CXL Fabric Management
Vincent Haché
Director of Systems Architecture, Rambus
Agenda

• CXL Overview
• System Management Requirements
• What is a Fabric Manager?
• Component Command Interface
• Management Architecture
• Management Command Sets
• MLD Management
There are 5 key resources covering Fabric Management:

1. CXL 2.0 Specification ([Link](#))
   - Introduces In-band device management via mailbox
   - Introduces FM API and MCTP transport details

2. CXL 2.0 Errata ([Link](#))
   - Critical fixes to asynchronous MCTP event notifications

3. Type 3 Management Using MCTP CCI ECN ([Link](#))
   - Enables MCTP-based device management, and generalizes key concepts that were formerly switch-specific

4. CXL FM API over MCTP Binding Specification ([Link](#))

5. CXL Type 3 Device CCI over MCTP Binding Specification ([Link](#))
CXL Delivers the Right Features & Architecture

**Challenges**

Industry trends driving demand for faster data processing and next-gen data center performance

Increasing demand for heterogeneous computing and server disaggregation

Need for increased memory capacity and bandwidth

**CXL**

An open industry-supported cache-coherent interconnect for processors, memory expansion and accelerators

**Coherent Interface**

Leverages PCIe® with 3 mix-and-match protocols

**Low Latency**

Cache and Memory targeted at near CPU cache coherent latency

**Asymmetric Complexity**

Eases burdens of cache coherent interface designs
CXL 2.0 Usage Models - Recap

**Caching Devices / Accelerators**
- Usages:
  - PGAS NIC
  - NIC atomics
- Protocols:
  - CXL.io
  - CXL.cache

**Accelerators with Memory**
- Usages:
  - GPU
  - FPGA
  - Dense Computation
- Protocols:
  - CXL.io
  - CXL.cache
  - CXL.memory

**Memory Buffers**
- Usages:
  - Memory BW expansion
  - Memory capacity expansion
  - Persistent Memory
- Protocols:
  - CXL.io
  - CXL.mem

Type 1 Device

Type 2 Device

Type 3 Device

*Compute Express Link™ and CXL™ Consortium are trademarks of the Compute Express Link Consortium.*
Existing Conventions:

System Management Requirements

- BMC
- SMBus
- JBOM
- CXL Switch
- Host A
- Host B
- Rack-mount Appliance
- Server
- Chassis
- CPU
- PCIe
- BMC
- T3
- MCTP via PCIe VDM
- Compute Express Link™ and CXL™ Consortium are trademarks of the Compute Express Link Consortium.
New Device Capabilities:

**Multi-Logical Device (MLD)**

- Host A
- Host B
- CXL Switch
- JBOM
- SMBus

**Multi-Headed Devices**

- Host A
- Host B
- Type 3

Compute Express Link™ and CXL™ Consortium are trademarks of the Compute Express Link Consortium.
What is a Fabric Manager?

Fabric Manager (FM) is a conceptual term that refers to the application-specific logic composing systems, allocating pooled resources, managing platforms, etc.

Can take many forms:
- BMC in a rack-mount appliance
- Management software running in a host
- Embedded FW in a CXL Switch

Compute Express Link™ and CXL™ Consortium are trademarks of the Compute Express Link Consortium.
Framework is flexible by design to enable a wide variety of applications (embedded, automotive, hyperscale...)

Most management capabilities are optional

FM is required for advanced system operations:

- Use of MLDs – FM is responsible for assigning LDs to hosts
- Memory pooling – FM is responsible for binding switch ports to host hierarchies

What is a Fabric Manager?
Component Command Interface
Intro to Component Command Interface

- Commands are processed by a Component Command Interface (CCI)
- Two types
  - Mailbox CCI – presented through memory registers
  - MCTP-based CCI – presented as an MCTP EP
- Not a queued interface
- Lengthy operations run as "Background Operations"
- A component may support multiple, with varying capabilities
- Command opcodes are 2B: 1B command set, 1B command
- Supported command list is reported through “Command Effects Log”
Mailbox CCI

Located in PCIe MMIO Space

Two types of mailbox:
- Primary – designed for use by driver; intended for privileged operations
- Secondary (optional) – designed for log/event record access; no interrupt or background operation support

Command inputs written to Command Payload Registers, outputs read from same region

Optionally generates MSI/MSI-X interrupts

Compute Express Link™ and CXL™ Consortium are trademarks of the Compute Express Link Consortium.
FM will first discover all MCTP EPs using MCTP specified discovery.

