MCTP over PCIe® Implementation

Eliel Louzoun
Hardware Architect
Intel Israel

Disclaimer

Presentation Disclaimer: All opinions, judgments, recommendations, etc. that are presented herein are the opinions of the presenter of the material and do not necessarily reflect the opinions of the PCI-SIG®.

Agenda

- What is MCTP?
- What is MCTP over PCI Express®?
- MCTP over PCI Express implementation
- Issues specific to PCI Express binding
- NC-SI usage model

MCTP OVERVIEW

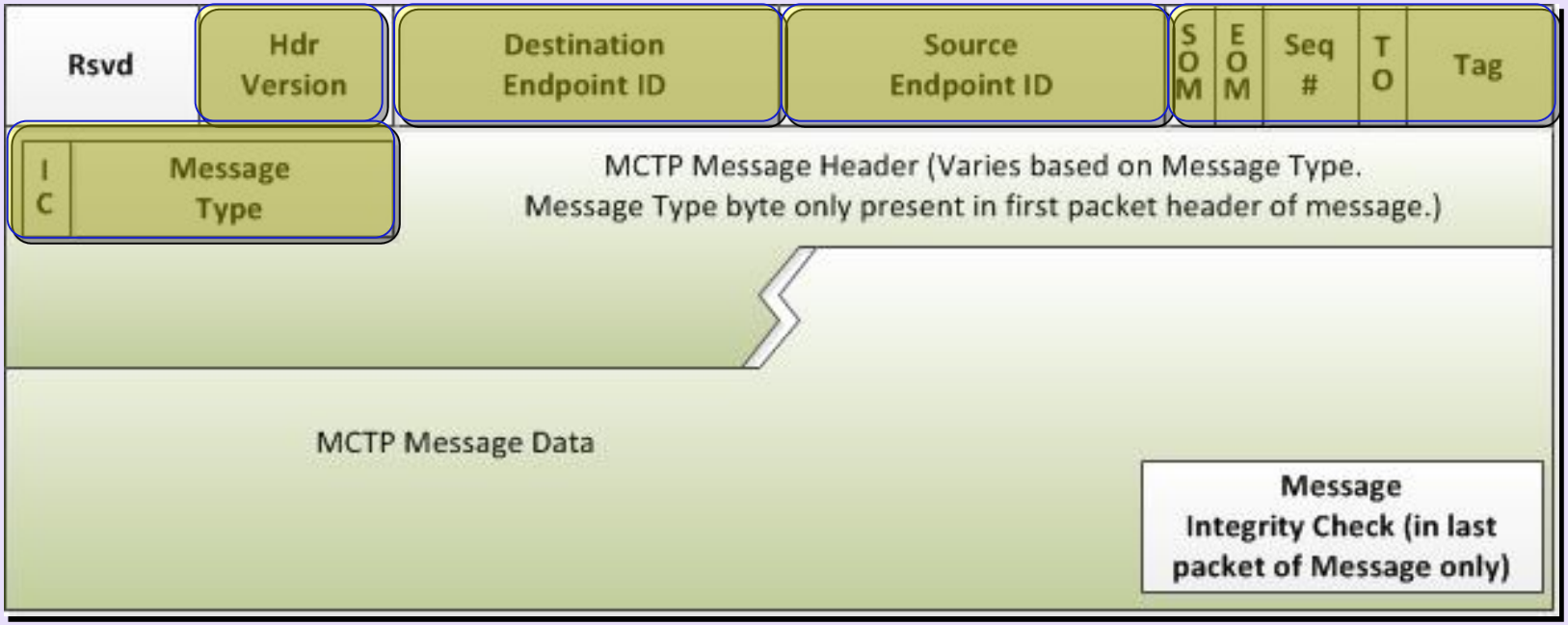
What is MCTP?

