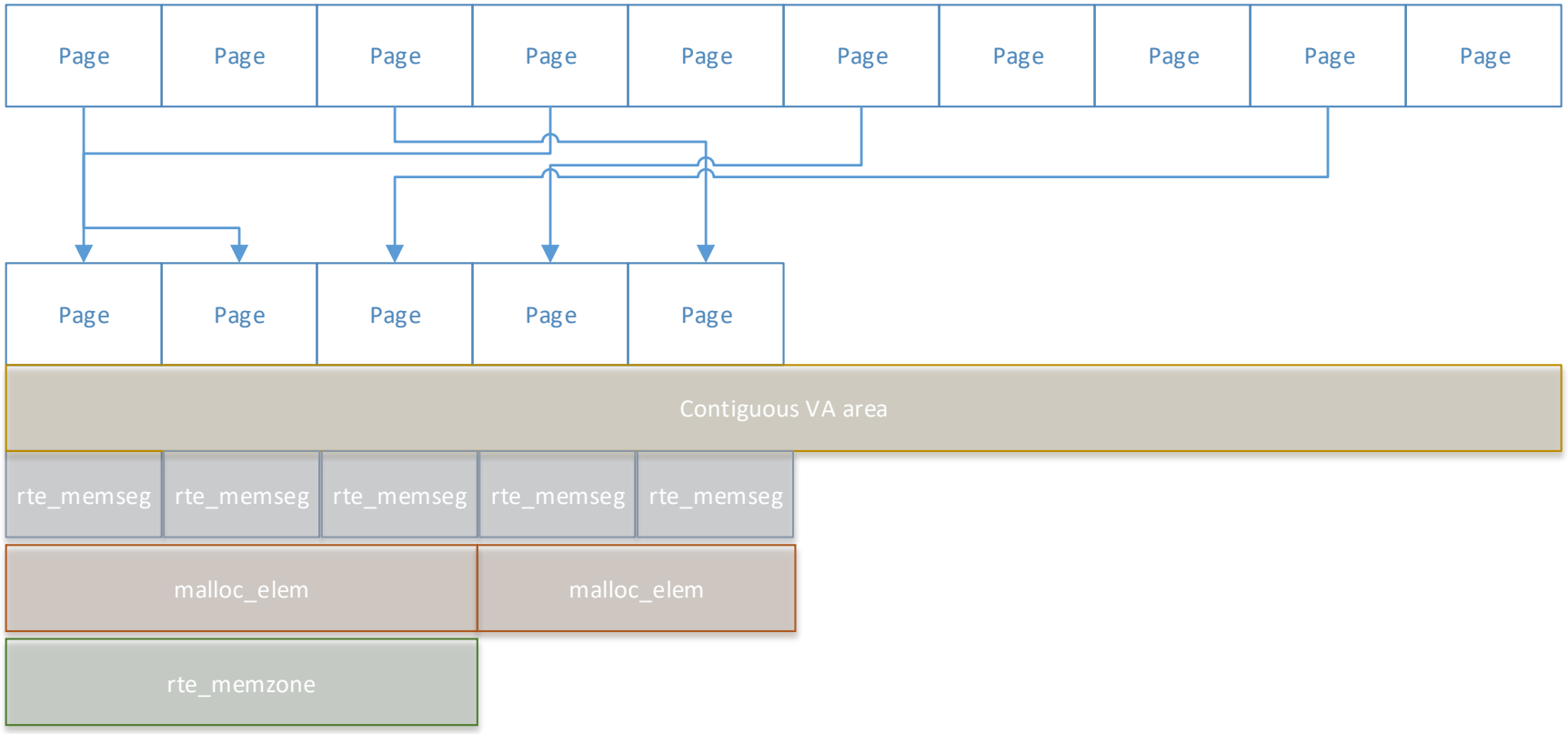
From Dublin Summit 2018, by Anatoly (2/2)