CCIs will advertise support for CXL Message Types:
- Type 07h for FM API commands
- Type 08h for General and Memory Device Commands

Supported over any physical interface for which an MCTP binding spec is defined.
Time-consuming management operations are defined as “Background Commands”

Up to one per CCI supported, but recommended to only support those commands on one interface

Command receives immediate response to indicate B0 has started
Example Rack-mount Appliance with Spec Terminology:

Management Architecture

Compute Express Link™ and CXL™ Consortium are trademarks of the Compute Express Link Consortium.
Management
Command
Sets
Command Set Categories

General Device Opcodes (0000h – 3FFFh)
- General Component Command Sets

Class-specific Opcodes (4000h – BFFFh)
- Memory Device Command Sets
- FM API Command Sets

Vendor-specific Opcodes (C000h – FFFFh)
- Vendor-specific Command Sets

Compute Express Link™ and CXL™ Consortium are trademarks of the Compute Express Link Consortium.
General Component Command Command Sets

Uses command opcodes 0000h to 3FF FH

Applicable to all classes of devices (Type 1/2/3 devices and switches)

Includes generic capabilities used in the discovery and management of all classes of devices

Information and Status (00h)
- Identify and CCI status

Events (01h)
- Read/clear event records and interrupt configuration

Firmware Update (02h)
- Program and activate

Timestamp (03h)

Logs (04h)
- Supported command list (Command Effects Log)
General Component Command Sets

<table>
<thead>
<tr>
<th>Table 173. CEL Entry Structure</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Byte Offset</strong></td>
</tr>
<tr>
<td>-----------------</td>
</tr>
<tr>
<td>0</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

Information and Status (00h)
- Identify and CCI status

Events (01h)
- Read/clear event records and interrupt configuration

Firmware Update (02h)
- Program and activate

Timestamp (03h)

Logs (04h)
- Supported command list (Command Effects Log)
Memory Device Command Sets

Uses command opcodes 4000h to BFFFH

Applicable to Type 2/3 devices

Includes all commands specific to management of memory media

Used by System FW during boot and kernel drivers after boot

- Identify (40h)  
  - Identify memory device capabilities
- Capacity Config and Label Storage (41h)  
  - Manage labels for persistent memory
- Health Info and Alerts (42h)  
  - Media state, temperature, health alerts
- Media and Poison Management (43h)
- Sanitize (44h)  
  - Secure clearing of memory
- Persistent Memory Data-at-Rest Security (45h)  
  - Set security parameters, lock, unlock, etc.
- Security Passthrough (46h)  
  - Passthrough for SFSC commands
- SLD QoS Telemetry (47h)
Uses command opcodes 4000h to BFFFh

Applicable to CXL switches and MLDs

Includes binding commands, LD assignment, and port control

Used by FM to manage switch-attached, disaggregated resources

Physical Switch (51h)
- Identify, port status, port resets

Virtual Switch (52h)
- Binding and unbinding in multi-VCS switches

MLD Port (53h)
- Command tunneling

MLD Component (54h)
- Capacity allocation and QoS management

Compute Express Link™ and CXL™ Consortium are trademarks of the Compute Express Link Consortium.
MLD Management
MLDs are accessed by LD-ID, with up to 16 LDs available to hosts (0h to Fh)

LD-ID FFFFh is mandatory and reserved as the ‘FM-owned LD’, a management target with no memory resources

FM-owned LD is only .io accessible, as .mem and .cache only include 4 LD-ID bits

MLD must implement a CCI for each LD plus one for the FM-owned LD

FM may tunnel commands to MLDs through switch, as needed
Configuring FM-owned LD through a switch:

- **FM** -> **Switch**
  - **Tunnel Management Command Request**
    - Port = X
    - Set LD Allocations Request
  - **MCTP-Capable I/F**
  - **Tunnel Management Command Response**
    - Length
    - Set LD Allocations Response

- **Switch** -> **MLD**
  - **Set LD Allocations Request**
  - **MCTP PCIe VDM**
  - **Set LD Allocations Response**
  - **FM-Owned LD**
    - Set LD Allocations Response

Compute Express Link™ and CXL™ Consortium are trademarks of the Compute Express Link Consortium.
Provisioning individual LDs via tunneling:
Provisioning individual LDs via tunneling through a switch:
MLD Management

Host A (Powered off)
Device Drivers

Host B (Powered off)
Device Drivers

MLD Management Diagram

Compute Express Link™ and CXL™ Consortium are trademarks of the Compute Express Link Consortium.
MLD Management

Host A (Powered off)
Device Drivers

Host B (Powered off)
Device Drivers

CXL Switch

Identify
Get Physical Port State

BMC

RM

T3 MLD

Device Drivers
MLD Management

Host A
(Powered off)
Device Drivers

Host B
(Powered off)
Device Drivers

Tunnel Management Command

CXL Switch

Identify
Set LD Allocations
Etc.

T3 MLD

LD-1D0
LD-1D1
LD-1D2
LD-1D3

BMC
RM

Compute Express Link™ and CXL™ Consortium are trademarks of the Compute Express Link Consortium.
MLD Management

**Host A**  
(Powered off)

**Host B**  
(Powered off)

Device Drivers

- BMC
- FM

**CXL Switch**

- **Bind (LD0, Host A)**

**T3 MLD**

Device Drivers

Compute Express Link™ and CXL™ Consortium are trademarks of the Compute Express Link Consortium.
MLD Management

Compute Express Link™ and CXL™ Consortium are trademarks of the Compute Express Link Consortium.
Summary
Summary

Key concepts:
- Flexibly-defined architecture to serve variety of applications
- Management available over many interfaces
- Fabric Manager (FM) – any logic initiating management commands
- Component Command Interface (CCI) – management command target in components
- Management Command Sets
- MLD Management
Q&A

Please share your questions in the Question Box.
Thank You