- MCTP is a DMTF standard defining a transport protocol between elements within a platform
  - ✓ **D**istributed **M**anagement **T**ask **F**orce
  - ✓ **M**anagement **C**omponent **T**ransport **P**rotocol
- MCTP can be sent over multiple media type:
  - ✓ PCI Express
  - ✓ SMBus/i^2c
  - ✓ UART
  - ✓ Host Interface
  - ✓ USB (planned)

MCTP messages

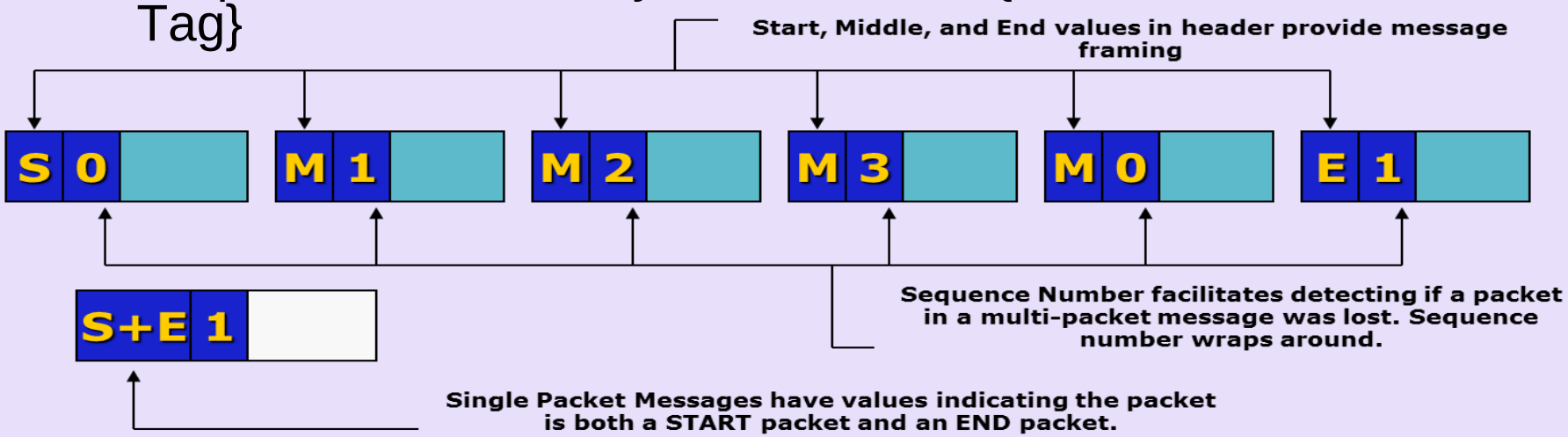
- MCTP can transport multiple type of messages over a Common Transport layer
  - ✓ Control
    - Messages used to establish the transport layer
  - ✓ Platform-Level Data Model (PLDM)
    - Conveys sensor and platform configuration data using DMTF defined messages
  - ✓ Network Controller Sideband Interface (NC-SI)
    - a control protocol used to establish a link between an MC (Management Controller) and the network
  - ✓ Ethernet
    - Pass-through traffic between MC and network.
  - ✓ OEM defined