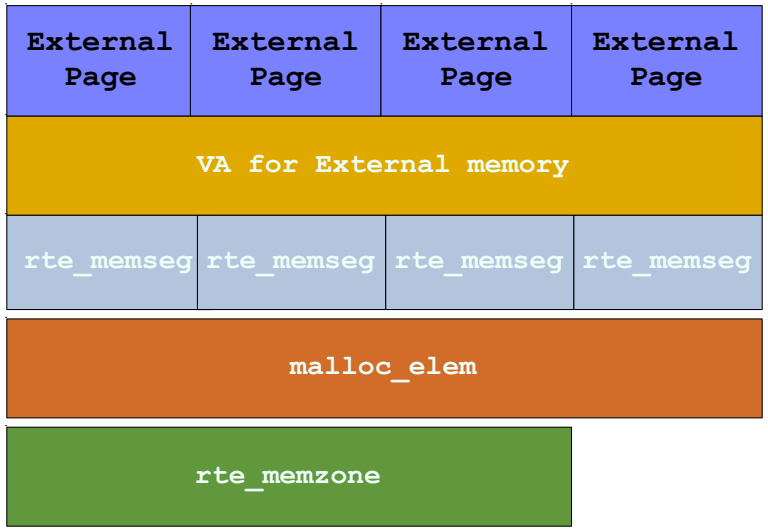
18.05+ DPDK Memory Architecture

- VA layout is independent from PA layout
- VA layout is fixed, PA layout is not



Support Externally Allocated Memory





Support Externally Allocated Memory (cont'd)

- Invalid socket ID is used to refer to external memory
  - #define EXTERNAL_HEAP_MIN_SOCKET_ID
    (CONST_MAX((1 << 8), RTE_MAX_NUMA_NODES))
- All the standard allocation APIs can work with the socket ID
  - rte_malloc_socket(…, socket_id)

Support Externally Allocated Memory (cont'd)

- Dynamically create a new memseg list
  - msl->external = 1
  - Keep track of IOVA addresses
  - If no IOVA is provided, RTE_BAD_IOVA is set
- Generate memory events
  - RTE_MEM_EVENT_ALLOC / RTE_MEM_EVENT_FREE
- Registration for DMA is automatically done via memory event callback
  - vfio_mem_event_callback() for VFIO
  - mlx4/5_mr_mem_event_cb() for Mellanox MLX4/5 PMD
    - Registered by lookup miss
    - Deregistered by free event

How to Use

- Create a named heap
  - rte_malloc_heap_create(heap_name)
- Add external memory to the heap
  - rte_malloc_heap_memory_add(heap_name, **addr**, len, **iova**, n_pages, pgsz)
- Get socket ID of the heap
  - socket_id = rte_malloc_heap_get_socket(heap_name)
- Allocate memory from the heap via standard DPDK APIs
  - rte_malloc_socket(…, socket_id)
  - rte_pktmbuf_pool_create(…, socket_id)
  - and much more.

Example Code

- Unit test
  - test/test/test_external_mem.c
- testpmd
  - --mp-alloc <native|anon|**xmem**|**xmemhuge**>
  - setup_extmem()

