REGIONAL SDC 25

BY Developers FOR Developers

Regional SDC Denver
April 30, 2025

UEC specifications : Productization & Value creation

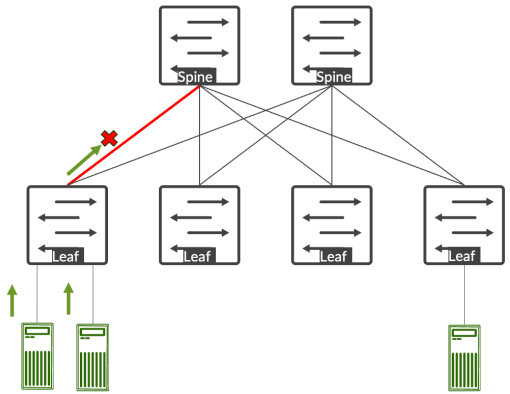
From an AI-Data Center Vendor perspective

Mahesh Subramaniam
Director of Product Management, AI Datacenter

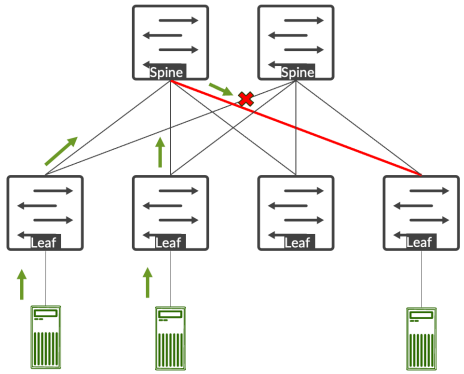
SNIA

AI Datacenter Challenge

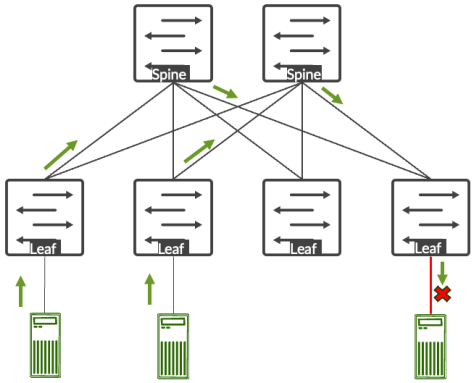
1. High BW flows + Low Entropy -> Elephant Flows & In-cast Congestion
2. (R)Evolution NICs on AI servers -> Greedy on BW & High Radix Switches
3. Failure of Static Load Balancing -> Need of Advance Load Balancing