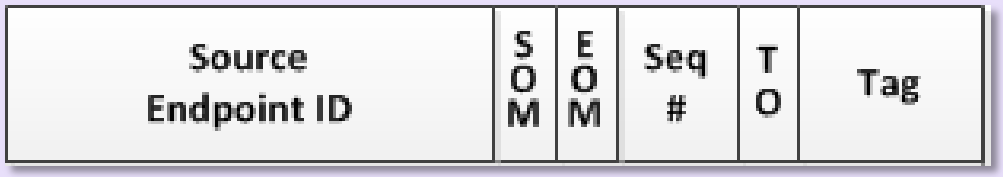
MCTP packet format



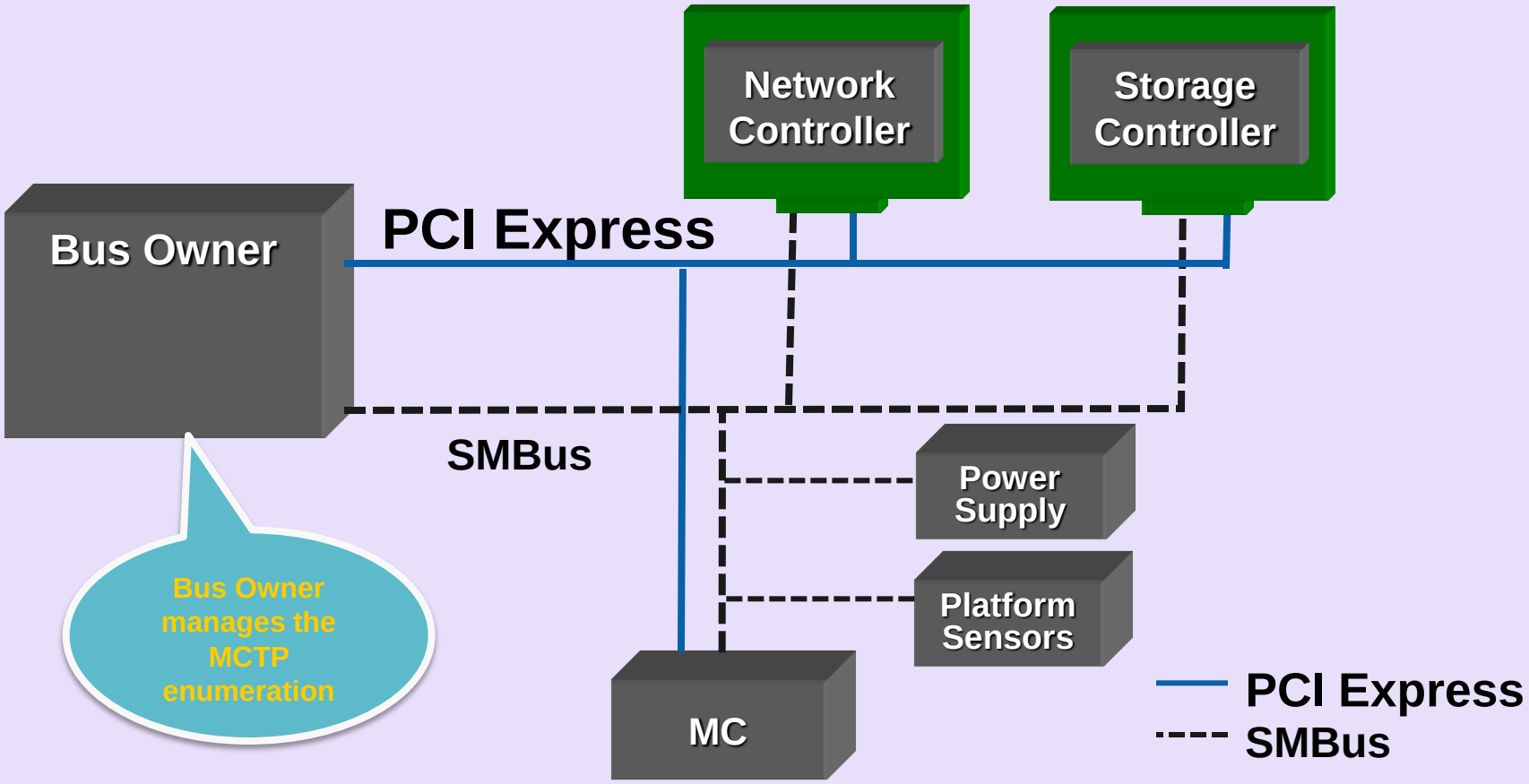
MCTP segmentation and Reassembly

- MCTP messages can be split to packets
  - ✓ Accommodates the medium payload size limitations
  - ✓ Minimum packet size is 64 byte
  - ✓ Requires segmentation and re-assembly process
  - ✓ Sequence identified by terminus ID = {Source EID, TO and Tag}