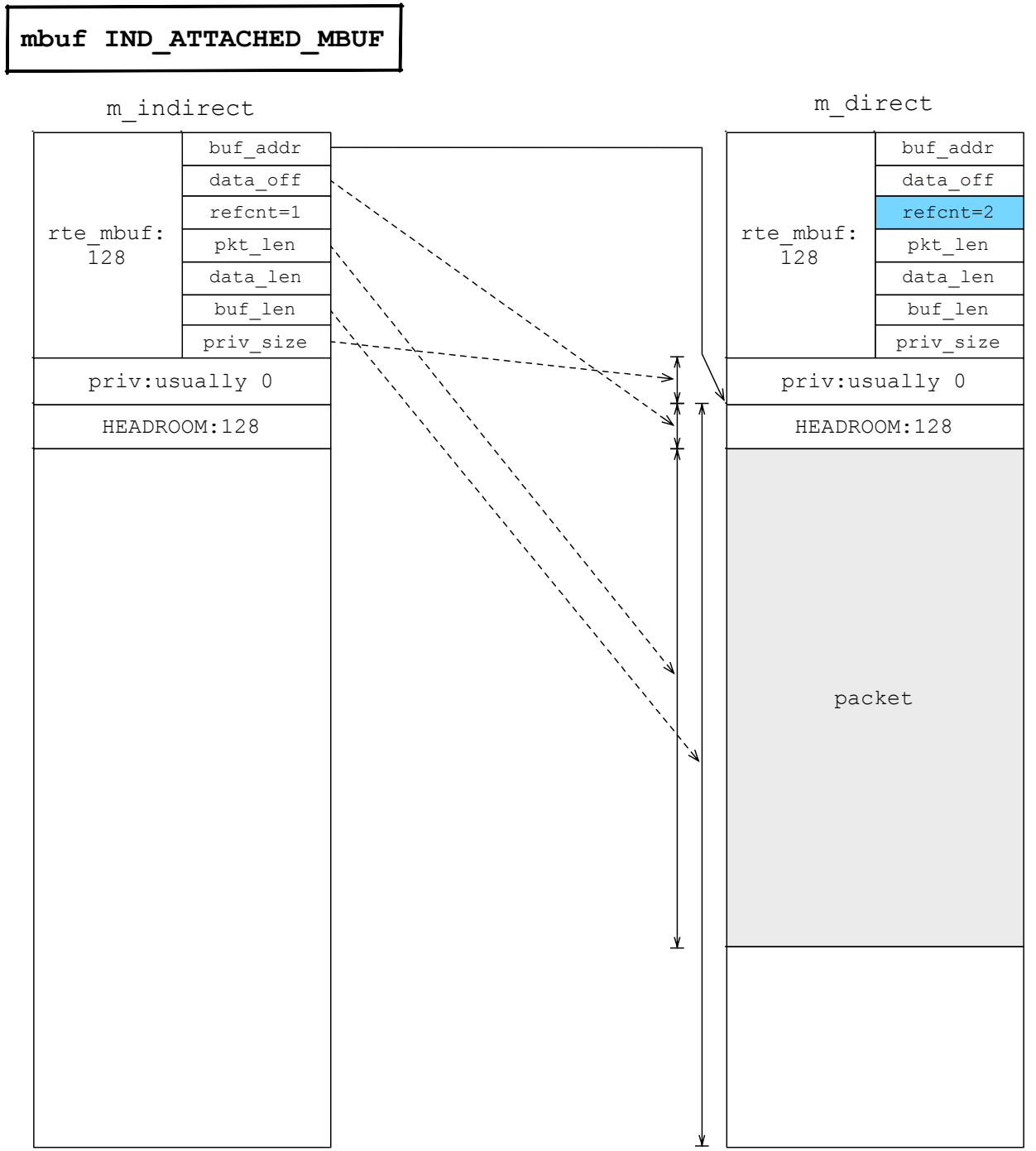
Case 3 – Transfer Device Buffer over Network

- Buffer for Storage/GPU device is generally:
  - Allocated externally with page granularity
  - Entire page is solely used for the device
    - Overhead for malloc_elem would not be allowed
    - Still need to register for DMA
      - rte_vfio_dma_map()
- Need to slice it for transferring over network
  - Indirect MBUF needs data copy
  - Could observe lots of ‘hacks’ to forge mbuf->buf_addr/buf_iova
  - MBUF having external buffer attachment can be used instead

MBUF Indirection

- Marked with IND_ATTACHED_MBUF
- MBUF pointing to another MBUF allocated from a mempool