In-cast at ingress



In-cast at transit

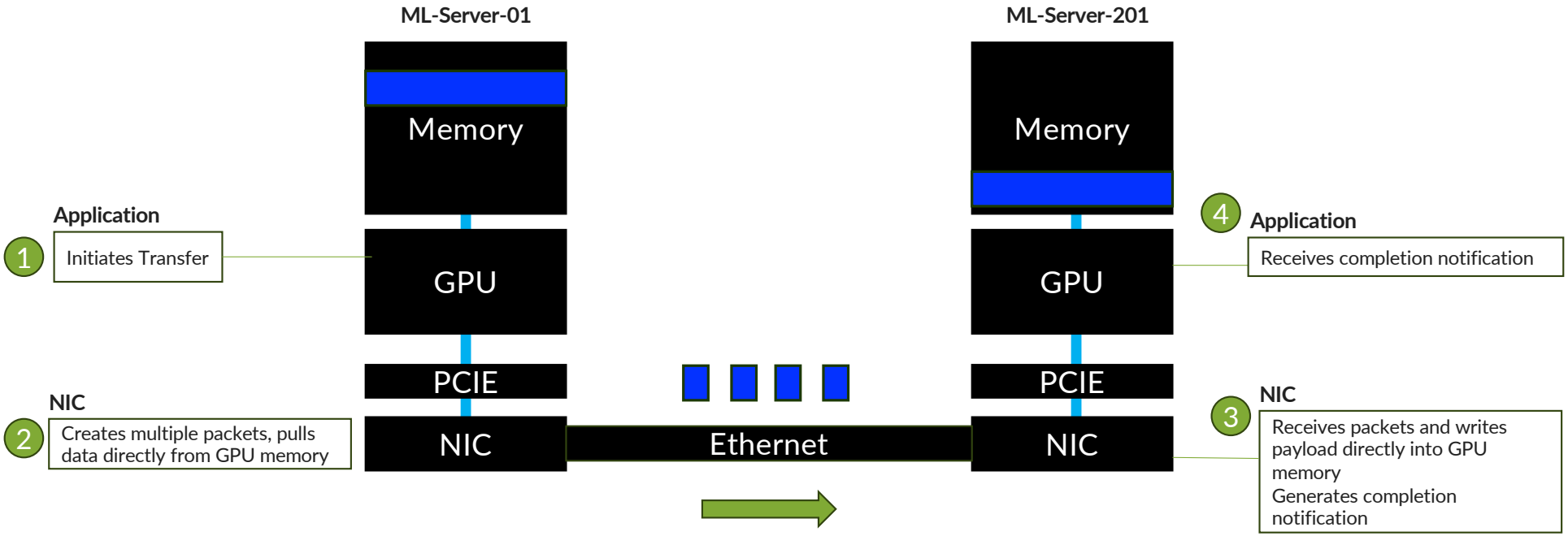


In-cast at egress

Congested Link

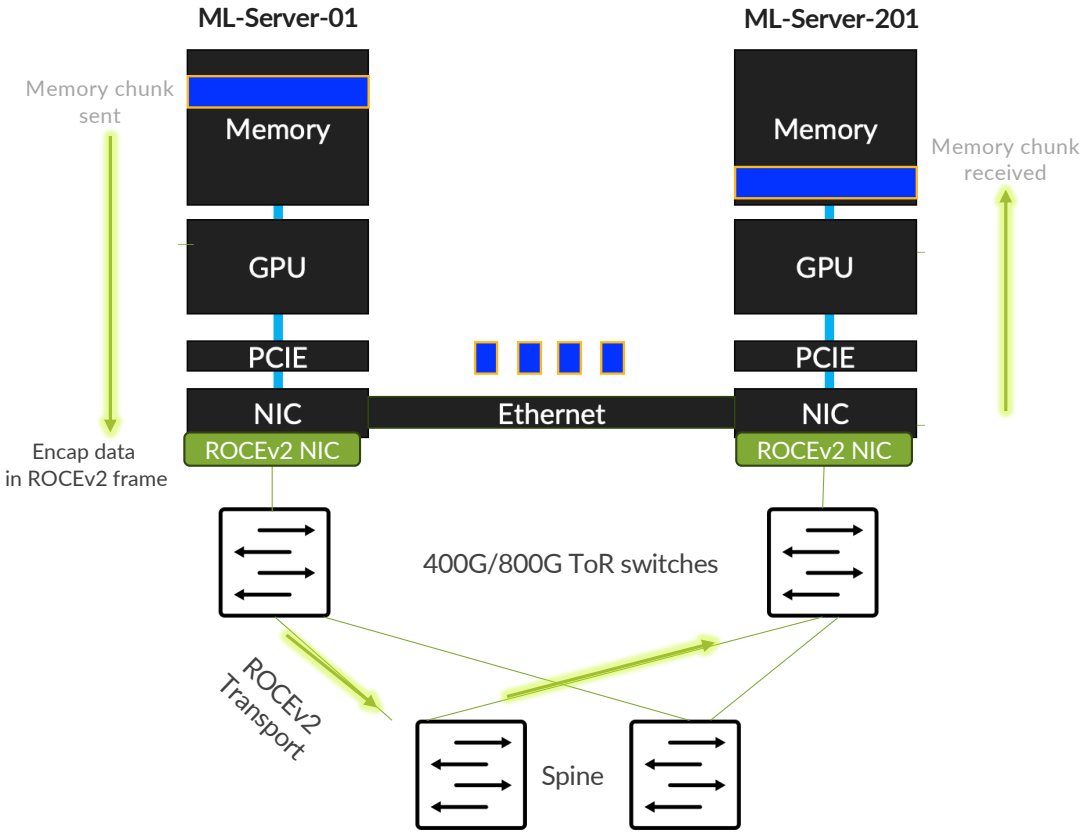
REGIONAL SDC

RDMA Transfer



REGIONAL SDC

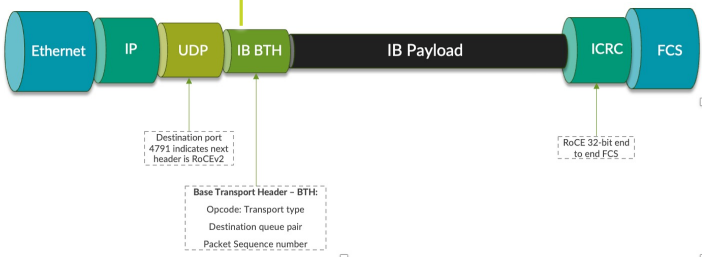
RDMA Transfer by RoCEv2




```
/* operations --- just used to define real constants */
IBV_OPCODE_SEND_FIRST                          = 0x00,
IBV_OPCODE_SEND_MIDDLE                         = 0x01,
IBV_OPCODE_SEND_LAST                           = 0x02,
IBV_OPCODE_SEND_LAST_WITH_IMMEDIATE            = 0x03,
IBV_OPCODE_SEND_ONLY                           = 0x04,
IBV_OPCODE_SEND_ONLY_WITH_IMMEDIATE            = 0x05,
IBV_OPCODE_RDMA_WRITE_FIRST                    = 0x06,
IBV_OPCODE_RDMA_WRITE_MIDDLE                   = 0x07,
IBV_OPCODE_RDMA_WRITE_LAST                     = 0x08,
IBV_OPCODE_RDMA_WRITE_LAST_WITH_IMMEDIATE      = 0x09,
IBV_OPCODE_RDMA_WRITE_ONLY                     = 0x0a,
IBV_OPCODE_RDMA_WRITE_ONLY_WITH_IMMEDIATE      = 0x0b,
IBV_OPCODE_RDMA_READ_REQUEST                   = 0x0c,
IBV_OPCODE_RDMA_READ_RESPONSE_FIRST            = 0x0d,