MCTP connections - example



MCTP over PCI Express

- VDMs are used to send MCTP messages over PCI Express links
  - ✓ Allows high speed communication between platform elements
  - ✓ E.g. MC to NIC Ethernet traffic
- Uses peer to peer communication
- MCTP Endpoint ID is mapped to a Requester ID {bus, device, function}

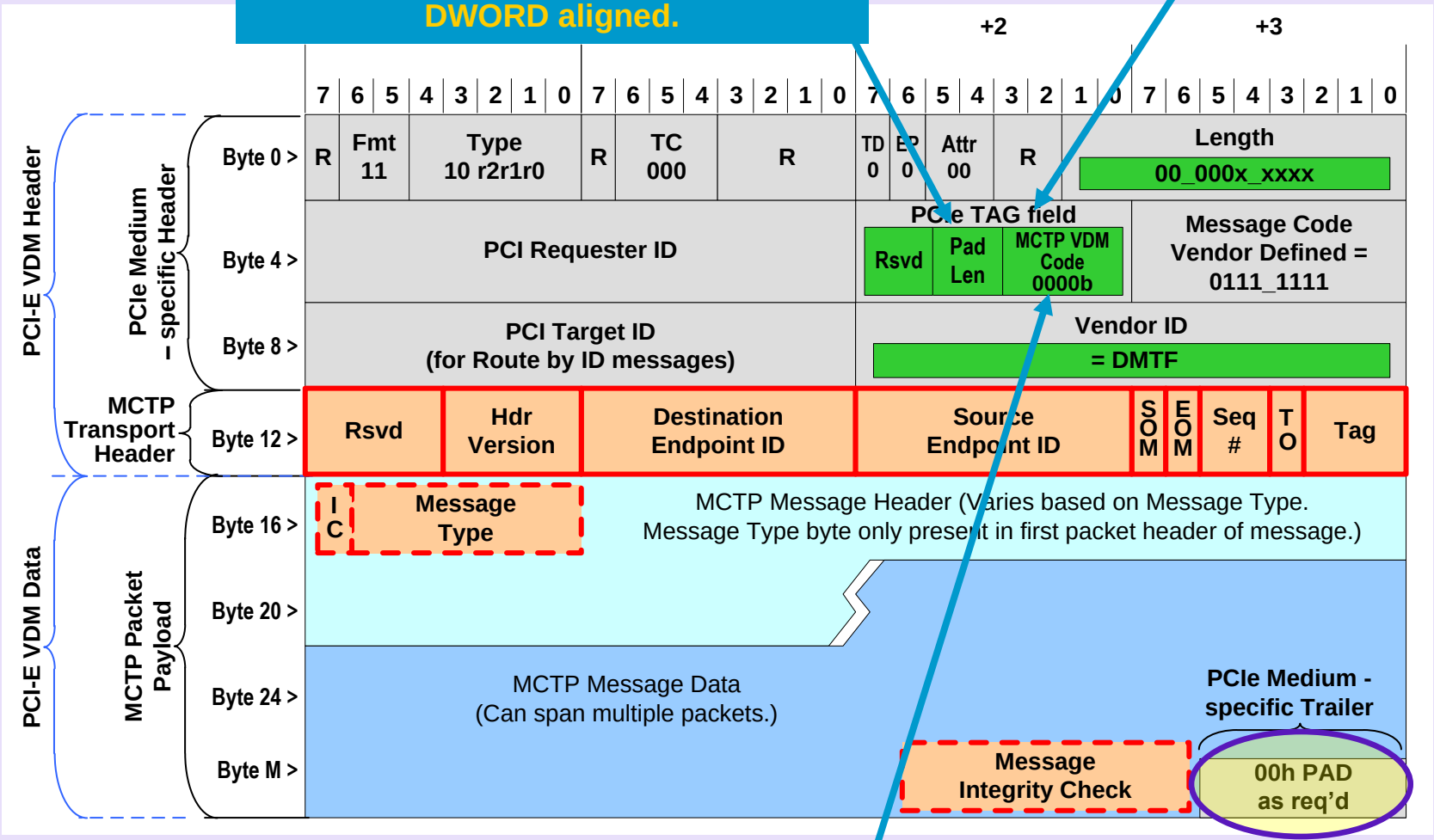
MCTP over PCI Express routing

- Route by ID for regular messages
- Broadcast from Root Complex for Bus owner's Discovery messages
- Route to Root Complex for endpoint's Discovery messages

PCI Express VDM Packet Format

Pad Length - (2-bits) indicates # of pad bytes to get PCI Express message DWORD aligned.