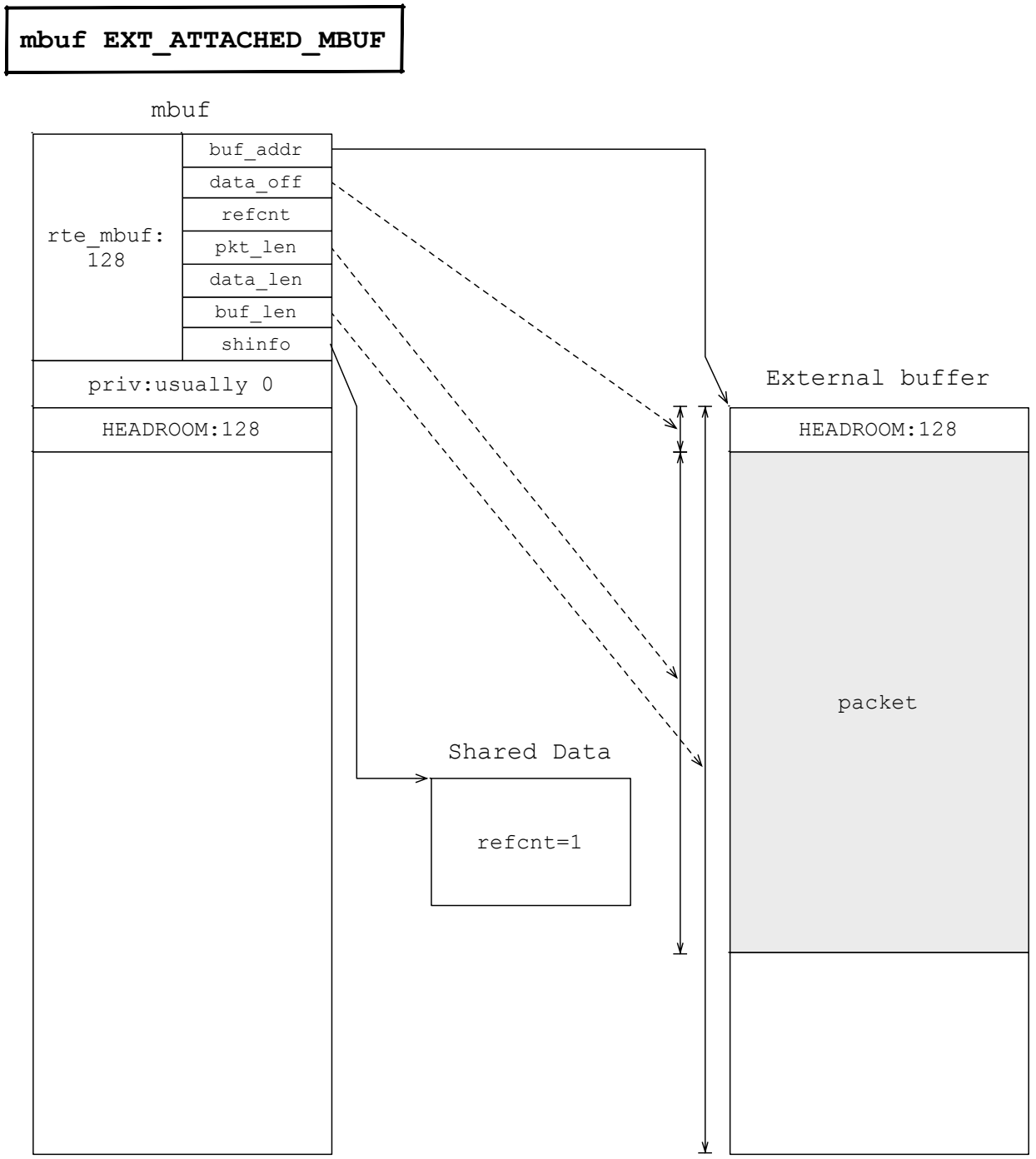
- rte_pktmbuf_attach()
- rte_pktmbuf_detach()