IBV_OPCODE_RDMA_READ_RESPONSE_MIDDLE           = 0x0e,
IBV_OPCODE_RDMA_READ_RESPONSE_LAST             = 0x0f,
IBV_OPCODE_RDMA_READ_RESPONSE_ONLY             = 0x10,
IBV_OPCODE_ACKNOWLEDGE                         = 0x11,
IBV_OPCODE_ATOMIC_ACKNOWLEDGE                  = 0x12,
IBV_OPCODE_COMPARE_SWAP                        = 0x13,
IBV_OPCODE_FETCH_ADD                           = 0x14,
```

Elephant Flow-lets

```
User Datagram Protocol, Src Port: 55470 (55470), Dst Port: 4791 (4791)
InfiniBand
  Base Transport Header
    Opcode: 129
    0... .... = Solicited Event: False
    .0.. .... = MigReq: False
    ..00 .... = Pad Count: 0
    .... 0000 = Header Version: 0
    Partition Key: 65535
    Reserved (8 bits): 64
    Destination Queue Pair: 0x0000f1
    0... .... = Acknowledge Request: False
    .000 0000 = Reserved (7 bits): 0
    Packet Sequence Number: 0
  Vendor Specific or Unknown Header Sequence
    Unknown/Vendor Specific Data: 00000000
  Invariant CRC: 0x00000000
```



REGIONAL SDC

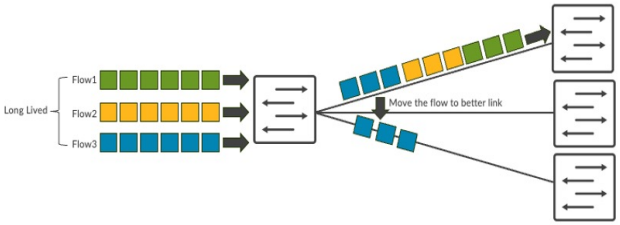
AI DC Software

Ready!!