MCTP defined usage of PCI Express TAG field



MCTP VDM Code uniquely identifies MCTP VDMs from other possible VDMs that may be defined under the DMTF Vendor ID

Intel MCTP over PCI Express Implementation

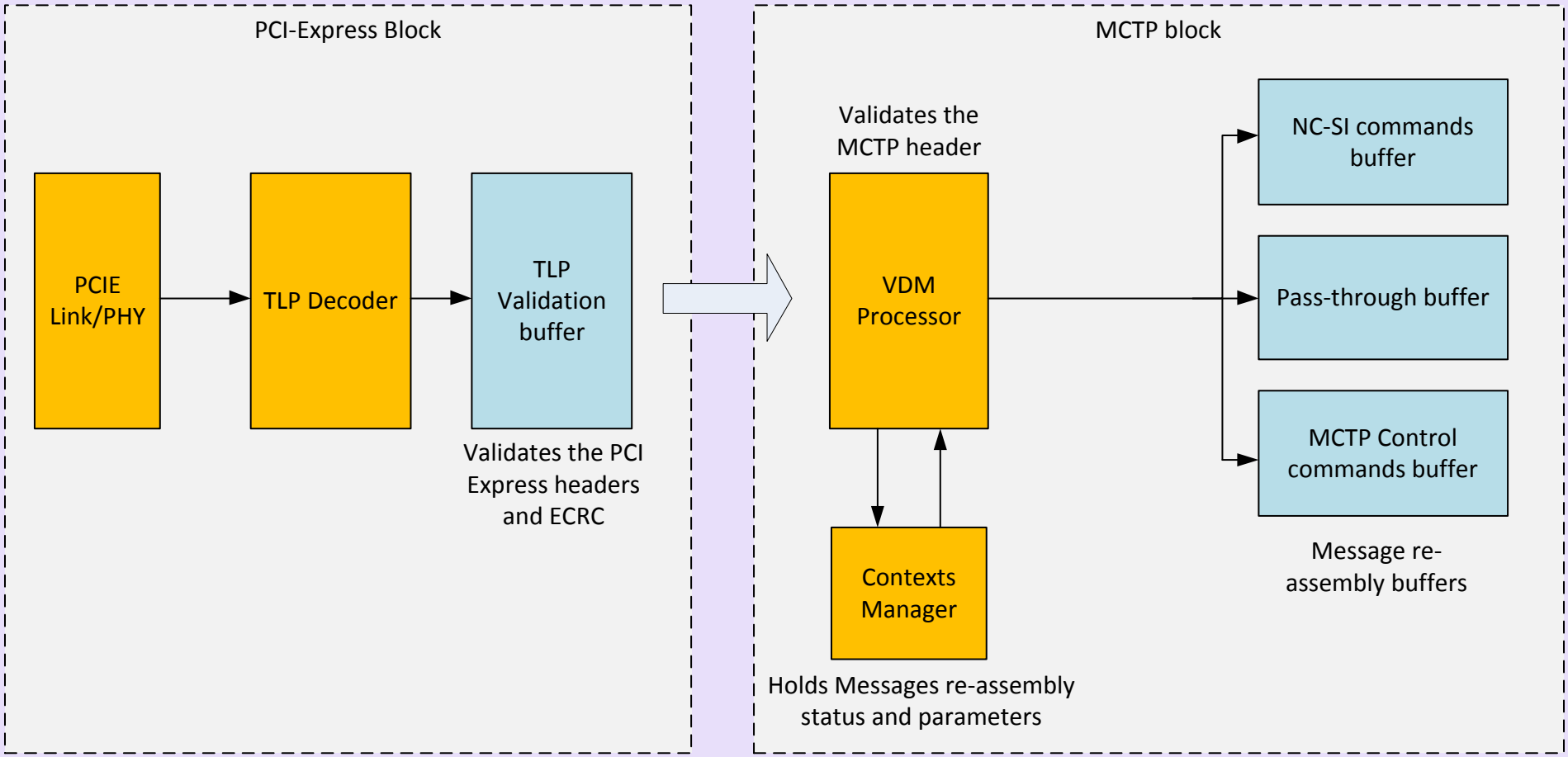
- MCTP over PCI Express is implemented in Intel server NICs
  - ✓ Starting from 2012 products
- Used to
  - ✓ Convey NC-SI and Ethernet traffic to a MC (a.k.a. pass-through traffic)
  - ✓ Allow the MC to control the NIC
- Partitioned between Hardware and Firmware