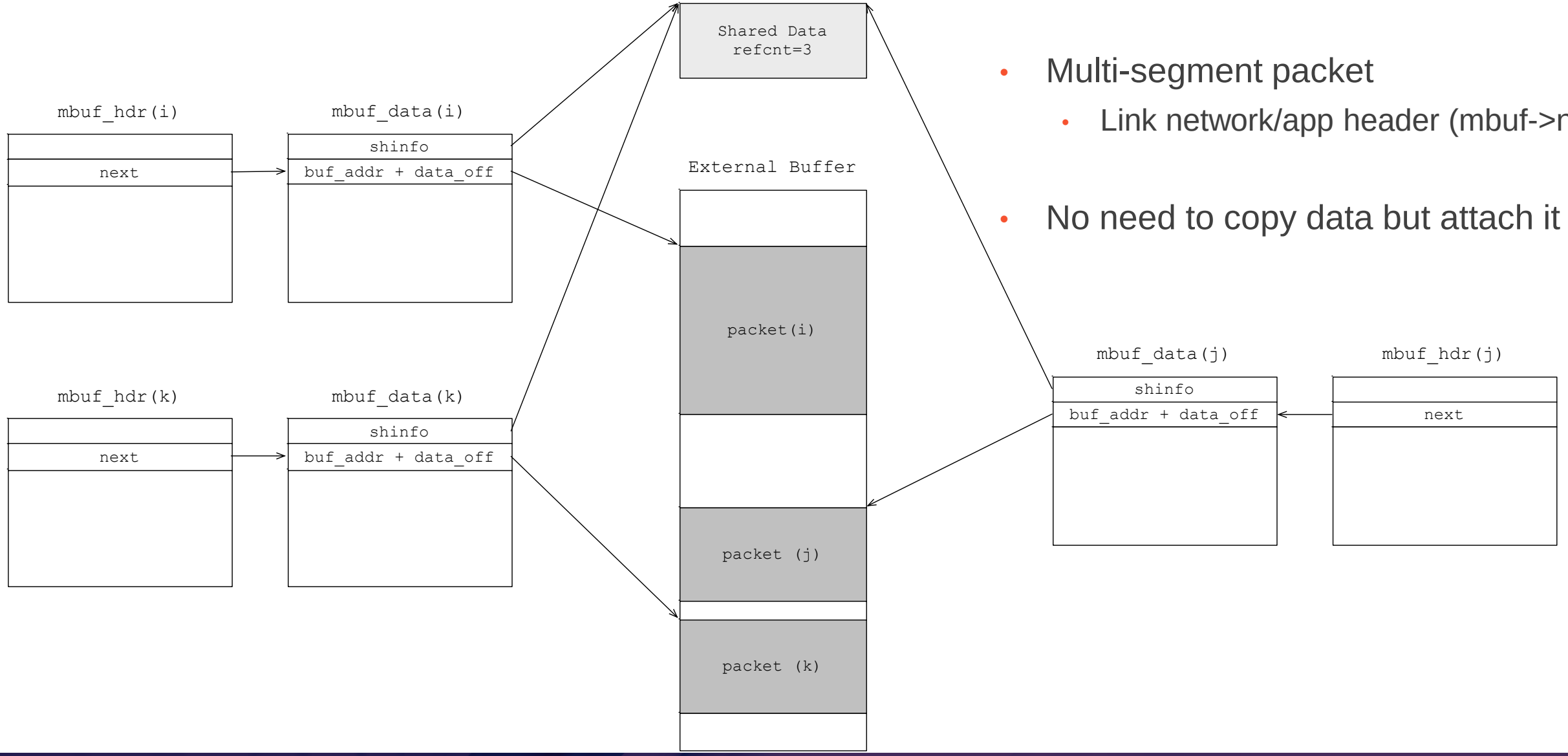
EXT_ATTACHED_MBUF

- Marked with EXT_ATTACHED_MBUF
- Attached buffer can be anonymous
- Need shared info (mbuf->shinfo)
  - refcnt_atomic
  - free_cb() and fcb_opaque

- rte_pktmbuf_attach_extbuf()
- rte_pktmbuf_detach_extbuf()
- Since v18.05



Transfer over Network



- Multi-segment packet
  - Link network/app header (mbuf->next)
- No need to copy data but attach it

rte_vfio_dma_map()

- Register DMA memory for VFIO
- Not every device uses VFIO
  - Mellanox MLX4/5 PMD has different way
    - Device has IOTLB-like translation table for better security
    - PMD uses VA in Rx/Tx descs
    - Registration by Verbs

[RFC] rte_dev_dma_map()

- Generic/vendor-agnostic API to register external memory for DMA
  - rte_vfio_dma_map() would be replaced

- Ongoing discussion in the mailing list
  - [RFC] ethdev: introduce DMA memory mapping for external memory by Shahaf

rte_extmem_register()

- rte_vfio_dma_map() doesn't create a memseg list, i.e. not managed by DPDK
  - Needed a way to create a memseg list for external memory without having overhead for malloc_elem

- Anatoly submitted a new patchset last week
  - [Patch 0/4] Allow using external memory without malloc
  - Suggesting two separate calls for using external memory w/o malloc heap
    - rte_extmem_register() -> rte_dev_dma_map()


DPDK
DATA PLANE DEVELOPMENT KIT
QnA


DPDK
DATA PLANE DEVELOPMENT KIT
Thank You!