Congestion Avoidance
(Effective utilization of fabric capacity)

Efficient Load Balancing Features

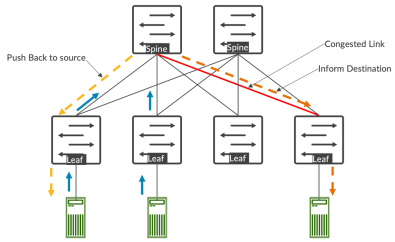
- DLB
- GLB
- Selective DLB
- Reactive Path Rebalancing
- BGP-DPF: GPU/QP pinning**



Congestion Control
Preventing packet drops (lossless)

Congestion Management Features

- DCQCN (PFC+ECN)
- Telemetry:
  - Interface level
  - Queue level
- PFC Watch-dog
- Per-Q Alpha Threshold value support



** In-Planning

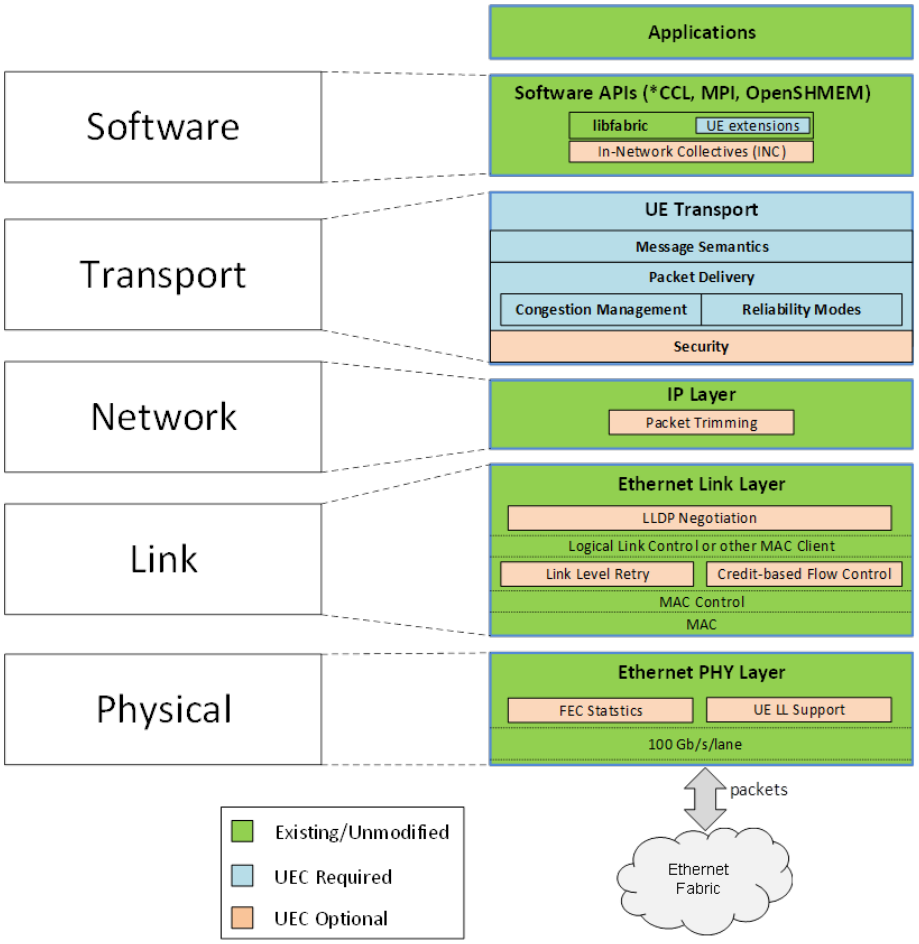
REGIONAL SDC

UEC – 10,000 Ft View

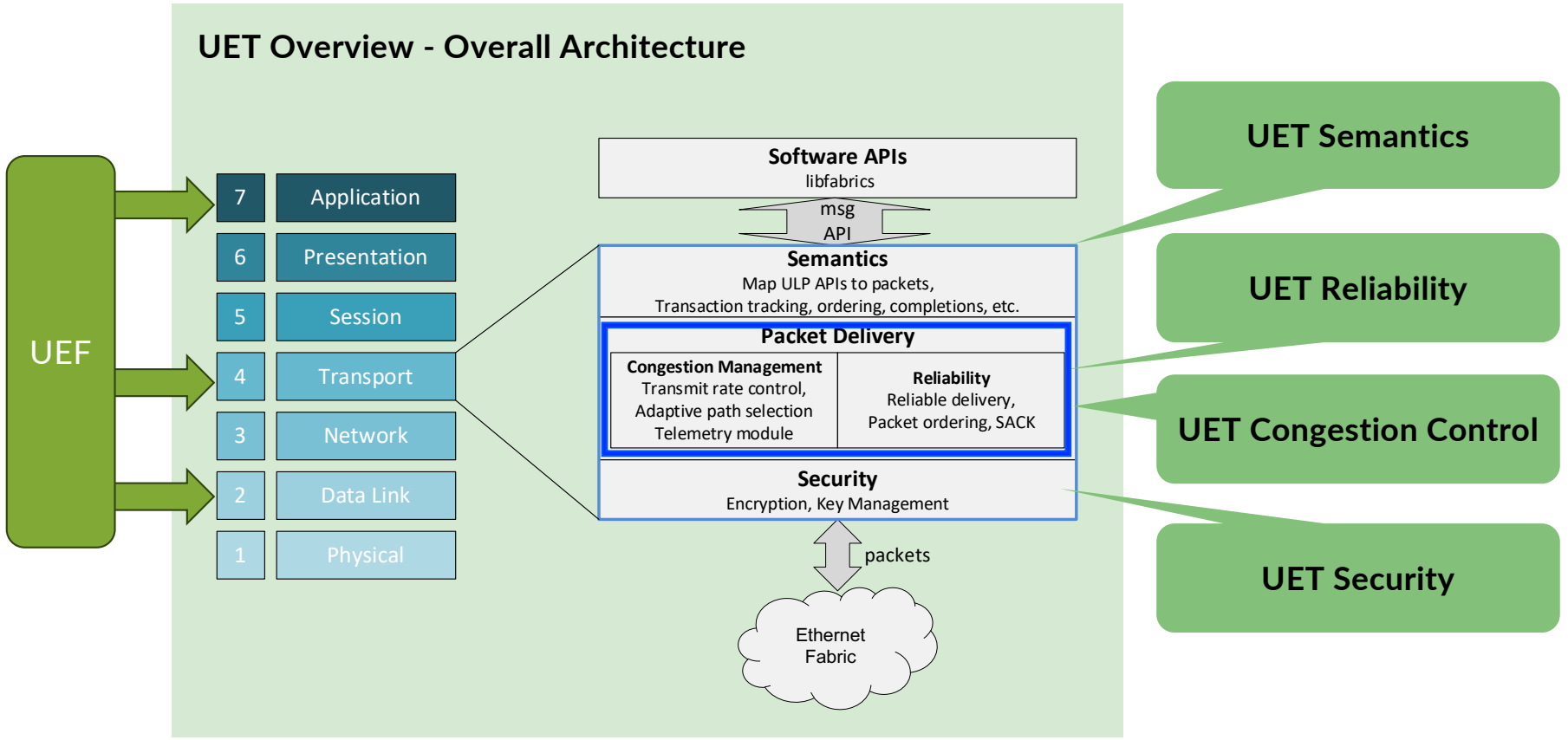
	InfiniBand	Lossless Ethernet Fabric (JUNIPER)	Ultra Ethernet Consortium
Congestion Avoidance Effective utilization of fabric capacity	Central control	• Distributed architecture • Dynamic Routing • Intelligent Load Balancing (DLB, GLB, sDLB etc.)	• Dynamic Routing • Packet Spraying • Flexible Re-ordering on NIC • In Network Collectives (INC)
Congestion Control Preventing packet drops (lossless)	Credit Request/Grant mechanism	DCQCN (PFC + ECN)	• Link-level credit mechanism (LLR) • Receiver credit mechanism (RCCC) • Sender Network Signal CC (NSCC) • Telemetry based CC (CSIG) • Packet Trimming (DCN)