MCTP DESIGN CONSIDERATIONS

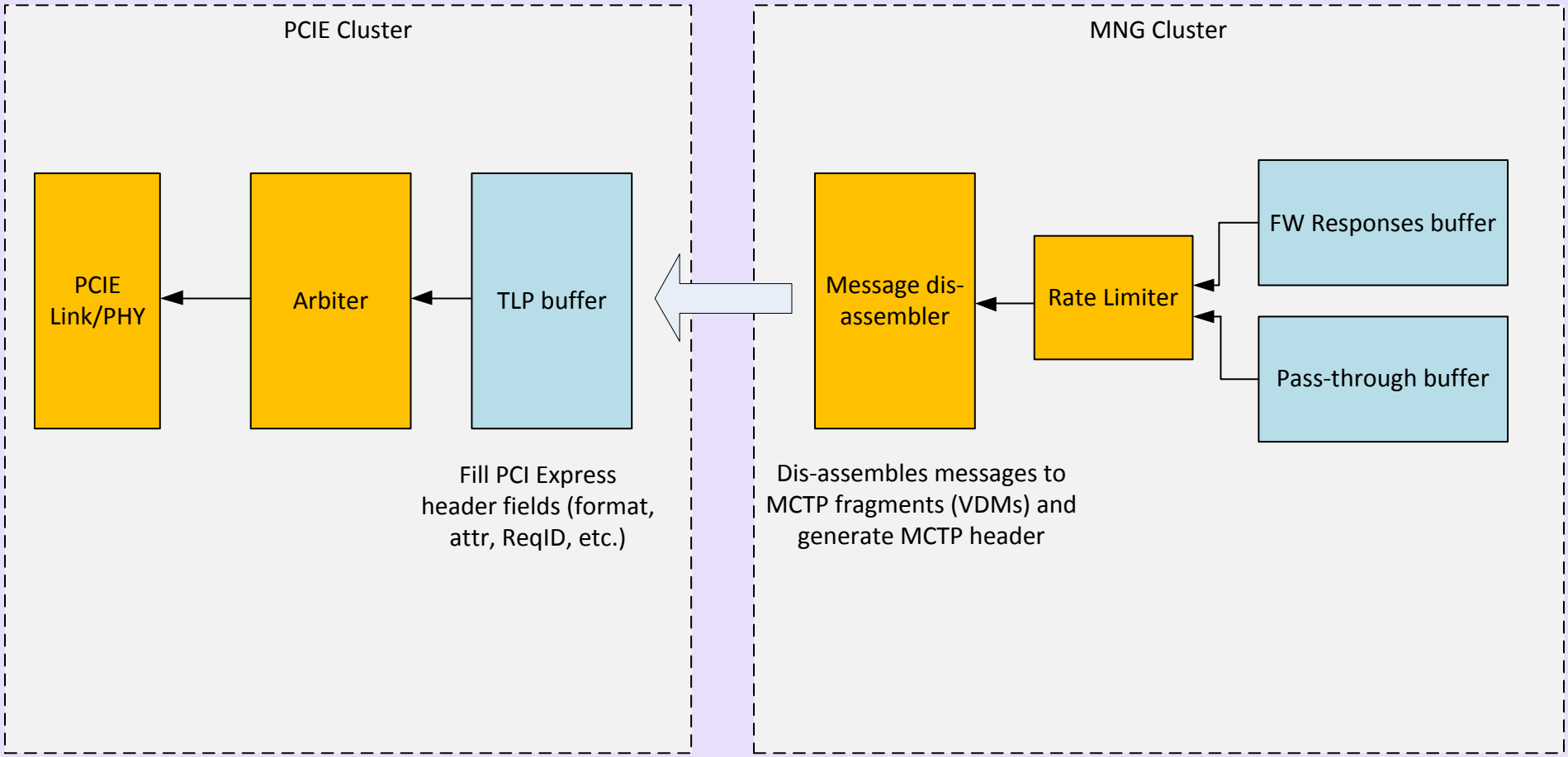
Design Considerations

- Block Diagram
- Number of Re-assembly context
- HW/FW demarcation
- TLP validation
- Flow Control
- ECRC generation
- Mapping of MCTP traffic to PCI Express functions
- Transition between medium
- Relationship with BIOS enumeration

Block Diagram - inbound



Block Diagram - outbound



Number of contexts

- MCTP requires re-assembly of packets into a message
  - ✓ Messages may be interleaved
  - ✓ Need a re-assembly context for each concurrent message - Terminus ID and Sequence Number
  - ✓ Number of contexts should be defined according to usage model
    - Example – an MC working with multiple NICs needs a context per NIC

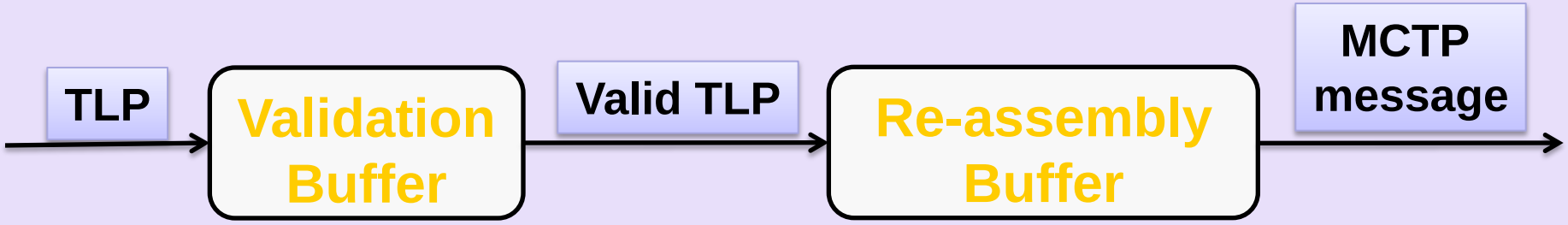
HW/FW demarcation

- PCI Express is fast …
- Cannot implement all MCTP over PCI Express handling in Firmware
- Data path in Hardware
  - ✓ Ethernet pass-through traffic between MC an LAN.
  - ✓ Segmentation and reassembly of all MCTP packets
- Control path in Firmware
  - ✓ Control Commands handling – both MCTP and NC-SI
- Allows line rate handling of traffic and flexibility for new control requirements

TLP validation

- TLPs need to be validated at PCI Express level before handling at MCTP level
- Implementation options:
  - ✓ Validate on reassembly buffer
    - Saves a buffer
    - Requires rollback in case of error
  - ✓ Validate on separate buffer

Option 2 – separate buffer, was chosen



Flow Control

- PCI Express is fast …
  - ✓ Endpoint may not be so fast
- Example:
  - ✓ A MC getting traffic from a high speed Ethernet connection. Needs to consume traffic in Firmware that may have limited buffering

Need to allow throttling of MCTP traffic

Flow Control

- Option 1: Use PCI Express credits
  - ✓ Credits are point to point. Will slow down the entire PCI Express fabric to the MC consumption rate
- Option 2: Use Ethernet flow control
  - ✓ Limited to the MC to NIC traffic. Not adequate for other type of messages
  - ✓ Understood from current implementations. Still “bursty” and requires buffering to absorb flow control reaction time
- Option 3: Use rate limiters
  - ✓ Define an acceptable rate for the receiver
    - Constrain transmitter to this rate
  - ✓ Can be either packet (PCI Express VDM) rate or message (Ethernet packet) rate

Option 3 – rate limiter, was chosen

ECRC generation

- PCI Express requires ECRC insertion on all or none of the TLPs

> ❑ If a device Function is enabled to generate ECRC, it must calculate and apply ECRC for all TLPs originated by the Function

- MCTP over PCI Express binding specification does not allow ECRC in TLP
- Specifications contradict
- NIC advertise ECRC support
- Added a dedicated control that defines if ECRC should be inserted to MCTP TLPs

Transition to other medium in low power state

- Management traffic is required in all S states
  - ✓ Pass-through traffic to MC
  - ✓ Sensors information
  - ✓ Pre-boot configuration
- Could use MCTP over SMBus always, but…
  - ✓ Can use the additional Bandwidth in S0
- Recommendation
  - ✓ Transition from MCTP over SMBus in Sx to MCTP over PCI Express in S0

Mapping of MCTP traffic to PCI Express functions

- MCTP Endpoint ID (EID) needs to be mapped to a Requester ID - {bus, device, function}
- In multi function devices, multiple RIDs are available
- Options:
  - ✓ One MCTP EID per RID
    - Fits when MCTP endpoints logically maps to functions
    - Example – integrated MC and Graphic controller
  - ✓ One MCTP EID for entire device – mapped to one RID
    - Fits when MCTP endpoint maps to device
    - Example – NC-SI package mapped to a single EID

Relationship with BIOS enumeration

- PCI Express spec allows to send VDMs immediately after reset (D0u state)
- At this state, enumeration may still be in progress
- NIC waits for Bus Master Enable (BME) setting before using VDMs
  - ✓ Consider MCTP traffic same as DMA
  - ✓ Non standard behavior
- This behavior should be coordinated with Bus Owner
  - ✓ Should wait for end of BIOS enumeration before MCTP discovery process

Relationship with BIOS enumeration – cont.

- NIC has two MCTP channels
  - ✓ Over SMBus and over PCI Express
  - ✓ NC-SI traffic transitions between them
- If BME cleared, keep NC-SI traffic over SMBus
  - ✓ Assumes PCI Express channel not established yet

NC-SI usage model

- Previous model:
  - ✓ Connection over dedicated RMII – 100 Mbps
- Current model
  - ✓ MCTP over PCI Express + SMBus
    - PCI Express in S0, SMBus in Sx.
  - ✓ Fits add on cards (NICs)
    - No need for dedicated connectivity
  - ✓ Higher bandwidth in S0
    - At the expense of Sx bandwidth

Binding transition State Machine

- NC-SI session should be independent of underlying binding
- Independent establishment of MCTP connection in SMBus and PCI Express
  - ✓ SMBus is always there
- Transition from SMBus to PCI Express initiated by MC by sending an NC-SI command to the NC on PCI Express
- Automatic fallback from PCI Express to SMBus when PCI Express becomes unavailable (BME disabled)

Summary

- MCTP over PCI Express provides an high bandwidth, standardized, channel between entities in the platform using existing connections.
  - ✓ Allows control of platform elements.
- Allows manageability traffic through add-in cards.
- Can be combined with SMBus to provide connectivity in all power states

Thank you for attending the PCI-SIG Developers Conference

Israel 2013

For more information please go to www.pcisig.com