REGIONAL SDC

UEC– Overall Architecture

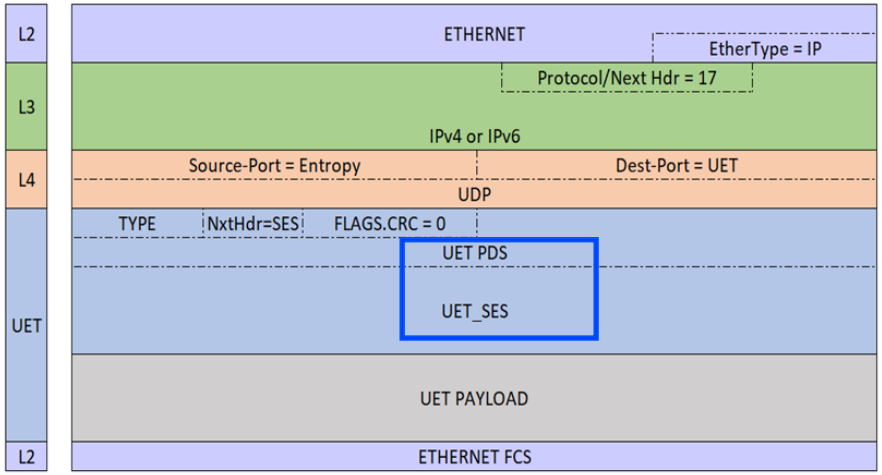


Ultra Ethernet Fabric – ordering and transactions

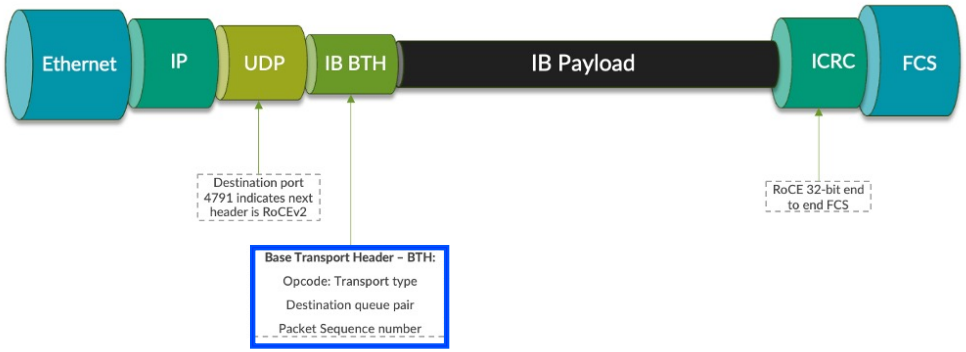


UEC vs ROCEv2 – Frame Format

UEC Transport – Packet Format



ROCEv2 Transport – Packet format



UEC : Network & NIC dependency

	UEC New Features	Any NIC Support Required ?	Network-Switch
1	Packet Level Multi-Pathing & UET forwarding (Co-existence of sprayed and unsprayed traffic)	Yes	No need of special-code (Need End-to-End validation)
2	Packet trimming (Drop Congestion Notification)	Yes	YES (Need Dev. Work)
3	Link layer retry (LLR)	No**	YES (Need Dev. Work)
4	Credit-based link-level flow control (CBFC)	No**	YES (Need Dev. Work)
5	In-Network Collectives	Yes	NIC Support require (Need End-to-End validation)
6	NSCC – assisted ECN and RTT pass-through	Yes	NIC Support require (Need End-to-End validation)
7	RCCC – credit pass-through & monitoring	Yes	YES (Need Dev. Work)
8	CSIG (Congestion Signalling)	Yes	YES (Need Dev. Work)

THANK YOU